Camera, Gabriele; Garratt, Rod; Monnet, Cyril

Working Paper

Truth by consensus: A theoretical and empirical investigation

Discussion Papers, No. 24-04

Provided in Cooperation with:
Department of Economics, University of Bern

Suggested Citation: Camera, Gabriele; Garratt, Rod; Monnet, Cyril (2024) : Truth by consensus: A theoretical and empirical investigation, Discussion Papers, No. 24-04, University of Bern, Department of Economics, Bern

This Version is available at:
https://hdl.handle.net/10419/296596

UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Truth by Consensus:
A Theoretical and Empirical Investigation

Gabriele Camera, Rod Garratt, Cyril Monnet

24-04

April, 2024

DISCUSSION PAPERS

Schanzeneckstrasse 1
CH-3012 Bern, Switzerland
http://www.vwi.unibe.ch

Truth by Consensus: A Theoretical and Empirical Investigation[†]

Gabriele Camera
Chapman University

Rod Garratt
BIS

Cyril Monnet
University of Bern
Study Center Gerzensee

April 25, 2024

Abstract

Truthful reporting about the realization of a publicly observed event cannot be guaranteed by a consensus process. This fact, which we establish theoretically and verify empirically, holds true even if some individuals are compelled to tell the truth, regardless of economic incentives. We document results from an experiment where subjects routinely misreported a commonly known event when they could monetarily gain from it. Relying on majority consensus did not help uncover the truth, especially if complying with the majority granted small personal monetary gains. This highlights the difficulties in relying on shared consensus protocols to agree on specific events, and the importance of institutions with trusted, impartial observers.

Keywords: coordination, experiments, DeFi, digital currency, dishonesty, trust.
JEL codes: C70, C90, E04, E05

[†]We thank for comments and suggestions participants at several conferences and seminars, including the 2024 Santa Barbara Conference on Experimental and Behavioral Economics, 2023 HEC blockchain conference, 2022 Berlin Summer workshop in experimental economics, Bank of Canada, Toulouse School of Economics, and the University of Bonn. We are indebted to Jeff Kirchner for programming the software used in the experiment, and Megan Luetje for laboratory help. G. Camera, has obtained Institutional Review Board (IRB) approval Protocol # 23-62 at Chapman University.

"*To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true; so that he who says of anything that it is, or that it is not, will say either what is true or what is false; but neither what is nor what is not is said to be or not to be.*"
—Aristotle, *Metaphysics*, Book 4, Ch.7

1 Introduction

Objective truth - a factual statement with a definite correspondence to reality - need not be the accepted truth. Economic and social systems "decide" on the truth via potentially complex arrangements. History is ripe with examples where objective truth was suppressed or ignored in favor of more convenient truths for those in power. Regardless of the objective truth and societies perception of it, there is no way to ensure that economic or social decisions are made in a manner that accurately reflects this truth. Some process has to be invoked to record the truth so that it can be used to determine contingent outcomes – economic or not. Historically this process was often controlled by a king or religious leader. In modern systems it could be a news source like Bloomberg, or a group of (academic) experts, which many trust to provide accurate information. The ability of these entities to maintain their role as purveyors of truth will depend on society's perception of their accuracy. It is arguable that the smooth functioning of economic and social systems requires trusted sources of truth. In fact, in this paper we contend that trusted sources are crucial for accepted truth to equal objective truth. The alternative to a trusted authority is a consensus process and, as we argue theoretically and demonstrate empirically, there is no way to ensure the truth will emerge from a consensus process.

In this study we develop a strategic analytical framework, which we use to investigate whether it is possible to uncover the truth about a contentious event from the independent accounts of those who witnessed it. This is not merely an academic question. In fact, it is an especially relevant issue now that technological innovation has the potential to radically transform the financial and economic landscape into a fully decentralized system that no longer relies on layers of trusted (and costly) institutions (Nakamoto (2008)). This raises important questions for global financial security. Consider for example the booming industry based on Decentralized Finance (DeFI) networks, where no golden source of truth is available and instead participants from across the globe must decide themselves what the truth is. Activity in the DeFi world is conducted through smart contracts — self-executing programmable contracts between two or more parties. Smart contracts do not require a vetting authority because their immanence relies on decentralized ledger technologies (Buterin (2014)). However the implementation of many potentially useful smart contract applications depends upon the verification of the realization of some real event (e.g., insurance contracts). This creates a problem. Given their fully decentralized nature, how does a smart contract select what the true state of the world is? More generally, how do fully decentralized systems function when their operation depends on the existence of a single, mutually accepted, record of the truth, but there is no single authority that can provide this record?

The answer presumably depends not only on how likely network participants are to lie, but also on the method used to aggregate the information they provide. Yet, a search of the empirical economics literature reveals few studies on this topic. The closest literature is experimental studies of how often individuals misreport the outcome of some random event when they are

confronted with a non-strategic choice. The main finding is that misreporting occurs but is not the norm: individuals who have *no* economic incentive to be honest lie much less than expected. This "aversion to lying" is ascribed to implicit costs thought to stem from emotional, social or psychological factors (e.g., see the survey in Abeler et al., 2019). This evidence is encouraging because it suggests that taking into account the reports of multiple witnesses should allow the truth to reveal itself in strategic settings such as DeFi networks. For instance, if only 4 out of 10 people usually misreport what they see, then the truth will surface if we collect and give equal weight to all reports and adopt a majority consensus rule.

Our strategic analytical framework advances existing experimental designs in three ways. In the typical design an honesty norm *cannot* be incentive compatible (it lowers earnings), lying is an *individual* act (a lone witness provides the only account of the event), and an *authoritarian* mechanism establishes what "the truth" is (by the lone witness, or by the computer). By contrast, our design is based on an analytical framework where an honesty norm *can* be incentive compatible, lying is a *social* act (many witnesses provide accounts of the event), and a *democratic* mechanism establishes what "the truth" is. We call this analytical framework the "consensus game," and we use it to put forward an impossibility result on the emergence of truthful equilibria when agents who are self-interested and rationally respond to economic incentives cannot rely on a trusted single authority to provide a record of the true state. This analysis allows us to derive hypotheses that we test with data collected in the laboratory.

To give an overview, our setup considers a situation where multiple individuals seek to enter into agreements based on the outcome of a real world event. However, there is no trusted party (i.e., contractible source) that can be

used to determine payoffs so payoffs must be based on some form of collective agreement on the true state of the world. By appealing to three basic properties, anonymity, neutrality, and monotonicity, we can restrict attention to a majority consensus protocol. In this environment, players report the "truth" that is most beneficial for them. Incentives to reward consensus do not necessarily make things better. Rather, they lead to a situation that is akin to a beauty contest à la Keynes (1936) in which players report what they think the majority of others will report. We conclude that the only way that individuals are willing to report the true state is if they are completely indifferent as to what the true state should be. That is, their payoffs cannot depend on their actions or their individual reports – a situation that is tantamount to autarky. This suggests that absent additional motivation (e.g., deontological preferences toward truth-telling) we might not see truth-telling.

The experimental data supports this hypothesis and reveals three key insights. First, relying on multiple and equally-informed witnesses to establish the true state of the world is bound to end up in frequent failures. In the experiment, when the group majority could benefit from misreporting an event, the truth seldom emerged—between 0% and 25% of the times. Second, promising a small compensation for conforming to the majority report makes matters worse. In the experiment it caused a significant increase in dishonest reports and significant increase in failures to establish what truly happened. Similarly, individuals who fear finger pointing may prefer to report dishonestly to conform with the majority report. Third, dishonesty is less than what pure economic considerations would suggest, but unfortunately it is also responsive to the size of prospective monetary gains. So when a lot is at stake even the minority can be easily bribed into distorting the truth. In the experiment minority players made about 60% false reports when they could earn only $1

by going along with the majority's lie, but dishonesty jumped to 80% when they could earn $5 by easily colluding on a lie.

Overall the data are consistent with the notion that humans suffer implicit costs from being, or being seen as, dishonest. Yet, in the experiment these implicit costs were insufficient to overcome the economic incentives favoring dishonesty. The aversion to lying was per se insufficient to ensure that a simple majority consensus mechanism could generate an accurate record of a public event.

Our analysis highlights potential weaknesses from relying on shared consensus protocols in the operation of automated markets on decentralized platforms. It also highlights the importance of institutions with trusted, impartial witnesses. This contributes to two different literatures. Our theoretical analysis contributes to the literature of peer prediction games (PPG). In these settings, there is typically one object that has an unknown characteristic and several agents can exert effort to obtain a signal about this characteristic. These signals can be correlated (i.e, come from the same distribution), but there is typically no known "ground truth." The literature studies peer prediction mechanisms that are used to make sure agents exert effort and that they communicate their signals truthfully. Unlike this literature, we consider a game where there is common knowledge, and players are heterogeneous in that some have a stake in the "truth" and some do not.[1] The empirical analysis in

[1]An aspect that makes both our game and PPG similar is that in both cases agents have to send signals about what they observe. Also neither the peer prediction mechanism nor ours can rely on the ground truth (even if it exists) to discipline agents. There are, however, important differences between PPGs and our game. Our game is one of common knowledge (everybody knows the true state, and everybody knows that everybody knows, etc.), while there is no common knowledge in PPG. This is in part because there may be no objective ground truth, rather the truth may be subjective. In addition, in PPGs, agents do not know the signals of others. Trivially, our agents do not need to exert effort to coordinate on what they commonly and freely observe. Our agents have an ex-ante stake in the "truth," while in PPG the only reward is the one agents obtain by "getting it right" relative to

our paper fills in an important gap in the experimental economics literature, which has largely focused on individual, non-strategic tasks where the only witness to an event has also full authority to establish facts and payoffs. We expand this area of study to situations where there are many witnesses to an event, their choices are strategic, and they all share in the authority to establish facts and payoffs. Our study provides much needed empirical evidence to determine if and how groups can establish the truth about some real world event, which is fundamental for understanding the operation of financial and payments systems with decentralized governance.

2 Related experimental literature

An increasing body of research in experimental economics suggests that dishonest behavior is not the norm (see the meta-study in Abeler et al., 2019). People *do* behave dishonestly when they have an economic incentive from doing so, but *less* than one might imagine in terms of frequency and the size of the lie. Several behavioral factors intervene in the decision to lie, including the fear of possible negative reputational consequences, cultural norms, social identity, and psychological or other implicit costs from knowing to be dishonest.

For instance, Fischbacher et al. (2013) report that about 40% of subjects lie when their payoff is proportional to their report about a private roll of a six-faced die. Abeler et al. (2019), which also use a random number design, report a strong preference for truth-telling—participants forgo on average about 3/4 of potential gains from lying. In a sender-receiver game, Erat and Gneezy (2012) report that the two participants are unwilling to lie even when do-

other agents. Finally, in our game, the reward is idiosyncratic and depends on "types."

ing so benefits both, while Tergiman and Villeval (2023) report that there is widespread dishonesty. In a task where subjects are paid for finding specific numbers in a sequence of matrices, Mazar et al. (2008) uncover a preference for truth-telling when participants are reminded of morality standards as compared to when they are not. Cohn et al. (2014) and Rahwan et al. (2019) rely on a subject pool of banking professionals and a social-identity priming intervention to study whether the banking culture causes dishonesty as measured by the self-reported outcomes of coin tosses. The first study offers evidence suggesting that the banking industry fosters a culture that undermines the honesty norm, but not the second study. Gneezy et al. (2018) also considers how social identity factors affect dishonesty and, in particular, the size of a lie.

Table 1 helps us situate our contribution to the extant experimental literature. There are two important aspects that define the various designs: (i) how many individuals witness and make a report about some event that affects payoffs, and (ii) whether a trusted authority exists, which can establish the truth about the event.

The top right cell captures the majority of experiments, which involve just 1 participant who has full authority to establish a fact and, consequently, her payoff. These designs have four important features. First, they consist of a non-strategic *individual* task in which the subject is the only witness to an event (usually a random number) and provides the only account of it. Second, the witness has *full authority* to establish "the truth" about the event and what her payoff will be. Third, the witness' account is *unverifiable*, so the individual act of lying is undetectable, which removes possible psychological costs from others' negative judgments.[2] Fourth, truth-telling is *not incentive-*

[2]With unverifiable events the experimenter can statistically infer dishonesty *of the subject*

compatible because there are economic incentives for dishonesty (e.g., when reporting a die roll earnings increase in the number reported). In this situation, a rational, self-interested individual should invariably lie. As the data reveal that dishonesty does not always prevail, this suggests that other intervening factors—cultural or behavioral—induce individuals to forego monetary gains, and make honest reports.

Table 1: Contribution to the experimental literature.

	Trusted authority	No trusted authority
Single witness	2 players, 1 informed Computer establishes truth (signaling game, matrix task,...)	1 player Witness establishes truth (die roll, coin toss,...)
Multiple witnesses	3 players, 2 informed Computer establishes truth (cheap-talk game)	N players, all informed Group establishes truth (our consensus game)

Some designs study dishonesty in strategic settings by adding one decision-maker who cannot witness the event; see the top-left cell in Table 1. A common design is based on a sender-receiver game where the witness (the sender) privately observes an event and gives her account of it to an uninformed decision-maker (the receiver). In Gneezy (2005), Erat and Gneezy (2012) and Tergiman and Villeval (2023), the receiver contributes to determine the outcome by selecting whether to trust the report (and going along with it) or not. Yet, payoffs also depend on the *actual* event and truth about it is not established by participants but instead by a "trusted authority." This authority is the computer program, which itself selects the event, and therefore can assign payoffs

pool by identifying possible departures from chance in the aggregate data. See Gneezy et al., 2018 or Mazar et al., 2008 for variants where the experimenter is privy to the event and how this affects behavior.

based on that. Ex-post, in some cases the individual act of lying cannot be detected or punished (e.g., Erat and Gneezy, 2012; Gneezy, 2005), while in others it can (Tergiman and Villeval, 2023). The common thread in these designs is the presence of just one witness and a trusted institution that by assumption can always establish "the truth" (the computer). This holds true also in Gino et al. (2013) where individuals compete for payoffs in a zero-sum game and, in some treatments, can choose to avoid being monitored by a central authority that by design always knows the truth (the computer).

In the above experiments, dishonesty affects at most one counterpart, and is an *individual* act that cannot be detected (one exception being Tergiman and Villeval, 2023 where dishonesty can be endogenously detected by adopting a pre-existing institution). However, in field settings dishonesty is primarily a *social* act. To see this, consider that in many field situations there typically are multiple witnesses to an event, and multiple accounts of it. It is also rare for a single individual to have the full authority to determine what a fact is and, based on that, what payoffs should be. In other words, in the field governance is rarely dictatorial: no single authority exerts full control on what "the truth" is or what the distribution of payoffs will be.

The experimental literature has largely shied away from studying these kinds of situations. There are some experiments about cheap-talk models where the number of senders is increased to two (bottom-left cell in Table 1), but even here the group is very small (3 people) and payoffs are determined based on the action of a single, and uninformed, individual in collaboration with a trusted authority (the computer) that can establish the truth about the event (e.g., Lai et al., 2015; Vespa and Wilson, 2016).

What is missing is an understanding of what happens when multiple individuals witness some event, but cannot rely on a trusted authority to establish

what really happened and how to allocate payoffs (bottom-right cell in Table 1. In particular, we are interested in settings where although everyone is equally informed about the event, no-one has sole authority to determine what events actually transpired and what payoffs should be. Here governance is truly egalitarian. What should we expect in this case? Could a "truth norm" prevail in a group where no individual can control the information that gets reported, or payoffs? Our paper aims to fill this gap.

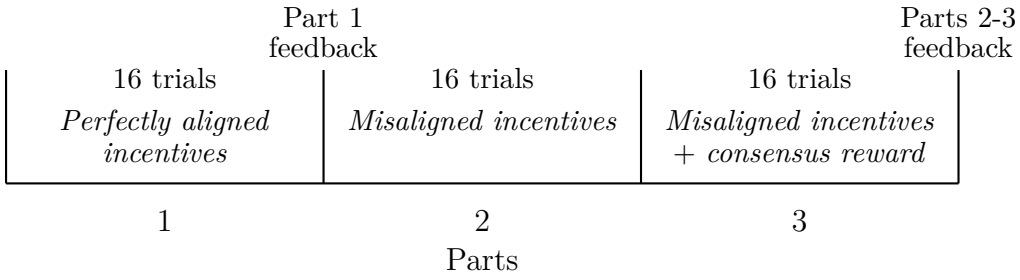
To address these questions we build a strategic analytical setting where a group of participants all witness the same event and give independent accounts about it. There is no asymmetric information and the determination of facts (and, hence, payments) do not depend on the arbitrary pronouncement of one witness or by relying on a trusted institution (e.g., the computer). They entirely depend on majority consensus: each participant's report constitutes an equally weighted piece of evidence about the event, which ultimately determines individual payoffs. We call this the "consensus game."

Three aspects differentiate ours from other design. First, *verifiability*: everyone in the group (as well as the experimenter) observes the true state and can detect individual dishonesty. Second, *incentive compatibility*: in the game truth-telling is a (Bayes) Nash equilibrium. Third, *egalitarian governance*: everyone in the group makes a report and no single individual has the power to control what the "truth" or the payoffs will be. These three aspects combined suggest that truth should more easily emerge than in settings were individuals are asymmetrically informed (e.g., one expert or a committee of experts within a group) or where a single individual has the power to determine payoffs. In other words, truth about a fact stands a better chance to surface when many observe that fact *and* are free to provide an independent account about what they see.

3 Experimental design

Each session involves a group of 15 subjects and is divided into three different parts. Each part consist of 16 repetitions (trials) of a strategic task which we call the *consensus game*. Parts differ according to the structure of economic incentives, as explained below.

Figure 1: Layout of a Session.



3.1 Part 1: perfectly aligned incentives

We start by describing the consensus game. In each trial, players see a binary event — the image of one side of a coin, marked either heads (H) or tails (T). The image is the same for everyone in the group. Each player must then independently make an anonymous report about the image being either H or T. This binary choice is made without consulting others or receiving information about their choices. We say that the report is *truthful* if it corresponds to the image displayed, and is *false* otherwise. Players receive no feedback on other's choices or earnings until all 16 trials have been completed. At that point, they see the realized distribution of choices in the group, in each trial.

Figure 2: Choice task in a trial of Part 1.

In the experiment, the image represented a side of the Kennedy Half Dollar; see Figure 2. Reports are made by clicking on a large letter, either H or T. The instructions (see Appendix B) informed players that they were free to make any report they wished, reports were made in private, and the task would be repeated 16 times after which outcomes would be revealed. Players were also informed that new instructions would be provided at the end of Part 1. Subjects had paper and pencil to keep records, if they wished.

Payoffs depend on majority consensus. In Part 1, there is an economic incentive to avoid making a minority report, similarly to a "beauty contest" game: a player earns \$6 if his report matches the majority report in *all 16 trials*, and otherwise earns \$1; see Table 2.

Table 2: Player's payoffs in Part 1: perfectly aligned incentives

Report is in the	
Majority (in all 16 trials)	Minority (in 1 or more trial)
\$6	\$1

Hence, the report affects the player's payoff, and it can affect the payoffs of others only if the player is pivotal, i.e., if his report flips the majority. In this situation, there are two Nash equilibria, however the coin serves as a powerful coordination device or sunspot (see Cass and Shell, 1983). Subjects could, in principle, coordinate on either the observed face of the coin or the opposite. A natural focal point à la Schelling (1960) is the observed face.

3.2 Part 2: misaligned incentives

The task remains the same but the payoff structure is different. Now, in each trial, every player is assigned a label, either H or T. The distribution of labels in the group is shown before reports are made. Here the choice task is presented slightly differently; see Fig.3. Players were informed that they would receive no feedback on outcomes until the end of the session; they would also be paid for just one trial. At the end of the session, a volunteer would publicly draw a ball from a bingo cage containing 16 balls numbered 1 through 16; the ball would identify the trial selected for payment in Part 2. Players were also informed that new instructions would be provided at the end of Part 2.

Figure 3: Choice task in a trial of Parts 2 and 3.

Payoffs depend on majority consensus. However, in Part 2 there is an economic incentive to report one own's label and not necessarily the coin: a player earns \$6 if his label matches the majority report *in the trial*, and otherwise earns \$1; see Table 3. If the player is pivotal, then his report affects his and others' payoffs, and otherwise it is inconsequential. Unlike in Part 1, the coin is no longer a natural coordination device because *some* players—those whose label differs from the coin—have an economic incentive to lie to maximize the chance that the majority report reflects their label.

Table 3: Player's payoffs in Part 2: misaligned incentives

		Report is in the	
		Majority	Minority
Label = {	Majority Report	\$6	\$6
	Minority Report	\$1	\$1

3.3 Part 3: misaligned incentives and consensus reward

The task is as in Part 2, with just a small variation in the payoff structure. Payoffs still depend on majority consensus. However, in Part 3 there is an economic incentive to avoid making a minority report and to not necessarily report the coin: a player earns \$5 if his label matches the majority report in a trial, and otherwise earns \$1; a player earns an additional \$1 if his report matches the majority report; see Table 4. There was no feedback until the end of the experiment, when one randomly selected trial of Part 3 would be paid. Now the player's report affects his payoff even if he is not pivotal, as siding with the majority pays a small amount. As in Part 2, the coin is no longer a natural coordination device because now only the distribution of labels as well as the player's own label matter for the player's payoff, and players whose label differs from the coin have an economic incentive to lie.

Table 4: Player's payoffs in Part 3: misaligned incentives + consensus pay.

		Report is in the	
		Majority	Minority
Label = {	Majority Report	\$6	\$5
	Minority Report	\$2	\$1

3.4 Parts 2 and 3: manipulations

Parts 2 and 3 have two sources of variation: within- and between-subject.

The within-subject manipulation concerns the structure of economic incentives in Parts 2-3. We structure the distribution of labels such that in each part, every participant experiences 6 different decisional situations—with or without an economic incentive to lie—based on the distribution of labels and

the coin. This is done to identify causal effects of economic incentives on behavior. Letting $\circledast$ denote the coin face, the breakdown of trials is illustrated in Table 5: there are 4 *same-labels* trials and 12 *mixed-labels* trials. In *same-labels* trials everyone has the same label so economic incentives are perfectly aligned: in 2 of these trials, the label is $= \circledast$ and in 2, it is $\neq \circledast$ (so there is an incentive to lie). In *mixed-labels* trials 2/3 of players had the same label and the rest the opposite label, so economic incentives were misaligned: (i) in 8 of these trials the player has a *majority label*, 4 trials with label $= \circledast$ and 4 trials with label $\neq \circledast$ (so there is an incentive to lie); (ii) in the remaining 4 trials the player has a *minority label*, 2 trials with label $= \circledast$ (there is an incentive to lie in Part 3) and 2 trials with label $\neq \circledast$ (there is an incentive to lie in Part 2).[3]

Table 5: Trials in Parts 2-3.

	Same-label		**Mixed-label**			
	Label		Min. label		Maj. label	
	$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$
	(1)	(2)	(3)	(4)	(5)	(6)
Part 2	2	2	2	2	4	4
Part 3	2	2	2	2	4	4

Notes: Each part has 16 independent trials (no feedback). The cells display the number of trials each subject played in that Part. The notation $= \circledast$ ($\neq \circledast$) means the player's label matches (does not match) the coin. Trials where the subject had an economic incentive to misreport are shaded. In Part 1 subjects had no label and played 16 trials. Overall each subject played 48 trials.

The between-subject manipulation concerns the size of groups in Parts 2-3. We randomly assigned 120 subjects to 8 sessions, each with 15 participants. In half of these sessions, participants were randomly re-matched into 3-player groups at the start of Part 2 and 3. All other design parameters are unaltered

[3]Table B1 in Appendix B shows the exact labels assigned to participants and coins shown.

and, in particular, the informational structure of the game remains the same. Table 6 below summarizes the main differences and similarities across sessions.

Table 6: Experimental design: summary

Variable	Large group	Small group
Sessions (subjects)	4 (60)	4 (60)
Group size in parts 1,2,3	15, 15, 15	15, 3, 3
trials in a part	16	16
Female subjects	0.76	0.78
Salient $ Earnings (avg.)	$13.98	$13.45
min, max	$3, $18	$3, $18

Notes: Salient earnings are rounded up to the next quarter, exclude a $7 fixed participation amount. The session dates are: 11/29, 11/30 (2 sessions), 12/2 (2 sessions), 12/6 (2 sessions), & 12/7 2022. Sex was voluntarily self-reported at the end of the decision task (3 participants out of 120 did not report an answer). The high share of females partly reflects the composition of the subject pool at Chapman University.

Discussion. In Part 1 there is an economic incentive to make a report that matches the majority report—as in a beauty contest game—and the coin is a natural and explicit coordination device to accomplish this. Here, there is no economic incentive to lie. This is no longer true in Part 2, where there can be an economic advantage from lying. Here, neither the coin nor the labels of others should influence the player's decision and only the subject's label matters. In Part 3 there is an economic incentive to make a majority report—as in Part 1—but now there are also economic incentives to lie because the distribution of labels, and not the coin, can serve as a coordination device. The theoretical basis for these statements in laid out in Section 4.

From an empirical perspective, Part 1 serves three distinct purposes. First, it allows subjects to establish whether there is a common rationality baseline and, if not, how significant is the departure from that assumption. This is something uncommon in laboratory experiments and can be methodologically valuable: every participants sees how others behave when there is no economic

incentive to lie. Second, it allows us to determine if—in a frictionless environment where there are no distorting economic incentives—subjects can coordinate on some report using the coin as an explicit coordination device. Third, by measuring the frequency of minority reports in Part 1, we can quantify possible underlying coordination frictions due to random errors, task complexity, psychological factors (e.g., a contrarian personality), or other confounding factors. This allows us to empirically test whether coordination frictions alone can prevent the emergence of truth in our design.

Parts 2 and 3 allow us to establish the causal effect of economic incentives on the choices to report the truth. In mixed-labels trials, majority label players have an economic incentive to lie when their label differs from the coin, because if they all do so, then they gain \$6 instead of \$1. In Part 3, the minority label player has also an economic incentive to lie in this case, because they can earn an extra \$1. Contrasting Large to Small groups allows us to measure whether the outcome of the consensus mechanism is influenced by the size of the group who observes and reports on a given phenomenon. Our prior is that there could be at least three possible effects associated with a reduction in the group size. First, lower strategic uncertainty, because there are less decision-makers; this might facilitate coordination on some report. Second, a majority-label player can ensure that the truth will emerge in mixed-label trials where the majority has an economic incentive to misreport the coin; here, the player can side with the lone minority-label player and report the coin to ensure the majority report corresponds to the coin. Hence, if a single individual is averse to lying, she can ensure that the lies of others cannot distort the truth. Third, in a 3-player group dishonesty entails possibly smaller psychological costs because a lie can only affect payoffs of few other players and is also observed by few others.

3.5 Experimental procedures

We recruited 120 undergraduate student subjects (23% males) through announcements at Chapman University. No participant had previous experience with the game. The experiment was conducted in the Economic Science Institute's laboratory in 2022. We ran 8 sessions, one on November 29, two on November 30 , December 2 and 6, and one on December 7. Each session had 15 participants. Subjects were randomly assigned to sessions.

Each subject had a private terminal; neither communication nor eye contact was possible among subjects. The experimenter publicly read the instructions at the beginning of the session; each subjects had a paper copy of the instructions. During the experiment, subjects could use pen and paper to create records of the outcomes and their choices. The experiment was programmed and conducted using Django (`https://www.djangoproject.com/`). The software is available for use upon request here: `https://esi-portal.chapman.edu/`. The source code for the software is available here: `https://github.com/jeffreykirchner/distributed_consensus`. On average, a session lasted approximately 40 minutes including instructions, a short anonymous survey after completing the experiment, and payments. Instructions are in Appendix B). At the conclusion of all three parts of the experimental session, one of the 16 decision tasks of Part 2 and one of Part 3 were selected for payment using a bingo cage randomization device operated by two volunteers—as explained in the instructions. Average earnings were \$20.70 per subject ($\min = \$10.00, \max = \$25.00$) including a \$7 fixed participation payment.

4 Theoretical model and impossibility result

Here we offer a general formulation of the model behind the experimental design, and use it to determine equilibrium outcomes of each treatment. The theoretical results are then used to derive hypotheses.

Suppose an odd number S of players bet on the outcome of a public coin toss. Players $i = 1, ..., S$ have already made their bets which corresponds to their label $L_i \in \{H, T\}$. Let $\mathbf{L} = \{L_1, ..., L_S\}$ be the profile of labels. Given $\mathbf{L}$, all players can compute the fraction of players whose label is heads, $\eta = \sum_i \mathbb{I}\{L_i = H\}/S$, where $\mathbb{I}\{x\}$ is the indicator function. Obviously, if $\eta > 1/2$ the majority label is heads and tails otherwise. Given their own label, all players observe $\mathbf{L}$ so they can compute η. It turns out that η is all that matters for our results, but we will assume players observe $\mathbf{L}$, as we do in the experiment.[4] Following the information on η, nature tosses a fair coin with realization $\theta \in \{H, T\}$. This realization is publicly observable by all players. In a standard setting, a trusted authority would input that realization in the outcome function of the game to give a payoff for each player according to whether their label matches or not the outcome of the coin toss. However, we assume there is no single authority that can do this. Rather, the outcome of the coin toss, for the purposes of paying off players, must be determined by some voting procedure involving all players.[5] Appealing to May's Theorem (May, 1952), we assume the voting rule is majority voting and voting decisions are simultaneous.

May's Theorem. *Consider any 2-candidate election with an odd number of*

[4]Knowing individual labels is inconsequential when players take simultaneous actions, but it matters a great deal if actions were sequential.

[5]Selecting a fraction of players, or even just one, would not change the result that the truth cannot emerge all the time, since that selection cannot depend on θ itself.

voters. If the voting system satisfies (1) equal treatment of voters (anonymity), (2) equal treatment of candidates (neutrality), and (3) monotonicity (if A wins with a votes, then A wins with a+1 votes), then the voting system is simple majority.

In the voting game, we denote by v_i the vote of player i. In principle, this vote is a function $v_i(L_i, \mathbf{L}, \theta)$ that maps her initial label (or bet), the profile of labels, and the public realization of the coin toss into $\{H, T\}$. Notice that players can use θ as a coordination device because it is common knowledge that they all observe it, although θ cannot be used to pay players directly. Given a complete set of votes $\mathbf{v} = \{v_1(L_1, \mathbf{L}, \theta), ..., v_S(L_S, \mathbf{L}, \theta)\}$ the recorded realization of the public coin toss is

$$r(\mathbf{v}) \equiv \begin{cases} H & \text{if } \sum_i \mathbb{I}\{v_i = H\} > N/2 \\ T & \text{otherwise} \end{cases}$$

In words, the recorded realization is H if the majority of players vote it is H, and the recorded realization is T otherwise. This recorded realization is all that matters for the purpose of paying off bets. We will say that player i is a winner whenever $L_i = r(\mathbf{v})$, and player i is a loser otherwise. Winners each get $W > 0$, while losers get $\ell \geq 0$ with $\ell \leq W$. In addition, players who vote with the majority earn π. Combining everything, the payoff of player i is

$$V_i = \begin{cases} W & \text{if } L_i = r(\mathbf{v}) \\ \ell & \text{if } L_i \neq r(\mathbf{v}) \end{cases} + \begin{cases} \pi & \text{if } v_i = r(\mathbf{v}) \\ 0 & \text{if } v_i \neq r(\mathbf{v}). \end{cases}$$

Notice that the player's *label* determines whether they are winners or losers, but the player's *vote* determine whether they agree with the majority. It should be clear that we can rescale payoffs, so that ℓ is more akin to a participation

fee as we do in the experiment because we can write

$$V_i = \ell + (W - \ell)\mathbb{I}\{L_i = r(\mathbf{v})\} + \pi\mathbb{I}\{v_i = r(\mathbf{v})\}.$$

We will therefore write the proof as if players were paid a participation fee ℓ, but they would only win $w = W - \ell \geq 0$ when their label matches the outcome of the vote, and nothing otherwise. The equilibrium concept is Bayesian Nash equilibrium.

Proposition 1. *Suppose* $\mathbf{L} = \emptyset$*,* $w = 0$*, and* $\pi > 0$ *(Part 1), there are two pure-strategy Nash equilibria where* $v_i = H$ *for all* i *or* $v_i = T$ *for all* i*.*

Proof. When players are not assigned any label and $w = 0$ and $\pi > 0$, players only care about voting in agreement with the majority. If a player believes that the majority will vote $c \in \{H, T\}$, then their best response is to vote c as well. □

Notice that the realization of the coin toss is payoff-irrelevant. However, the realization of the coin toss is a focal point in the sense of Schelling (1960), and therefore it is natural to expect that players in the experiment use it as a coordination device.

Proposition 2. *Suppose* $w > 0$*. If* $\pi = 0$ *(Part 2) there is a Bayesian Nash equilibrium in weakly dominant strategies where* $v_i(L_i, \mathbf{L}, \theta) = L_i$ *and hence* $r(\mathbf{v}) = H$ *if* $\eta > \frac{1}{2}$ *and* T *otherwise. In this case, voting strategies do not depend on the publicly observed outcome of the coin toss. If* $w > 0$ *and* $\pi > 0$ *(Part 3) there are two Bayesian Nash equilibria, one in which all players vote* H *and one in which all players votes* T*. In this case, voting strategies do*

not depend on individual labels (i.e., betting positions) or the publicly observed outcome of the coin toss.

Proof. If $w > 0$ and $\pi = 0$, payoffs depend entirely on whether the player's label matches the outcome of the vote $r(\mathbf{v})$, the state decided upon by majority voting. Consider player i whose label is L_i, and who considers voting H or T. The only time the player's vote matters for their own payoff is when it is pivotal and beneficial; that is when their vote changes the outcome from $r(\mathbf{v}) \neq L_i$ to $r(\mathbf{v}) = L_i$. In such a circumstance the player strictly prefers to vote their label, and in all other circumstances the player's vote does not matter. Hence $v_i(L_i, \mathbf{L}, \theta) = L_i$ for all i and all triples $(L_i, \mathbf{L}, \theta)$ is a weakly dominant strategy. Given that players vote their own label, the unique equilibrium will have $r(\mathbf{v}) = H$ when $\eta > 1/2$ and $r(\mathbf{v}) = T$ otherwise, as stated in the proposition.

If $w > 0$ and $\pi > 0$, then players can earn a positive payoff by *voting* with the majority, even if their *label* does not match the majority vote. If a player's label is T, then they would strictly prefer to vote for T whenever they expect the number of others voting for T to be greater than or equal to $\frac{S-1}{2}$. Otherwise, the player with label T would strictly prefer to vote for H. Likewise, if an player has label $L_i = H$, then they would strictly prefer to vote for H whenever they expect the number of others voting for H to be greater than or equal to $\frac{S-1}{2}$. Otherwise, the player with label $L_i = H$ would strictly prefer to vote for T. Given that voting now depends on expectations regarding how others are voting there are two equilibria: $v_i(L_i, \mathbf{L}, \theta) = H$ for all i and all triples $(L_i, \mathbf{L}, \theta)$ and $v_i(L_i, \mathbf{L}, \theta) = T$ for all i and all triples $(L_i, \mathbf{L}, \theta)$. □

Intuitively, the result for $\pi = 0$ holds because voting according to the player's own label, can never lower the player's payoff and sometimes it might help them win. If all players vote according to their label, then the record will match the label of the majority. If $\pi > 0$, then players no longer have a weakly dominant strategy to vote according to their own label, because they will leave money on the table if the majority votes differently.

Propositions 1-2 show that we there is strategic uncertainty in all Parts, as we have multiple equilibria, with and without truth-telling. In Parts 2 and 3, Proposition 2 shows conditions under which the truth emerges by luck, i.e., when the profile of labels happens to correspond with the realization of the coin toss. Also, if $\pi > 0$, more players will vote with the majority so that the consensus on the "wrong truth" will be stronger. This analysis can be extended to equilibria in mixed strategies, and situations where there are player "types," some who always tell the truth (deontological individuals) and others who are instead rational payoff maximizers. These observations are organized in two remarks.

Remark 1. Propositions 1 and 2 focus on pure-strategy Nash equilibrium. In Appendix A, we show that mixed strategy equilibria also exist. We do so by studying (type-symmetric) mixed-strategy Nash equilibrium for the case with 3 players. When everyone has the same label, a mixed strategy equilibrium only exists when $\pi > 0$, otherwise it is always a dominant strategy for a player to vote his label. However, when labels are mixed in the group, the only equilibrium with mixed strategies has the property that majority-label players use mixed strategies: it is never optimal for a minority-label player to use a mixed strategy, even if majority-label players are mixing. Instead, it is always individually optimal for a minority-label player to vote his own label. Overall,

the only players who have an economic incentive to use mixed strategies are the ones with the majority label.

Remark 2. Suppose a fraction $\tau \in (0,1)$ of players always tells the truth. Proposition 2 holds as long as τ is sufficiently small relative to the share of majority labels. To see this, without loss of generality, suppose a fraction $\eta > 1/2$ of players has label H. If the outcome of the coin flip is T, a measure $\tau\eta$ of H players will vote T because these players always tell the truth. The remaining H players (who do not necessarily feel compelled to tell the truth) have an incentive to vote H whenever their H vote represents the majority, i.e., $\eta(1-\tau) > 1-\eta+\tau\eta$ or $\tau < 1-1/(2\eta)$. Otherwise, they will vote T. If η is large, the fraction τ of truth-tellers has to be large to reverse the incentives. If η is arbitrarily close to 1/2, we obtain truth-telling for all $\tau > 0$, because in that case the truth-tellers are pivotal.

4.1 Testable hypotheses

Based on Section 4, we put forward five testable hypotheses, each associated to an alternative hypothesis.

We say that "truth emerges" in a trial if the majority report corresponds to the coin shown in that trial.

H 1. *In Part 1, the truth will emerges in 100% of trials.*

Proposition 1, shows that players can coordinate on two possible Nash equilibria, truthful-report or a false-report in every trial of every Part. In Part 1, we hypothesize that the majority report should be truthful because (i) the coin is a natural and explicit coordination device and (ii) subjects have nothing to gain from coordinating on reporting the opposite of the coin. Deviating to make a minority report can only lower the subject's payoff without affecting

the payoff of others. Hence, we expect 100% of reports will be truthful in Part 1. An alternative hypothesis is that the coin does not allow subjects to coordinate on a truthful-report equilibrium and, hence coordination frictions will sometimes prevent the emergence of the truth even if subjects have no economic incentives to lie.

H 2. *In Parts 2-3, the truth will emerge up to 50% less frequently than in Part 1.*

In Parts 2 and 3, subjects have an economic incentive to make a false report when their label differs from the coin, as Proposition 2 shows. In this case, the majority report will be false when the coin differs from the label of the group's majority. Since in Parts 2 and 3 the majority label differs from the coin in 8 of the 16 trials (see Table 5), we expect up to a 50% decline of truthful majority reports relative to Part 1. The behavioral alternative hypothesis is based on the idea that there is an intrinsic cost from being or being known to be dishonest, which dominates the economic incentive to misreport. If so, then the coin should be reported in all parts and, hence there should be no difference in false reports.

Further, we conjecture that, in Parts 2-3 the distribution of labels will serve as the coordination device. We thus put forward the following hypothesis:

H 3. *The majority report will reflect the majority label.*

H 4. *Players will report their label except when they are minority-label in Part 3.*

These hypotheses are derived based on the theoretical observation that majority label players always have an economic incentive to report their label. In Part 2 doing so is a weakly dominant action also for minority label players, but not in Part 3. There, minority label players have an economic incentive

to report the majority label given that majority label players report their own label; this incentive is smaller as compared to majority label players ($1 vs $5). A behavioral alternative is that players are averse to make false reports, in which case the alternative to H 3 is that the majority report and majority label match only when the latter corresponds to the coin image; the alternative to H 4 is that there is no difference in behavior when the subject has a majority or a minority label. The main implications of H 3-4 is that the majority report will be truthful only if the majority label matches the coin.

Conjecturing that participants are rational payoff maximizers, we put forward two final hypotheses:

H 5. *Consider trials where a player has an economic incentive to lie. The relative frequency of false reports will be unaffected by the player's label, minority or majority.*

H 6. *Group size will not affect choices and outcomes.*

H 5 follows from noting that minority label players have a small economic incentive to make a false report in half of the trials: in Part 2, when the coin face differs from their label (in the hope to flip the majority report) and in Part 3, when it matches their label (conform to the majority report and earn $1). Majority-label players also have a significant economic incentive to make a false report in exactly half of the trials—when the coin face differs from their label (falsely report to ensure the majority report is false and earn $5). To the extent that players try to maximize their income, the frequencies of false reports should not differ for majority and minority-label players. H 6 follows from noting that the structure of economic incentives is unaffected by size in Parts 2 and 3.

Alternative behavioral hypotheses are based on the possibility that subjects suffer psychological costs from false reporting. Since minority-label players

have weaker economic incentives to misreport as compared to majority-label players, the economic benefit from lying might be dominated by its psychological cost and, hence, the frequency of false reports will be lower than majority-label subjects in contrast to H 5. In 3-player groups a majority-label player is also pivotal: she can flip the majority report from false to truthful. Again, if there are psychological cost from a lying action, we expect that truthful reports will be more frequent in small as compared to large groups, in contrast to H 6.

5 Results

Here we report the main results that emerge from the analysis of the experimental data. Our empirical strategy utilizes data at the individual level (1 obs. = 1 subject, $N = 60$ per Part, each treatment), at the group level (1 obs. = 1 group, $N = 4$ in Large for every Part, $N = 4, 20, 20$ in Small for Parts 1, 2, and 3 respectively) and, when necessary, at the session level (1 obs. = 1 session, $N = 4$ per Part, each treatment). To be conservative in our analysis, we compare outcomes across Parts using Wilcoxon signed-rank tests on matched observations (session- and individual-level data); we determine possible size effects using a two-sample Mann-Whitney tests with exact statistics (group- and individual-level data). We complement these tests with logit panel regression analyses of choices where one observation corresponds to one subject in a trial, controlling for group size, Part fixed effects and the subject's decision time in the trial. We calculate the size of an effect using Cohen's d (the standardized mean difference).

Recall that report r is false if it does not correspond to the coin image shown in a trial, and it is otherwise truthful. False reports are assigned value 1, and

0 otherwise. Consequently, $\frac{1}{16}\sum_{t=1}^{16} r_{ij}(t) \in [0,1]$ measures the unconditional (relative) frequency of false reports for subject i, in part $j = 1, 2, 3$ of a session.

5.1 Aggregate view: did the truth emerge?

The data support the hypothesis that truth emerged without difficulty in Part 1. Table 7 provides mean and standard deviation of the share of false reports in a Part, using a subject as the unit of observation (col. 1-2). Col. 4, shows the fraction of subjects who made a truthful report in every trial of the Part. Col. 2a shows the percentage of trials where the majority report was truthful, in that it corresponded to the coin shown to subjects. Col. 3 reports the fraction trials in which a group achieved full consensus (in parentheses we have at least 90% consensus).

Table 7: Overview of reports in the experiment.

		False Reports		Truth emerges		Full	Always
		⊛=Maj	⊛=Min	⊛=Maj	⊛=Min	consensus	honest
Part	Size	(1a)	(1b)	(2a)	(2b)	(3)	(4)
1	Large	0.05	–	100%	–	0.50	0.82
2	Small	0.16	0.63	97%	27%	0.36	0.08
	Large	0.19	0.61	100%	16%	0.11	0.10
3	Small	0.03	0.76	100%	13%	0.67	0.10
	Large	0.05	0.72	100%	0%	0.22	0.13

Notes: Col. (1a-1b): 1 obs. = 1 subject ($N = 60$ per size & Part), relative frequency of false reports (mean); Col. (2a-2b): 1 obs. = 1 group ($N = 4$ per Part when Large, $N = 20$ when Small), mean share of trials where "majority report = coin." Col. (3): 1 obs. = 1 group, mean share of trials with 100% consensus in the group on some report. Col. (4): 1 obs. = 1 subject, fraction of subjects who made truthful reports in all 16 trials. ⊛=*Maj*: trials of Parts 2-3 where the coin matched the label of the group majority (possibly everyone); here, the group majority (possibly everyone) had an economic incentive for truth-telling so few could gain from misreporting; ⊛=*Min*: trials of Parts 2-3 where the coin matched the label of the group minority (possibly no-one); here, the group minority (possibly no-one) had an economic incentive for truth-telling so many could gain from misreporting. In Part 1 there were no labels so we report data under the column (1a) and (2a) (everyone had an economic incentive for truth-telling). Very few subjects submitted truthful reports in the entire session: 4 in *Small* and 3 in *Large*.

Based on the data reported in first row of Table 7 we state the following.

Result 1. *In Part 1 the average subject made a false report in 5% of trials and 82% of subjects never made a false report. The majority report always revealed the truth.*

Recall that the group size in Part 1 is always 15 subjects and subjects had no labels. In col. 1a, the average share of false reports is overwhelmingly small: only 5% of reports were false (this number is statistically different than 0, one-sided t-test, p-value <0.001, N=120). As a consequence, the majority report was always truthful (col. 2a). Hence, we cannot reject H1.

Though subjects overwhelmingly coordinated on the truthful equilibrium, coordination was not perfect: a very small number of subjects could not (or

preferred not to) coordinate *on any* equilibrium, truthful or not. Out of 120 subjects, 98 consistently reported the coin image in all 16 trials of Part 1. Of the remaining 22 subjects, 8 made a false report with a frequency less than 25%, 9 between 25% and less than 50%, and 5 between 50% and 56.25%.[6] This is interesting because we expected 100% consensus on reporting the coin due to multiple reasons: (i) the economic incentives to coordinate on *some* report (true or false) are perfectly aligned, (ii) the choice is binary, and (iii) the coin is an explicit coordination device. Reporting the coin seems a natural choice because the coin explicitly anchors beliefs, and it represents the truth and, hence, avoids possible psychological cost from knowing to be lying. In fact, making a false report has no upside for someone who believes that people report what they see: it simply creates a $5 economic loss without affecting others' payoffs (e.g., lowering them to spite them).[7] We conclude that, in our setup, if the economic incentives to coordinate on *some* report (not necessarily the truthful one) are perfectly aligned, then random errors in choice are sufficiently small and did not prevent the truth from always emerging.

This performance sharply contrasts with that in Parts 2 and 3, where economic incentives were misaligned in 12 out of 16 trials.

Result 2. *In Parts 2-3 the average subject made a false report in 40% of trials, and 10% of subjects never made a false report. The majority report revealed the truth in 57% of trials.*

Result 3. *In Parts 2-3, group size had no significant effect on choices and outcomes. The truth emerged less frequently in Part 3 as compared to 2.*

[6] As a result, only 1 out of 8 groups fully coordinated on truthful reporting in every trial of Part 1.

[7] It is an economic loss because subjects simply forfeited a gain in the experiment. This economic loss occurred as soon as the subject's report was in the minority. It is thus possible that someone made an early reporting mistake and realized it, and hence simply reported at random in their subsequent Part 1 trials as their choice no longer mattered anyway.

Result 2 reports averages unconditional on the type of trial and group size. Table 7, instead, reports evidence conditional on the type of trial and group size. In Parts 2-3 in exactly half of the 16 trials most (or all) labels matched the coin, so many (or all) players had an economic incentive for truth-telling; these data are in col. 1a and 2a. Instead, col. 1b and 2b refer to the other half of trials, when few (or no) labels matched the coin, so few (or no) players had an economic incentive for truth-telling.

Consider col. 1a, when few in the group had an incentive to misreport. Pooling data from Large and Small, the average subject made a false report in about 17% of trials in Part 2 and 4% in Part 3. The difference relative to Part 1 is statistically significant only for Part 2 (1 obs. = 1 subject, p-value$<$ 0.001 for Part 2, p-value=0.863 for Part 3, N=60 matched obs.), and the size of the effect is small (0.42).

By contrast, when many in the group had an incentive to misreport, the average subject made false reports in 62% of trials in Part 2 and 74% in Part 3 (col. 1b, pooling data from Large and Small). The differences relative to Part 1 are highly significant (1 obs. = 1 subject, p-value$<$ 0.001 for Part 2 and 3, N=60 matched obs.); the size of the effect is also very large, approximately 1.69 and 2.20 for Parts 2 and 3. The differences between col. 1a and 1b are also highly significant(1 obs. = 1 subject, p-value$<$ 0.001 for Parts 2 and 3, N=60 matched obs. pooling Large and Small). In other words, subjects routinely lied when a majority of them could monetarily gain from doing so.

The frequency of false reports is also statistically different in Parts 2 vs. 3 both when we consider col. 1a or col. 1b (1 obs = 1 subject, p-values$<$ 0.01 for all pairwise comparisons in Small and Large, N=60 matched obs.). That is to say, the frequency of false reports increased when there was an economic incentive to comply with the majority report. There were more false reports

when the majority had an incentive to make a false report (col. 1b, Part 2 vs 3) and there were less false reports otherwise (col 1a, Part 2 vs 3). Finally, there are no statistical differences between Small and Large within a Part (1 obs. = 1 subject, N=60 per group size). This suggests that the probability of being pivotal did not affect subjects' choices in our experiment.[8]

This overwhelming tendency to misreport the coin face when this could generate monetary benefits, had a severely negative effect on the emergence of truth in Parts 2-3. When the majority had an economic incentives for truth-telling, the majority report was nearly always truthful (the only departure is 3% of trials in Small groups of Part 2); see col. 2a. These frequencies are statistically indistinguishable from those observed in Part 1. Now consider col. 2b, when few (or no) players had an economic incentive for truth-telling. Here the majority report was seldom truthful, from 0% to 27% of trials. Pooling data for Parts 2-3, this decline relative to col. 3a data is highly significant in both large and small groups (1 obs. = 1 group, p-values <0.001, for each comparison and group size, N=8, 40 matched obs. for Large and Small), while group size has no significant effect on the frequency of truthful majority reports (1obs. = 1 group, p-value=0.429, 0.506, for col. 2a and 2b, N=8, 40 obs. for Large and Small). If we pool data for Small and Large groups for the case where the majority of players had an economic incentive to misreport (col 2b), the truth emerged 25% and 11% in, respectively, Parts 2 and 3: this decline is statistically significant (1obs. = 1 session, p-value=0.031, N=8 obs.), which is evidence that not only the truth emerged infrequently when there were economic incentives for false reporting, but even less frequently when there

[8]This is in line with the experiment in Robbett and Matthews (2018), where partisan voters' responses to a multiple-choice questionnaire were not significantly affected by group size when useful information was freely available, even if group size affected the probability of being pivotal.

was an economic incentive to comply with the majority report.

Col. 3 shows that, overall, full consensus on a report—truthful or not—was as hard to reach in Parts 2-3 as it was in Part 1. We cannot reject the hypothesis that it was as frequent in Parts 2-3 as in Part 1 (1 obs = 1 session, p-values =0.875,0.125 in Small, p-values=0.125, 0.250, in Large, N=4 matched obs.). However, the group's size did act as a coordination friction: it interfered with frequency of full consensus that, in Parts 2 and 3, was three times more frequent in 3-player groups as compared to 15-player groups (1 obs. = 1 session, p-value=0.029 for both Parts 2 and 3, N=4 per group size). This finding holds if we consider a consensus of at least 90% of reports. Hence, we cannot reject H2.

These results are confirmed by a logit panel regression with random effects at the individual level and robust standard errors (SE) adjusted for clustering at the session level. We first check for influence of group size on truth-telling by regressing a subject's choice of report (false or true) in a trial of Parts 2-3 on the categorical variable Large size (Small is the base case). We include the standardized continuous variable *Decision time*, which controls for the time spent making the choice. Marginal effects are in Table 8. The first two columns consider data from respectively, Parts 2 and 3. There is no significant effect of group size in any Part, meaning that having a design in which players can be pivotal or not, did not affect behavior. Hence, we cannot reject H6.

We also regress a subject's report in a trial on the categorical variable *Part* and report marginal effects in the last two columns in Table 8 for Small and Large groups. Since Part 1 had only large groups Part 2 is the base case in col. (3), while Part 1 is the base case in Col. 4. Consider col. 4. There is a significant increase in the probability of making a false report as we go from Part 1, to Parts 2 or 3; the increase is between 34 and 36 percentage

points. However, these increments are statistically similar in Parts 2 and 3 (Wald test, p-value=0.216). Col. 3 also shows that the probability of making a false report in Small groups was statistically similar in Parts 2 and 3 (the coefficient on Part 3 is statistically insignificant). This is evidence that size did not affect the choice to make false reports.

Interestingly, decision time is positively associated with the probability of false reports. A one standard deviation increment in decision time is associated with a (small but statistically significant) increase in the probability of making a false report in Small groups and in Part 2—about 2 percentage points. To the extent that strategic misrepresentation of a public event (the coin) requires a cognitive effort, this finding suggests that false reports are intentional and not the result of random choice errors.

Table 8: False reports: Marginal Effects.

Dep. variable= 1 if false report	Part 2 (1)	Part 3 (2)	*Small* (3)	*Large* (4)
Large	0.008	-0.014		
	(0.016)	(0.048)		
Part 2				0.360***
				(0.016)
Part 3			0.004	0.342***
			(0.017)	(0.009)
Decision time	0.019**	0.021	0.020**	0.004
	(0.009)	(0.018)	(0.009)	(0.004)
N	1920	1920	1920	2880

Notes: Marginal effects from panel Logit regression on reports. One obs.=one subject in a round. Robust standard errors (in parentheses) adjusted for clustering at session level. Col. 1: data for Part 2 only (both sizes); Col. 2: data for Part 3 only (both sizes); Col 3: data for Small only (Parts 2 and 3 only); Col 4: data for Large only (all Parts). The dependent variable is 1 if the individual submits a false report (0, otherwise). *Large* is an indicator functions taking value 1 for the Large treatment, and 0 otherwise (Small is the base of the regression in cols. 1-2). *Part* is an indicator functions taking value 1 for the corresponding part, and 0 otherwise (Part 2 is the base of the regression in col. 3, Part 1 is the base of the regression in col. 4). We also include the standardized *Decision time* variable, the time in milliseconds it took the subject to make the report. Marginal effects are computed at the regressors' mean value (at zero for indicator variables). Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5% and 10% level, respectively. The coefficients on Part 2 and Part 3 are statistically similar in col. 4 (Wald tests, p-value=0.216).

Summing up we have the following take-aways. First, truth always emerged in groups where economic incentives were perfectly aligned (Part 1). Subject could have earned an identical amount of money by coordinated on *false* reports; yet, they did not choose to do so. By contrast, truth seldom emerged when economic incentives were misaligned (Parts 2-3). In fact, adding a small payment for complying with the majority report increased the frequency of false-reporting and decrease the number of trials in which the majority report was truthful. Second, relying on more witnesses and more reports did not make truth emerge more often (Large vs. Small comparisons). Third, some

coordination frictions were present even if subject had an explicit coordination device (the coin) and a simple binary choice; yet, these frictions were small enough to never prevent the emergence of truth (Part 1). We conclude that in the experiment false majority reports were intentional and a consequence of the structure of economic incentives. To understand what caused the large decline in truthful reports in Parts 2 and 3, we turn to study individuals behavior based on the different decisional situations they faced.

5.2 The effect of economic incentives on truth-telling

In Parts 2-3 each participant had a label corresponding to one side of the coin. These labels, as well as the majority report, affected their payoffs and, consequently, the economic incentives to report the coin; see Tables 3-4. To understand how economic incentives influence reporting choices, we varied the distribution of labels from trial to trial (see Table B1). In 4 of 16 trials everyone had the same label (homogeneous group): in these same-label trials, the economic incentives were perfectly aligned in the group. Instead, in the remaining trials we had a mixed group, where a 2/3 majority had one label (majority-label players) and the rest the opposite label (minority-label players): in these mixed-label trials, the economic incentives were misaligned.

We start by establishing the following.

Result 4. *In Parts 2 and 3, the majority report overwhelmingly matched the majority label in both large and small groups.*

Table 9 shows how often the majority report matched the majority label, in the average group (1 obs = 1 group). The mean relative frequency is reported separately for trials in which the coin was equal to the minority or majority label (col. 1 vs. 2), of which we have an equal amount of observations (6 trials each). Instead, data from the same-label trials is in square brackets (4 trials).

Table 9: How often the majority report reflected the majority label.

		Coin is equal to: Minority label (1)	Majority label (2)
Part 2	Small ($N = 20$)	0.66 [0.95]	0.97 [0.97]
	Large ($N = 4$)	0.79 [1.00]	1.00 [1.00]
Part 3	Small ($N = 20$)	0.86 [0.90]	1.00 [1.00]
	Large ($N = 4$)	1.00 [1.00]	1.00 [1.00]

Notes: 1 obs. = 1 group in a Part (16 trials total). The cells report the average fraction of the 12 mixed-label trials where the majority report matched the majority label; in 6 of these trials the coin was equal to the minority label (col. 1) and in 6 it was equal to the majority label (col. 2). Results for the 4 same-labels trials are in square brackets (all labels differed from the coin in col. 1, while all matched it in col. 2). The rows separate the data by Part and group size.

In same-labels trials the majority report nearly always corresponded to the subjects' labels (no less than 90% of trials, see square brackets). As a result, truth consistently emerged only when this was economically beneficial for all participants (col. 2), and otherwise it almost never emerged (col. 1). Introducing heterogeneity in labels altered this outcome only when the coin differed from the majority label, i.e., when majority-label participants had to lie in order to receive \$5; here the frequencies range between 66% and 100% (col. 1). These frequencies are still large but significantly below the corresponding ones in col. 2 (1 obs. = 1 group, p-value $< 0.001, 0.016$ for Parts 2 and 3, N=24 matched obs. pooling data for both group sizes), which suggests the presence of costs from lying—the majority report reflected the majority label less frequently when this kind of report was false (col. 1). In fact, we also observe a statistically significant difference between Parts 2 and 3 in col. 1 (1 obs = 1 session, p-value =0.031, N=8 matched obs. pooling data for both group sizes), which suggests a greater incentive to make a false report existed in Part 3 as compared to 2. Finally, we cannot reject the hypothesis

that, in each Part, the frequencies are equal in Small and Large groups at the 10% level or better (1 obs. = 1 group, N=20 for small and N=4 for Large). Based on this evidence, we fail to reject H 3.

Summing up, the truth consistently emerged when this was economically beneficial for everyone (square brackets in col. 1-2) or for the group's majority (col. 2). Otherwise, truth infrequently emerged (col. 1). To provide more details about behavior, and to show how economic incentives affected truth-telling, we study individual behavior in a trial distinguishing decisional situations according to whether or not everyone had not the same label in the group, and whether or not the decision maker had the majority label.

5.2.1 Behavior in same-label trials

In 4 trials of Parts 2 and 3, every group member had the same label, hence economic incentives were perfectly aligned in the group. Truth-telling was incentive compatible only in 2 of those trials—where the label was equal to the coin. Subjects overwhelmingly reported their own label in all circumstances, although less frequently when doing so was false reporting.

Result 5. *In same-label trials, participants routinely reported their label: in 99% of trials when it matched the coin, and in 82% of trials when it did not. Group size had no significant effect.*

Evidence comes from col. 1-2 in Table 10, showing the relative frequency of false reports for the average subject in trials where everyone had the same label. Reporting the label is consistent with truth-telling when the label matches the coin. The data are strikingly similar across group sizes and Parts: subjects reported their own label 98.5% of the trials when this was a truthful report (the complement of col. 1), and otherwise only about 82% of the times (col. 2). This difference in frequency is highly significant (1 obs. = 1 subject, p-

value < 0.001 for each group size and Part, N=60 matched obs.) and the effect size is ≈ 0.6, pooling sizes and Parts 2-3. In other words, when subjects had to lie to secure an economic gain, they lied much less (about 18.5 ppts). This switch in behavior is an indication that not everyone was comfortable with lying to secure an economic gain. About 15% of subjects never submitted a false report in same-label trials (13% and 15% in Part 2 for, respectively, Small and Large, and 15% in Part 3 for both group sizes). This is too small a proportion to affect outcomes and, hence, to ensure the emergence of truth in the group.[9] Although the data reveals behavior consistent with aversion to lying for the average participant, this trait was not sufficiently widespread among participants for truth to consistently emerge. Hence, we cannot reject H 4 in groups where economic incentives were perfectly aligned.

[9]This is perhaps why some subjects always made truthful reports: they expected that it would have no effect on their earnings. A subject who believes not to be pivotal can benefit from submitting a truthful report because doing so negates the utility decline associated with lying, without any adverse monetary consequences.

Table 10: Average share of false reports by decisional situation.

		Same-label trials		**Mixed-label trials**			
		All labels		Minority player		Majority player	
		$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$
Part	**Size**	(1)	(2)	(3)	(4)	(5)	(6)
2	Small	0.03	0.82	0.12	0.57	0.03	0.79
	Large	0.02	0.80	0.10	0.67	0.04	0.78
3	Small	0.00	0.83	0.59	0.12	0.00	0.81
	Large	0.01	0.82	0.47	0.20	0.00	0.81

Notes: 1 obs.: 1 subject in a Part 2 or 3 (N=60). The average shares reported are conditional on the type of trial: (i) same-labels (4/16 per Part, in col. 1-2) vs. mixed-labels (12/16 per part, in col. 3-6); (ii) minority-label subject (col. 3-4) vs. majority-label (col. 5-6), and (iii) the subject's label correspondence to the coin ($= \circledast$ means the player's label matches the coin, or else we use $\neq \circledast$). Note that the values in col. (2), (4), and (6) correspond to the average frequency with which a subject reported their own label in that decisional situation; in col. (1), (3), (5), this frequency corresponds to the complement of the values that are shown. Table B2 reproduces the calculations in this Table when we exclude subjects who never lied in a Part (5 and 6 in Parts 2-3, respectively, for Small, and 6 and 8 for Large—4 subjects in Small and 4 in Large never lied either in Part 2 or 3). Table B3 in Appendix reports the decision times for the average subject.

Did the truth emerge in trials where economic incentives were misaligned?

5.2.2 Behavior in mixed-label trials

In 12 trials of Parts 2 and 3, participants had unequal labels, hence economic incentives were misaligned in the group. In 6 of these trials the coin matched majority label—so the majority had an economic incentive to report the truth. In the other 6 the coin matched the minority label, so the economic incentives were reversed with truth-telling being incentive-compatible for the minority. In each case, every subject made choices both as a majority- and minority-label player (4 and 2 trials, respectively), either when her label matched the coin and when it did not. These four decisional situations correspond to col. (3)-(6) of Table 10. We report three distinct results.

Result 6. *Majority-label players overwhelmingly reported their label in both large and small groups. This behavior is similar to same-label trials.*

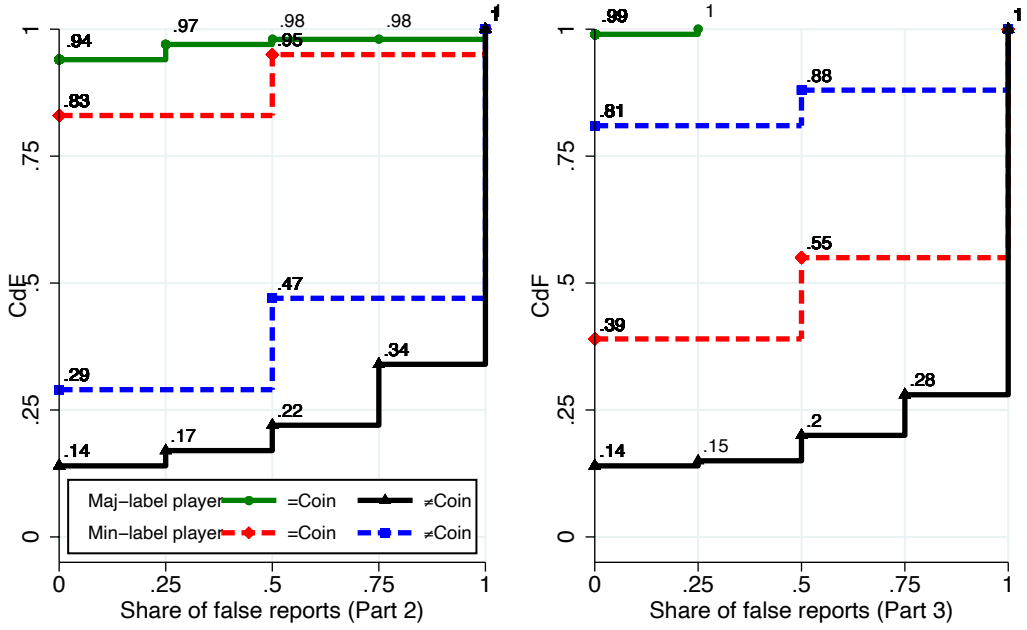
Evidence is in Table 10, showing the share of false reports for the average subject. The data are reported conditional on the four decisional situations in a trial: with a minority or a majority label, and with a label that matches the coin or not. Hence, the shares tells us how often the average player reported the majority label.[10]

Col. 5-6 refer to choices when subjects had a majority label. When their label matched the coin, they nearly always reported their own label (hence, made a truthful report); the frequencies are the complement of the values reported in col. 5, i.e., 97% and 96% in Part 2 for Small and Large , and 100% in Part 3. Col. 6 tells us how often majority-label subjects reported their label when it differed from the coin (hence, made a false report); the frequencies drop by about 20 percentage points as compared to col. 5. This decline is highly significant (1 obs = 1 subject, p-values < 0.001 in Small and Large, Part 2 and Part 3, N=60 matched pairs per Part and group size) and also large (pooling Parts 2 and 3 and group sizes, Cohen's d is about 0.62) suggesting the presence of aversion to lying. Notice also that majority-label players behaved identically to players in same-label trials—they never lied when the coin matched their label, and they lied about 80% of the times otherwise; we can neither reject the hypothesis of similar frequencies in col. 1 and 5, nor in col. 2 and 6 at the 10% or better level (1 obs = 1 subject, N=60 matched pairs per Part and group size). Finally, reporting behavior is statistically similar in Parts 2 and 3 (1 obs = 1 subject, p-values < 0.001 in Small and Large, N=60 matched pairs per group size). Hence, we cannot reject the first part of H 4 for mixed-label trials.

[10]The numbers are reported in Fig.B1 in Appendix B.

Additional evidence comes from Fig. 4, which reports the C.d.F of the share of false reports in mixed-label trials, pooling data for Small and Large groups. Consider the solid lines in each panel. The one with round markers first order stochastically dominates the on with triangle markers. In particular, 94% of majority-label players never made a false report in Part 2 when their label matched the coin (round markers); this number grows to 99% in Part 3. By contrast, this number drops to 14% when their label differed from the coin (triangle markers), which is when a large number of majority-label players always reported their label (66% in Part 2 and 72% in Part 3).

Figure 4: CDF of share of false reports (mixed-label trials).



Notes: C.d.F of share of false reports. 1 obs. = 1 subject in a Part, Small and Large pooled together ($N = 120$). Data only for mixed-label trials; these trials are separated based on the decisional situation (the subject has a majority or minority label, the coin image equals or differs from the subject's label). The values displayed above the markers are approximated to the second decimal.

Summing up, when participants had a majority-label, they invariably made truthful reports *only if* the coin matched their label. Otherwise, they had an economic incentive to lie and as a result routinely made false reports—though not always, which betrays an aversion to lying. This aversion to lying was not enough for the truth to emerge using our consensus mechanism, however. The main message is that truth is routinely obscured when the majority can gain from doing so. This is the same behavior and same outcomes observed in same-label trials, which is expected because the economic incentives for misreporting are the same ($5). Now we show evidence that participants did not mindlessly report their own label and behaved differently when they had weaker economic incentives to lie.

Result 7. *Minority-label players did not routinely report their label in both large and small groups. The average subject made false reports less frequently when she had a minority label as compared to a majority label.*

Evidence is in Table 10, col. 3-4. The average frequency of reporting the player's label is the complement of the values shown in col. 3, and can be read directly in col. 4. We do not find significant differences in behavior in Small vs. Large groups; we can reject the hypothesis of identical frequencies at the 10% level or better (1 obs. = 1 subject, p-value $\geq$ 0.130 for all four comparisons, N=60 matched obs.). Pooling data for both group sizes, in col. 3 the frequencies are 89% and 47% for Parts 2 and 3, and in col. 4 are 62% and 16%. For each Part, the differences between col. 3 and 4 are highly significant (1 obs.=1subject, p-values $\leq$ 0.01, N=120 matched obs. per Part). Finally, we see that minority-label subjects reported their label less often in Part 3 than 2, consistent with the switch in economic incentive; pooling data from both group sizes, we reject the hypothesis the frequencies of Part 2 and 3 are statistically similar in each column (1 obs. = 1 subject, p-value$\leq$ 0.001 for

respectively col. 3 and col. 4, N=120 matched obs. per column).[11] Hence, we cannot reject H 4.

Regarding the second part of Result 7, compare false reports in trials where subjects had an economic incentive to lie, using Table 10. Majority-label players have an incentive to lie when the coin differs from their label (it ensures they earn \$5); this is col. 6 for Parts 2-3. Instead, minority-label players have a weak economic incentive to lie in Part 2 col. 4, when the coin differs from their label (they might hope to flip the majority report in their favor), and a stronger incentive to lie in Part 3 col. 3, when the coin matches their label (they can earn \$1 by conforming to the majority). Consider Part 2, col. 4 vs. 6. The average frequencies of false reports are 0.57 and 0.67 for minority-label players in Small and Large groups; it is 0.79 and 0.78 for majority-label players. These differences are statistically significant at the 1% level or better (1 obs. = 1 subject, N=60 matched obs. per group size). Now consider Part 3, col. 3 vs. 6. The frequencies of false reports are 0.59 and 0.47 for minority-label subjects vs. 0.81 for majority-label, again highly statistically significant differences (1 obs. = 1 subject, N=60 matched obs. per group size).[12] Hence, we can reject H 5.

Since majority-label players have the strongest economic incentives for misreporting, Result 7 suggests that the frequency of false reporting is elastic to

[11] There is also evidence of aversion to lying. In Part 3 col. 3, subjects reports the majority label no more than 59% of the times despite the fact of being fairly certain to earn \$1 by doing so. When we exclude from the data subjects who never misreported the coin we still see evidence of aversion to lying. The data are reported in Appendix, in Table B2.

[12] If, instead, we pool all mixed-label trials (independent of the coin), the average share of false reports for minority-label subjects is 0.34, 0.38, 0.35 and 0.33 in, respectively, Part 2, Small and Large, and Part 3, Small and Large; they are 0.41, 0.41, 0.4, and 0.41 for majority-label subjects. Not all differences between minority- and majority-label subjects are statistically significant (1 obs. = 1 subject, p-value= $0.002, 0.182, 0.039, 0.0615$ for respectively Parts 2 in Small and Large, and Parts 3 in Small and Large, N=60 matched obs. per groups size and Part) and the effect size is truly small (pooling group size, it is 0.09 in Part 2 and 0.13 in Part 3.

prospective economic gains. Further evidence comes from the C.d.F in Fig. 4. Consider trials where the subject had an economic incentive to lie. In Part 2, compare the blue dashed line with square markers to the black solid line, while in Part 3 compare the red dashed line with circle markers to the black solid line. In both cases the C.d.F. of majority-label players first order stochastically dominates the C.d.F of minority label players. In particular, the share of majority-label players who *always lied* is close to 70% vs. 50% for minority-label players.

These results are confirmed by a logit panel regression with random effects at the individual level and robust standard errors (SE) adjusted for clustering at the session level. The regressions consider data from respectively, Parts 2 and 3, restricted to trials where the labels were mixed. We regress a subject's choice to make a false or true report in a trial on the categorical variables treatment (Small is the base case) and subject's label (majority label is the base case). We include the standardized continuous variable *Decision time*, which controls for the time spent in making the choice. Marginal effects are summarized in Table 11.

Table 11: False reports in mixed-label trials: marginal effects.

	All mixed-label trials		Minority-label		Majority-label
Dep. variable= 1	Part 2	Part 3	=coin	≠coin	≠coin
if false report	(1)	(2)	(3)	(4)	(5)
Minority label	-0.071***	-0.053**			
	(0.018)	(0.027)			
Part 3			0.479***	-0.464***	0.017
			(0.079)	(0.032)	(0.016)
Large	0.005	-0.006	-0.010	0.087	-0.014
	(0.029)	(0.042)	(0.028)	(0.054)	(0.015)
Decision time	0.032	0.024	0.008	-0.019	-0.000
	(0.021)	(0.023)	(0.006)	(0.032)	(0.008)
N	1440	1440	480	480	960

Notes: Marginal effects from panel Logit regression on reports. One obs.=one subject in a mixed-label trial. Robust standard errors (in parentheses) adjusted for clustering at session level. Col. 1: data for Part 2 only; Col 2: data for Part 3 only; Col. 3-4: data for trials in which the subject had a minority label, only. Col. 5: data for trials in which the subject had a majority label, which differed from the coin. The dependent variable is 1 if the individual submits a false report (0, otherwise). *Minority label* is an indicator function taking value 1 if the subject had a minority label in the trial, and 0 otherwise (majority label is the base of the regression). *Part 3* is an indicator functions taking value 1 in the corresponding part, and 0 otherwise (Part 2 is the base of the regression). *Large* is an indicator functions taking value 1 if the group is Large, and 0 otherwise (Small is the base of the regression); it is interacted with Part and Minority Label. We also include the standardized *Decision time* variable, the time in milliseconds it took the subject to make the report. Marginal effects are computed at the regressors' mean value (at zero for indicator variables). Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5% and 10% level, respectively.

There are two main observations. Col. 1 and 2 show that the probability of making a false report drops by 7 percentage points in Part 2 (5, in Part 3) when the player had a minority as compared to majority label. This suggests that misreporting was elastic to the size of possible economic gains. Minority label subjects had the smallest incentive to make a false report ($1, in Part 3), while majority-label players stood to gain the most in both Parts ($5) and hence their frequency of misreporting is the highest.

Col 3-4 show that the effect of introducing economic incentives for comply-

ing with the majority, in Part 3. Because of these incentives, minority-label players switched from reporting their own label to the majority label; in col. 3, the probability of making a false report increases by about 48 percentage points in Part 3 vs. 2, when their label matched the coin (they had an economic incentive to lie in Part 3 but not 2). The size of the effect is similar and the sign opposite in col. 4, where the minority label did not match the coin (there is an economic incentive to lie in Part 2 but not 3). Finally, col. 5 shows that majority-label players acted similarly in both Parts, in trials where they had an incentive to lie (their label differed from the coin).

6 Discussion and conclusion

This study theoretically demonstrates and empirically documents the problems societies encounter in establishing accurate records of social or physical phenomena. This problem manifests itself when records must rely on external evidence, which must be provided by human observers. In our laboratory societies, when only a minority had an economic stake in the truth, what was established as fact routinely misrepresented their actual experience. This suggests that one cannot generally trust that historical records reflect actual facts, which has deep implications for the kind of institutions that allow societies to function. If, for example, we take the view that a monetary system is simply a way to preserve and make publicly available records of economic activities (see Kocherlakota, 1998), then the system must somehow assure us that what is preserved in its records is factual.

In our consensus game, individual earnings depend on what the group majority report. Truth-telling and false-reporting where always a Nash equilibrium, independently of whether participants had different or similar stakes in

the truth. Part 1 reveals that truthful reporting is the salient choice when individuals simply benefit from aligning their reports. Here, coordinating on false-reporting provides no economic advantage and, although full consensus was not always achieved, we observe only minor coordination frictions.[13] In Part 1, almost all participants routinely reported actual facts, even if false-reporting could support identical payoffs. This established an empirical truth-telling baseline, which demonstrated to participants that their laboratory society was willing and capable to coordinate on reporting the truth. This experience should in principle anchor beliefs around truthful reporting in subsequent trials, biasing play in Parts 2 and 3 toward truth-telling.

Yet, behavior drastically changed in Parts 2 and 3 as compared to Part 1. When the majority had a stake in pushing a lie, the consensus protocol routinely failed to generate accurate records: the truth emerged in at most 1 out of 4 trials and sometimes never (Result1–2). Furthermore, when those who had a stake in the truth were promised a small compensation for conforming with the majority report, the frequency of false reports increased even more. The data reveal that the frequency of false-reporting is elastic to the size of prospective economic gains. By tracking individuals across all trials we find that the average participant made about 80% false reports when she could earn \$5 by easily colluding on a lie with the group's majority, while this frequency drops to about 60% when she could only earn \$1 by going along with the majority's lie (Results 4–7).

Overall these findings are consistent with the notion that people suffer

[13]We expected full consensus since the experimental design minimizes the possibility of coordination mistakes by adopting (i) a simple binary choice and (ii) an explicit and unmistakable coordination method (the public projection of the coin image). Because these two features are unlikely to characterize most field situations – choices are more complex and observations more opaque – this points to another weakness of simple consensus mechanisms.

implicit costs from the act of lying or from knowing that others are aware of their lying. We do not see 100% false reporting from those who had a stake in burying the truth. Yet, if this implicit cost was present, it was also insufficient to overcome the economic incentives against truth-telling. Out of 120 participants, only 7 never made a false report in any trial. It is possible that, due to the majority consensus mechanism adopted, the psychological aversion to lying was somewhat compensated (instead of being multiplied) by the soothing sense of being one among many who contributed to push a lie.

The paucity of unconditional truth-tellers in the experiment seems to be a major obstacle to the emergence of the truth when what is accepted as "fact" reflects the reports of many independent and anonymous witnesses. Interestingly, the number of reports – the group size – does not seem to matter, as it did not significantly affect behavior and outcomes. Relying on less witnesses and less reports, so that some participant could be pivotal in pushing the truth through, did not significantly alter outcomes. An interpretation is that if there is a psychological cost from lying, then it is not very elastic to the number of individuals who observe the lie.

The question is thus: what economic mechanism could maximize incentives for accurate and factual reporting?

One lesson from the experiment is that the truth always surfaced when participants had no stake in any specific equilibrium. This points to the importance of relying as much as possible on independent observers to provide the external evidence necessary to determine what should be recorded. This raises questions about the wisdom of current practice, whereby e.g. rating agencies – which are compensated by the corporate entities seeking a rating – are entrusted with the job of establishing the financial state of corporations. Similarly problematic is the practice of relying on industry experts in deter-

mining the safety of chemical or other compounds in consumers' products (e.g., glyphosate) when the experts' current or future remuneration may be tied to the industry's fortunes. To address situations such as these, our results suggest the need to create and maintain a pool of vetted independent observers in society, and draft them to populate reporting panels in a variety of industries. Though this is not guaranteed to remove all possible judgment distortions, it might help dilute the risk of a mismatch between what is known and what is ultimately accepted to be the "truth."

Another lesson is that only very few participants (about 6%) always provided factual reports, irrespective of the monetary incentives to do otherwise. It seems that the aversion to lying was often overpowered by the economic incentives to distort the record of an event. A natural question is thus: could a mechanism designer exploit the known presence of a few truth-tellers to elicit truth-telling from everyone else? Possibly, this could be accomplished using a sequential and public reporting mechanism to foster coordination on the known truth. If truth-tellers can be identified by the mechanism designer, and included in a reporting panel by allowing them to make the initial reports, then the truth would emerge early in the reporting sequence, which might makes it more costly to steer later reporting away from the truth-telling equilibrium.

This is not simply an academic question. Understanding how to let the truth about world events come to the surface is in many respects the challenge of our times. Consider for instance the large impact that the constellation of social media platforms exert on the truth that is ultimately accepted by the public about social, economic and even physical events.[14]

[14]Interestingly, the X platform (formerly Twitter) recently introduced a fact-checking tool called "Community Notes," which frequently appears below messages that get a very large audience, to give it context. These notes can be written and voted on by any user, and which notes are shown is decided entirely by an open source algorithm. Unlike algorithms

Participants in our experiment all knew the truth about the event, but still overwhelmingly chose to adopt the truth that economically benefited them the most. But in field situations we typically have to consume scarce resources –financial or otherwise – to uncover what the truth is. This suggests another important consideration: assuming that with enough external evidence the truth about some event is indeed verifiable, is it socially worthwhile to uncover it?

Global warming is a case in point: The uncertainty regarding that physical phenomenon has presumably fallen in the last 25 years thanks to increased scientific scrutiny. Still, did this greater body of knowledge have a meaningful impact on society? Governments seemingly do not select policies in regards to this subject guided by evidence-based arguments, but instead based on the truth that has been accepted by society's majority of which they are an expression. If so, should we allocate resources to uncover more aspects about global warming if it is up to the majority – and their political or ideological leanings – to ultimately decide what is or is not happening?

To be sure, this problem is even more pronounced in authoritarian societies. The Roman Catholic Inquisition persecuted Galileo Galilei in 1633 for providing empirical evidence supporting Copernicus' mathematical theory of heliocentrism, which at the time was the exact opposite of accepted truth. In the eyes of the Church – and mainstream scholars of the time – disproving that truth was hardly worth the effort, it seems. Heliocentrism was declared a heresy by a vocal minority of society—the ruling ecclesiastical authorities—

that simply calculate some index based on users' ratings, Community Notes attempts to prioritize notes that receive positive ratings from people across a diverse range of perspectives (see Buterin, 2023). That is, if people who usually disagree on their ratings agree on a particular note, then that note is scored especially highly. In a way, Community Notes acts as a consensus-building tool when everyone is free to provide their own "truth."

and disappeared from public view until about two centuries later those same authorities ruled that it could find its way into printed materials. We have come a long way from those dark days, but the experiment suggests that we still have a long way to go.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. Preferences for truth-telling. *Econometrica*, 87(4), 1115-1153.

Buterin, Vitalik. 2023. What do I think about Community Notes? Retrieved on Dec 14, 2023 from https://vitalik.eth.limo/general/2023/08/16/communitynotes.html

Buterin, Vitalik. 2014. Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform, https://ethereum.org/en/whitepaper/

Cass, David and Karl Shell. 1983. Do Sunspots Matter? *Journal of Political Economy* 91 (2), 193-227.

Cohn, A., E. Fehr, And M. A. Maréchal. 2014. *Business Culture and Dishonesty in the Banking Industry. Nature*, 516 (7529), 86-89.

Erat, Sanjiv, and Uri Gneezy, (2012) White Lies. *Management Science* 58(4), 723-733.

Fischbacher, U., And F. Föllmi-Heusi. 2013. Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association*, 11 (3), 525-547.

Gino, E., E. L. Krupka, R. A. Weber. 2013. License to Cheat: Voluntary Regulation and Ethical Behavior. *Management Science* 59(10), 2187-2203.

Gneezy, U., 2005. Deception: the role of consequences. *American Economic Review* 95(1), 384-394.

Gneezy, U., A. Kajackaite, And J. Sobel. 2018. Lying Aversion and the Size of the Lie. *American Economic Review*, 108 (2), 419-453.

The General Theory of Employment, Interest, and Money. London: Macmillan (reprinted 2007).

Kocherlakota, N.R., 1998. Money is memory. *Journal of Economic Theory*, 81, 232-251.

Lai, Ernest K., Wooyoung Lim, and Joseph Tao-yi Wang. 2015. An experimental analysis of multidimensional cheap talk. *Games and Economic Behavior* 91, 114-144.

Mazar, Nina, On Amir, and Dan Ariely. 2008. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research* 45 (6), 633-644.

May, Kenneth O. 1952. A set of independent necessary and sufficient conditions for simple majority decisions. *Econometrica* 20 (4),680-684.

Nakamoto, Satoshi. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. Whitepaper. https://bitcoin.org/bitcoin.pdf

Rahwan, Zoe, Erez Yoeli, and Barbara Fasolo. 2019. Heterogeneity in banker culture and its influence on dishonesty. *Nature* 575, 345-349

Robbett, Andrea and Peter Hans Matthews. 2018. Partisan bias and expressive voting? Journal of Public Economics 157, 107-120.

Thomas C. Schelling. 1960. The Strategy of Conflict. Cambridge, Mass.

Tergiman, Chloe and Marie Claire Villeval. 2023. The Way People Lie in Markets: Detectable vs. Deniable Lies. *Management Science*.

Vespa, Emanuel and Alistair Wilson. 2016. Communication With Multiple Senders: An Experiment. *Quantitative Economics* 7, 1-36.

A Appendix: Proofs

A.1 Mixed equilibria with 3 players

Below, we refer to the players, as 1, 2 and 3. First, suppose all players have the same label, and assume it is Tails (T) without loss of generality. Also, suppose player 1 and 2 use a mixed strategy with probability τ on T (and $1-\tau$ on H). We solve for the optimal strategy of player 3, using τ_3 for the probability of choosing T. The payoff is as specified in the main body of the paper.

The expected payoff to player 3 is as follows (we break it down to ease the reading of the payoff below).

- If player 3 votes T (with probability τ_3)
  - he gets $w+\pi$ when the majority's vote is T. This happens when one or both player 1 and 2 vote T. The probability of this event is $\tau^2 + 2\tau(1-\tau)$.
  - he gets 0 when the majority's vote is H. This happens when both players 1 and 2 vote H, with probability $(1-\tau)^2$.
- If player 3 votes H (with probability $1-\tau_3$)
  - he gets w when the majority vote is T. This happens when both of players 1 and 2 vote H, with probability τ^2.
  - he gets π when the majority vote is H. This happens when one or both player 1 and 2 vote H. The probability of this event is $(1-\tau)^2 + 2\tau(1-\tau)$.

Hence the payoff of player 3 is

$$\tau_3 \left\{ \left[\tau^2 + 2(\tau(1-\tau)) \right] (w+\pi) + (1-\tau)^2 \times 0 \right\} + (1-\tau_3) \left\{ \tau^2 w + 2(\tau(1-\tau))\pi + (1-\tau)^2 \pi \right\}$$
$$= \tau_3 \left\{ \tau (2-\tau) (w+\pi) \right\} + (1-\tau_3) \left\{ \tau^2 w + (1-\tau^2)\pi \right\}.$$

Player 3 mixes iff he is indifferent between voting H or T,

$$\begin{aligned} & \tau (2-\tau) (w+\pi) = \tau^2 w + (1-\tau^2)\pi \\ \rightarrow \quad & 2\tau (w+\pi) = 2\tau^2 w + \pi, \\ \rightarrow \quad & \tau = \frac{(w+\pi) \pm \sqrt{w^2+\pi^2}}{2w}. \end{aligned}$$

If $\pi > 0$, since $\tau \leq 1$, the only solution is

$$\tau = \frac{w + \pi - \sqrt{(w^2+\pi^2)}}{2w}$$

and $\tau \in (0,1)$ iff $\pi > 0$. If $\pi = 0$, then there is no (symmetric) mixed-strategy equilibrium.

Next, suppose players do not have the same label. Suppose two (players 1 and 2, without loss of generality) have label T and the other one (3) has label H. We want to show conditions for the existence of a (type-symmetric) mixed strategy Nash equilibrium. Suppose T-players vote T with probability τ and vote H with $1-\tau$. Suppose the H-player votes T with probability η and H with probability $1-\eta$. We write the expected payoff using the same logic as above.

The expected payoff of player 3 (label H) is

$$\begin{aligned}
&\eta\left\{\left[\tau^2\pi + 2(\tau(1-\tau))\pi\right] + (1-\tau)^2 w\right\} \\
&+ (1-\eta)\left\{\left[\tau^2 \times 0 + 2(\tau(1-\tau))(w+\pi)\right] + (1-\tau)^2(w+\pi)\right\} \\
&= \eta\left\{\tau(2-\tau)\pi + (1-\tau)^2 w\right\} + (1-\eta)(1-\tau^2)(w+\pi).
\end{aligned}$$

Hence player 3 uses a mixed strategy $\eta \in (0,1)$ iff

$$\begin{aligned}
&\tau(2-\tau)\pi + (1-\tau)^2 w = 2(\tau(1-\tau))(w+\pi) + (1-\tau)^2(w+\pi) \\
\rightarrow \quad &\tau = \tau^* := \frac{w - \pi + \sqrt{w^2+\pi^2}}{2w}.
\end{aligned}$$

It is easy to check that $\tau^* < 1$ $(=)$ whenever $\pi > 0$ $(=)$. Also, it is useful to note that $(1-2\tau^*)w + \pi < 0$ when $w/\pi > \sqrt{3}$.

If $\tau \neq \tau^*$, player 3 will optimally select $\eta = 0$ or $\eta = 1$. We have $\eta = 1$ iff

$$\tau^2 2w - 2\tau(w-\pi) - \pi \quad > \quad 0.$$

We now turn to the optimal strategy of a player with label T, i.e., a majority-label player. Given that the other T-player uses τ and the H-player uses η, the payoff of a T-player i using τ_i is

$$\begin{aligned}
&\tau_i\left\{\left[\tau\eta + \tau(1-\eta) + \eta(1-\tau)\right](w+\pi) + (1-\tau)(1-\eta) \times 0\right\} \\
+&(1-\tau_i)\left\{\left[\tau\eta w + \tau(1-\eta)\pi + \eta(1-\tau)\pi\right] + (1-\tau)(1-\eta)\pi\right\}
\end{aligned}$$

So the majority-label player is indifferent between playing H or T whenever

$$\begin{aligned}
&\left[\tau\eta + \tau(1-\eta) + \eta(1-\tau)\right](w+\pi) = \tau\eta w + \tau(1-\eta)\pi + \eta(1-\tau)\pi + (1-\tau)(1-\eta)\pi \\
\rightarrow \quad &\eta = \hat{\eta} := \frac{(1-\tau)\pi - \tau w}{(1-2\tau)w + \pi}.
\end{aligned}$$

Now note that:

- If $\hat{\eta} \in (0,1)$, we substitute $\tau = \tau^*$ to get the mixed equilibrium. Then $\hat{\eta} > 0$ whenever

$$\frac{\tau^* w - (1-\tau^*)\pi}{(2\tau^* - 1)w - \pi} > 0 \tag{1}$$

We will prove that $\hat{\eta} \in (0,1)$ is impossible. Start by considering $w/\pi > \sqrt{3}$. Here, the denominator of $\hat{\eta}$ is positive (as noted earlier). In this case $\hat{\eta} < 1$ requires $\tau^*\pi < \tau^* w - w$ which can never be the case. So the minority player mixing in this case is not an equilibrium when $w/\pi > \sqrt{3}$.

Now, consider $w/\pi \leq \sqrt{3}$. Here, the denominator of $\hat{\eta}$ is negative. So $\hat{\eta} > 0$ whenever $\tau^* w < (1-\tau^*)\pi$, which can never be the case when $w > \pi$. Therefore, we cannot have an equilibrium where both type of players use a mixed strategy, and in particular where the minority-label player will mix.

Next we turn to the equilibrium where the minority-label player uses a pure strategy, $\eta \in \{0,1\}$, while the majority-label players mix.

- If $\eta = 0$ then, using the indifference expression derived above, the majority-label player is indifferent if $\tau = \hat{\tau} := \pi/(w+\pi)$. This is an equilibrium iff the minority-label player prefers to choose $\eta = 0$, which requires

$$\begin{aligned} \hat{\tau}^2 2w - 2\hat{\tau}(w-\pi) - \pi &< 0 \\ \frac{\pi^2}{(w+\pi)^2} 2w - 2\frac{\pi}{w+\pi}(w-\pi) - \pi &< 0 \\ 2w\pi - 2(w^2 - \pi^2) &< (w+\pi)^2 \end{aligned}$$

which is always the case since by assumption $w > \pi$. Hence the minority-label player using a pure strategy (vote his label), and the majority player mixing, voting their label with probability $\pi/(w+\pi)$, is an equilibrium.

- If $\eta = 1$ then the majority-label player is indifferent only if $\tau = w/(w - \pi) > 1$, which cannot be the case.

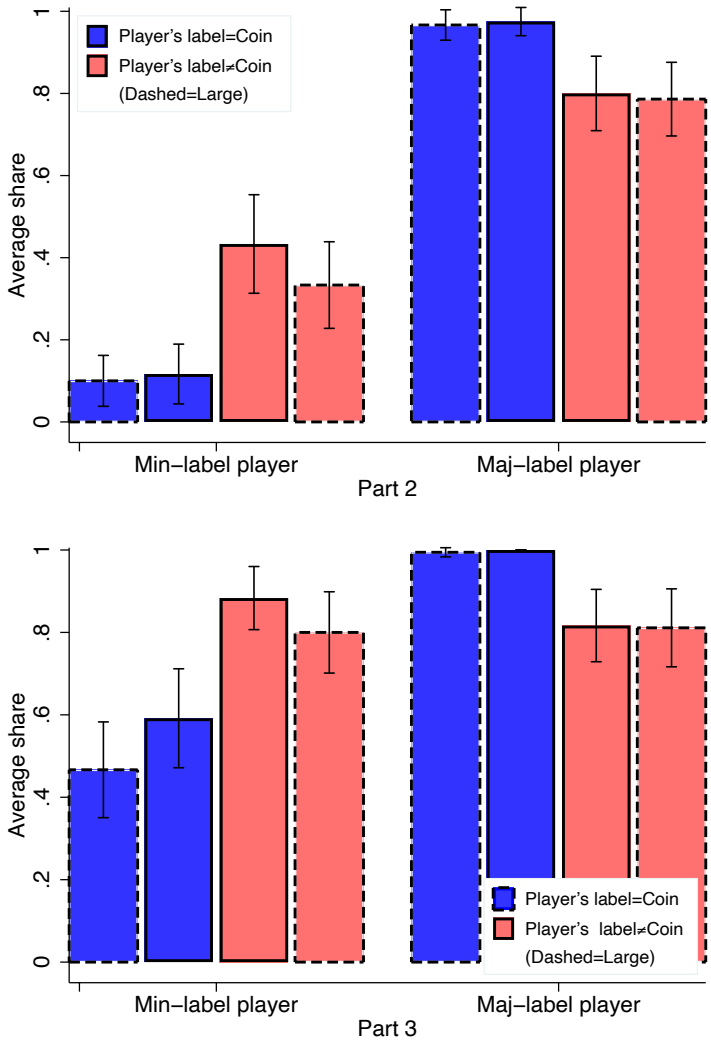
B Supplementary Material (for online publication)

Table B1: Distribution of Labels and Coin Images in Parts 2-3

	Labels in Part 2-3			**Coin**		
Trial	1/3	1/3	1/3	Part 1	Part 2	Part 3
1	T	H	H	H	T	H
2	H	T	H	T	T	H
3	H	H	T	H	H	T
4	H	H	H	H	H	T
5	T	T	H	T	H	T
6	H	T	T	T	T	H
7	T	H	T	H	T	H
8	T	T	T	H	H	T
9	H	T	H	T	T	H
10	H	H	H	T	T	H
11	T	T	H	T	T	H
12	H	T	T	H	H	T
13	T	T	T	T	T	H
14	T	H	T	T	T	H
15	T	H	H	H	H	T
16	H	H	T	H	T	H

Notes: Players had no labels in Part 1. The three "Labels" columns show the 16-trials sequence of labels for each player in the group in Parts 2 and 3. In each trial, 1/3 of players have one label (minority-label players), and 2/3 the opposite label (majority-label players). Hence, in a 15-player (3-player) group, there are 5 (1) minority-label and 10 (2) majority label players. The three "Coin" columns show the 16-trials sequence of coins shown to players: in each Part, each coin face was appeared in half of the trials.

Figure B1: Share of reports corresponding to the majority label.



Notes: 1 obs. = 1 subject in a Part 2 or 3 ($N = 60$). The average share of reports matching the majority label is shown conditional on the decisional situation of the subject: (i) with majority or minority label, (ii) where the coin image equals or differs from the subject's label. A dashed border distinguishes Large from Small groups.

Table B2: Average share of false reports (excluding fully honest).

		Same-label trials		**Mixed-label trials**			
		All labels		Minority player		Majority player	
		$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$
	Size	(1)	(2)	(3)	(4)	(5)	(6)
Part 2	Small ($N = 55$)	0.03	0.90	0.13	0.62	0.03	0.86
	Large ($N = 54$)	0.02	0.89	0.11	0.74	0.05	0.87
Part 3	Small ($N = 54$)	0.00	0.93	0.66	0.13	0.00	0.90
	Large ($N = 52$)	0.01	0.94	0.54	0.23	0.00	0.93

Notes: 1 obs.: 1 subject in a Part 2 or 3, excluding subjects who never misreported in that Part. The average shares reported are conditional on the type of trial: (i) same-labels (4/16 per Part, in col. 1-2) vs. mixed-labels (12/16 per part, in col. 3-6); (ii) minority-label subject (col. 3-4) vs. majority-label (col. 5-6), and (iii) the subject's label correspondence to the coin ($= \circledast$ means the player's label matches the coin, or else we use $\neq \circledast$). Note that the values in col. (2), (4), and (6) correspond to the average frequency with which a subject reported their own label in that decisional situation; in col. (1), (3), (5), this frequency corresponds to the complement of the values that are shown.

Table B3: Average decision time by decisional situation.

		Same-label trials		**Mixed-label trials**			
		All labels		Minority player		Majority player	
		$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$	$= \circledast$	$\neq \circledast$
Part	**Size**	(1)	(2)	(3)	(4)	(5)	(6)
2	Small	3.93	4.29	6.39	5.75	4.28	5.77
	Large	4.04	4.59	5.68	5.40	4.42	5.20
3	Small	3.24	3.87	5.08	4.73	3.92	4.62
	Large	4.15	5.25	6.51	6.57	5.71	5.69

Notes: 1 obs.: 1 subject in a Part 2 or 3 (N=60). The average decision time is reported in seconds and is conditional on the type of trial: (i) same-labels (4/16 per Part, in col. 1-2) vs. mixed-labels (12/16 per part, in col. 3-6); (ii) minority-label subject (col. 3-4) vs. majority-label (col. 5-6), and (iii) the subject's label correspondence to the coin ($= \circledast$ means the player's label matches the coin, or else we use $\neq \circledast$). Trials where the subject had an economic incentive to misreport are shaded.