Theory and Decision (2024) 97:67–88
https://doi.org/10.1007/s11238-023-09965-5

Suspicious minds and views of fairness

Øivind Schøyen[1,2]

Accepted: 29 October 2023 / Published online: 6 January 2024

Abstract
Do people with different views of what is fair attribute different intentions to actions? In a novel experimental design, participants were significantly more likely to attribute a no-redistribution vote to selfishness if they considered redistribution as being fair. I define this—attributing actions that do not adhere to one's own fairness view to selfishness—as suspicious attribution. I develop a theory of intention attribution to show how suspicious attribution arises from two other findings from the experiment: the participants underestimate the number of people with fairness views differing from their own and overestimate the selfishness of participants with other fairness views. I discuss how the findings can help explain political polarization.

Keywords Attribution · Projection bias · Redistribution · Fairness view · Morality · Polarization

1 Introduction

The present paper examines how peoples' interpretations of intentions are affected by their morality.[1] Understanding how differences in moralities affect our interpretations of intentions is central to understanding how attitude polarization occurs (Haidt, 2012). Moralities other than one's own are often difficult to understand or empathize with (Haidt, 2007). Empathizing with other moralities implies acknowledging objections to one's convictions, which is something people often have little incentive or interest in doing (Piketty, 1995). Thus, selfishness is often attributed to behavior not adhering to one's own morality. I define this type of intention attribution—

[1] A "morality" can be understood as a vector of beliefs and values that are internalized and embedded in a person, such as political ideologies or religious identities (Greif & Tadelis, 2010) More broadly, morality can be defined as "prescriptive, judgements of justice, rights, and welfare pertaining to how people ought to relate to each other" (Turiel 1983 in Kesebir and Haidt (2010)).

✉ Øivind Schøyen
oivind.schoyen@gmail.com

[1] The School of Business and Economics, UiT The Arctic University of Norway, Tromsø, Norway

[2] FAIR Centre of Excellence, NHH Norwegian School of Economics, Bergen, Norway

attributing behavior not adhering to one's own morality to selfishness—as suspicious attribution.

Suspicious attribution distorts perceptions of the number of people who have selfish intentions. Thus, it increases the legitimacy of choosing confrontation over dialog in encounters with out-group members. A recent example of such a process is the polarization of American politics. Since the 1990s, supporters of the Republican and Democratic parties have found increasingly less common ground. Both sides interpret the intentions of the other party with increasing suspicion, causing them to support off-center candidates. This has led to a vicious cycle of polarization (Haidt, 2012) and political deadlock (Binder, 2015).

The present paper focuses on suspicious attribution of voting in a simplified redistributive game. The game has two types of players: workers and predictors. Workers vote for either full or no redistribution after observing their own income. Predictors observe the worker's vote but not their income. The predictor then guesses the intention behind the workers' votes—fairness or selfishness. In this game, fairness and selfishness motives vary in a discernible way. Half of the workers who have above-median incomes have monetary incentives for no redistribution, whereas the other half have monetary incentives for full redistribution. Views of what constitutes fair redistribution also vary: experiments have shown that, when participants have no monetary incentive for the outcomes, people hold mutually excluding fairness views about redistribution (Cappelen et al., 2007; Roemer, 2009; Cappelen et al., 2020). Some find it fair to redistribute earnings to compensate for income differences arising from luck, effort, or performance, who are hereby referred to as "egalitarians," whereas others prefer differences in earnings to be reflected in income differences, who are hereby referred to as "libertarians."[2]

Suspicious attribution can contribute to polarization, that is, a dynamic of increasingly different attribution of voting. Suppose a Bayesian egalitarian observes a surprisingly high number of votes against redistribution. Under suspicious attribution, the egalitarian will attribute these votes to selfishness. This may lead the egalitarian to update his prior belief of how many selfish individuals exist in the population. If so, votes against redistribution will become stronger signals of selfishness. Fryer et al. (2019) shows how Bayesian updating under different priors can lead to polarization when the evidence is open to interpretation. Similarly, suspicious attribution interacting with associative learning could lead to polarization. Consider an egalitarian predictor attributing votes against redistribution to selfishness. Suppose this leads the predictor to develop an association between votes against redistribution and selfishness. Upon repeated observation, the association and the attribution of votes against redistribution to selfishness will strengthen. Thus, the egalitarian predictor's susceptibility to suspicious attribution will increase. This can lead to a similar dynamic of polarization.

To empirically test for suspicious attribution of redistributive votes, I have developed a novel laboratory experiment design. The participants were randomly assigned either the role of a predictor or worker. Workers completed tasks and earned money according to whether their output was above or below the output of the

[2] My experimental design is not dependent on whether people's views conform to these labels. All that is assumed is that either the participants find it fair to redistribute or not.

median worker. The workers then voted for full or no redistribution after observing how large their earnings were compared with that of the median worker. The predictors estimated how groups of workers casting different redistribution votes would score on a selfishness measure.

The selfishness measure is as follows: A worker flips a coin 10 times and is paid by the number of self-reported heads, hereby referred to as favorable coin tosses. The coin tosses are made in a setting where the worker is not being monitored. The measure of selfishness was how many favorable coin flips the predictors expected the worker to report. Suppose the predictor has a suspicious attribution of a worker's vote. In this case, the predictor will find it more likely that the worker will lie about the number of favorable coin tosses to increase their payment.

The theory section models the predictors' belief of whether the workers' redistribution vote is motivated by selfishness or fairness concerns. The prediction is based on prior beliefs about the prevalence of fairness views and selfish types in the population of workers. I show two types of prior beliefs that can give rise to suspicious attribution: projection bias in fairness views—inflated beliefs about how many people share one's fairness view—and out-group stereotypes—biased beliefs about how many people with opposing fairness views are selfish.

The present paper offers four main findings. *First*, both egalitarians and libertarians interpret votes for no redistribution as signals of selfishness. *Second*, egalitarians attribute votes for no redistribution to selfishness significantly more than libertarians. *Third*, both libertarians and egalitarians display projection bias in fairness views; they overestimate the prevalence of their own fairness views. *Fourth*, egalitarians have an out-group stereotype bias against libertarians; their estimates of the prevalence of selfish types among libertarians are upward biased and significantly higher among egalitarian than libertarian predictors. I find supportive evidence linking the suspicious attribution of no-redistribution votes to out-group stereotypes and projection bias; predictors' suspicious attribution biases are significantly correlated with their out-group stereotypes and prevalence estimates of fairness views. An interpretation section finds that the data can be best explained by suspicious attributions among egalitarians but not libertarians.

The present paper develops a novel experiment and a novel theoretical model to study attribution differences arising because of fairness views. The theoretical model combines several strands of the behavioral science literature—moral reasoning (Haidt, 2012), social cognition (Alicke et al., 2005), inference under the projection bias (Gagnon-Bartsch, 2017; Madarász, 2015; Fryer et al., 2019), and fairness views (Cappelen et al., 2007; Roemer, 2009)—by developing a specific model of inference of the intentions behind votes under projection bias and out-group stereotypes. In the study closest to this Graham et al. (2012), participants were asked about the moral reasoning of a typical out-group to study perceptions of the moral reasoning of political conservatives, moderates, and liberals. The present study differs from Graham et al. (2012) in several ways. Most importantly, it considers the attribution of intentions behind actions rather than moral reasoning. This is done as a first step to

investigate if attribution differences can generate or change group stereotypes. That is, to explain how one group identity, such as "conservative", becomes associated with selfishness among members of another social group, such as "liberals".[3]

The present paper proceeds as follows: Part 2 presents the theoretical framework of the experiment, Part 3 presents the experimental design, Part 4 interprets the results of the experiment, and Part 5 concludes the paper. An online appendix covers further data analysis, extra findings from the experiment, and experimental instructions (Schøyen, 2022).

2 Theoretical framework

This section presents the theoretical framework that guided the experimental design. The game and belief structure correspond to the experiment.

2.1 A redistributive game

The game has two types of players: workers and predictors. They play the following game:

Stage 0: Nature randomly draws an odd number of workers and predictors. Player i's, fairness preferences differ along two dimensions: *type*, which can be either egalitarian or libertarian, $F_i \in \{E, L\}$ and *strength*, $\beta_i \geq 0$ how much weight i puts on realizing their ideal behavior relative to their own extrinsic gain.[4] A portion $\overline{L} \in (0, 1)$ of the population is libertarian, and $1 - \overline{L}$ is egalitarian. Nature draws β_i from a nondegenerate distribution with support on $[0, \infty)$.

Stage 1: A worker flips a coin 10 times, observes oh_i heads, and reports fh_i heads. The worker is paid some amount $p > 0$ for each reported head.

Stage 2: Each worker produces an output $o_i \in [0, K]$.

Stage 3: Each worker casts a vote, v_i, for either full redistribution, r, or no redistribution, nr. The outcome is decided by a simple majority vote.

[3] This study further differs from Graham et al. (2012) in that it models and elicits predictors' beliefs about other participants making interested and disinterested choices, which allows disentangling the reasoning behind the attribution differences. The general experimental method of asking subjects about how outgroup reasoning was originally developed by Dawes et al. (1972). Projection bias was first demonstrated by Ross et al. (1977) and was subsequently shown to hold across many fields and situations (Alicke et al., 2005; Blanco et al., 2014; Rubinstein & Salant, 2016) Experimental studies have previously found outgroup stereotypes and projection bias. People generally attribute negative intentions to people with differing opinions (Reeder et al., 2005) and exaggerate the differences connected to political opinions in particular; I find that this bias holds for fairness views. Note that Dawes (1989) challenges whether observing people systematically overestimating the number of people sharing their preferences is sufficient evidence of projection bias. The effect proposed by Dawes (1989) could arise from a sampling issue; that is, people know their own preferences and should rationally have different estimates based on their observed sample. The discussion by Alicke et al. (2005) concludes that, in considering the sum of empirical evidence as a whole, the emergence of projection bias solely from Dawes (1989)'s sampling issue seems unlikely.

[4] The present paper distinguishes between intrinsic and extrinsic utility. The intrinsic utility from an activity is the inherent reward from the activity, e.g., self-expression or enjoyment of the activity. The extrinsic utility of an activity is utility gained through the activity that can serve other purposes, e.g., money or status. This definition is taken from Kreps (1997).

Stage 4: The predictors observe the structure of the game and the support for type distributions. They do not observe the workers' output, fairness view, or reported coin flips. Each predictor receives a strictly positive amount of money if they correctly predict the exact number of reported coin flips conditional on workers' cast vote.

2.2 Workers

We first introduce a general utility function to model an agent's trade-off between the intrinsic utility of living up an ideal or extrinsic gains. The distance between the ideal behavior, I_i, and chosen behavior, A, is defined by a function $v(A - I_i)$. The behavior A is increasing in extrinsic reward. The weight given on following the ideal relative to the extrinsic reward is given by β_i. The general utility function is now given by the following:

$$u(A) = A - \beta_i v(A - I_i). \tag{1}$$

We assume that intrinsic utility decreases from zero in the distance between the ideal behavior and chosen behavior

$$v'_A(A - I_i) > 0, v(0) = 0. \tag{2}$$

We now apply this general utility function to analyze the link between workers' voting and coin-flip reporting behavior.

We first consider workers' voting behavior in stage 3 of the game. By definition, half of the workers have output above-median output $o_i : \frac{\Sigma_j o_j}{\#\text{workers}} < o_i \equiv \overline{o}$. The other half have an below-average output, $o_i : \frac{\Sigma_j o_j}{\#\text{workers}} > o_i \equiv \underline{o}$. If the full redistribution option receives a majority, everyone is paid an amount $\overline{y}$. If the no-redistribution option receives a majority, workers with above average output, $\overline{o}$, receive an amount y_h and workers with below-average output, $\underline{o}$, receive an amount y_l, where $y_h > \overline{y} \equiv \frac{y_h + y_l}{2} > y_l \equiv 0$.

Fairness views, F_i, determine whether workers intrinsically find redistribution, E, or no redistribution, L, as the most fair in this game. The fairness preferences can grow out of different beliefs about whether the output in stage 2 of the game reflects a choice of effort or luck. The variation in fairness views could also reflect differences in attitudes about what constitutes legitimate sources of inequality.[5]

[5] Roemer (2009) defined a *strict egalitarian* as a person who believes that an equal distribution is a fair outcome regardless of the source of inequality, a *choice egalitarian* as a person who finds it fair to let income differences reflect effort differences but seeks to redistribute inequality arising from luck, and a *libertarian* as a person who never redistributes income differences arising from luck or effort. Here, I discuss how the E and L types in my theory relate to the typology of egalitarians and libertarians in Roemer (2009). Assuming beliefs in our game do not vary and that both types believe the output draw is random, the E type can be a choice or strict egalitarian believing that workers should be held accountable for choices of effort. Assuming both L and E types believe the draw of the output is not random, but a reflection of a choice of effort, the E type must be a strict egalitarian and the L type either a choice egalitarian or strict libertarian. Finally, allowing the beliefs about the draw of output to vary, the E type can be either a strict or choice egalitarian believing that the draw is not random, and the L type can be a choice egalitarian believing the draw is not random or a libertarian.

Note that by grouping workers along their output and fairness ideal, only two groups have incentives to vote against their fairness ideal.[6] Libertarians with a below-average output, $\underline{o}$, can get an median payment $\overline{y}$, rather than no payment, by voting for redistribution, r. Correspondingly, egalitarians with an above-median output, $\overline{o}$, can get a high payment y_h, rather than median payment $\overline{y}$, by voting against redistribution, nr. Thus, when choosing how to vote both these groups of voters face an identical trade-off; they must weigh the monetary reward of $\overline{y} - 0$ against a loss of intrinsic utility from deviating an equal distance from their ideal, $v(\overline{y} - 0)$.

We denote the difference in utility of a worker i with fairness view F_i and output o_i for voting v' rather than v as $\Delta u_{F_i,o_i,v'}$. We assume that the probability of being pivotal is equal for voting either for or against redistribution, such that it does not enter into the utility differential between the voting options. We denote an indicator function indicating whether the agent actually voted differently than their fairness view of $\mathbb{1}_{v_i}$. Inserting this into the general utility function, (1), the utility for low out-output libertarians is given by the following:

$$\Delta u_{L,\underline{o},R} = \overline{y} + \beta_i v(\overline{y} - 0)\mathbb{1}_R - 0 + \beta_i v(0 - 0)\mathbb{1}_{NR}. \tag{3}$$

Applying the assumption that $v(0) = 0$ from (2), this equals the following:

$$\Delta u_{L,\underline{o},R} = \overline{y} + \beta_i v(\overline{y})\mathbb{1}_R. \tag{4}$$

Correspondingly, the utility for high-output egalitarians is given by the following:

$$\Delta u_{E,\overline{o},NR} = \overline{y} - \beta_i v(\overline{y})\mathbb{1}_{NR}. \tag{5}$$

We have established that high-output egalitarians and low-output libertarians have an equal distance to their ideal distribution and equal monetary incentives to deviate from their ideal in (4) and (5). It follows that there exists a common level of prioritizing one's ideals, β_i, at which workers will vote according to their incentives rather than their fairness view. We define any worker with $\underline{\beta}_i < \beta^* \equiv \frac{\overline{y}}{v(\overline{y})}$ as a selfish worker, while any worker with $\overline{\beta}_i > \beta^*$ as unselfish.

We now consider workers' coin flip reporting in Stage 1 of the game. Workers' report head coin flips meriting payment, referred to as the number of favorable coin flips. We assume that, for both egalitarians and libertarians, the ideal is honest reporting of the actually observed favorable coin flips, oh_i. We denote the actual reporting of favorable coins rh_i, giving profit $fh_i \times p$. Assuming that both libertarians and egalitarians consider truthfully reporting the ideal action, that is, $I_i = oh_i$, then applying (1) to the coin-flip task gives the following:

$$u_i = rh_i \times p - \beta_i v(rh_i - oh_i). \tag{6}$$

It follows from (6) that unselfish workers to a larger extent than selfish workers prioritize their ideal truthful reporting coin flips. We can now establish the following theorem of worker behavior in Stages 1 and 3 of the game:

[6] The ideal action of libertarians with below-median output and egalitarians with above-average output is equal to the action maximizing their respective monetary incentives. Consequently, their choice of vote is independent of whether they put high or low weight on living up to their ideal, β_i.

Theorem 1 *Common threshold of selfishness Workers of both fairness views will vote according to incentives if and only if they are selfish. Selfish voters will in expectation report more favorable coin flips than unselfish workers.*

Proof Assume a worker of either fairness view voted against their ideal. From (4) and (5), it follows that $\frac{\overline{y}}{v(\overline{y})} < \beta_i \leq \overline{\beta}$, that is, the voter is selfish. Assume a voter is selfish, then $\frac{\overline{y}}{v(\overline{y})} < \beta_i \leq \overline{\beta}$, and it follows from (4) and (5) that they will vote against their ideals when they have a monetary incentive to do so.

To see that the expected coin reporting decreases in β_i, insert $\beta_i \in \{0, \beta_i \to \infty\}$ into (6). This yields $u_{i,f,o}(\beta_i = 0) = rh_i \times p$ with $\arg\max_{rh_i^*} u_{i,f,o}(\beta_i = 0) = 10$ and $u_{i,f,o}(\beta_i \to \infty) = rh_i \times p - \infty v(rh_i - oh_i)$ with $\arg\max_{rh_i^*} u_{i,f,o}(rh_i \to \infty) = oh_i^* \to 0$, respectively. For any $\beta_i \in (0, \infty)$, it holds that the reported number of favorable coins decreases in β_i, that is, $\frac{\partial rh_i^*}{\partial \beta_i} = v(rh_i - oh_h) \geq 0$. Because selfish workers have lower β_i by definition, it follows that the expected reported coin flip is higher among the selfish than unselfish voters; $E[rh_i | \beta_i \leq \overline{\beta}] > E[rh_i | \beta_i > \overline{\beta}]$. □

A portion $(S|F) \equiv \frac{\#F \text{ workers with} \beta_i < \beta^*}{\#F \text{ workers}} \in (0, 1)$ of players with fairness views F are selfish types, S and $1 - (S|F)$ are nonselfish types, NS. This follows from the nondegenerate distribution of selfishness $\beta_i \in (0, \infty]$ and fairness views, $\overline{L} > 0$. Theorem 1 on selfish workers implies that workers will vote according to the matrix in Table 1.

Table 1 The choice of actions for the different types of workers. The prevalence of libertarians is $\overline{L}$ and that of egalitarians is $1 - \overline{L}$. The prevalence of selfish types among egalitarians is $(S|E)$ and among libertarians is $(S|L)$

	Choice Matrix Libertarian Workers	
	Portion $\overline{L}$ of the population	
	Above average output($\overline{o}$)	Below average output($\underline{o}$)
Selfish $(S\|L)$	nr	r
Non-Selfish $(1 - (S\|L))$	nr	nr
	Choice Matrix Egalitarian Workers	
	Portion $1 - \overline{L}$ of the population	
	Above average output($\overline{o}$)	Below average output ($\underline{o}$)
Selfish $(S\|E)$	nr	r
Non-Selfish $(1 - (S\|E))$	r	r

2.3 Predictors

Predictors observe only the worker's vote v and predict the worker's reported favorable coin flips. As established in Theorem 1, selfish types will report more coin flips. To analyze the predictors' estimate, we consider an egalitarian predictors' beliefs that a no-redistribution vote, nr, is cast by a selfish voter, $P(S|nr)$[7]

$$P(S|nr) = \frac{\text{Number of selfish types casting } nr}{\text{Number of selfish types casting } nr + \text{ Number of nonselfish types casting } nr}. \quad (7)$$

The predictors form their belief of $P(S|nr)$ based on their knowledge of the structure of the game, the structure of the preferences of different types given in Table 1, and their prior belief about the number of workers who are libertarians, $\overline{L}$, selfish, $\overline{S}$, and selfish libertarians, $(S|L)$.[8] I apply Bayes' rule to (1) to find the following expression mapping from the prior beliefs $(S|L), \overline{L}, \overline{S}$ to $P(S|nr)$:

$$E[(S|nr)] \equiv P(S|nr) = \frac{\overline{S}}{\overline{S} + 2(1 - (S|L))\overline{L}}. \quad (8)$$

We now analyze how the prior beliefs of the predictor affect their assessment of worker selfishness (8). Votes against redistribution become a stronger signal of selfishness with fewer libertarian fairness views, regardless of the distribution of selfish types among libertarians and egalitarians. This follows from the fact that only libertarians vote against redistribution if they are nonselfish. Projection bias implies an egalitarian predictor will underestimate the prevalence of libertarians, $\overline{L}$. Projection bias will always increase the degree to which a predictor takes a vote not adhering to his fairness view as a signal of selfishness. We establish this as a theorem:

Theorem 2 *Intention attribution and projection bias in fairness views Take any two fairness views n and m* $\in \{E, L\}$, *such that* $m \neq n$. *If holders of fairness view m and n have common beliefs about the portion of selfish types,* $(S|E)^m = (S|E)^n$ *and* $(S|L)^m = (S|L)^n$, *projection bias implies a predictor with view m taking a vote adhering to view n as a stronger signal of selfishness than a predictor with view n.*

Proof The proposition implies that $\frac{\partial E[(S|nr)]}{\partial L} < 0$ and $\frac{\partial E[(S|r)]}{\partial(1-\overline{L})} < 0$ must hold for any $(S|E)$, $(S|L)$. For $\frac{\partial E[(S|nr)]}{\partial L} < 0$ to hold (8) implies the following strict inequality must hold:

$$[(S|E)L + (S|E)(1 - \overline{L})] > [(S|E) - (S|L)]. \quad (9)$$

Inequality (9) states that the weighted average of the prevalence of selfish types for both fairness views must be strictly larger than the difference in the prevalence of

[7] All statements hold *mutatis mutandis*: for a libertarian predictor observing a redistribution vote.

[8] Because predictors only get paid for exact reporting and all players are assumed to believe that the probability distribution of selfish types contingent on fairness views and redistribution votes are single peaked, the game equilibrium is given by workers voting according to their type and predictors honestly reporting their beliefs, here considering the case where an egalitarian observes a no-redistribution vote believing that libertarians and egalitarians are equally likely to draw low output, i.e., $P(\underline{o}|L) = P(\underline{o}|E) = \frac{1}{2}$.

selfish types between the two groups. Because the average selfishness is a convex combination of the selfishness of the two groups and $[(S|E),(S|L)] \gg 0$, this strict inequality always holds. $\frac{\partial E[(S|r)]}{\partial(1-\overline{L})} < 0$ follows mutatis mutandis. □

Out-group stereotypes, defined as an upward bias in the perceived selfish libertarians, $(S|L)$, increase the degree to which a no-redistribution vote is taken as a signal of being a selfishness type. This holds for any prevalence of libertarians, $\overline{L}$. This can be verified by taking the derivative of (8) with respect to $(S|L)$: $\frac{\partial P(S|nr, P(\underline{o}|L)=P(\underline{o}|E)=\frac{1}{2})}{\partial(S|L)} = \frac{2\overline{L}\overline{S}}{(\overline{S}+2(1-(S|L))\overline{L})^2} > 0$ and observing that it is defined and positive for any $\overline{L} \in (0,1)$.

We have established that projection bias and out-group stereotypes against libertarians influence egalitarian predictors' belief of a worker voting against redistribution selfishness, $E[(S|nr)]$.[9] I now define suspicious attribution in this game. Denoting the belief of a predictor holding fairness view m as E^m, I define suspicious attribution as follows:

Definition 1 *Suspicious attribution in the redistributive game* Predictors with suspicious attribution will find workers casting votes not adhering to their own fairness view more likely to be selfish types than other workers: $E^E[(S|r)] < E^E[(S|nr)]$ and $E^L[(S|nr)] < E^L[(S|r)]$.

3 Experiment

3.1 Experimental design

The participants who signed up for the experiment were randomly allocated roles. Each participant had only one role: a *worker* or *predictor*. *Predictors* were paid according to the accuracy of their estimates. The *worker* role included studying predictions about concrete outcomes rather than hypothetical scenarios. By separating these roles, the experiment elicited beliefs from predictors who themselves made none of the choices they were being questioned about. This was done to reduce the risk of upward bias in estimates caused by predictors' norm-seeking rationalization of choices. All participants were asked about their age, gender, risk preferences, and whether they voted for political parties that actively pursued higher or lower levels of redistribution. A detailed overview of the experiment design and roles is shown below

[9] The vote is uninformative, and no learning occurs about the selfishness of the worker—that is, $E[P(S|r)] = E[P(S|nr)]$—if the predictor believes that there are exactly equal portions of selfish and nonselfish types among libertarians and egalitarians and there are an exactly equal number of egalitarians and libertarians. This can be seen by inserting $\overline{L} = \frac{1}{2}$ and $S_E = S_L$ into (8) to attain $P(S|nr) = \overline{S}$. This point is part of a set of priors at which the vote is uninformative of the workers' type. The set forms a hyperplane containing $\overline{L} = \frac{1}{2}$ and $S_E = S_L$ in the $\overline{L}, S_E, S_L$ space. The informativeness of a redistributive vote as a signal of the workers' type increases in the distance of the priors from this hyperplane.

Table 2 The overview of the different experimental stages for the two experimental roles, workers, and predictors

Workers				
Stage	W1	W2	W3	W4
	Workers flip a coin 10 times and report the number of favorable flips; $S_i \in [0..10]$	Workers copy words from text.	Workers are informed whether their number of correctly copied words is above or below the median	Workers vote for either full redistribution or no redistribution

Predictors					
Stage	P1	P2	P3	P4	P5
	Predict the selfishness of workers contingent on the vote: $E[\overline{S}]$, $E[(S\|r)]$, and $E[(S\|nr)]$	Predict the selfishness of workers with no monetary incentives over outcomes fairness views and selfishness contingent fairness views: $E[1-\overline{L}]$, $E[(S\|E)]$, and $E[(S\|L)]$	Predictors report the options they find fair, which determines their fairness views; $F \in \{E, L\}$	Predictors flip a coin 10 times and report the number of favorable flips; $S_i \in [0..10]$	Predictors receive payments relative to the prediction accuracy

The participants had only one of two roles. The workers completed their work before the predictors. Worker selfishness was measured as the expected reported favorable coin flips, $E[\overline{S}]$. The predictors first report expected selfishness contingent on workers' vote, $E[(S|r)]$, $E[(S|nr)]$. They then predicted the prevalence of fairness views, $E[1-\overline{L}]$, and then selfishness contingent on fairness views $E[(S|E)]$, $E[(S|L)]$ a subsample of workers. This subsample of workers did not complete the work round in W2 and had no incentives when it came to the outcome of their vote

3.1.1 Workers

The workers started by flipping a coin ten times. Each worker was told to report the number of times the coin was flipped on the tails side and that they would be paid four Norwegian kroner per tails outcome. There was no monitoring during the coin-flipping phase, an absence of which the participants were aware. Consequently, the participants could misreport the number of times the coins landed on the tails side to increase their earnings. The average reported coin flips by the workers were the empirical measures of selfishness.

The workers were then given 15 min to copy as many words as possible from a passage of text. The number of correctly copied words for each worker and the median number of copied words for all workers were recorded. The half of the workers who had above-median numbers of correctly copied words earned a wage of 100 kroner, whereas the other half earned nothing. The workers observed their own earnings and then cast votes for either full redistribution, so that the workers were paid identically or no redistribution, so that workers' payments equaled their earnings. The option that received a simple majority was implemented.

To elicit workers' fairness views, I elicited the votes of workers with no monetary incentives over the outcome of the vote. To this end, a group of workers was randomly selected from the same subject sample. These workers did not work or receive payment, but they voted for or against redistribution.

3.1.2 Predictors

The predictors were presented with all the details of the worker's role. They were also informed that the worker participants were randomly assigned to the role and drawn from the same subject pool as themselves. The predictors were then incentivized with the quadratic scoring rule to report their expected beliefs of the behavior of the workers.

First, the predictors were asked about the average number of reported favorable coin flips (the degree of selfishness). This was done to elicit their beliefs about the general selfishness of workers. Second, the predictors were asked about the selfishness of workers voting for redistribution and workers voting against redistribution. This was done to test for suspicious attribution. Third, the predictors

Table 3 Fairness views $(1-\overline{L})$ and reported favorable coin flips S_m

Predicted favorable coin flips				
	N	Egalitarians 1-$\overline{L}$	$(S\|E)$	$(S\|L)$
Predictors	210	56%	6.11	6.55
			(2.0)	(2.13)

The average reported number of favorable coin flips out of ten tosses by predictors' fairness views and number of participants with egalitarian fairness views

Table 4 Predicted prevalence of fairness views in a group of 100, $E[L]$, and predicted selfishness, $E[S]$, among workers according to their votes, $E[(S|r)]$, $E[(S|nr)]$, and according to their fairness views, $E[(S|E)]$, $E[(S|L)]$

Predicted favorable coin flips						
	N	E[1-$\overline{L}$]	$E[(S\|r)]$	$E[(S\|nr)]$	$E[(S\|E)]$	$E[(S\|L)]$
Average predictions	210	60	5.42	6.83	5.75	6.58
		(24)	(1.29)	(1.70)	(1.63)	(1.29)
Average predictions by egalitarians (E)	126	67	5.33	6.94	5.72	6.72
		(21)	(1.25)	(1.69)	(1.26)	(1.63)
Average predictions by libertarians (L)	84	50	5.56	6.65	5.79	6.37
		(25)	(1.33)	(1.70)	(1.38)	(1.61)

were asked about the selfishness of worker participants according to their fairness views. This was done to elicit their beliefs about any correlation between holding a particular fairness view and being selfish, that is, out-group stereotypes. Fourth, the predictors were asked about the number of workers holding each fairness view. This was done to test for projection bias in fairness views.

Thus, in the notation of the theoretical framework, we have elicited both the selfishness contingent on vote, $E^i[(S|v_i)]$, which has been modeled as a Bayesian probability in (8), and the parameters of the theory model this inference to be made upon, that is, $P^i(S|E), P^i(S|L)$.

The predictors then did the coin-flip task. This was done to control for the correlation between their own coin-flipping behavior and their beliefs about others' reporting. Finally, the predictors were asked about their own fairness views.

3.2 Experiment sample and procedures

To show suspicious attribution, a redistributive vote must be taken as a signal of being a selfish type, not simply a sign of membership in another social group. Minimal group distinctions have been shown to change social inference from projection bias to stereotyping (Alicke et al., 2005). Thus, the experiment was conducted on a sample with few or no salient social group cleavages. I did this laboratory experiment on law students from the University of Bergen. The Faculty of Law at the University of Bergen has only one field of study, so the sessions only involved students from the same discipline. If the data were drawn from, for example, the social science faculty in which students had different specializations, such as sociology and economics, different redistributive votes would likely reflect stereotypes about different social groups.

The sample size was 210 predictor participants and 18 worker participants. The pre-experiment hypothesis and analysis plan is available in the online pre-analysis plan (Schøyen, 2017). This plan was posted online before the author had access to the data from the main experiment. I used the redistribution option that the predictors

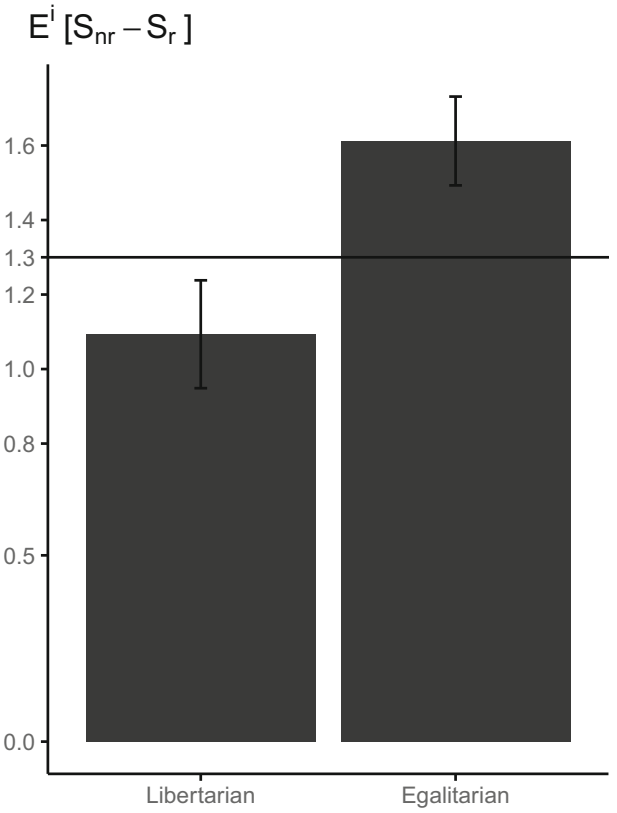

Fig. 1 Suspicious attribution of votes against redistribution. Differences in the predictors' reported expected average number of coin flips between workers voting for full or no redistribution, $E^m[(S|nr) - (S|r)]$. The difference is reported by the predictors' fairness views, $m \in \{E, L\}$. The solid horizontal line indicates the differences between workers voting for no redistribution and for redistribution, $(S|nr) - (S|r)$, 1.3 in the worker sample

reported finding fair as a measure of their fairness view rather than the predictors' cast vote. About 15% of the predictors voted for another redistribution alternative than the one they reported as fair. Except for projection bias in the fairness views, all the results were insignificant when applying the predictors' behavior (casting the third-party vote) as the measure of fairness view. This indicates that beliefs regarding others' intentions correlate with what the predictors reported as being fair, but they had a weaker correlation with the vote that the predictors cast. The experiment was conducted on September 3, 2015.

4 Data

4.1 Descriptive statistics

As visible in Table 3, most workers and predictors were egalitarians. The self-reported number of coin flips showed that the participants significantly exaggerated the number of favorable coin flips.[10]

The group-level data in Table 4 present the summary statistics of the predictors' expectations of worker behavior. The predictors of both egalitarian and libertarian fairness views correctly expected other participants to over-report the number of favorable coin flips.

The workers voting for no redistribution were expected to be more selfish than those voting for redistribution, $E[(S|r)] < E[(S|nr)]$, and libertarian workers were expected to be more selfish than egalitarian workers, $E[(S|E)] < E[(S|L)]$.

As I show in the results section, the magnitudes and significance of the differences in expectations of selfishness between libertarian and egalitarian predictors increased when comparing the differences within one predictor's estimates rather than with group averages.

4.2 Results

We now study suspicious attribution in the experimental data. The relevant statistic is the differences in the predictors' beliefs about workers casting votes for or against redistribution, not the general level of attribution of selfishness. This follows from the definition of suspicious attribution in the redistributive game; there should be a difference between the attribution of votes, here as following the definition of suspicious attribution in the redistributive game developed in the theory section. Thus, I focus on the within-predictor difference in the estimated average difference in selfishness between workers voting for or against redistribution, $E^i[(S|nr) - (S|r)]$.

Figure 1 shows the differences between egalitarian and libertarian predictors' selfishness estimates for workers' voting for or against redistribution, $E^i[(S|nr) - (S|r)]$. The solid line shows the actual observed difference, $(S|nr) - (S|r)$ in the selfishness measure, that is, reported coin flips. The predictors of both fairness views predicted that workers voting for no redistribution were more selfish. Predictors with egalitarian fairness views predicted larger differences between the perceived selfishness of workers voting for no redistribution and for redistribution than libertarian predictors. This effect is significant when comparing within-predictor beliefs.

This finding can also be shown in a regression framework. Letting the binary variable D^E indicate a predictor's fairness view, which is equal to 1 if a predictor

[10] The reported number of flips had 0.000 % probability of being observed under honest reporting, i.e., observing $(S|E) = 6.11 \times N_E$ or $(S|L) = 6.55 \times N_L$ from a binomial distribution with $p = \frac{1}{2}$ and $N' \times 10$ trials. This held for the participant group as a whole and all subgroups of participants.

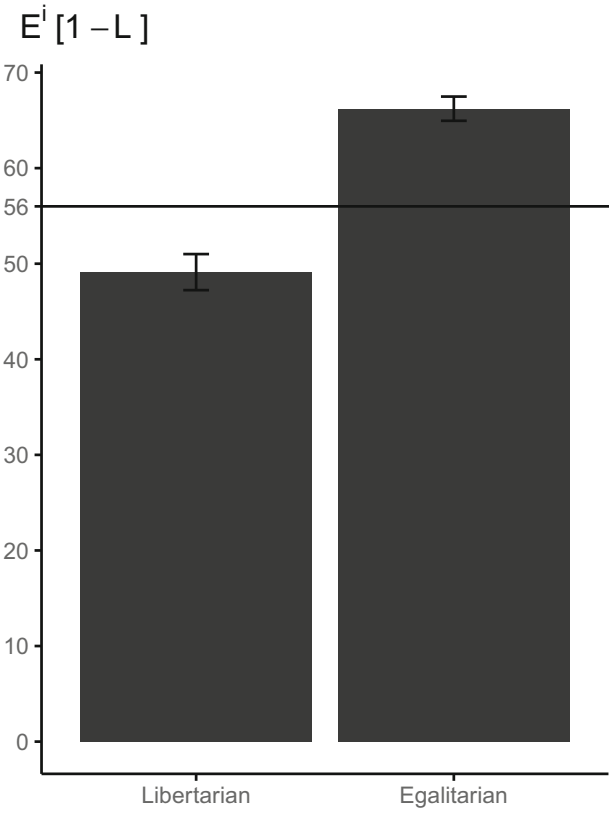

Fig. 2 Projection in fairness views. Distribution of the predictors' estimates of the number of workers who find redistribution fair, $E^i[1-\overline{L}]$, as reported by predictors' fairness view. The solid horizontal line indicates the sample value for the 210 predictors; $(1-\overline{L}) = 56$ % of the predictors cast an unincentivized vote that workers should redistribute

finds the redistribution option the fair option and 0 otherwise, yields the following equation:

$$E^i[(S|nr) - (S|r)] = \beta_0 + \beta_1 D^E + \epsilon_i. \tag{10}$$

The difference in the predictors' expectations of selfishness contingent on voting according to the predictors' fairness views, β_1 in (10), can be estimated as $0.51(p = 0.063)$, whereas β_0 can be estimated as $1.093(p = 0.000)$. Egalitarians' predictions of the differences in selfishness between workers voting for redistribution or no redistribution are about one-third larger than those of libertarians.[11] The differences in the predictors' attributions based on their fairness views are the main finding of the experiment: egalitarians considered a vote for no redistribution to be a

[11] The probability of observing the estimate given a zero effect, p value, is indicated in parentheses throughout the main text. The standard deviations of the estimates can be found in Appendix (Schøyen, 2022)

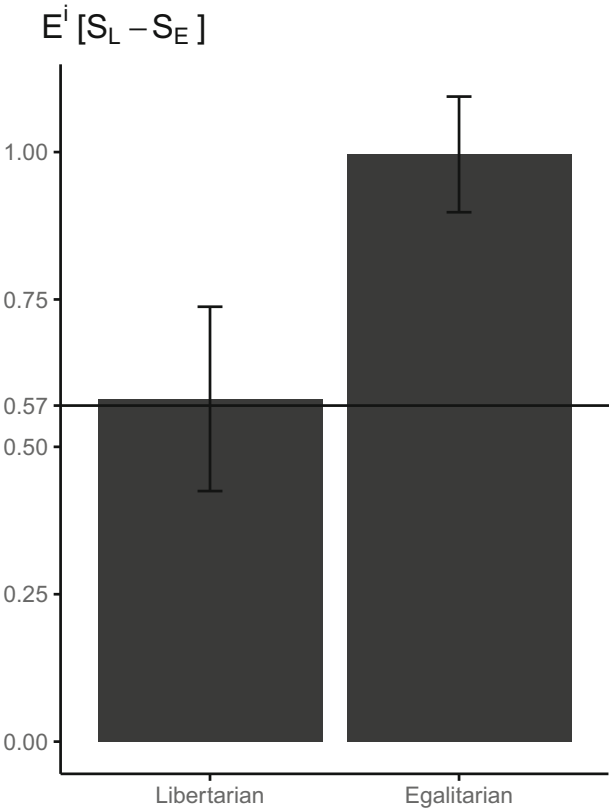

Fig. 3 Out-group stereotypes against libertarians. Difference in the predictors' reported expected average number of coin flips between workers casting an unincentivized vote for no redistribution or full redistribution, $E^i[(S|L) - (S|E)]$, as reported by predictors' fairness view. The solid horizontal line indicates true differences in the reported coin flips between the participants casting an unincentivized vote for no redistribution or full redistribution in the sample of 210 predictors; $(S|nr) - (S|r) = 0.57$ coin flips

significantly stronger signal of selfishness than libertarians. This suggests that one's fairness view affects one's intention attribution. This finding is robust when controlling for the participants' genders, political preferences, expected coin flips of an average participant, and the participants' own reported coin flips (Schøyen, 2022).

Based on (10), I establish $\beta_0 > 0, \beta_1 > 0$ as the first result, and based on $\beta_1 > 0$, I establish the second result.

Result 1 The participants interpreted votes for no redistribution as signals of selfishness.

Result 2 The participants' interpretations of the intentions behind redistribution choices differed according to their fairness views.

The predictors displayed significant projection bias in their fairness views; the libertarian predictors reported that they thought half the workers were libertarians,

$E^L[1-\overline{L}]=0.5$, whereas egalitarians thought that about two-thirds of the workers were egalitarians, $E^E[1-\overline{L}]=0.67$. The actual prevalence among the 210 predictors was $1-\overline{L}=0.56$, which is significantly different from either group's estimate at the 0.01 level. This is shown in Fig. 2. Thus, both egalitarian and libertarian predictors overestimated the commonness of their fairness views, and both displayed projection bias. This result is highly robust even when including controls (Schøyen, 2022). Based on this, the result of projection bias is as follows:

Result 3 The participants displayed projection bias in fairness views.

Out-group bias against libertarians, as measured by the difference between real and expected selfishness, were present for egalitarian predictors. Egalitarian predictors expected the difference in selfishness between participants casting unincentivized votes against and for redistribution, $E^i[(S|L)-(S|E)]$, to be equal to 1. Libertarian predictors expected this difference to be equal to 0.58. The difference for the sample $(S|L)-(S|E)$ for the 210 predictors was 0.57. Thus, as can be seen from Fig. 3, while egalitarian predictors had upwardly biased perceptions of selfishness, the libertarians did not.

I now turn to analyze the differences in out-group stereotypes. I do this by regressing the expected differences in workers' selfishness, $E^i[(S|L)-(S|E)]$, according to the predictors' fairness views, D^E

$$E^i[(S|L)-(S|E)]=\beta_0+\beta_1 D^E+\epsilon_i. \tag{11}$$

The effect of a predictor being egalitarian regarding out-group stereotypes, β_1, is estimated at $0.41(p=0.110)$. Controlling for gender, political preference, and reported coin flip, the estimate of the equivalent of β_1 is slightly less at $0.36(p=0.160)$, whereas including the participants' prevalence estimates of egalitarians raises the estimate and significance of the equivalent of β_1 to $0.50(p=0.060)$ (Schøyen, 2022).

Both egalitarians and libertarians expected libertarians to be more selfish, but egalitarians did so more than the actual observed difference. This is the fourth result.

Result 4 The egalitarians had biased out-group stereotypes against libertarians.

Both projection bias and out-group stereotypes were significantly correlated with differences in predictors' perception of the selfishness of workers, contingent on their vote (Schøyen, 2022).

5 Interpretation of results

The data support significant differences compatible with suspicious attribution among the egalitarians in the sample. Predictors were paid according to their predictions' accuracy, and the setting encouraged them to provide their best guesses of the behavior of fellow group members. The combination of material incentives and framing as a competition for accuracy makes it reasonable to assume that the measured effect does not reflect differences in the participants' points of view.

The effect of participants' fairness views on the intention attribution of the votes ($E^i[(S|nr) - (S|r)]$) is larger than age, gender, and political preference (Schøyen, 2022). Regression analysis of $E^i[(S|nr) - (S|r)]$ points to that both projection bias and out-group stereotypes contribute to the difference intention attribution, supporting the proposed theory developed in Sect. 2. The analysis points to out-group stereotypes as the main driver of $E[(S|nr) - (S|r)]$ (Schøyen, 2022).

One could interpret the coin-flipping task as a signal of dishonesty, which can lead to a more minimal interpretation of the results as an indication that voters for nonredistribution are seen as more dishonest. Implicitly, the coin-flip measure of selfishness assumes that a salient reason to deviate from a norm of not lying is the monetary incentive; that is, that people are selfish and that the norm of honesty is universal. However, when asked how many workers found no redistribution fair with no incentives over the outcome, egalitarians believed, on average, that one-third of participants had libertarian fairness views, weakening this interpretation. Further, the above-mentioned survey finding attitudes about lying to be equal weakens this explanation. Thus, although being a less minimal explanation, it accounts for more of the data. Note that the more minimal interpretation of nonredistributors as more dishonest could also generate polarization. If one believes that one's opposition is generally dishonest, it makes it less interesting to engage them in constructive discourse. This breakdown of communication can lead to polarization, because, as the crowd becomes internal, the message becomes more radical and confrontational toward the opposition.

The asymmetry in suspicious attribution between egalitarians and libertarians (Result 1) could be sample specific. Studies of fairness norms show that Norway is a country with a majority of egalitarians (Almås et al., 2019), which could affect several macro-level factors, such as school curricula and media perspectives.[12] One indication of this tendency is the common assessment of Norway as having a strong social democratic tradition, which is also part of its national identity (see, e.g., Sejersted (2011)). This could lead egalitarians to strengthen their suspicious attributions. Conversely, the social democratic tradition might weaken and even reverse any suspicious attributions among libertarians. If culture is the explanation for this asymmetry, redoing the experiment on a sample with a more libertarian culture might give suspicious attribution among both egalitarians and libertarians or only libertarians.

The asymmetry in suspicious attribution between egalitarians and libertarians (Result 1) could also be connected to the topic of the vote: redistribution. An egalitarian concern for equal reward regardless of effort could be more easily understood than a more nuanced libertarian mental model connecting effort to reward. Some support for this explanation is found in studies showing that, as children age, they develop increasingly nuanced mental models of what constitutes fair distribution, hence reducing the portion of egalitarians (Almås et al., 2010).

[12] Supportive of this explanation is that the participants' political preferences were generally not correlated with their fairness views for this sample (Schøyen, 2022) This finding is in contrast to a large sample online study of German respondents finding a significant correlation between fairness views and political preferences (Müller & Renes, 2021). The difference can be explained by larger differences between political parties in Germany than in Norway.

However, this cannot account for the significant differences between the predictions of egalitarians and libertarians (Result 2).

Interpreting the data in support of suspicious attribution builds on the assumption that both egalitarians and libertarians find lying to be equally selfish. For example, if libertarians would find lying more selfish than voting according to incentives, they would map an equal assessment of worker selfishness to fewer coin flips.[13] To verify that beliefs of lying are orthogonal to the fairness view, I conducted an online study on Amazon's Mechanical Turk platform (MTurk) of 99 US-based respondents to ensure that our selfishness measure was uncorrelated with ideological preferences. This confirmed that the measure of selfishness was orthogonal to fairness views and that lying was seen as immoral by 88 of the 99 respondents (Schøyen, 2022).

The workers were asked to report coin flips before voting and could thus be prone to moral licensing (Blanken et al., 2015; Simbrunner & Schlegelmilch, 2017). Moral licensing for the workers would imply making a trade-off between ethical behavior in the coin-flip task and the voting task. There are no available data to test whether the workers were prone to do this trade-off; libertarian and egalitarian workers would do moral licensing toward different votes and the fairness views of the workers are not known. The differences in attribution of votes (Result 2) could be affected by predictors believing the workers to be prone to moral licensing. The available data can neither confirm nor reject this. Since workers of different fairness views would consider different votes as the morally right action, any effect of moral licensing on the differences in attribution of votes (Result 2) would be mediated through predictors' prevalence estimates.

Finally, the theory considers the case of no correlation between worker fairness view and output level, i.e., $P(\underline{o}|L) = P(\overline{o}|E) = \frac{1}{2}$. This case is compatible with our experimental data, and there is an estimated zero correlation (estimated at -0.092 with a p value of 0.729) between a worker's output and fairness view.

The experiment applied a combination of well-established elicitation methods of beliefs in experimental studies.[14] However, these elicitation methods are not without issues.[15]

[13] This follows from that if the predictors expect the workers' utility loss of deviating from honest reporting differs across fairness views (e.g., $v_E^{'}(rh_i - oh_i) \neq v_E^{'}(rh_i - oh_i)$), Theorem 1 of a common separating threshold of selfishness ($\beta^* \equiv \frac{\bar{y}}{v(\bar{y})}$) does not hold.

[14] The approach of using choices made by a party with no monetary incentives in the outcomes to signify fairness views, which was first used by Harsanyi (1962), and a coin-flipping task as a measure of selfishness have previously been used as an unobtrusive measure of dishonesty (Cohn et al., 2014) It also applies the quadratic scoring for eliciting predictors' beliefs following (Blanco et al., 2010). Finally, the predictors are incentives to predict an in-experiment task rather than a vignette design, reducing the risk of inattention and decreasing the distance to real-world behavior.

[15] The belief elicitation is made using the quadratic scoring rule. The rule is $b = A - K(\text{ estimate - true value})^2$, where b is how many predictors are paid. For all estimates, A = 20 kroner. For estimates in the [0,100] range $K = \frac{1}{11.25}$ and for estimates in the [0,10] range $K = \frac{1}{1.125}$. If the value of b is negative, the respondent will be paid nothing for their estimate. There is little evidence that more elaborate elicitation methods improve the accuracy of belief elicitation (Blanco et al., 2010). Eliciting beliefs with incentives for correct estimates sequentially correlated measures could create incentives for hedging (Blanco et al., 2010; Crosetto et al., 2020).

Although, bias in the elicitation methods could theoretically bias the level estimates, for example, $E[S_r]$, the empirical analysis has investigated the difference between the expectation of redistribution and no-redistribution votes between the estimates of libertarians and egalitarians, that is, $E^E[S_{nr} - S_r] - E^L[S_{nr} - S_r]$. Thus, there is no evidence or indication that these issues would account for any of the established results (Results, 1,2,3 and 4).

6 Conclusion

The current paper has documented significant differences in how participants understand the intentions behind votes against redistribution according to the participants' fairness views. These differences correlate with the two other findings of projection bias and out-group stereotypes against libertarians among participants. Suspicious attribution among egalitarian, but not libertarian, participants is the best available explanation of the results.

The experimental design could be altered to see if suspicious attribution generalizes to intention attribution across other cleavages, such as intention attribution between individuals with different ethnicities, genders, national identities, religions, and lifestyle choices. One example is religious practices. Assuming that dedication to faith and beliefs about correct practice vary, do more orthodox believers suspiciously attribute the intentions of liberal statements to selfishness? Another application of the model could be to understand how people attribute the intentions underlying stated views on immigration policies. Assume people hold varying beliefs about the social desirability of open borders and varying degrees of willingness to contribute to social welfare. Do people who believe in the social desirability of an open-border immigration policy have a suspicious attribution for votes against open borders?

Suspicious attribution bias could be an enduring challenge for societies with a large plurality of moralities. The bias implies that larger heterogeneity of behavior could lead to more behavior being attributed to selfishness. This will lead members of heterogeneous societies to have a bias in their estimates of the number of selfish types. Moving beyond the attribution of nonselfish intentions to hostile intentions, the mechanism can create cycles of distrust and eventual conflict. Hence, perhaps, the most important direction of future research is investigating what can decrease our tendency for suspicious attribution. Such research may contribute to identifying interventions that can reduce prejudice and conflict.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11238-023-09965-5.

Acknowledgments The author is grateful to the FAIR Centre, Department of Economics, Norwegian School of Economics (NHH) for organizational support in conducting the experiment. Special thanks to Xianwen Chen for excellent research assistance on conducting the experiment. The author would like to thank my PhD supervisor Bertil Tungodden for excellent comments, support, and encouragement. The author would also like to thank Garry Charness, Ole-Andreas Elvik Næss, Tristan Gagnon-Bartsch, Andrea Mannberg, Jonas Tungodden and seminar participants at Department of Strategy and Management, Norwegian School of Economics, Hitotsubashi Institute for Advanced Study, Oslo Metropolitan University, the 2018 ASREC Conference at Chapman University, Cologne University, and the 2016

Norwegian Meeting of Economists for their comments. Schøyen gratefully acknowledges the warm hospitality at Hitosubashi University's Institute of Economic Research and Stanford University Department of Economics where part of this work was done.

Funding Open access funding provided by UiT The Arctic University of Norway (incl University Hospital of North Norway). The funding has been received from Norges Forskningsråd with Grant nos. 236995, 262675.

Data availability The data that support the findings of this study are available upon reasonable request from the author.

References

Alicke, M. D., Dunning, D. A., & Krueger, J. (2005). *The self in social judgment*. Psychology Press.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science, 328*(5982), 1176–1178.

Almås, I., Cappelen, A., & Tungodden, B. (2019). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *NHH Dept. of Economics Discussion Paper*, (4).

Binder, S. (2015). The dysfunctional congress. *Annual Review of Political Science, 18*, 85–101.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H.-T. (2010). Belief elicitation in experiments: Is there a hedging problem? *Experimental Economics, 13*(4), 412–438.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H.-T. (2014). Preferences and beliefs in a sequential social dilemma: A within-subjects analysis. *Games and Economic Behavior, 87*, 122–135.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41*(4), 540–558.

Cappelen, A. W., Falch, R., & Tungodden, B. (2020). Fair and unfair income inequality. *Human Resources and Population Economics: Handbook of Labor* (pp. 1–25).

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review, 97*(3), 818–827.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature, 516*(7529), 86–89.

Crosetto, P., Filippin, A., Katuščák, P., & Smith, J. (2020). Central tendency bias in belief elicitation. *Journal of Economic Psychology, 78*, 102273.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25*(1), 1–17.

Dawes, R. M., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude. *Journal of Personality and Social Psychology, 21*(3), 281.

Fryer, R. G., Jr., Harms, P., & Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association, 17*(5), 1470–1501.

Gagnon-Bartsch, T. (2017). Taste projection in models of social learning. *Working Paper*.

Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PloS One, 7*(12), e50092.

Greif, A., & Tadelis, S. (2010). A theory of moral persistence: Crypto-morality and political legitimacy. *Journal of Comparative Economics, 38*(3), 229–244.
Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.
Harsanyi, J. C. (1962). Bargaining in ignorance of the opponent's utility function. *Journal of Conflict Resolution*, 29–38.
Kesebir, S., & Haidt, J. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed.).
Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. *The American Economic Review, 87*(2), 359–364.
Madarász, K. (2015). Projection equilibrium: Definition and applications to social investment and persuasion. *Working paper*.
Müller, D., & Renes, S. (2021). Fairness views and political preferences: Evidence from a large and heterogeneous sample. *Social Choice and Welfare, 56*(4), 679–711.
Piketty, T. (1995). Social mobility and redistributive politics. *The Quarterly Journal of Economics, 110*(3), 551–584.
Reeder, G. D., Pryor, J. B., Wohl, M. J., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin, 31*(11), 1498–1510.
Roemer, J. E. (2009). *Equality of opportunity*. Harvard University Press.
Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301.
Rubinstein, A., & Salant, Y. (2016). Isn't everyone like me?: On the presence of self-similarity in strategic interactions. *Judgment and Decision Making, 11*(2), 168.
Schøyen, Ø. (2017). Pre-analysis plan: Attributing intentions under projection bias. American Economic Association RCT Registry at: https://www.socialscienceregistry.org/trials/2310.
Schøyen, Ø. (2022). Suspicious minds and views of fairness: Online appendix. *Theory and Decision*.
Sejersted, F. (2011). *The age of social democracy: Norway and Sweden in the twentieth century*. Princeton University Press.
Simbrunner, P., & Schlegelmilch, B. B. (2017). Moral licensing: A culture-moderated meta-analysis. *Management Review Quarterly, 67*(4), 201–225.