Can Online Juries Make Consistent, Repeatable Decisions?

ANONYMOUS AUTHOR(S)

A jury of one's peers is a prominent way to adjudicate disputes and is increasingly used in participatory governance online. The fairness of this approach rests on the assumption that juries are consistent: that the same jury would hand down similar judgments to similar cases. However, prior literature suggests that social influence would instead cause early interactions to cascade into different judgments. In this paper, we report an online experiment that changes participants' pseudonyms as they appear to collaborators, temporarily masking a jury's awareness that they have deliberated together before. This technique allows us to measure consistency by reconvening the same jury on similar cases. Counter to expectation, juries are equally consistent as individuals, a result that is "good for democracy." But this consistency arises in part due to group polarization, as consensus develops by hardening initial majority opinions. Furthermore, we find that aggregating groups' perspectives without deliberation erodes consistency.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: juries, deliberation, groups, individuals, governance

ACM Reference Format:
Anonymous Author(s). 2021. Can Online Juries Make Consistent, Repeatable Decisions?. In *Yokohama '21: ACM CHI Conference on Human Factors in Computing Systems, May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In 2019, just two weeks apart, the United States 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 without the imam, but the Buddhist man's could not [48]. Legal outrage ensued; how could the court provide such starkly different judgments on two nearly identical cases?

A similar sort of indignation erupts from judgment disparities in online social computing systems, such as when Facebook chose not to take down a U.S. Congressman's post appearing to incite violence against Muslims (it read, in part, "kill them all"), but to remove a post from a Black Lives Matter activist that read, "all white people are racist" [6]; in another case, YouTube demonetized videos if they included the word "transgender," but did not demonetize them when "transgender" was removed and all other aspects of the video remained constant [4]. Juries are gaining momentum online as a mechanism for resolving such disputes. Multiple platforms offer online juries for pre-testing arguments or adjudicating small claims [8, 31], and juries increasingly regulate anti-social behavior [18, 25]. Facebook, for its part, has adopted the jury concept into its new five-member panels of Oversight Board members, widely dubbed its "Supreme Court" [2]. However, as juries gain prominence online, we have no guarantees whether they will be seen as inconstant and volatile.

Manuscript submitted to ACM

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. As Hastie et al. write in their seminal book, "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" [22].

This issue inspires our primary research question: *are jury decisions consistent*? In this paper, we report an online experiment that compared group jury decisions to a baseline of individually-made decisions, since this is frequently the realistic alternative: rather than being judged by a 12-person jury, a single judge could decide; rather than using a digital jury, a moderator or community manager could unilaterally decide.

Prior work is inconclusive on whether group decisions would produce similar outcomes against similar cases. One theory suggests that groups regularly align with the perceived majority opinion [7, 16, 21, 22, 33, 43, 47], predicting that groups would be consistent. This tendency is so strong that group members align with the majority view even when it opposes their own decision [41]. However, a second line of literature argues that group decisions may be quite inconsistent: members of a group do not have global knowledge, and must judge the perceived consensus based on the available social signals [27]. These social signals can be arbitrary, as groups tend to follow the positions of those who speak first [45] and polarize to more extreme views [44] — giving rise to much more volatile outcomes.

Our research question has previously been unanswerable because past studies have largely focused on in-person juries [16, 17, 22, 28, 37]. In-person juries are unable to directly measure the consistency of a jury on two similar cases: over the course of deliberation, the jury would build a shared social context that would influence subsequent deliberations. Therefore, tests of consistency have been confined to comparing a jury's non-deliberative predispositions to its final verdict [37], comparing a decision to the "true" court result [14], or comparing one jury's conclusion to an entirely separate "shadow jury" [28]. A direct study of consistency therefore requires a within-subjects experiment, in which the *same* jury deliberates and decides twice.

This paper resolves this theoretical contradiction in the literature. We contribute an experimental method that directly measures the consistency of group judgments in online juries, comparing them with the judgments of individual decision-makers. While this sort of within-subjects experiment is impossible with traditional juries, pseudonym masking can cause online juries to temporarily mask their social context [50]. Our experiment implemented a platform that pseudonymously convenes juries online. When the same jury meets for a second time, we change the pseudonyms of their perceived collaborators. As a result, the same jury can deliberate twice, without realizing that the group members have remained unchanged. Juries in our experiment adjudicated a pair of correlated cases from a large online community that debates interpersonal conflicts, where pairs of cases were pre-tested to have correlated outcomes. In a within-subjects manipulation, we also asked participants to adjudicate another pair of correlated cases alone. We measured consistency based on the proportion of matching decisions amongst pairs in each condition. Since deliberations are reliant on social signals that vary from interaction to interaction, we preregistered a hypothesis that groups would be less consistent than individuals.

Across $N = 259$ trials, we find that groups and individuals are equally consistent, contrary to both our hypothesis and study participants' own predictions. After deciding a case, a group gave the same judgment for its paired case 62.92% of the time (95% CI: 0.558, 0.700); individuals working alone gave the same judgment 63.88% of the time (95% CI: 0.607, 0.670). Our results suggest that deliberation is a far more stable process than we expected; arbitrary social influences did not play as large of a role in shaping the juries' decisions. This result is highlighted when we directly compared group decisions with those made by nominal groups — that is, the votes of individuals working alone, aggregated as if they were a jury. Only 46.07% (95% CI: 0.387, 0.534) of nominal groups decided consistently. Since

nominal groups did not deliberate, these results suggest that deliberation is a key ingredient to decision consistency. Finally, an important factor in decision consistency may in fact be the tendency to align with the perceived majority view [7, 21, 33]. More than a quarter (27.5%) of participants changed their minds following deliberation. Just under half (41.7%) of the pre-existing majority factions grew in size after deliberation, compared to 7.4% for non-deliberating nominal groups.

Overall, we find that deliberation enables groups to develop consistent outcomes, a result that is "good for democracy." However, this consensus often takes the form of leaning harder in the direction of pre-existing opinions, supporting prior studies in group polarization [44]. Ultimately, our results suggest that online platforms can proceed with increased confidence that their jury decisions will not be random or mercurial, but largely stable when conditioned on the membership of the jury.

2 RELATED WORK

Within HCI, the study of how juries decide is part of a larger conversation on the study of groups. This literature has been dominated by concepts like collective intelligence [52], the wisdom of crowds [46], social choice [26], and online deliberative governance [3, 19, 53]. However, the computationally mediated context has also introduced a trove of novel experimental scaffolds for understanding groups, such as evaluating the impact of autonomous agents in social situations [24]; augmenting group formation [38], dynamics [54], or memory [49, 50]; and predicting group outcomes [1, 12]. Here we draw on these methods, in particular 2 way-pseudonym masking [50], but we also engage a broader literature on the underlying nature of how groups make decisions. As a result, this paper refers to *juries* as a specific instance of decision-making groups, and uses the term *groups* to describe the general behavioral patterns we observe.

A group decision carries implicit legitimacy. In the offline world, juries in the United States of America are viewed as so vital that a trial by jury 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) [5]. Scholars have described group deliberation as a "special form of speech structured according to democratic principles...designed to transform private prejudice into considered public opinion and to produce more legitimate solutions" [34]. Because of groups' implicit legitimacy, society also hands them substantial power — from the offline jury to Facebook's "Supreme Court."

We focus in this paper on *consistency*, the property of similar cases being decided in similar ways. Consistency is a key assumption of jury-based decisionmaking: if juries decide capriciously, it undermines the legitimacy of the enterprise. This motivates our core research question:

Research question 1 (RQ1). *Do groups or individuals make decisions more consistently?*

Literature in social influence indicates that group deliberation may reduce consistency because the social process of deliberation may erode individual opinions through path dependence. Groups often fail to account for insights that only a few members know about, even when it is important [42]. Group members are far more likely to follow the perspective of the person who happens to speak first [45]. Group dynamics often reflect larger societal biases: for example, juries tend to choose white, middle-class males of high status to serve as the foreperson [22, 30], which in turn enables such individuals to wield disproportionate influence [16], potentially mitigating benefits of working as a group. Finally, Shah et al. found support for the "messy middle model," concluding that 30% of papers submitted to NIPS 2016 (now NeurIPS) were inconsistently decided in a parallel committee, as they were in between acceptance or

rejection [40]. Based on this literature, we pre-registered a hypothesis[1] that groups' social influence will cause their outcomes to vary substantially.

Hypothesis 1 (H1). *Individuals' decisions will be more consistent than those of groups.*

A critical consideration brought about by this hypothesis is that how groups and individuals are compared may impact its validity. In section 3.5 we further discuss 3 differing comparison strategies drawing from existing measures: *comparing groups to individuals directly* (our main measure), *comparing groups to nominal groups* [35] (individuals' votes aggregated into a group vote, as a pseudo-group), and *comparing individuals who worked alone to individuals who worked in groups*.

The mechanism of decision-making differs sharply between groups and individuals. An individual's decision consistency is most heavily dependent on three 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 person is situated in. 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 [36]. Though at times, individual judges may be swayed by arbitrary factors, such as lunch breaks [15], 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 [39].

In contrast, groups' decision-making processes tend to be more complex. Group dynamics include two major types of countervailing social forces when deciding together. The first type of process, which Moscovici and Faucheux 1972 describe as "conformity" [32], is a tendency to submit to a majoritarian view; the results of group deliberation tend to predictably favor the pre-existing majority [47], with minority faction members less likely to speak up [7, 33]. In jury studies, the size of the majority faction is among the strongest predictors of a jury's final verdict [16, 21, 22, 47]. This tendency to adhere to the majority view is often so powerful that it overrides an individual's private judgments and causes a polarization effect. Individuals will tend to strongly conform to the perceived truth, even at the expense of honestly representing their own beliefs [43, 45]. 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 [44]. Similarly, a large-scale survey of 3,500 jurors by Son et al. 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" [41]. Groups members are often so keen on aligning themselves with the perceived majority view that they systematically overestimate the prevalence of the majority opinion [27] and shift opinions towards the majority even when the platform is anonymous [23].

The contrast between individuals' heuristic-based process and groups' tendency to conform informs our second hypothesis:

Hypothesis 2 (H2). *While deliberating in groups, members are more likely to diverge from their initial predispositions in order to align with the majority viewpoint.*

The second major type of social force generates the opposite effect: deliberation may yield a novel or unexpected perspective. Combining Moscovici and Faucheux's categories of "innovation" (acceptance of the minority view by the majority) and "normalization" (compromise) [32], this second force drives the process of deliberation to generate perspectives that cannot be explained from merely aggregating the jurors' pre-deliberation views. In Diamond et al.'s 1998 study, the authors used actors to videotape different versions of a civil suit, with details altered in each version.

[1]We pre-registered our hypothesis on 20 April 2020, prior to the collection of our data, at the following URL: https://aspredicted.org/blind.php?x=9py7ab

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 [17]. 33% of verdict variation, however, could not be explained by either features of deliberation or the features of jurors.

Thus, we hypothesize that, while individual members of groups will change their minds in order to align with the group's opinion (H2), the group as a whole will, via exposure to novel perspectives, diverge from the pre-deliberation majority viewpoint more often than an aggregation of individuals who do not deliberate.

Hypothesis 3 (H3). *Overall, groups are less likely than individuals to decide in exactly the same way as their initial predispositions.*

Furthermore, if H3 is true, the process of generating new perspectives from deliberation would lead to less consistency for groups, thus bolstering H1, our core consistency prediction.

3 METHOD

To test our hypothesis, we designed a within-subjects experiment that measures how consistent an online jury would be when adjudicating two similar cases. A legitimate jury decision should be replicable: two similar cases should have the same outcome. We enable this comparison by introducing an experimental method that resets social dynamics within the online jury, such that jury members do not realize that they are deliberating with the same group a second time. To provide a point of comparison, we also asked individuals to adjudicate two similar cases alone.

3.1 Design

Figure 1 summarizes the experimental design. We begin with an overview and then explain each component in more detail. Each online jury consisted of 5–9 members and four rounds of adjudication. Participants adjudicated four unique cases in total — two as a group and two individually, with the ordering of cases and individual/group rounds randomized. The cases were chosen as pairs, such that each pair of cases had highly correlated judgments: a specific judgment on one case should, in most cases, be associated with a specific judgment on the other. We then calculated consistency from the rates at which individuals and groups judged the similar cases the same way. Finally, since the experiment uses a within-subjects design, we controlled for differences in opinion among subjects sampled.

Within group rounds, we asked participants to discuss with jury members and attempt to agree before the final vote, but we did not enforce unanimity and stressed that jurors should vote according to their genuine convictions. Within individual rounds, we asked participants to self-elaborate by writing out the reasoning for their decision. This elaboration controlled for the influence of simply stating one's thoughts aloud and isolated the effect of social influence. After running several dozen pilot studies and inspecting the results, we determined that most of the online deliberations concluded within seven minutes. Thus, all rounds were seven minutes long.

We calculated consistency from a final in-round vote that participants submitted at the end of seven minutes' deliberation or self-elaboration. For individuals, a decision was consistent if their final vote for the first case matched the vote of its correlated pair. For groups, the verdict supported by the majority of group members was considered to be the final group decision. Groups were thus consistent if the final verdict of the first case matched the verdict for its correlated pair.

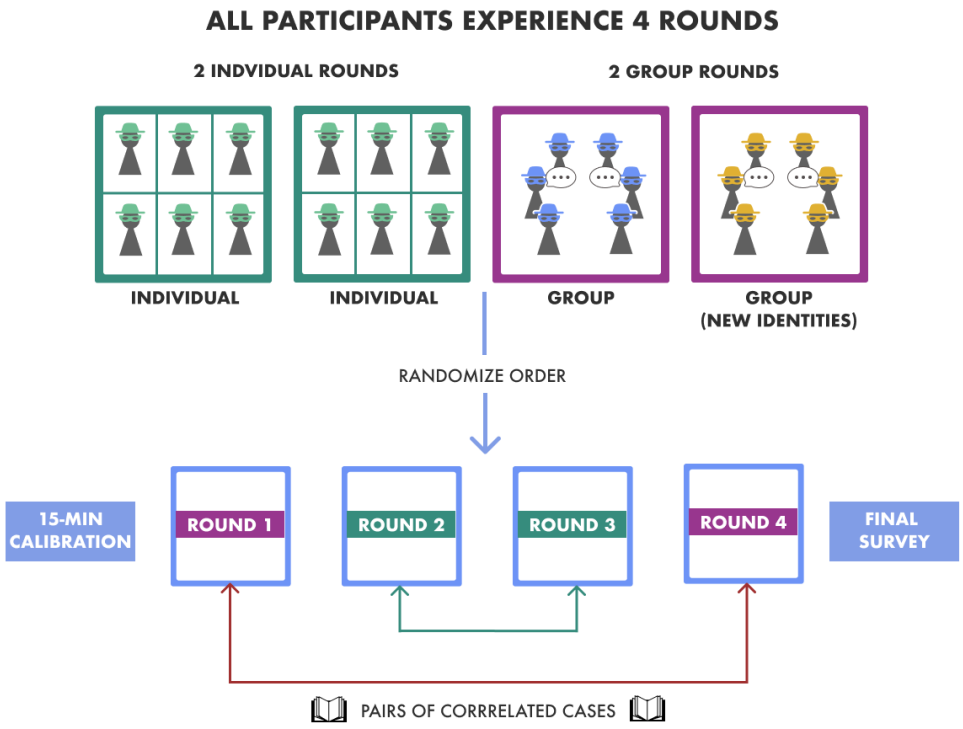

Fig. 1. An experiment consists of four rounds: two individual tasks and two group tasks. Participants remain the same throughout all 4 rounds. In the second group round, the changed mask color indicates a new pseudonym identity. The ordering of these individual and group rounds are randomized: in the figure, the first and fourth rounds are shown as group rounds, while the second and third are individual. Each pair of individual rounds and pair of group rounds adjudicates two cases that are known to have correlated outcomes. Individuals and groups are randomly assigned to adjudicate one of two possible case pairs in a given experiment. We measure consistency by comparing the rates at which individuals and groups give the same judgment for the two paired cases.

In addition to the in-round vote, participants filled out pre-surveys and post-surveys for each of the four rounds. The pre-survey created a baseline for understanding subjects' initial predispositions, and the post-survey measured whether subjects privately disagreed with their in-round vote.

3.2 Parallel Teams Infrastructure

A key requirement of our experiment is that the online juries do that realize that they are working with the same participants twice; we achieve this using a *parallel worlds* technique of *two-way pseudonym masking*, which has been established in prior work [49, 50].

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

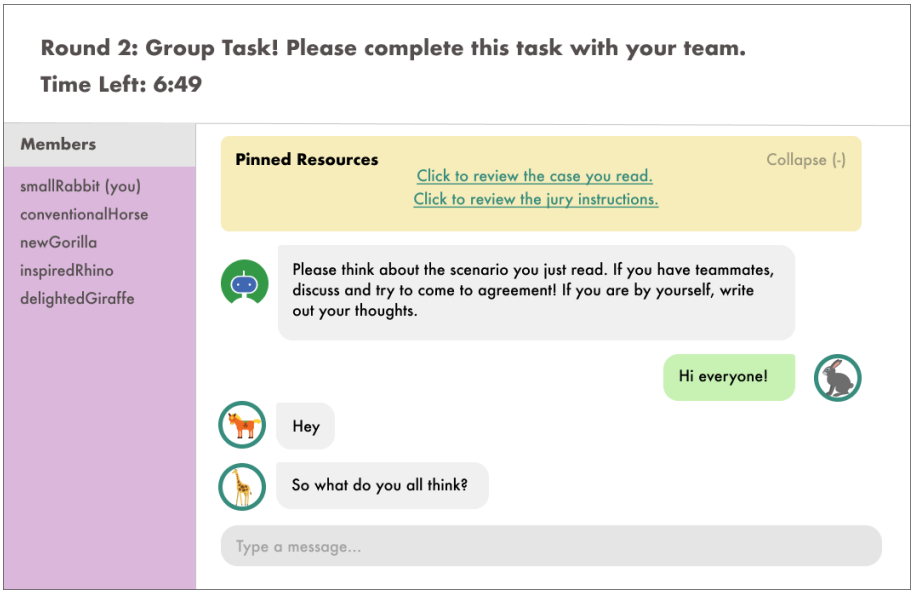

Fig. 2. Participants were shown the above user interface in group rounds. The "Members" sidebar showed the pseudonyms of a participant's collaborators. By changing the names displayed in this sidebar, we were able to create an illusion that users were working with new collaborators when, in fact, collaborators remained the same. The chatroom sidebar appeared green when users worked alone and purple when users worked in a group. Resources and polls (where users submitted the final verdict) appeared in a persistent banner at the top of the screen.

'inspiredDolphin' in all four rounds, and merely believes that she is working with new team members. Figure 2 illustrates the pseudonymous user interface.

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. Prior work [49, 50] using this method has 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.

3.3 Participants

Drawing on previous work [49, 50], participants were sourced from Amazon Mechanical Turk. We selected only workers located in the United States to ensure a shared language and cultural background. We also selected only individuals who had completed at least 100 tasks, which allowed us to avoid short-term Mechanical Turk users and increase the quality of participants. We recruited nine individuals for each jury. This procedure slightly over-recruited in order to account for the perils of online experimentation — nine members insulated against inevitable Internet connectivity issues and other sources of potential drop-off. Subjects participated in four rounds of deliberation. We only analyzed data from trials in which at least five members were active throughout the deliberation. Our 5–9 member juries draw upon Facebook's five-member moderation panels [2], as well as six- to eight-member civil juries [22].

Accounting for time to read prompts and answer surveys, the task lasted 78 minutes in total. Workers were paid via bonus at a rate of $15/hour, per standards of fair payment [51].

3.3.1 Exclusion Criteria. 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. We build upon prior work that uses the pseudonym masking technique [49, 50], implementing 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 select 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.

Samples in which the number of participants who successfully identified their teammate was greater than 2 standard deviations (σ = 1.102) above the mean (1.029) were excluded from the study for potentially detecting the manipulation. Additionally, we manually inspected chat logs and removed samples where participants explicitly recognized each other. These two exclusion criteria were included in our experimental pre-registration.

Post-registration, we noticed a few participants typing spam content into the chats. Since our results depend on high-quality deliberation, we used a combination of visual inspection of chat logs and algorithmic filtering to identify individuals who gave especially terse, irrelevant, or spam remarks. Rounds that contained two or more such participants were also excluded from analysis. We identified and removed six such rounds.

3.4 Case Materials

3.4.1 Selection of case content. Prior work has conducted studies in which subjects were exposed to various versions of the same case, with some facts and details altered [17]. Diamond et al.'s study used two cases, both of which described a civil dispute between an employer and employee over a workplace lung injury. Details between the cases, such as the number of cigarettes smoked by the employee, were altered between the case versions, and both cases were balanced (approximately half of jurors leaned towards voting for the defendant, and half leaned towards the plaintiff). However, most of the cases in such experiments take an hour or longer, which was infeasible with our experimental design where we required four cases per trial.

To address this issue, we drew on a popular online community that posted and deliberated who is at fault in complex interpersonal conflicts. We scraped 23,055 posts from Reddit's 'Am I the Asshole?' (AITA) subreddit, an online discussion board where members describe a controversial personal situation, and the community then adjudicates who is at fault. AITA posts are thematically similar to content adjudicated by other online juries, which often include non-legal claims, personal disputes, and the airing of grievances [31]. Additionally, AITA posts are brief enough for use online, and they posit largely binary outcomes ("yes, the person is at fault", versus "no, they are not"). For the purposes of our study, we eliminated the option to assign fault to both parties or neither.

We used two criteria to select cases for the study:

(1) Balance: similar to Diamond et al., we selected cases in which the percentage of votes "for" and "against" each party in the dispute were relatively even. Drawing on the "messy middle model" [40], we assumed that imbalanced cases would appear to have a clear answer, and therefore all respondents would be likely to answer consistently regardless of the individual or group condition.

(2) Consistency: since our research question requires measuring consistency, we selected cases with a strong baseline of individual consistency in order to create an effective contrast with group consistency. Thus, for pairs of cases, most respondents should decide the dispute in favor of the same party.

In order to meet the balance criterion, we filtered the scraped posts by the number of Reddit votes they received, finding 56 cases with similar numbers of votes on each side. After a pretest on Amazon Mechanical Turk with 136 participants, we isolated the eight most balanced cases.

3.4.2 Baseline consistency for selected cases. To create case pairs, we rewrote five of the eight balanced cases from a different perspective, replacing all named entities. This processes created the 'opposite side' of the dispute. A wife's complaint about her husband, for instance, would be rewritten as a husband's complaint about his wife. We defined a consistent decision as one that rules in favor of the same party twice: if someone answers "yes, the wife is at fault" for the original dispute, answering "no, the husband is not at fault" for the rewritten version is considered consistent.

On Amazon Mechanical Turk, we collected $N = 128$ judgments on the five case pairs, and only two pairs of cases met both criteria of balance and consistency. The final statistics for the pairs are presented in the appendix.

Additionally, we ensured that viewing the rewritten cases did not prime individuals' judgments 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 was 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.

3.5 Calculation of Consistency

Our core dependent variable is consistency, but 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" [14] to "consistency with one's predispositions" [37] to "consistency with predicted results" [10]. Our novel approach of comparing groups to themselves therefore required its own standard of measurement.

We use a total of three comparisons to determine consistency, with one primary measure and two auxiliary comparisons. All of these analyses were pre-registered, and address different aspects of our research question:

(1) Comparing groups to individuals (the primary measure).
(2) Comparing groups to nominal groups (individuals' votes aggregated into a group vote, as if they were a pseudo-jury).
(3) Comparing individuals who worked alone to individuals who worked in groups.

3.5.1 Comparing groups to individuals. This metric is the most straightforward. Group consistency is computed via the *aggregate consistency* of all members in a team, in which the group's decision is calculated via simple majority vote. If the outcome is identical for both group decisions, the group is "consistent." Meanwhile, individual consistency is calculated at the *participant level*: that is, an individual is consistent if their two in-round votes match. The overall rate consistency for groups can then be directly compared to the overall rate for individuals. This core metric is beneficial because it offers a direct way to answer our research question. However, this metric has three drawbacks, which our auxiliary metrics address:

(1) This metric compares two samples of different sizes, since there are far fewer groups than individuals.

(2) A flat consistency rate is blind to changes in voting patterns between group and individual conditions. If two people in a group are inconsistent, changing their votes from "for" to "against," and another two are inconsistent in the other direction (changing their votes from "against" to "for," these changes would make no difference to overall group consistency, but they should be considered.
(3) This metric compares two different ways of calculating consistency: individual consistency is calculated on a vote-by-vote basis, but group consistency is based on majority rule. It is possible that the mechanism of majority rule inflates consistency by smoothing over individual differences.

3.5.2 Comparing groups to nominal groups. This first auxiliary metric addresses limitations (1) and (3). Drawing on previous literature [20, 35], we treat individuals as "nominal" groups for this analysis. A nominal group simply aggregates individual votes into a non-deliberating group decision, in which the outcome is also calculated by majority vote. Thus, six people who worked alone and independently voted for Party A would be aggregated as a nominal group decision for Party A. This comparison enables us to directly isolate the effect of deliberation: other than the deliberation aspect, groups and nominal groups are identical.

3.5.3 Comparing individuals to group members. The second auxiliary metric addresses limitations (1), (2), and (3). In this analysis, we consider each member of a group to be a separate individual, calculating participant-level consistency for each member. We then conduct a paired analysis between individuals (who worked alone) and group members. This comparison allows us to isolate individual behavior: that is, *are the people who vote in groups or the people who work alone more consistent in making decisions?*

3.5.4 Accounting for Ties. Since group size ranged from 5–9, a simple majority vote for groups occasionally yielded a tie. Ties were logged as a third, special case, and groups were considered consistent if they tied twice, and inconsistent otherwise. Additionally, we report statistics for the data with all ties removed.

4 RESULTS

In our study, $N = 1403$ active participants completed 259 online jury trials, conforming to pre-registered values. We collected data between April and July 2020. Recruited participants skewed female (57.2%) and were on average 36.5 years old ($\sigma = 11.86$). 73.2% of participants self-identified as White, 13.6% as Black, and 8.9% as Asian. On a 7-point scale ranging from "Extremely Conservative" (assigned a score of -3) to "Extremely Liberal" (assigned a score of +3), participants were, on average, between "Moderate" and "Slightly Liberal" (0.46). After conducting our preregistered data exclusion filters, 178 trials remained in the analysis, with $N = 1121$ total subjects.

4.1 Comparing Consistency (H1)

4.1.1 Primary metric: Comparing groups to individuals. First, per pre-registration, we conducted an analysis of our overall consistency. Our results indicate that groups were 62.92% consistent, and individuals were 63.88% consistent (Figure 3). The 95% confidence interval for individual consistency is entirely contained within that of groups, and a two-sided two-proportion Z test yielded $p = 0.808$. Thus, the data do not support H1, as group and individual consistency differences are too similar to be discernible.

4.1.2 Comparing groups to nominal groups. Aggregate consistency, however, does not tell a full story. Next, per pre-registration, we conducted a paired analysis that compared group consistency with the consistency of the same individuals when they were deciding alone (nominal groups). This analysis tests whether group consistency can be

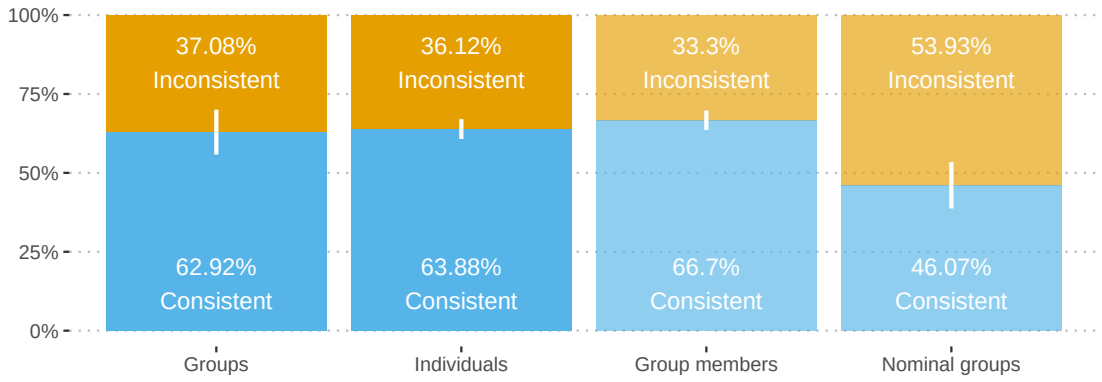

Fig. 3. Groups and individuals are equivalently consistent in their decisions, contrary to our prediction. Nominal groups — effectively, individuals voting without deliberation — are inconsistent and indistinguishable from random chance. The chart shows the proportion of repeated decisions that were consistent and inconsistent based on in-round voting, with 95% CI at the boundary. Darker colors show the primary metrics of group and individual consistency. Lighter colors show auxiliary metrics: the consistency proportion of the votes of individual group members, and of the individuals aggregated into nominal groups.

Table 1. Groups and Individuals are similarly consistent. Group members are more consistent than individuals, while nominal groups are inconsistent.

	Groups	Individuals	Group Members	Nominal Groups
Mean Consistency	62.92%	63.88%	66.70%	46.07%
95% Confidence Interval	(0.558, 0.700)	(0.607. 0.670)	(0.636. 0.698)	(0.387. 0.534)

attributed to the unique factor of deliberation, or whether aggregating independent opinions could achieve the same effect. In other words, this test draws a comparison between a deliberative democracy (such as a jury or task force) and blind voting. H1 would predict that blind voting is more consistent than group deliberation.

The results tell a dramatically different story. We find a paired mean difference of −0.169 (95%CI: −0.281, −0.073), indicating that nominal groups (blind voting) are much *less* consistent than deliberation. Both a two-tailed two-proportion Z test and McNemar's Chi Squared ($X^2 = 9.557$) yield significant results for the comparison between groups and nominal groups ($p < 0.01$). The data suggest that, even though individuals are very self-consistent with judgments that they make alone, making decisions by aggregating independent votes is about as consistent as a coin flip.

This outcome is likely linked to the balanced cases we selected for the task; since approximately half of pre-tested individuals leaned toward each side of the cases, the votes for nominal groups were incredibly close. In non-tied nominal groups, the winning outcome had a mean of 62.44% of the votes (SD = 17.07) — thus, having just one person switch their vote would have altered the group's overall outcome, causing nominal groups' consistency to appear random.

In contrast to nominal groups, the process of deliberation drove more consensus, and groups tended to win by a larger majority (the mean was 74.02% of the votes, with SD = 22.57). Groups were consequently able to decide these contentious cases more consistently by preventing a few individuals from diminishing the collective decision.

4.1.3 Comparing individuals to group members. Finally, per pre-registration, we focused on the individual level, and conducted a paired analysis comparing the consistency of each group member with the consistency of the same person while working alone. This participant-level analysis tests whether deliberating in a group causes individuals to become more erratic voters — that is, H1 would predict that, in the group condition, allowing social influence to easily sway one's vote could lead the person to be less self-consistent across decisions. Again, the results support the opposite conclusion. A McNemar's Chi-Squared test yields significant results ($X^2 = 4.78$; $p < 0.05$), suggesting that members of groups are significantly more likely to decide the case consistently when deliberating than when simply self-elaborating and voting twice. Contrary to H1, which predicted that members of groups could blindly follow others' opinions at the cost of adhering to consistent opinions, the data suggest that, in fact, members of groups are more self-consistent with their opinions than they are when deciding alone.

We therefore find that H1 is not supported across any of the three metrics of consistency.

4.1.4 Subjects Underestimated the Consistency of Groups. In a post-study survey, we asked our participants to predict group and individual consistency. Of all participants surveyed, 70.47% believed that individuals would be consistent (based on a binary question), while just 54.17% believed that groups would be consistent. The stark contrast between confidence in individual consistency and group consistency stood out against the closeness of the empirical group and individual consistency that we found (62.92% and 63.88%). In other words, even as juries turned out quite consistent, the jurors did not trust the consistency of the very juries they had just participated in.

4.2 Individuals' Divergence from Predispositions (H2)

For each case deliberated, group members did tend to more frequently change their minds from their initial predispositions. Compared to the opinion expressed in pre-deliberation surveys, more than a quarter (27.48%, 308 participants) of group members expressed an opposing opinion by the end of deliberation. By contrast, only 8.21% of individuals (92 participants) changed their own minds by self-elaboration. A two-tailed, two-proportion Z test is strongly significant ($p < 10^{-32}$). Thus, H2 is supported: group members are more likely to diverge from their initial predispositions in order to align with the majority viewpoint.

4.3 Groups' Conformity to the Pre-Existing Majority (H3)

The data show that a smaller proportion (72.75%) of deliberating groups voted with the pre-existing majority compared to nominal groups (76.12%). However, a two-sided, two-proportion Z test misses significance ($p = 0.303$). Thus, H3 is not supported.

When the majority does win for groups, it wins more decisively. Groups that vote in line with the pre-existing majority view tend to expand the size of the majority faction, pulling more people in line with the majority view. Among groups, this expansion occurred in 41.69% of cases, compared to just 7.38% for nominal groups. The final size of the majority faction is also larger (the winning outcome had 74.02% of votes in deliberative groups, compared to 62.44% in nominal groups), but had a greater spread (SD = 22.57, compared to 17.07). This increased variance reflects the unpredictability of deliberations — some social forces within deliberation push towards increased consensus with the existing majority, while others push towards generating new perspectives.

Table 2. Our primary results with all ties removed are not significantly different. Note that individual and group member results are unchanged, since ties only apply at the group level.

	Groups	Individuals	Group Members	Nominal Groups
Mean Consistency	62.71%	63.88%	66.70%	45.14%
95% Confidence Interval	(0.556, 0.698)	(0.607. 0.670)	(0.636. 0.698)	(0.378, 0.525)

4.4 Results with Ties Removed

In our main analysis, we accounted for ties by allowing tied juries to be consistent if the same jury tied twice. For completeness, we also present our primary results with all ties removed in Table 2, demonstrating that ties did not play a significant role in our results. None these quantities are outside of the confidence interval of their analogous measure without ties removed.

4.5 Manipulation Check

Finally, we confirm that the teams in our analysis did not recognize each other during the repeated deliberations.

Within each team, the accuracy of answering the manipulation check question ($\mu = 0.164, \sigma = 0.149$) was, in fact, lower than chance ($\mu = 0.198, \sigma = 0.041$), and not significantly different from chance ($p = 0.532$). Thus, because participants were unable to do better than random guessing in identifying former partners, we consider the use of two-way pseudonym masking to have successfully "erased history" and enabled juries to reconvene independently.

4.5.1 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"

Several 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, many 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. Furthermore, a keyword search found that 43% of responses contained keywords such as *guess, random, unsure, no idea,* and *don't know*. Of course, wording of natural-language responses varied widely, so this value provides a very rough lower bound for the true number of random guessers.

5 DISCUSSION

Contrary to H1, groups and individuals in our experiment were equally consistent. Only one of our hypotheses (H2) was supported.

5.1 Deliberating Groups are Much More Consistent than Non-Deliberating Groups

Our core finding is that, by and large, groups stick to a conclusion and rule the same way together. The upshot is that society's trust in groups is well founded — but only when they deliberate. Whereas deliberative decisions led to consistent rulings across similar cases, mere voting (aggregating the decisions that individuals made alone) led to randomness.

One explanation for this outcome may be that individual votes tended to be decided with much narrower margins. If the first vote was won by just one ballot, and at least one person decides the second scenario inconsistently, then the two decisions become inconsistent, especially with a majority-rule method of calculating results. This phenomenon is reminiscent of the way that modern-day elections are often decided by just a few votes, with candidates scrambling for the attention of a small sector of swing voters. It also reflects the Founders Fathers' worry of the "tyranny of the majority": indeed, democracy (by pure voting) is capricious.

This result strengthens prior work, which found that users on digital platforms prefer a deliberative jury over a blind voting jury based on the attributes of *legitimacy, trust, equality, fairness*, and *care* [18]. That blind voting is less consistent than deliberation may have contributed to users' perceptions that voting is less legitimate and less trustworthy. Indeed, with blind vote judgments consistent only half the time, they are not just *optically* less legitimate; they *are* less legitimate.

5.2 Individuals are More Self-Consistent After Participating in Deliberation

We also found that individuals were more consistent as members of groups than when deciding alone. While we were correct that group members tend to change their minds more often (H2), the process of deliberation appeared to catalyze a learning process for many of our participants. This process resulted in strong participant-level consistency in the group condition, for, as members learned, they may have developed stronger intuitions or insights about the issues at hand. Quotes from the post-study survey frequently highlighted this aspect of the experience:

> "If you're in a group you have people to bounce ideas off of. It makes it easier to see both sides of it."
> "If you are by yourself you only have your thoughts and opinions. Sometimes hearing another person's point of view may sway your original opinion."

5.3 Groups Tend to Affirm the Pre-Existing Majority

Following deliberation, group votes are won by a larger majority, leaning harder in the pre-existing direction and reflecting some attainment of consensus. The strong presence of majority rule in our data supports prior findings that groups coalesce around the perceived majority opinion [7, 16, 21, 22, 33, 47]. Since group decisions are also heavily influenced by social cascade effects [43], in which members are loath to contradict the first opinion that emerges [45], this result is not surprising. Contrary to our hypothesis (H1), in which we noted that the arbitrariness of who speaks first and the whims of social dynamics could drive groups to inconsistency, our results indicate that these phenomena actually drive group decisions to be more consistent overall. However, one catch may be that they also cause groups to be too quick to take the path of least resistance. As some participants noted:

> "Groups will tend to choose the easiest answer that the first person introduces. Other ideas will not emerge as often, which makes them more predictable. Individuals are given more room[.]"
> "Once a group has reached a choice, it has already been hashed out by multiple people. So, they tend to use that group idea pretty quickly thereafter."

Thus, paradoxically, groups enable people to learn more about both sides of the issue, but they also lead group members to be more selective about what they learn, and unwittingly settle for overly simplistic answers. One participant captured this tension in their reflection:

> "As an individual, respondents are not being swayed by outside opinions or considerations. I know my thoughts changed in the rounds where we were chatting in a group. [However ,]...people also fall victim to hive mind mentality."

5.4 Participants' Own Predictions

While we were surprised to find that individuals and groups are equally consistent, we also discovered that our participants had also underestimated the consistency of group deliberation, while overestimating the consistency of individuals. In a post-survey, 70.5% of participants believed that an individual encountering the same situation twice will "answer the same way," whereas only 54.2% believed that groups would answer the same way. The very individuals who participated in the online juries did not seem to realize that they had themselves made consistent decisions as a group. This 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 [29].

These perceptions have concrete implications for the use of group decision-making in real-world settings. 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 [13]. As a result, attempts to introduce deliberation into governance procedures, both online and offline, must grapple with the need to justify the decision-making process to skeptics who see group decisions as fundamentally inconsistent.

5.5 Design Implications

Online governance continues to shift toward more democratic practices, from Facebook's "Supreme Court" [2] to online jury platforms [8, 31]. Meanwhile, juries continue to play a prominent role in the offline world too, deciding 154,000 cases each year in the United States of America [11]. This study confirms that the use of these democratic decision-making bodies results in consistent decisions. In particular, our work has clear implications for online platforms: proceed with confidence. Groups will not hand out capricious decisions, and they can be trusted to deliver thoughtful, legitimate decisions.

Our results also show that platforms should specifically preference deliberation over blind voting, even if blind voting appears to be a simpler, and therefore more attractive, solution. Polling without deliberation leaves the final decision to just a few swing voters, making the outcome volatile rather than considered. In cases where deliberation is too time-consuming or costly to implement, having a solo moderator or judge is preferable to mere voting. Individuals reliably make repeatable decisions, so trust in solo moderators is not misplaced.

Finally, platforms have an ethical obligation to select the composition of these deciding groups carefully. Our results are conditioned on the membership of the jury; since group decisions tend to strengthen the pre-existing majority view,

it is crucial that the group reflects the constituent population of the platform as a whole. Otherwise, the design risks alienating those whose voices were never represented in the jury — or who were so marginalized in the discussion that they were unable to state their case. As is the adage in technical fields: garbage in, garbage out; when the majority of a group holds an inaccurate belief, deliberation decreases accuracy [9]. Our work further suggests that, in these cases, the results of deliberation would be *consistently inaccurate*. Deciding with an unrepesentative jury may therefore be worse than no jury at all.

5.6 Limitations and Future Work

A limitation of our work is that our deliberations were brief (seven minutes long). In some cases, time constraints may have artificially cut off deliberation or pushed 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. The time limit also makes the study less ecologically valid for offline juries, since trials typically take place over the course of several days, with jurors spending hours at a time in a shared deliberation room.

This work may also be less generalizable to offline juries because the richness of signals — both within the deliberation room and for each case — are attenuated on an online platform. Offline jurors can raise their hands, write on slips of paper, and whisper to the person next to them; lawyers can accompany their cases with multimedia exhibits. These elements were stripped down in our text-based platform.

Another limitation lies in the cases used for the study. We asked participants to decided cases sourced from Reddit, which, while representative of the types of disputes frequently resolved online, are not representative of offline civil or criminal jury cases.

We also made the assumption that only balanced cases would be effective for testing consistency, since imbalanced cases would have an "obvious" answer that would make most subjects consistent. However, this assumption severely limited the diversity in our case selection. Of the tens of thousands of threads that we scraped from Reddit, only a small handful were balanced, and, ultimately, only two pairs were both balanced and consistent. Future work should replicate this study with a wider range of case pairs, including cases with content representative of the content moderation scenarios in which online juries may be called serve.

Replication with a more diverse array of cases will also be useful in noting how group and individual behavior patterns differ depending on the content of the cases. In our study, individual and group consistency were nearly identical for Case 1, with a Chi-Squared test returning $X^2 = 0.019, p = 0.8912$. In Case 2, however, groups were actually substantially more consistent than individuals, with a Chi-Squared test returning $X^2 = 3.886, p = 0.049$. These differences indicate that the case content is a factor underlying group consistency that should be explored further.

Additionally, our subjects knew that they were participating in a study, and may have felt that their decisions have no real-world consequences. This issue is, in fact, a known limitation of mock jury studies, with previous research finding that mock juries are sometimes harsher and sometimes more lenient than true juries [28].

We also recognize that, while our study resets social dynamics, we do not erase the reasoning that participants used to decide each case. Thus, cases could be decided consistently because the person had already thought through a similar case. We note, however, that the overall consistency rates are not extraordinarily high, so this explanation appears unlikely.

Finally, this study reports the quantitative results of jury deliberation, primarily based on the concrete artifacts of deliberation — verdicts and survey data. Future work should inductively explore the qualitative data within the

conversation transcripts, which may reveal insights about the types of conversations and deliberation styles that lead to greater consistency.

6 CONCLUSION

We conducted a field experiment with 259 online juries to test whether groups or individuals make more consistent decisions. Our study found that groups and individuals are equally consistent. Indeed, groups are a much stabler form of decision making than voting, which results in nearly random judgments. Groups decisions are, in the words of one participant, "more rational and have less volatility than individual judgments. Groups serve to moderate each other, quelling extremes and representing a common consensus." Although group consensus leads consistent judgments overall, some group decisions can be overly simplistic — leading members to quickly submit to the majority view with less thought. Ultimately, this work expands our understanding of group decision-making processes, enriching our theories about and tempering our expectations for the groups to which our society entrusts so much.

REFERENCES

[1] 2020. My Team Will Go On: Differentiating High and Low Viability Teams through Team Interaction. *Proceedings of the ACM on Human-Computer Interaction* 5 (2020).

[2] 2020. *Oversight Bord Bylaws.* Retrieved September 3, 2020 from https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf

[3] Tanja Aitamurto and Helene E Landemore. 2015. Five design principles for crowdsourced policymaking: Assessing the case of crowdsourced off-road traffic law in Finland. *Journal of Social Media for Organizations* 2, 1 (2015), 1–19.

[4] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[5] Albert W Alschuler and Andrew G Deiss. 1994. A brief history of the criminal jury in the United States. *The University of Chicago Law Review* 61, 3 (1994), 867–928.

[6] Julia Angwin and Hannes Grassegger. 2017. *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children.* Retrieved September 3, 2020 from https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

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

[8] Federico Ast. 2017. *Kleros, a Protocol for a Decentralized Justice System.* Retrieved October 31, 2019 from https://medium.com/kleros/kleros-a-decentralized-justice-protocol-for-the-internet-38d596a6300d

[9] Joshua Becker, Abdullah Almaatouq, and Emőke-Ágnes Horvát. 2020. Network Structures of Collective Intelligence: The Contingent Benefits of Group Discussion. *Working Paper* (2020).

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

[11] Keith Burghardt, William Rand, and Michelle Girvan. 2019. Inferring models of opinion dynamics from aggregated jury data. *PLOS ONE* 14, 7 (2019). https://doi.org/10.1371/journal.pone.0218312

[12] Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. *arXiv preprint arXiv:2005.04246* (2020).

[13] Kevin M Clermont and Theodore Eisenberg. 1991. Trial by jury or judge: Transcending empiricism. *Cornell L. Rev.* 77 (1991), 1124.

[14] Lee J Curley, Jennifer Murray, Rory MacLean, and Phyllis Laybourn. 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 (2017), 211–219.

[15] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.

[16] Dennis J Devine, Laura D Clayton, Benjamin B Dunford, Rasmy Seying, and Jennifer Pryce. 2001. Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, public policy, and law* 7, 3 (2001), 622.

[17] Shari Seidman Diamond, Michael J Saks, and Stephan Landsman. 1998. Juror judgments about liability and damages: Sources of variability and ways to increase consistency. *DePaul L. Rev.* 48 (1998), 301.

[18] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Vol. 3.

[19] Jenny Fan and Amy X Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[20] Joshua T Gyory, Jonathan Cagan, and Kenneth Kotovsky. 2019. Are you better off alone? Mitigating the underperformance of engineering teams during conceptual design through adaptive process management. *Research in Engineering Design* 30, 1 (2019), 85–102.

[21] J Richard Hackman and Nancy Katz. 2010. Group behavior and performance. (2010).

[22] Reid Hastie, Steven Penrod, and Nancy Pennington. 1983. *Inside the Jury.* Harvard University Press.

[23] Kokil Jaidka, Alvin Zhou, Yphtach Lelkes, Jana Egelhofer, and Sophie Lecheler. 2020. The "majority illusion" in social networks. *Under Review* (2020).

[24] Malte F Jung, Nikolas Martelaro, and Pamela J Hinds. 2015. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 229–236.

[25] Yubo Kou and Bonnie Nardi. 2013. Regulating anti-social behavior on the Internet: The example of League of Legends. (2013).

[26] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1035–1048.

[27] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The" majority illusion" in social networks. *PloS one* 11, 2 (2016), e0147617.

[28] Robert J MacCoun. 1987. Getting inside the black box: Toward a better understanding of civil jury behavior. (1987).

[29] Robert J MacCoun and Tom R Tyler. 1988. The basis of citizen's perceptions of the criminal jury. *Law and Human Behavior* 12, 3 (1988), 333–352.

[30] Nancy S. Marder. 2005. *The Jury Process.* Foundation Press.

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

[32] Serge Moscovici and Claude Faucheux. 1972. Social influence, conformity bias, and the study of active minorities. In *Advances in experimental social psychology*. Vol. 6. Elsevier, 149–202.

[33] Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication* 24, 2 (1974), 43–51.

[34] Beth Simone Noveck. 2004. *Democracy Online.* Psychology Press, Chapter Unchat: Democratic Solution for a Wired World, 21–46. https://doi.org/10.1007/3-540-09237-4

[35] Vincent E Owhoso, William F Messier, Jr, and John G Lynch, Jr. 2002. Error detection by industry-specialized teams during sequential audit review. *Journal of accounting research* 40, 3 (2002), 883–900.

[36] John W Payne, John William Payne, James R Bettman, and Eric J Johnson. 1993. *The adaptive decision maker.* Cambridge university press.

[37] Robert T Roper. 1980. Jury Size and Verdict Consistency:" A Line Has to be Drawn Somewhere"? *Law and Society Review* (1980), 977–995.

[38] Niloufar Salehi and Michael S Bernstein. 2018. Hive: Collective design through network rotation. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–26.

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

[40] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and analysis of the nips 2016 review process. *The Journal of Machine Learning Research* 19, 1 (2018), 1913–1946.

[41] Jae-Young Son, Apoorva Bhandari, and Oriel FeldmanHall. 2019. Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific reports* 9, 1 (2019), 1–15.

[42] Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology* 48, 6 (1985), 1467.

[43] Cass R Sunstein. 1999. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper* 91 (1999).

[44] Cass R Sunstein. 2000. Deliberative Trouble? Why Groups Go to Extremes. *The Yale Law Journal* 100, 71 (2000), 71–119.

[45] Cass R Sunstein and Reid Hastie. 2014. Making dumb groups smarter. *Harvard Business Review* 92, 12 (2014), 90–98.

[46] James Surowiecki. 2005. *The wisdom of crowds.* Anchor.

[47] Linda Thorne, Dawn W. Massey, and Joanne Jones. 2004. "An Investigation of Social Influence: Explaining the Effect of Group Discussion on Consensus in Auditors' Ethical Reasoning. *Business Ethics Quarterly* 14, 3 (2004), 525–551. https://doi.org/10.5840/beq200414321

[48] Nina Totenberg. 2019. Supreme Court Sees 2 Similar Death Penalty Questions Very Differently. (2019). https://www.npr.org/2019/03/30/708238203/supreme-court-sees-2-similar-death-penalty-questions-very-differently

[49] Mark E Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S Bernstein. 2019. Did It Have To End This Way? Understanding the Consistency of Team Fracture. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[50] Mark E Whiting, Irena Gao, Michelle Xing, Junior Diarrassouba N'Godjigui, Tonya Nguyen, and Michael S Bernstein. 2020. Parallel Worlds: Repeated Initializations of the Same Team To Improve Team Viability. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020).

[51] Mark E Whiting, Grant Hugh, and Michael S Bernstein. 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. 197–206.

[52] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[53] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: Building Governance in Online Communities. *arXiv preprint arXiv:2008.04236* (2020).

[54] Sharon Zhou, Melissa Valentine, and Michael S Bernstein. 2018. In search of the dream team: temporally constrained multi-armed bandits for identifying effective team structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

Table 3. Consistency results from a pre-survey with $N = 128$. These final two case pairs were the only two pairs that were both consistent and balanced.

	Case 1	Case 2
Consistency for case pair	68.2%	65.1%
Consistency 95% CI	[0.594, 0.761])	[0.562, 0.733]

Table 4. Balance results from a pre-survey with $N = 128$. These final two case pairs were the only two pairs that were both consistent and balanced.

	Original Case	Reverse Case
Case 1 Balance	50.4%	41.4%
Case 2 Balance	44.5%	55.4%

A CASE PAIR STATISTICS

Table 3 and Table 4 show statistics for consistency and balance for the four cases used for the study.

- Consistency for case pair = % of responses that ruled in favor of the same party
- Balance = % of 'yes' responses for each case

B REDDIT CASE PAIRS

Case 1 and Case 2 are the original cases. We authored the reversed versions. We changed and added personal details so as not to make the cases identical. We also introduced deliberate typos and grammatical errors to make them seem like authentic Reddit cases, rather than contrived examples.

B.1 Case 1

I live in an area in California that is mostly mexican and people of other latin american descent. My wife is Guatemalan and so is her family. Im a white guy and a self taught spanish speaker. Many people in my wife's family don't speak english that well and her mom doesn't speak any other than a few phrases. She can barely order food at a restaurant and can't really do much without a family member with her. Generally i speak with my wife's family in spanish. My wife's mom sometimes calls me when she can't get ahold of other family members asking me to translate when she's doing basic errands or has some sort of issue with someone who can't speak spanish. It's getting pretty annoying having to deal with it so i told my mother in law that she needs to learn how to speak english. I told her I'm sick of dealing with her shit because she can't speak english, she's been here for 8 years she should be able to speak english by now, in comparison i taught myself spanish and portuguese within a year and russian in 2 without even living in the countries that speak them. This honestly shouldn't be that hard. I gave her sources that i've used, duolingo, mango and provided dictionaries. I've even told her about ESL classes she can take.

Many people in my family are upset with what i said. They said what i did was incredibly disrpesectful and that I'm flaunting my white privlege by doing this shit and I'm acting racist. Somehow telling someone you should learn the language of the country you live in is racist. They went on about how I'm a racist trump supporter and whatever. Nobody in my family is on my side on this issue not even my wife and I'm wondering where i went wrong. I was incredibly

polite to my mother in law and i explained that i'd help her along the way but she needs to become independent. I love her and my family they're all decent people but she needs to learn to speak english as do the rest of the family.

EDIT: i am aware that the US does not have an official language, stop pointing that out everyone fucking knows that. I'd also like to rephrase something, when explaining the situation to my MIL i didn't tell her "I'm sick of dealing with your shit" i told her that it's getting frustrating that I'm being used as a translator at random parts throughout the day and i want you to be independent and be able to do what you need to do without one of us as a translator in a really polite and respectful tone

B.2 Case 1 Reverse

I immigrated to the U.S. from El Salvador when I was 15. Spanish is my first language, and my family lives in a mostly Latinx part of California. I'm the oldest child and have always lived with the immigrant mentality that I should be the one to blaze the path and put my family first. This is especially true since my parents speak very limited English, and I have had to help them translate everything from menus to legal documents my whole life.

A few years ago I married a caucasian guy who has learned Spanish in school. We make fun of him for being gringo but he has also been incredibly helpful in bridging cultural gaps.

Currently I'm working full-time while also going back to school for my Master's. So when my mom calls for translation-related stuff, I pass it onto my husband, since he also speaks Spanish and currently has more free time. After a while she's basically just started calling him directly whenever she has a problem.

My husband has gotten increasingly irritated and says I'm TA for expecting him to translate for my parents and put his life on hold whenever my mom calls. But he's married me, so my mom is his family too, and he should be able to step up to the plate and help out. My husband says that my parents should have just learned English on their own, that there's resources like duolingo or rosetta stone. But that's such a ridiculous an privileged assumption — America is a country of immigrants. My parents have been working hard to make it in this country their whole lives, they don't need this shit on top of that and now is the time that we should be taking care of them. I don't see how he can't spend like 10 minutes of his time every once in a while to help out; isn't that the respectful thing to do?

so tl;dr: AITA for making my husband translate stuff for my mom?

B.3 Case 2

My son(25) and his partner(20) are staying at my house for a week. My son's partner is French, but he can speak English, he just has an accent. Since he got there he has said few sentences in English. He mainly talks French.

Yesterday we were all in living room, they started talking, French of course. I stopped them and nicely told them that I find it disrespectful that they're speaking language I don't understand and excluding me from conversation. They both stopped talking, my son's husband went upstairs pretty soon after that. My son told me that I was making big deal out of nothing and making everyone feel awkward.

B.4 Case 2 Reverse

I'm American (25M) and recently married a German guy (23M). We now live part-time in Germany and part-time in the U.S.

We're currently staying at my mother's house for the weekend, and she's been gracious enough to host us and provide us meals. It's been nice seeing family again.

Though my husband speaks English he can be shy about it — he says he hates that he sometimes doesn't know the right words. And for me in particular he just prefers German. So we almost always speak German to each other.

This now poses a dilemma for me becuase we're now at my mother's house. It means I have to choose between including my mother in the conversation or making my husband comfortable. I chose the latter, but this means that over the weekend I've talked mostly in German.

My mom says I'm the asshole for not including her, that it's disrepectful and she wants to be caught up to my life after I been away for so long. I think the language I speak is my own choice, and it was also my decision to make sure my husband feels comfortable here. AITA