PNAS RESEARCH ARTICLE | SOCIAL SCIENCES

Incentivizing free riders improves collective intelligence in social dilemmas

Ofer Tchernichovski[a,1], Seth Frey[b,c], Nori Jacoby[d,2], and Dalton Conley[e,1,2]

Edited by David Melamed, The Ohio State University, Columbus, OH; received July 6, 2023; accepted October 5, 2023 by Editorial Board Member Mark Granovetter

Collective intelligence challenges are often entangled with collective action problems. For example, voting, rating, and social innovation are collective intelligence tasks that require costly individual contributions. As a result, members of a group often free ride on the information contributed by intrinsically motivated people. Are intrinsically motivated agents the best participants in collective decisions? We embedded a collective intelligence task in a large-scale, virtual world public good game and found that participants who joined the information system but were reluctant to contribute to the public good (free riders) provided more accurate evaluations, whereas participants who rated frequently underperformed. Testing the underlying mechanism revealed that a negative rating bias in free riders is associated with higher accuracy. Importantly, incentivizing evaluations amplifies the relative influence of participants who tend to free ride without altering the (higher) quality of their evaluations, thereby improving collective intelligence. These results suggest that many of the currently available information systems, which strongly select for intrinsically motivated participants, underperform and that collective intelligence can benefit from incentivizing free riding members to engage. More generally, enhancing the diversity of contributor motivations can improve collective intelligence in settings that are entangled with collective action problems.

collective intelligence | crowd wisdom | social feedback | social dilemmas | computational social science

Significance

Collective intelligence processes such as voting and crowd-sourcing select for prosocial participants through their voluntary nature. In large-scale online experiments, we induced participation among free riding participants by manipulating incentives. We found that, in contrast to prior intuitions, recruiting such participants increases collective intelligence due to their better-quality ratings. Our results were robust across diverse populations and across virtual environments that simulate different real-world scenarios. We suggest that recognizing the collective action dimensions of collective intelligence can improve outcomes in systems by embracing the diversity in contributor motivations.

Collective intelligence (1) is driven by communication and cooperation (2) through increasingly technologically mediated environments (3–7), especially at larger scales. This is evident in peer-production systems (8, 9), collaborative filtering (10), social ratings (11), voting (12, 13), tagging (14), and other social mechanisms ubiquitous on the Internet (15). In such crowd-sourced systems, information is a public good vulnerable to free riding (16): that is, most people exploit information often but contribute evaluations rarely, if ever. The collective action problem of incentivizing information-sharing is typically passed over in studies of collective intelligence, despite the fact that all but a fraction of people free ride on crowd-sourced information goods in real world applications. Here, we merge these problems into a joint collective action/collective intelligence paradigm to investigate what features of technologically mediated environments (17–19) may facilitate or hinder collective intelligence when they are vulnerable to free riding (20–22). We use the term "free rider" without implying any antisocial motive; rather, we deploy the term simply to refer to participants with lower contributions to the group. Regardless of its drivers, free riding has the effect of under-provisioning the public information good. The challenge here is that even simple manipulations, such as introducing incentives, can affect the amount of information but also can change the population characteristics of responders in terms of diversity of motivation, skills, and personalities of responders. The critical question is how these changes in the amount, accuracy, and biases of evaluations add up to affect collective intelligence overall.

We evaluated the role of social preferences in securing efficient collective learning (23) by using incentives to measure the effect of participants who are intrinsically less motivated to contribute to the evaluation pool. Intuitively, because of their intrinsic motivation to serve the group, a cohesive population of cooperative, intrinsically motivated agents should outperform a mixed population including less motivated agents. Alternatively, a diversity of motivational styles may increase the diversity of information sources, which is known to increase collective intelligence (24) and stability (25). To study the interactions between collective intelligence and (intrinsic versus extrinsic) motivation, (26) we explicitly modeled the public good aspect of frequency and quality of information sharing. In contrast to prior studies, we did not provide performance-based incentives: The incentives we provide are to contribute, not necessarily to contribute well (27). Providing (and removing) incentives for mere participation within a public good provisioning scenario allows us to study the

Author contributions: O.T., S.F., N.J., and D.C. designed research; O.T., S.F., and N.J. performed research; O.T. contributed new analytic tools; O.T., S.F., N.J., and D.C. analyzed data; and O.T., S.F., N.J., and D.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: tchernichovski@gmail.com or daltonclarkconley@gmail.com.

[2]N.J. and D.C. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2311497120/-/DCSupplemental.

Published November 6, 2023.

direct effect of incentives on subjects in a collective action framework, while testing for the compositional effects on crowd wisdom due to variation in who participates. While acknowledging that many factors that we cannot control (or exclude) might drive these behaviors, we attempt to narrow down mechanistic options by supplementary experimental manipulations.

Results

Joint Collective Action/Collective Intelligence Games. We gauged collective intelligence by presenting participant groups with multiarmed bandit (MAB) problems, which have known optimal behavior (28–31). In each turn of our explicit MAB game, the participant selects one of two displayed arms (Fig. 1*A*) that are randomized from a pool of nine arms (*Explicit MAB GAME*: *artificial design* and *SI Appendix*, Figs. S8 and S9). Each arm dispenses a fixed (but initially unknown) number of coins once the lever is pressed. A dashboard next to each available lever presents the mean rating score. The goal is to select the arm with higher reward, which requires exploration and memory. The optimal strategy for finding the best arm is well understood (30, 31) but here we combine it with a public good problem: Participants can opt in to share ratings of bandit arms for the next five rounds (turns). Ratings were posted on a common dashboard, a shared public information good (Fig. 1 *A*, *Top*). In this manner, participants can opt in to help each other by sharing their experience (as with crowd-sourced recommendation systems) or avoid interacting with the information system and act based on experience alone. Once opted in (joining a club or guild of players, see *SI Appendix*, Figs. S9*B* and S11*C*), participants can either cooperate or "free ride," namely decline to provide publicly beneficial ratings. The choice of whether to rate or free ride was repeated in every turn, with no penalty for free riding (*SI Appendix*, Fig. S11*D*). In this manner, we can distinguish a decision to dissociate from the information system (opt out from the club and from the dashboard) from a decision to exploit the dashboard and free ride (use the information without contributing; *SI Appendix*, Fig. S1*C*). We estimate cooperation as the overall proportion of

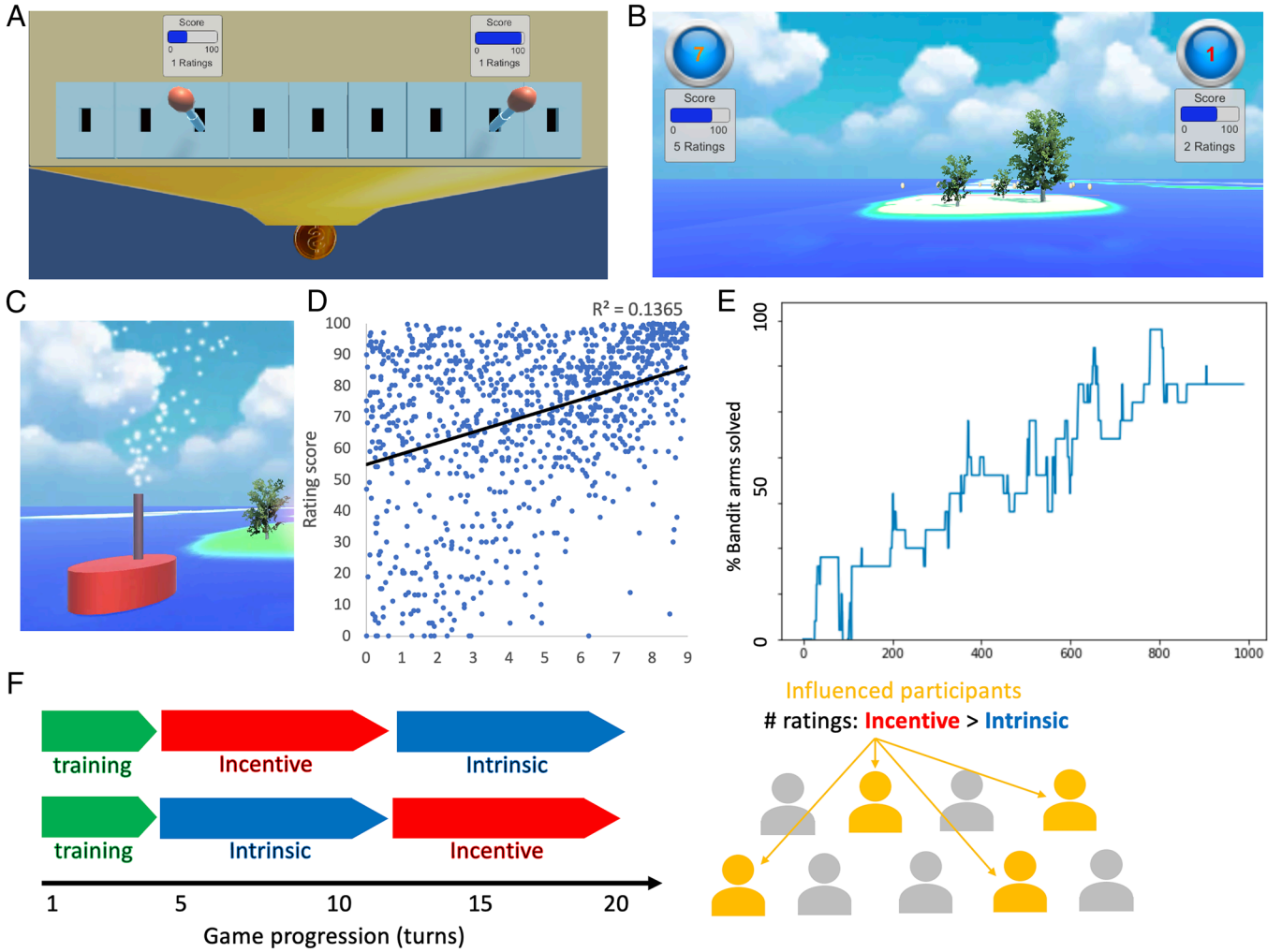

Fig. 1. Public good collective intelligence games expose the collective action dimension of crowd intelligence. (*A*) An explicit 9-arm-bandit game (MAB) with a voluntary recommendation system. In each turn, the participant selects one of two displayed arms (levers in *A*). Each arm dispenses a fixed (but initially unknown) number of coins once the lever is pressed. A dashboard next to each available lever or ferry presents the mean rating score. (*B*) A 9-arm MAB problem embedded in a 3D naturalistic game. In each turn, the participant selects one of two displayed arms (ferry type in *B*). (*C*) Participants collect coins on islands and select between two ferry types to commute between islands and then ride the ferries. The game includes a total of nine ferry types (=bandit arms) that vary in their speed. (*D*) Ratings scores vs. arm profit. Data are jittered to improve visualization. The line shows the linear trend. (*E*) Estimation of collective intelligence via accumulation of voluntary ratings. Time course of correct % bandit pairwise comparisons (above chance) by an observer of the rating scores presented in the dashboard over iterations. (*F*) Game stages schema (balanced order). At the end of the game, we identify influenced participants as those who rated more often once incentivized.

choosing to rate (#ratings/(#ratings + #declined + #opted-out). We could then test whether collectives improve faster (choose arms with higher reward) when it is encouraged, but not profitable, to provide information (the default in collective intelligence) or profitable—which may reduce the relative contribution of intrinsically motivated agents.

For the selection of intrinsically motivated agents to be ecologically valid, we developed a naturalistic 3D game (*MAB game*: *naturalistic design* and *SI Appendix*, Figs. S10 and S11) that does not explicitly reveal the collective intelligence problem and, instead, provides an experience similar to that of collective action problems in the real world (Fig. 1*B*), where people are not obliged to contribute evaluations (32). In the naturalistic game, participants choose and evaluate service utilities while exploring and gaining profit in the virtual world (these virtual gains were paid to participants as real money at the end of the game). In each turn, participants navigate in virtual islands to search and collect coins (that count toward their monetary reward for participation, Fig. 1 *B* and *C*). They commute between islands by riding simulated ferries. Each ferry type corresponds to an arm in the embedded MAB problem with a total of nine ferry types (=bandit arms). As in the explicit MAB game, in each turn, players choose between two arms (ferry types) selected from the overall nine, with a dashboard indicating aggregated rating scores for each ferry. Here, the arm reward is not monetary but the speed of the ferry: faster ferries save time that can be used for collecting more coins and thus increase subjects' revenues. In all other respects, the two games were identical.

We evaluate the rate of convergence of evaluations towards ground truth from the viewpoint of an observer of the dashboard who needs to choose between a random pair of arms. Rating scores were noisy (Fig. 1*D*), but as the game proceeded, the mean score gradually approximated the bandit arms' rewards (*Analysis variables*). We calculate the probability of choosing a higher-reward arm based on presented dashboard ratings that are available in each game iteration (turn). For example, after 10 turns, most arms are unknown, but after 100 turns, an observer who chooses an arm with a higher mean score is likely to gain more. Our estimate of the collective intelligence is the curve of the proportion of arms solved (selected correctly above chance, Fig. 1*E*).

Prior to the games, we prescreened participants to exclude bot usage and extreme inattentiveness (*Prescreening and bot detection*). In both games, participants voluntarily rated about a third of the turns (37%, n = 184 valid participants in the explicit MAB game and 35%, n = 115 valid participants in the naturalistic game). This proportion of rating behavior is comparable to the proportion of altruistic behavior reported in a long-term prisoner's dilemma game (2). In order to estimate the effect of this volunteer (nonrandom sample) on collective intelligence, we incentivized participants to rate by offering a monetary bonus in exchange for submitting each rating score. In one group (Fig. 1*F*), the incentive was presented in the first part of the game and removed in the second half, and in the other group, it was presented in the second half. This allowed us to evaluate the incentive effect on cooperation and on rating accuracy within-subject. We identified participants who responded to the incentive and contributed more ratings as "influenced" participants (Fig. 1*F* and *Subgroups definitions*). Therefore, the incentive influence can be used to identify ratings contributed by participants whose rating frequency was uninfluenced by rewards and ratings contributed mostly by influenced participants, who frequently free ride or opt out unless incentivized. We first test for an incentive effect regardless of order or type of disengagement and then explore the dynamic effect of adding and removing the incentive on different behaviors.

Cooperation vs. Rating Accuracy. In the naturalistic game (Fig. 2*A*) the incentive increased the proportion of ratings (pooled across participants) from 35% [CI: 30 to 42] to 70% [CI: 64 to 77] (*Bootstrap confidence intervals for pooled means*). At the individual level, about 46% of participants were influenced by the incentive and contributed more ratings. Interestingly, incentivizing less intrinsically motivated participants did not decrease but increased the pooled rating accuracy computed as R^2s of rating scores (pooled across participants) against the ground truth (see intrinsic vs. incentivized participants in Fig. 2*B*). We evaluate this outcome in a subset of 64 subjects who rated frequently enough to obtain a robust estimate of their accuracy, which is the within-subject R^2s of rating scores with the ground truth (*Analysis variables*). Rating accuracy was significantly higher in the influenced participants (Fig. 2*C*, mean and SEM R^2 = 0.44 [0.39 to 0.50] in influenced vs. R^2 = 0.20 [0.15 to 0.25] in uninfluenced participants, P = 0.002, Mann–Whitney U test, n = 38 and 28 uninfluenced). Further, ratings from less intrinsically motivated participants, who rated less than 50% of ferry rides, were more accurate compared to ratings from participants who rated more than 50% regardless of the incentive (Fig. 2*D*, P = 0.013, Mann–Whitney U test, n = 52 and 33 valid participants). In sum, participants who were reluctant to contribute to the public good provided higher quality ratings.

These outcomes were all replicated in a new group (n = 250 participants) who played the explicit MAB game (Fig. 2 *E* and *F*), although due to differences in game and reinforcement design (*Explicit MAB GAME: artificial design* and *MAB game: naturalistic design*), the incentive (bonus) was less effective and increased ratings contributions to only 51% [CI: 44 to 56] and only 29% of participants were measurably influenced by the incentive and rated more often. Consistently with the results of the naturalistic game, here too, incentive influence and low proportion of ratings were both associated with higher rating accuracy within subjects (Fig. 2 *G* and *H*). Quantitatively, we observed mean and SEM accuracy R^2 = 0.65 [0.58 to 0.73] in influenced vs. R^2 = 0.31 [0.27 to 0.35] in uninfluenced participants (P = 0.0039, Mann–Whitney U test, n = 53 and 31 uninfluenced). Similarly, participants who rated less than 50% were more accurate (R^2 = 0.47 [0.4 to 0.54] vs. R^2 = 0.34 [0.29 to 0.39] in participants who rated more than 50%, P = 0.054 Mann–Whitney U test).

Finally, in two additional groups who played the naturalistic game (n = 170 participants), we tested for an effect of stronger and weaker incentives and obtained similar results with no effect of incentive magnitude (within the range we tested) on rating accuracy within influenced and uninfluenced participants (*SI Appendix*, Fig. S2).

The experimental design allows us to distinguish between less intrinsically motivated participants who opt out from the information system and those who joined but (at least occasionally) declined to contribute ratings (free riders). We tested it by performing meta-analysis across groups of naturalistic game participants (n = 444 valid participants and 7,596 rating scores). The incentive affected both the decision to opt out and the decision to free ride and exploit the dashboard: Incentives increased participation (opting in joining the club decisions) from 66 to 74%, and reduced free riding (within participants who opted in) from 23 to 10%. Interestingly, rating accuracy was higher in participants who free ride (Fig. 3*A*, t = 2.7, P = 0.008) but slightly lower in participants who opted out (Fig. 3*B*, NS). This suggests that the positive effect of the incentive on rating accuracy may stem from incentivizing the free riders—who were already engaged with rating information but were reluctant to contribute to it without a reward for their ratings.

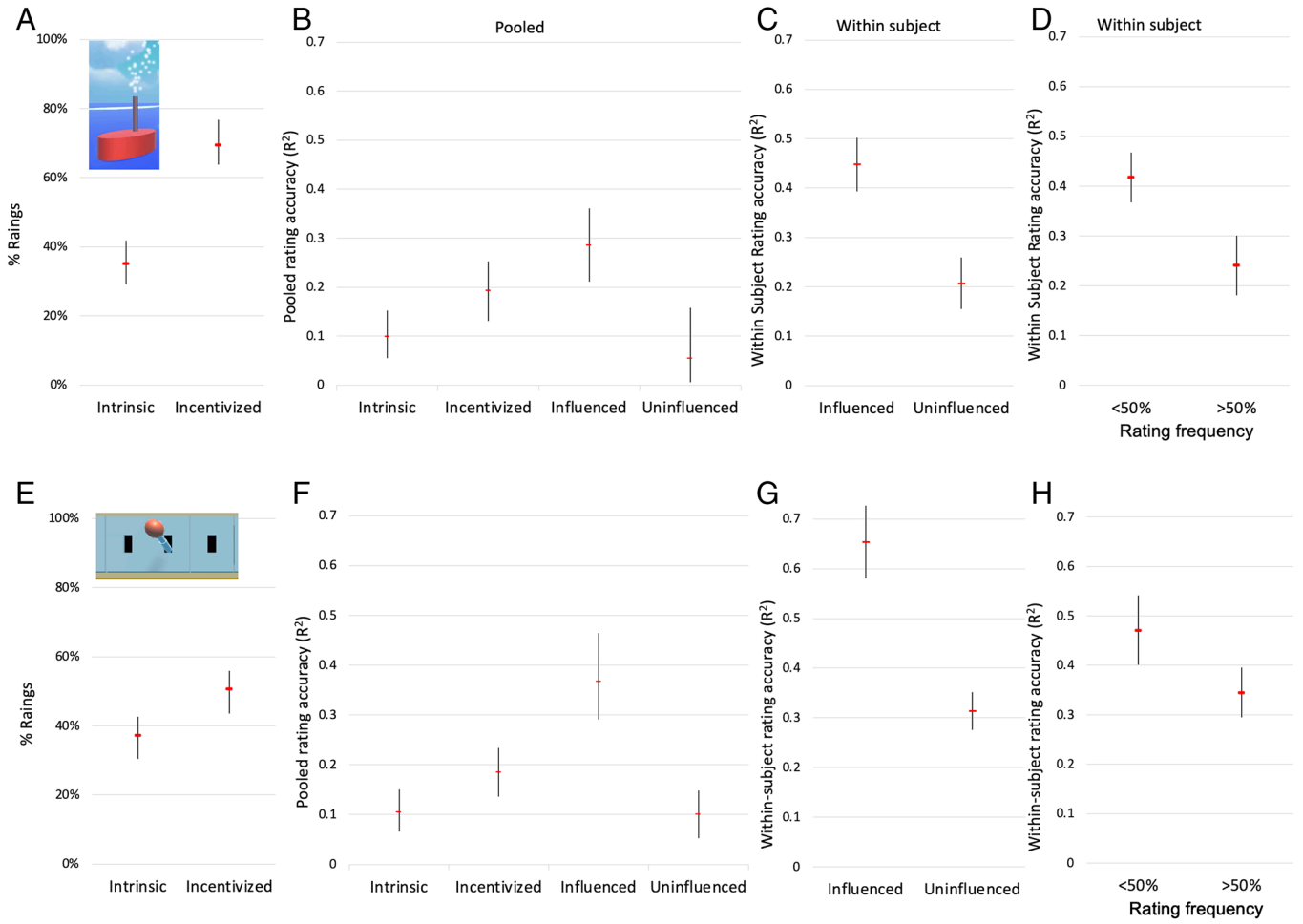

Fig. 2. In both the naturalistic and formal conditions, intrinsically motivated participants provided less informative ratings than those who were responsive to incentives. Naturalistic game: (*A*) Intrinsic vs. incentivized percentage of ratings. (*B*) Pooled rating accuracy (R^2s with ground truth) with 90% bootstrap error bars comparing intrinsic vs. incentivized participants. We also present rating accuracy separately for participants that were influenced by the incentives compared to those who did not rate more often once incentivized. (*C*) Within subject rating accuracy comparing incentive-influenced to uninfluenced participants. (*D*) As in (*C*), comparing participants of low (<50, less intrinsically motivated) vs. high (>=50, more intrinsically motivated) proportions of rating (mean and SEM). Explicit MAB game: *E*–*H*, same as *A*–*D* for the explicit MAB game.

Finally, we look at the dynamic effects of adding and removing incentives by performing a meta-analysis across groups of naturalistic game participants (*SI Appendix*, Fig. S1*D*). Removing the incentive during the second half of the game strongly decreased the number of ratings from influenced participants (from 803 to 276), but it did not affect the pooled rating accuracy of these influenced participants (R^2 = 0.21 [0.16 to 0.27] during incentive vs. R^2 = 0.17 [0.11 to 0.24] after incentive removal, NS). Similar results were obtained when the incentive was added in the second part of the game (*SI Appendix*, Fig. S1*E*). Comparing participants who joined the game early or late did not reveal any differences in rating accuracy either (*SI Appendix*, Fig. S1*F*). Overall, results suggest that the incentive simply boosted the number of ratings from influenced participants (indicated as numbers in *SI Appendix*, Fig. S1). Differences in the pooled rating accuracy can be explained by changes in the composition of participants who contributed to the public good, with stronger influence from participants who were less intrinsically motivated during the incentive.

Rating Accuracy and Collective Intelligence. The incentive accelerated the MAB improvement because rating scores became more abundant and more accurate. To isolate the relative contribution of rating accuracy, we used the naturalistic game data to simulate the progression of the MAB solution. In the simulation, we run the same number of turns as in the real data, bootstrapping game turns with replacement. The simulated agent chose arms with higher median scores or randomly if ratings were not available for both arms or were equal. We ran the simulation in two manners to assess the relative contribution of rating quantity (cooperation level) and accuracy. In the first simulation, the agent based its choice on the objective reward of the arms (perfect rating). We used this type of information that is not available to the real participant to evaluate the effect of ratings quantity. In the second way, we run the simulation, the agent's choices were based on the bootstrapped real (actual) ratings. The difference in performance between the two simulations estimates the effect of ratings accuracy. That is, it estimates the residual contribution of rating accuracy to the solution within a given amount of rating scores (level of cooperation). Based on cooperation level alone (Fig. 4, blue curves), the higher cooperation rate in the incentivized group could have allowed them to solve half of the bandit arms within nine iterations, compared to 20 iterations in the control (intrinsic) group. But with real ratings, it would have taken about 80 iterations in the incentivized group (Fig. 4*A*, green curve), and more than 500 in the control group (Fig. 4*B*, red curve). Note that the nonsimulated data from the experiment (Fig. 4*C*) show similar trends to the simulated data (Fig. 4*C*, green curve and Fig. 4*B*, green curve; Fig. 4*C*, red curve and Fig. 4*A*, red curve),

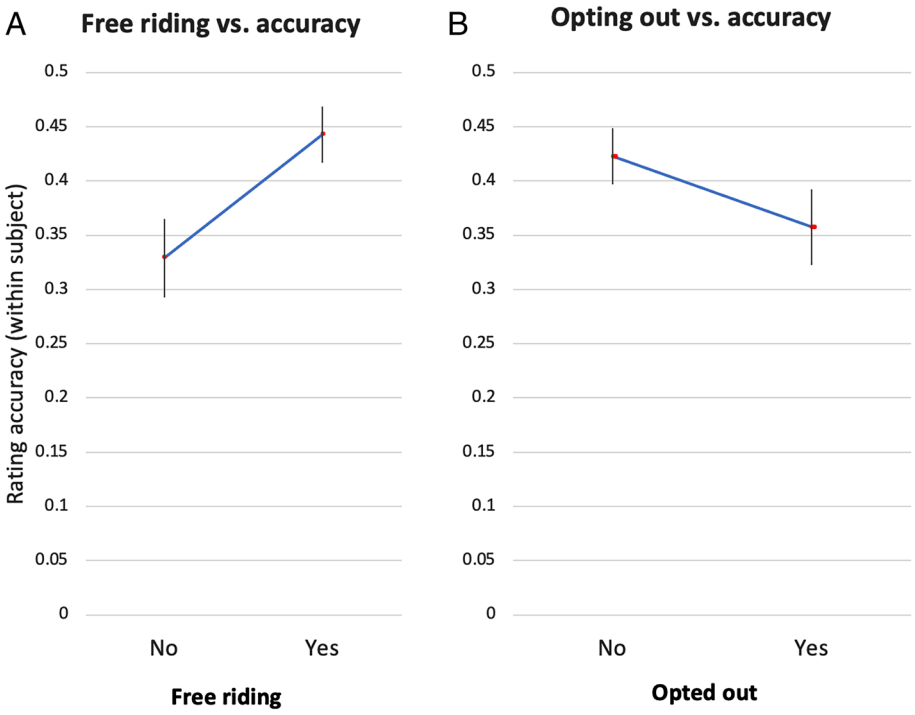

Fig. 3. Rating accuracy in free riders and opted-out participants: (*A*) Rating accuracy in participants who never free ride (n = 109) vs. those who did free ride (n = 190). Free riding was associated with more accurate ratings (*t* test, t = −2.7, *P* = 0.008). (*B*) Participants who never opted out were slightly more accurate than those who did (t = 1.5, *P* = 0.13, NS).

with some nonmonotonic behavior indicating temporal effects that are not captured by the simulation. These results suggest that rating accuracy is a major impeding factor to collective intelligence in public good problems.

Numerosity vs. Rating Accuracy. A possible explanation of our results so far is that participants who rated less frequently are rational players who, on the one hand, are more skilled in estimating (and memorizing) the bandit arm rewards and, on the other hand, decide to either opt out or to free ride (*Subgroups definitions*) as contributions do not increase their monetary compensation. To test for that, we introduced a simple quantity estimation test (Fig. 5*A* and *SI Appendix*, Fig. S7) prior to the public good game. We hypothesize that if participants who rated less frequently are indeed rational and more skilled players, they should also perform better on the quantity estimation test. In this manner, performance in the estimation test would be predictive of accurate rating in the public good MAB game. We therefore compared the evaluation accuracy across the quantity estimation and public good games. Since ratings in the public good games are voluntary, we could only include data from participants who rated public goods enough times (>5, see *Analysis variables*) to obtain a stable estimate. In the explicit MAB game, the quantity estimation task predicted only 6.8% of the variance in rating accuracy within subjects (Fig. 5*B*, n = 84 valid participants with sufficient data to estimate accuracy, *P* = 0.02, Pearson correlation). In the naturalistic game, it predicted a negligible portion of the variance in rating accuracy (about 1%, Fig. 5*C*, n = 78 valid participants, NS). Overall, the quantity estimation task poorly predicted rating accuracy in our public good games.

To test whether improving the quality of participants can improve prediction, we replicated the experiments with a cohort

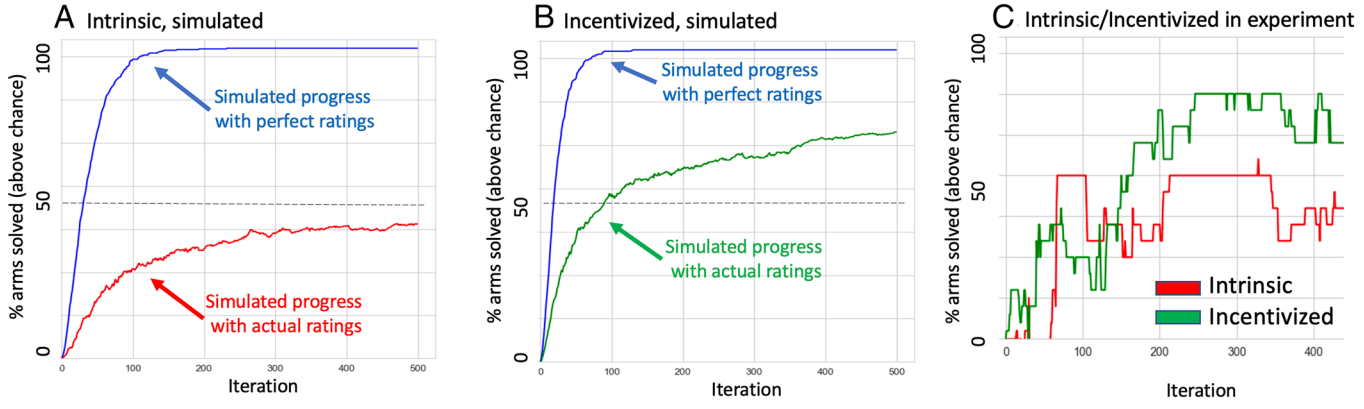

Fig. 4. Simulations show that relying on intrinsic motivation of intrinsically motivated participants slows the collective's convergence on the game solution. Bootstrap simulation of collective intelligence comparing intrinsic (*A*) vs. incentivized (*B*) cohorts. Data is from the naturalistic game. Blue curve simulates the progression based on % ratings with an agent that uses the objective arms reward. The red/green lines simulate the progression based on the actual ratings (sampled with replacement from n = 287 intrinsic ratings and 519 incentivized ratings). The gap between the lines represents the marginal effect of rating accuracy, given the amount of rating scores. (*C*) Direct comparison of the real time experimental results.

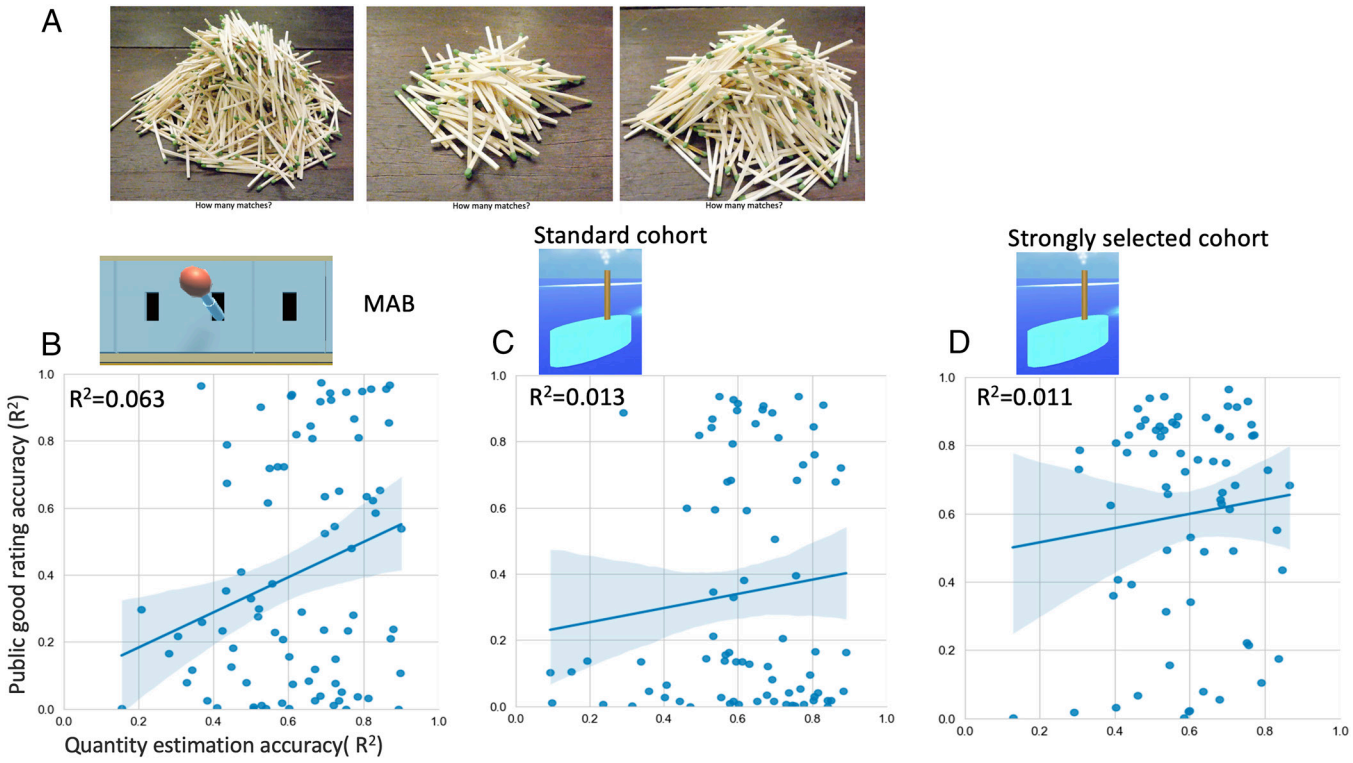

Fig. 5. Numeracy does not account for the lower accuracy of ratings by intrinsically motivated participants. (*A*) Prior to the public good game participants were presented with a quantity estimation task, where they estimated the number of matches in 12 images presented in a pseudo-random order. (*B*) Scatter plot of participant R^2 in the quantity estimation task against participant's R^2 in the MAB game (n = 84 valid participants) shows a positive correlation between rating accuracy and basic estimation skills. (*C* and *D*) The same effect as in *B* for the ferry games with 95% approval (*C*, n = 78 valid participants) vs. 99% approval (*D*, n = 69 valid participants) cohorts.

of strongly selected participants (MTurk workers with >99% approval over more than 5,000 completed tasks). Ratings provided by the strongly selected cohort were much more accurate (mean and SEM within-subject $R^2 = 0.60 \pm 0.04$ in the strongly selected cohort vs. $R^2 = 0.34 \pm 0.04$ in the standard cohort, *SI Appendix*, Fig. S3), but here too, rating accuracy in the quantity estimation task did not predict rating accuracy in the public good game ($R^2 = 0.01$, Fig. 5*D*, n = 69 valid participants, NS). Further, there were no statistically significant differences between uninfluenced and incentive-influenced participants in quantity estimation accuracy of matches in a pile (mean $R^2 = 0.57$ in uninfluenced vs. $R^2 = 0.64$ in influenced, NS). Why would a participant who performed well in the quantity estimation task provide inaccurate ratings in the public good game? One option is that different subpopulations of participants differ in their rating biases in the public good game and that rating biases can potentially interact with rating accuracy.

Rating Bias vs. Rating Accuracy. To evaluate whether differences in rating biases can explain the higher accuracy in the influenced participants, we tried to identify idiosyncratic differences in mean rating scores across participants. We tested for a correlation between participants' mean rating scores (an estimate of participant's positive or negative rating bias) and rating accuracy. In the explicit MAB game, the within-subject mean rating score was negatively correlated with rating accuracy, explaining 19% of the variance (*SI Appendix*, Fig. S4, $R^2 = 0.19$, $P < 0.001$, Pearson correlation). A similar effect was detected in the naturalistic game: Mean rating scores were negatively correlated with rating accuracy, explaining about 27% of the variance (Fig. 6*A*, n = 78 valid participants, $R^2 = 0.27$, $P < 0.001$, Pearson correlation). These results show that ratings from more accurate people were more likely to be biased downward, thus suggesting that rating bias can explain at least some of the variance we observed in rating accuracy between the influenced and uninfluenced participants. Indeed, the incentive-influenced participants showed a negative rating bias (*SI Appendix*, Fig. S6*D*). We therefore suggest that social bias rather than cognitive factors mediate the differences in rating accuracy across influenced vs. uninfluenced participants in the public good game.

We considered that the near-zero rating accuracy in participants who rated very high could be a trivial ceiling effect. It could either indicate a strong positive bias in attitude (toward both participation and evaluation), or mirror a "lazy" strategy of rating high by default, driven perhaps by a moral self-licensing effect in which the prosocial act of being willing to rate gives license for the self-interested act of rating poorly (33). We tested whether this is the case using two approaches: First, looking at the cohort of strongly selected workers (with 99% approval), rating accuracy was much higher and mean rating scores were lower compared to our standard cohort, but the trend remained similar (Fig. 6*B*). It persisted even when considering only participants whose mean rating scores did not exceed 80% (Fig. 6*B*). A second approach to testing for a "lazy ratings" effect is by imposing extra effort (time costs) for submitting high scores. These time-costs were shown to be salient in a previous study (34). To test the effects under time costs, we ran additional experiments (N = 100 valid participants) where participants submitted ratings using a costly slider (34). The costly slider requires pressing continuously on the buttons of a "crawling" slider. Note that the time cost is proportional to the deviation from the center, requiring an increasing investment of time to report extreme scores. As expected, the time cost reduced the mean rating scores (histogram in Fig. 6*C*), but it did not eliminate the negative trend (Fig. 6*C*, n = 52 valid participants,

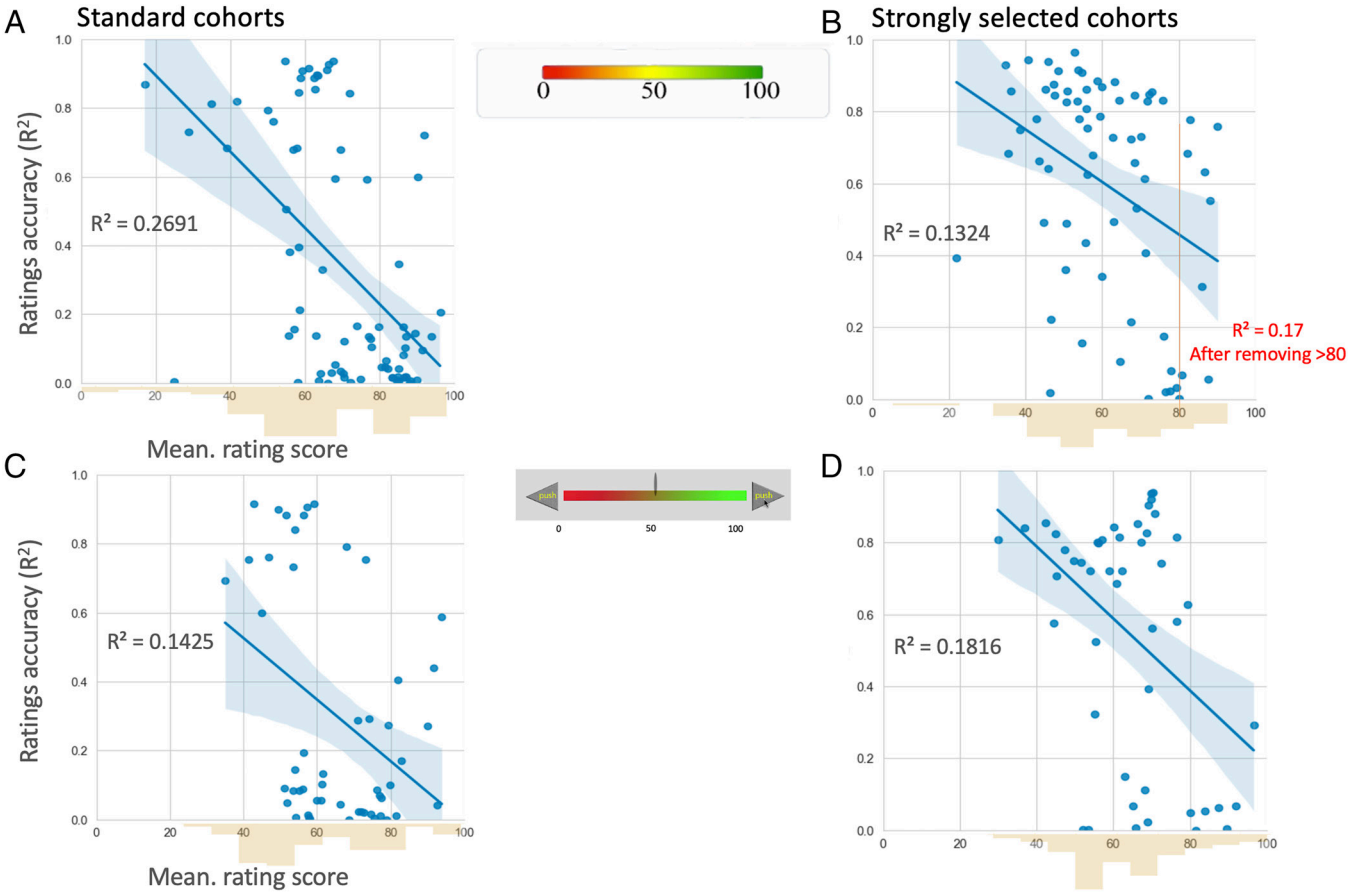

Fig. 6. Mean rating scores and rating accuracy are negatively correlated across conditions. Scatter plot of mean rating scores vs. mean rating accuracy. Each marker presents a mean rating score and accuracy for one person. Correlations in all panels are all statistically significant ($P < 0.001$); histograms are of mean rating scores. We compare two treatments: cost-free ratings (clickbar, *A* and *B*) and costly ratings (slider, *C* and *D*). For each treatment, we recruited standard cohorts and strongly selected cohorts (99% approval, right). For costly signaling, participants rated using a slider with two buttons (center). Pressing a button continuously moves the cursor slowly, with a time cost of up to 3 s for reporting extreme scores.

$R^2 = 0.14$, $P < 0.001$). Finally, we combined both treatments (highly selected group + costly signaling), and again the negative correlation was replicated (Fig. 6*D*, n = 48 valid participants, $R^2 = 0.18$, $P < 0.001$, Pearson correlation).

As an additional validation, we wanted to confirm that the lower mean scores by influenced participants did not stem from their choosing less profitable MAB arms. On the contrary, we found that influenced participants chose slightly better MAB arms (*SI Appendix*, Fig. S6 *A* and *C*, NS). Although the influenced participants chose slightly more profitable arms, their mean rating scores were still lower, confirming a negative bias, one that was even stronger in the naturalistic games (*SI Appendix*, Fig. S6 *B* and *D*).

Rating Accuracy vs. Survey Responses. So far, we found that our incentive improved rating accuracy by recruiting ratings from a subpopulation of participants who rate more accurately. We also showed that participants who tend to free ride, rate more accurately. However, all these estimations are intrinsic to the game and subject to collider bias, e.g., free riders might also be more rational or more attentive. We therefore tested whether rating accuracy is associated with specific personality traits. In a new group (n = 131 participants), we performed a Big Five personality test (35) and altruism assessment test (36) after participants played the naturalistic game. We cannot exclude the possibility that the game might have biased the survey responses to a certain extent. Within these limitations, we found that rating accuracy was negatively correlated with altruism score (r = −0.62, $P < 0.05$, False Discovery Rate (FDR) adjusted, *SI Appendix*, Fig. S5*A*), but not with quantity estimation accuracy (*SI Appendix*, Fig. S5*B*). We also found a negative correlation between altruism score and incentive influence (n = 68 valid participants, R = −0.33, FDR adjusted $P = 0.03$, *SI Appendix*, Fig. S5), confirming that the incentive recruited ratings from participants with lower altruistic tendency—who tend to rate more accurately. We also found a positive correlation between openness score and incentive influence (n = 68 valid participants, R = 0.35, FDR adjusted $P = 0.02$), which suggest that personality traits other than altruism might also explain accuracy in rating score. All other personality measures were not significantly correlated with the incentive influence.

Discussion

The joint collective action/collective intelligence games revealed strong interactions between incentives, free riding behavior, and collective intelligence. For example, free riders who were responsive to incentives provided higher quality evaluations and balanced out the over-optimistic ratings of more intrinsically motivated contributors. Recruiting ratings from less intrinsically motivated participants decreased the average rating score, which, most likely, brings the information system average closer to the population mean (the ground truth). A recent study shows that incentives can be used to reinforce accurate evaluations (27) within subjects. Here, we show further that even an incentive that is not designed to (and did not) alter evaluation accuracy within subjects can still

increase collective intelligence, simply by adding more ratings from differently motivated participants into the pool of evaluations. This is significant because it is often impractical to reinforce accurate evaluation, especially when there is no ground truth against which to anchor—which is the case in most real-world applications. More generally, manipulating signaling incentives and costs can be used for identifying features of technologically mediated environments that either hinder or facilitate collective intelligence, even when the ground truth is not accessible in real time. Such manipulations can also potentially reveal hidden sampling endogeneity effects (37) that might compromise the reliability of crowd-sourced evaluations.

Our study design allows for estimation of collective intelligence in a public-good game that is vulnerable to free riding. However, our intrinsic measures (e.g., free riding) are merely descriptive, and we cannot infer motivation. Although free riding behavior was correlated with higher rating accuracy, and although rating accuracy is also negatively correlated with altruism score, we cannot infer causality, and our evidence regarding motivation should be interpreted cautiously. In particular, free riding is a rational behavior, and rational behavior is likely to be associated with rating bias. Other mechanisms cannot be excluded either. For example, some competitive participants might submit many inaccurate ratings in order to gain relative advantage. Conversely, some intrinsically motivated participants may submit many evaluations in order to better memorize which arms were more profitable. The results of the numerosity skills tests and altruism survey suggest that such cases are likely to be the minority, but they cannot be excluded. Within these limitations, our results suggest that manipulating incentives can improve information quality and correct biases that otherwise hinder collective intelligence in a collective action setting. We don't know whether future studies with stronger extrinsic validations will be able to positively identify a single motivation and exclude all others, but they can hopefully help in further narrowing down the list of viable candidate mechanisms.

Differences in rating biases between participants who submit ratings more or less frequently may vary in different scenarios, but even in our simple virtual environment, where priors are presumably weak, biases, and differences in accuracy across subgroups were surprisingly strong. We suspect that intrinsically motivated participants underperform because of their stronger bias. If true, adding evaluations from less intrinsically motivated participants is likely to be beneficial in many complex real-world scenarios. Since many of the currently available information systems strongly select for prosocial and intrinsically motivated participants through their voluntary nature, they are likely to underperform. In these situations, collective intelligence may benefit from incentivizing even a fraction of free riders to engage.

An additional open question is whether the improved information quality we observed is sustainable. According to previous studies, external incentives may "crowd out" evaluations from intrinsically motivated participants (38) with potentially long-term negative effects. However, a recent massive study (39) did not detect any negative unintended consequences of offering a payment incentive for getting COVID vaccine over prolonged periods and across multiple domains. One limitation of that study is that the payment was low and increased vaccination rate by only about 5%. In our study, the incentive increased participation by about 30%, also with no apparent side effects. We replicated this outcome in a range of incentive magnitudes (*SI Appendix*, Fig. S2) and found that all incentive magnitudes increased rating accuracy. However, we cannot exclude that in natural conditions incentives might reveal negative outcomes over time.

Understanding how incentives may affect collective intelligence is particularly important for testing new voting systems, such as liquid democracy (40). Campbell et al. (13) show that collective intelligence decreases with liquid democracy because people tend to delegate too much, reducing the effective sample size for crowd wisdom. This negative effect overwhelmed the positive effect of delegation to good experts. Our findings suggest a different perspective to this: In a voluntary rating system where people tend to exploit, incentivizing people to vote may have a positive effect because people who vote less are often more like "experts"—they rate more accurately. Perhaps an optimal level of incentives (41) in liquid democracy can reduce delegation to a level that maximizes the proportion of accurate evaluations in the pool while maintaining an effective sample size.

In the wild, collective intelligence platforms like online crowd-sourced ratings and peer production systems are public goods that come with the risk of collective action problems like free riding. By capturing this explicitly in an experimental design, we observe the interactions between collective action and collective intelligence to find that mixed populations of participants that diverge in terms of their motivation outperform (show greater group intelligence than) entirely intrinsically motivated populations, suggesting that motivational diversity may be just as important as diversity in knowledge and experience in leveraging crowds towards optimal collective outcomes. Our results demonstrate the promise of large-scale online experiments to contribute to the understanding of collective intelligence challenges that are entangled with collective action problems.

Materials and Methods

All data and code for the paper can be found in the Open Science Foundation (OSF) repository: https://osf.io/d953m/.

Participants.

Recruitment and participants. Participant recruitment was managed by *PsyNet*, a framework to develop and deploy large-scale online experiments. We recruited a total of 1,281 participants. For each experiment, participants were recruited automatically until we collected the desired sample. Sample size was determined by a sample-size analysis (see *Sample-size power analysis* below). After excluding participants that are suspected to have used bots (see below) and participants who did not finish the games, we included 721 participants who made 15,028 decisions. All participants provided informed consent in accordance with the Max Planck Society Ethics Council approved protocol (2021_42). All participants were recruited online using Amazon Mechanical Turk (MTurk). We required three conditions to take part in our experiments: i) be at least 18 y old, ii) be in a quiet environment (e.g., a room with low background noise), and iii) use an up-to-date Google Chrome browser. These requirements guaranteed compatibility with our testing platform, PsyNet (see ***PsyNet*** below). Participants were paid at a US $9/h rate according to how much of the experiment they completed (e.g., if participants left the experiment early, they were still paid proportionally for their time).

Prescreening and bot detection. Relying solely on highly curated participants would have defeated the purpose of the current study, which aims to understand both motivated and unmotivated participants. In MTurk, worker quality, demographic composition and motivation of participants are highly variable, which can properly simulate a public good problem in a natural and diverse environment, with some proportion of participants that try to minimally engage or even exploit the experimental system and compensation mechanisms. Still, we needed to exclude participants who use bots or were otherwise completely disengaged. We also needed to test whether the quality of recruits affected our findings. To address that, we implemented the following prescreening task, which we also used as a treatment:

All participants performed a simple quantity estimation task prior to the game (see procedures below for details). We used the quantity estimation task as a bot/attention detector: Our Javascript code exposed the correct answers, such that bots (but not ordinary participants, at least without trying to inspect the javascript

code manually) could fetch those numbers and match them with each image. In a similar manner, we detected and excluded data from 11 cases that we suspected as either less sophisticated bots (most excluded participants) or inattentive participants that provided very low quality of quantity estimations $R^2 < 0.2$ (less than 1% of the excluded participants). In this manner, we detected 11.3% of bot usage in standard quality cohorts but only 1.6% of bot usage in the strongly selected cohort. Once a bot usage was identified, we excluded the bot-generated data. However, to prevent adaptation, we allowed suspected Bots to finish the task and fully compensated them.

Other exclusions. Each MTurk worker could only participate in each study once.

Standard participant cohorts. Participants with at least 2,000 previously submitted tasks on MTurk with a 95% approval rate on average.

Strongly selected participant cohorts. Participants with at least 5,000 previously submitted tasks on MTurk with a 99% approval rate on average.

Details of participants' groups. Overall attrition rate was 560 out of 1,281 participants. Most attrition was due to technical failure to run the WebGL game on the browser. We only included participants who performed the full experiment in the analysis. The details of the participant groups in each experiment are provided in Table 1.

Asynchronous experimental design. We run all experiments asynchronously, such that participants join the game, do their task, and leave the system independently. This means that participants who joined early experienced a different information system than those who joined late. We did not detect strong differences in outcomes between participants who joined early, and those who joined late within treatment arms. For example, comparing within-subject rating accuracy across uninfluenced and influenced participants shows similar differences, when restricting the analysis to participants who joined early, or to those who joined late in the game (*SI Appendix*, Fig. S1*F*). Most other measures including #Ratings, #Opted-out, %Free ride, Incentive influence, and #Coins collected were within the margin of error. However, the mean rating score was somewhat higher in participants who joined late (63.8 ± 1.1 in early vs. 67.4 ± 1 in late participants), probably due to priming.

Procedure.

PsyNet. The experiments were all implemented using PsyNet (https://psynet.dev/) (42), a framework for complex experiment design which builds on the Dallinger (https://dallinger.readthedocs.io/), a platform for online automatic participant recruitment.

Game programming and design. Online games were deployed by combining PsyNet and Unity®. We developed an interface for embedding WebGL compiled 3D Unity® games in our PsyNet experiments. All games were designed and developed by our team. Participants interact with the experiment via a web browser, which communicates with a back-end Python server cluster responsible for organizing the experiment and communicating with WebGL Unity games. In our experiments, this cluster was managed by Heroku (https://www.heroku.com/), supporting the experiment management and stimulus generation workload, as well as a Postgres database for sorting results. All groups were presented with detailed video instructions to decrease the reliance of specific terms in the experience of participants. Code for the implemented experiments can be found in the OSF repository: https://osf.io/d953m/.

Quantity estimation task. Following Tchernikovski et al., participants were sequentially presented with 12 images showing piles of matches (34). For each image, we prompted the participant to estimate the number of matches. To avoid responding without paying attention to the image, participants had to wait at least 12 s before responding. After submitting an estimation, we presented the correct answer.

> 1. Task instructions: You will play two games. This first is a game of guessing the number of matches in an image. We will present 12 images, and pay you 5 cents for each image you estimated.
>
> 2. We require that you spend at least 12 s estimating each image. This was enforced by enabling the Next button (*SI Appendix*, Fig. S7) after 12 s.

Explicit MAB GAME: artificial design.

Game design. Bandit arms provide a fixed reward (one to nine coins). After instruction slides, participants play 20 turns, where in each turn, they choose an arm of the two available options, and receive a reward and feedback based on their choice.

Table 1. Group composition

Group	Participants	Valid	Evaluations	Ratings
1. Explicit MAB	250	182	3,640	1,550
2. Naturalistic, stand selection	212	114	2,436	1,121
3. Naturalistic, strongly selected	122	85	1,793	946
4. Naturalistic, CostSlider, stand selection	274	107	2,256	963
5. Naturalistic, CostSlider, strong selection	122	70	1,470	825
6. Naturalistic, small/large bonus	170	95	2,003	953
7. Naturalistic, personality tests	131	68	1,430	712
Total	1,281	721	15,028	7,070

#participant is the number of recruited participants. #valid is the number of participants who finished the experiment and were not excluded as bots. #Evaluation is the number of human decisions if to rate the MAB/ferry. #Ratings is the actual number of rating scores submitted.

Game progression. The first step in each experiment is an informed consent. After the informed consent, participants see general instructions (*SI Appendix*, Fig. S8). Then, they are introduced to a waiting room which introduces a delay of 5 to 20 s, which is used here only for compatibility with future synchronized games (*SI Appendix*, Fig. S9*A*). Then, they are proposed to join the club (or guild) (*SI Appendix*, Fig. S9*B*), and in this case they will be further asked to perform rating, or not join the club and in this case they will not see ratings and not be asked to rate. Next, we present the status of the incentive by either presenting "We now pay for rating the arm" or removing incentive "We no longer pay for rating." Finally the game starts (*SI Appendix*, Fig. S9*C*), and the turn starts with a choice based on the rating information (if available) and based on the incentive (if available; *SI Appendix*, Fig. S9*D*). In all cases, participants received reward and feedback after the end of each trial (*SI Appendix*, Fig. S9*E*), and in case they opted in (participate in the club), they are further asked to choose to rate or skip rating. In case that the participant chooses to rate, they are introduced with a rating device, such as a clickbar (*SI Appendix*, Fig. S9*F*) which they can move. In the experimental condition "costly slider" the participants are not free to move the slider directly but instead they need to press continuously on the push buttons (*SI Appendix*, Fig. S9*G*).

Incentive design. For each five turns (before turns 1, 6, 11, and 16), the participant was prompted to choose if to join a player's club or play singly (*SI Appendix*, Figs. S9 and S13). Once joined a club, the player should decide in each turn whether to rate arm or not (*SI Appendix*, Fig. S8*C*). There was no penalty for free riding. Incentive to rate was introduced either for turns 1 to 10 (early incentive) or for turns 11 to 20 (late incentive).

MAB game: naturalistic design.

Game design. Bandit arms (different ferries) provide a fixed reward, with ride duration (speed) ranging between 2 and 20 s. The game included 20 turns. After instruction slides, participants played 4 turns of training, followed by two sessions of eight turns each (Fig. 1*F*).

Game progression. The first step in each experiment is an informed consent. After the informed consent participants were presented with the instructions for the game via a short video and slides (*SI Appendix*, Fig. S10). Then, participants went through a guided training phase where they experienced the game under a controlled environment that makes sure they understand the task and calibrates their expectations. In the first trial, they collected coins (*SI Appendix*, Fig. S10*A*) and experienced the fastest ferry (*SI Appendix*, Fig. S10 *B–D*). Then, they experienced riding the ferry and were prompted about the second trial (*SI Appendix*, Fig. S10 *C–D*). In the next two trials, the participant was also introduced with

selecting a MAB arm and rating it after the ferry ride (*SI Appendix*, Fig. S10*E*). Participants were informed about the end of the training phase. Next, we presented the status of the incentive by either presenting "We now pay for rating the ferry" (*SI Appendix*, Fig. S11*D*) or removing the incentive: "We no longer pay for rating the ferry." Next, participants make a choice of whether to join the ferry club (this opt-in/opt-out choice repeat every 5 turns) (*SI Appendix*, Fig. S11 *A–C*) or without incentive (*SI Appendix*, Fig. S12 *C* and *F*).

If participant joined the club (opted in), the dashboard was presented. After riding, the player could still decline rating without sanctions (*SI Appendix*, Fig. S11*D*). If the player opted out, she can play alone and not receive requests for feedback for five turns (*SI Appendix*, Fig. S11*F*).

Incentive design. For each five turns (before turns 1, 6, 11 and 16), the participant was prompted to choose whether to join a player's club or play singly. The incentive included two components: free ferry ride (one cent saving per ride) and a reward for submitting ratings, which we varied between games (1 to 2 coins, 2 to 4 coins, 6 to 12 coins per rating). Once she joined the "ferry club," the player was prompted to decide in each turn whether to rate the ferry or not. There was no penalty for free riding. Incentive to rate was introduced either for turn 4 to 12 (early incentive) or for turns 13 to 20 (late incentive).

Personality tests. We conducted a shorter version of the Big Five personality test (35) and the altruism assessment test (36) after participants played the naturalistic game. *SI Appendix*, Fig. S12 shows screenshots from the Big Five test.

Quantification and Statistical Analysis.

Subgroups definitions. We focused on comparing rating accuracy in two subgroups of participants: Those who responded to the incentive by rating more often, and those who did not. In the naturalistic game, participants play 16 rounds: 8 with incentive and 8 without incentive (in a balanced order).

Influenced participants. We defined incentive-influenced participants where #ratings(incentivized) - #ratings(intrinsic) > 1. This threshold was determined by evaluating the distribution of net incentive influence which is the number of ratings with minus without incentive (*SI Appendix*, Fig. S13). As shown, the distribution is bimodal, and with a sharp peak around zero and a broad mode above it. Given the steep drop from 138 to 19 on the right side, we considered a threshold of >1 as conservative enough. All other participants were called uninfluenced (including rare cases of apparently negative influence, see below). A posteriori, setting a more liberal threshold #ratings(incentivized) - #ratings(intrinsic) > 0 gives very similar results:

With >1 we got R2(influenced, uninfluenced) = [0.468, 0.29].

With >0 we get R2(influenced, uninfluenced) = [0.453, 0.29].

Other thresholds can be tested by editing line 64 in the script "Initial data processing" in the deposited data.

Opt in and Opt out (from the participant guild or club). Participant decision if to opt in or out was repeated every five turns. See *SI Appendix*, Figs. S9*B* and S11 *B* and *C*.

Free riding. Participants who opted in to the participant club and further decided each turn if to rate or not (free ride). See *SI Appendix*, Figs. S9*E* and S11*D*.

Sample-size power analysis. We used data from the explicit MAB game (n = 182 valid participants) to estimate sample size needed for all other groups. We focused on within-subject comparison between incentive-influenced and uninfluenced participants. We calculated the mean rating accuracy for each group:

Rating accuracy within subject

Given:

Influenced	Uninfluenced	Diff means	Pooled SD
0.56	0.32	0.24	0.35
0.36	0.31		

Then, with alpha=0.05 (Type I error rate), beta = 0.2 (Type II error rate.), q0=q1=0.5 (proportion of trials in incentivized conditions, and expected proportion of influenced participants), we obtained N(sample) = 33, N(total) = 66. Therefore, we recruited 212 participants for our standard cohort to obtain 114 valid participants. For all other groups, we aimed at obtaining about 70 to 100 valid participants to test for replication in different conditions.

Bootstrap CIs for pooled means. We performed bootstrap analysis by randomly sampling participants with replacement. Bootstrap was performed 100 times, and 90% CI was computed.

Analysis variables.

Ground truth/arm reward. Number of coins per MAB arm in explicit MAB game. In the naturalistic game, we considered ferry ride duration (ranging from 2 to 20 s) as a negative reward. For each participant, we computed the mean rating score and the mean arm reward.

Rating scores and % rated. Our rating devices had the range of 0 to 100. In the game, participants were presented with mean scores. In figures, we present raw scores or mean scores, or % rated (Fig. 2), which is the proportion of turns where a rating score was submitted.

Pooled rating accuracy. The coefficient of determination R^2 between the ground truth and pooled rating scores for each treatment or subgroup.

Within-subject rating accuracy. The coefficient of determination R^2 between the ground truth and rating scores submitted by a participant. For the analysis of rating accuracy, we included only participants who voluntarily rated the MAB arms at least five times (out of 16 turns) which we considered as the minimum sample size for obtaining a meaningful within subject R^2: The proportion of participant the fit this criterion varied across groups, ranging between 47% in the raw MAB game to 81% in the naturalistic game with strongly selected cohorts. A posteriori we tested a range of threshold between 4 and 7 and found that our results are insensitive to the specific choice:

With n > 4 we got R2(influenced,uninfluenced) = [0.46, 0.30]

With n > 5 we got R2(influenced,uninfluenced) = [0.46, 0.29]

With n > 6 we got R2(influenced,uninfluenced) = [0.47, 0.29]

With n > 7 we got R2(influenced,uninfluenced) = [0.49, 0.29]

%bandit arms solved. Input vectors are ground truth and mean score for each arm. For each pair of arms (9 arms → 36 pairs), if both have rating data, and if mean ratings are not equal, test if score and ground truth are aligned. For example if reward i > reward j and score i > score j, we increment the score. If they are misaligned we decrement the score. A score of 36 means that scores and ground truth are aligned for all pairs of arms. We present the score as percentage (that is score/36 * 100). See ComputeBanditScore() code in deposited data: (https://osf.io/d953m/).

Random jitter for display purpose. We used a random jitter for display purpose only in Fig. 1*D*. The jitter was done by adding a uniformly distributed 1 unit noise to the integer MAB reward.

Simulation code. See ComputeSolution() in deposited data: (https://osf.io/d953m/).

Data, Materials, and Software Availability. Experimental results data have been deposited in Osf.io (https://osf.io/d953m/) (43).

ACKNOWLEDGMENTS. We thank Peter Harrison for PsyNet and programming support.

Author affiliations: [a]Department of Psychology, Hunter College, The City University of New York, New York, NY 10065; [b]Department of Communication, University of California, Davis, CA 95616; [c]Ostrom Workshop, Indiana University Bloomington, Bloomington, IN 47408; [d]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics, Frankfurt 60322, Germany; and [e]Princeton and National Bureau of Economic Research, Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ 08544

1. J. Becker, D. Brackbill, D. Centola, Network dynamics of social influence in the wisdom of crowds. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E5070–E5076 (2017).
2. A. Mao, L. Dworkin, S. Suri, D. J. Watts, Resilient cooperators stabilize long-run cooperation in the finitely repeated Prisoner's Dilemma. *Nat. Commun.* **8**, 13800 (2017).
3. N. E. Leonard, S. A. Levin, Collective intelligence as a public good. *Collect. Intell.* **1**, 263391372210832 (2022).
4. D. Ha, Y. Tang, Collective intelligence for deep learning: A survey of recent developments. *Collect. Intell.* **1**, 263391372211148 (2022).
5. T. Segaran, *Programming Collective Intelligence: Building Smart Web 2.0 Applications* ("O'Reilly Media, Inc.", 2007).
6. D. Miorandi, V. Maltese, M. Rovatsos, A. Nijholt, J. Stewart, *Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society* (Springer, 2014).

7. H. Shirado, N. A. Christakis, Network engineering using autonomous agents increases cooperation in human groups. *iScience* **23**, 101438 (2020).
8. T. W. Luke, "Perversities or problems in the rise of peer production with knowledge socialism: Collegiality, collaboration. collective intelligence" in *Knowledge Socialism*, M. A. Peters, T. Besley, P. Jandrić, X. Zhu, Eds. (Springer, 2020), pp. 61–80.
9. M. A. Peters, T. Besley, S. Arndt, Experimenting with academic subjectivity: Collective writing, peer production and collective intelligence. *Open Rev. Educ. Res.* **6**, 26–40 (2019).
10. Y. Koren, S. Rendle, R. Bell, "Advances in collaborative filtering" in *Recommender Systems Handbook*, L. Rokach, B. Shapira, F. Ricci, Eds. (Springer, 2022), pp. 91–142.
11. E. Escrig-Olmedo, M. Fernández-Izquierdo, I. Ferrero-Ferrero, J. Rivera-Lirio, M. Muñoz-Torres, Rating the raters: Evaluating how ESG rating agencies integrate sustainability principles. *Sustainability* **11**, 915 (2019).
12. M. Revel, D. Halpern, A. Berinsky, A. Jadbabaie, "Liquid democracy in practice: An empirical analysis of its epistemic performance" in EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization Arlington VA USA (Association for Computing Machinery, New York, NY United States, 2022) (available at: https://daniel-halpern.com/files/liquid-in-practice.pdf).
13. J. Campbell, A. Casella, L. de Lara, V. Mooers, D. Ravindran, Liquid democracy. Two Experiments on delegation in voting, w30794 (National Bureau of Economic Research, 2022), 10.3386/w30794.
14. Q. Kong *et al.*, Weakly labelled audioset tagging with attention neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **27**, 1791–1802 (2019).
15. C. Fiesler, J. Jiang, J. McCann, K. Frye, J. Brubaker "Reddit rules! characterizing an ecosystem of governance" in *Proceedings of the International AAAI Conference on Web and Social Media* (2018), https://ojs.aaai.org/index.php/ICWSM/article/view/15033, vol. 12.
16. U. Fischbacher, S. Gächter, Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *Am. Econ. Rev.* **100**, 541–556 (2010).
17. C. W. R. Webster, C. Leleux, Smart governance: Opportunities for technologically-mediated citizen co-production. *Inform. Polity.* **23**, 95–110 (2018).
18. Q. Zhong, S. Frey, Institutional similarity drives cultural similarity among online communities. *Sci. Rep.* **12**, 18982 (2022).
19. S. Frey, N. Schneider, Effective voice: Beyond exit and affect in online communities. *New Media Soc.* **25**, 2381–2398 (2023).
20. J. Lorenz, H. Rauhut, F. Schweitzer, D. Helbing, How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9020–9025 (2011).
21. B. Jayles *et al.*, How social information can improve estimation accuracy in human groups. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12620–12625 (2017).
22. C. Hess, E. Ostrom, *Understanding Knowledge as a Commons: From Theory to Practice* (MIT Press, 2011).
23. K. R. McKee *et al.*, Social diversity and social preferences in mixed-motive reinforcement learning. arXiv [cs.MA] [Preprint] (2020). http://arxiv.org/abs/2002.02325 (Accessed 12 February 2020).
24. R. P. Mann, D. Helbing, Optimal incentives for collective intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5077–5082 (2017).
25. S. S. Levine *et al.*, Ethnic diversity deflates price bubbles. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 18524–18529 (2014).
26. E. A. Locke, K. Schattke, Intrinsic and extrinsic motivation: Time for expansion and clarification. *Motivation Sci.* **5**, 277–290 (2019).
27. S. Rathje, J. Roozenbeek, J. J. Van Bavel, S. van der Linden, Accuracy and social motivations shape judgements of (mis)information. *Nat. Hum. Behav.* **7**, 892–903 (2023), 10.1038/s41562-023-01540-w.
28. A. Sankararaman, A. Ganesh, S. Shakkottai, Social learning in multi agent multi armed bandits. *Proc. ACM Meas. Anal. Comput. Syst.* **3**, 1–35 (2019).
29. S. Shahrampour, A. Rakhlin, A. Jadbabaie, "Multi-armed bandits in multi-agent networks" in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP, 2017), pp. 2786–2790.
30. V. Kuleshov, D. Precup, Algorithms for multi-armed bandit problems. arXiv [cs.AI] [Preprint] (2014). http://arxiv.org/abs/1402.6028 (Accessed 25 February 2014).
31. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, ed. 2, 2018).
32. O. Tchernichovski, S. Eisenberg-Edidin, E. D. Jarvis, Balanced imitation sustains song culture in zebra finches. *Nat. Commun.* **12**, 2562 (2021).
33. W. Lasarov, S. Hoffmann, Social moral licensing. *J. Bus. Ethics.* **165**, 45–66 (2020).
34. O. Tchernichovski, L. C. Parra, D. Fimiarz, A. Lotem, D. Conley, Crowd wisdom enhanced by costly signaling in a virtual rating system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7256–7265 (2019).
35. B. Rammstedt, O. P. John, Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Pers.* **41**, 203–212 (2007).
36. E. Manzur, S. Olavarrieta, The 9-SRA Scale: A Simplified 9-Items Version of the SRA Scale to Assess Altruism. *Sustain. Sci. Pract. Policy.* **13**, 6999 (2021).
37. J. J. Heckman, Sample Selection Bias as a Specification Error. *Econometrica.* **47**, 153–161 (1979).
38. K. Underhill, When extrinsic incentives displace intrinsic motivation: designing legal carrots and sticks to confront the challenge of motivational crowding-out. *Yale J. Regul.* **33**, 213 (2016).
39. F. H. Schneider *et al.*, Financial incentives for vaccination do not have negative unintended consequences. *Nature.* **613**, 526–533 (2023).
40. M. Brill, T. Delemazure, A.-M. George, M. Lackner, U. Schmidt-Kraepelin, Liquid democracy with ranked delegations. *AAAI* **36**, 4884–4891 (2022).
41. R. Koster *et al.*, Human-centred mechanism design with Democratic AI. *Nat Hum Behav.* **6**, 1398–1407 (2022).
42. P. Harrison *et al.*, Gibbs sampling with people. *Adv. Neural Inf. Process. Syst.* **33**, 10659–10671 (2020).
43. O. Tchernichovski, N. Jacoby, S. Frey, D. Conley, Rating accuracy in collective information. Open Science Framework. https://osf.io/d953m/. Deposited 5 October 2023.