Chapman University Digital Commons

ESI Working Papers

Economic Science Institute

6-2024

Moral Content Diminishes Preference Falsification

Maxine Bonneau
Imperial College London

Tanya O'Garra
Imperial College London

Praveen Kujal
Chapman University

Follow this and additional works at: https://digitalcommons.chapman.edu/esi_working_papers

Part of the Econometrics Commons, Economic Theory Commons, and the Other Economics Commons

Recommended Citation

Bonneau, M., O'Garra, T., & Kujal, P. (2024). Moral content diminishes preference falsification. *ESI Working Paper 24-11*. https://digitalcommons.chapman.edu/esi_working_papers/407/

This Article is brought to you for free and open access by the Economic Science Institute at Chapman University Digital Commons. It has been accepted for inclusion in ESI Working Papers by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

Moral Content Diminishes Preference Falsification

Comments

ESI Working Paper 24-11

This article is available at Chapman University Digital Commons: https://digitalcommons.chapman.edu/esi_working_papers/407

Moral Content Diminishes Preference Falsification [1]

Maxime Bonneau, Tanya O'Garra[2], Praveen Kujal[3]

Abstract:

We examine how the moral (or neutral) content of an issue influences the tendency to falsify attitudes, given varying social and monetary incentives to engage in such 'preference falsification'. We conduct an incentivized two stage online experimental study where, in a prior first stage, attitude strength over moral and neutral issues is elicited. Then, in the second stage participants in groups of ten were asked to express their preferences regarding the moral or neutral issues for each possible combination of supporters and opposers in their group, each associated with varying monetary payoffs. More than half of the participants falsify their preferences between the two phases for both moral and neutral frames. The rate is significantly lower for the moral (vs neutral) issues. Participants' average monetary cost to avoid falsifying preferences is higher in moral treatments, and increases with the level of attitude strength.

[1] Acknowledgements: We would like to thank Gary Charness, Michalis Drouvelis, Andrej Angelovski, and seminar participants at Middlesex University, BSB Burgundy School Of Business, the Sciences Po LIEPP.

[2] Centre for Environmental Policy, Imperial College London, South Kensington Campus, U.K

[3] Accounting, Finance and Economics Department, Middlesex University Business School, The Burroughs, London, NW4 4BT, U.K.

1. Introduction

In his book, Private Truths, Public Lies, Kuran (1995) explores the causes and consequences of holding a set of opinions privately while expressing another in public. He terms this phenomenon Preference Falsification, i.e., when individuals express attitudes that conflict with their private beliefs. He argues that there are many situations in which individuals face a dilemma between extrinsic rewards[4] from voicing opinions they may not privately agree with, and intrinsic rewards from asserting their personal views. Divergence between private and public preferences may explain the persistence of widely-disliked behaviors or policies, such as abusive work cultures highlighted by the #metoo movement, or authoritarian political regimes (Kuran, 1995; Jiang and Yang, 2016; Kalinin, 2018). This suppression of private preferences has important implications for social policy because contexts in which a majority are falsifying their preferences are potential sources of abrupt political and social change (Kuran 1995; Ross et al 2023). In such circumstances, small changes in, e.g. information about private preferences in the public domain, or in the extrinsic rewards from falsifying one's preferences, may tip the social equilibrium (Ross et al. 2023; Frank 1996).

How Kuran's (1995) theory may have important implications is best illustrated by an example from Frank (1996). Frank (1996) points out that the interesting dynamics emerge from the dependence of reputational utility *on the distribution of people* who favour the two alternatives, i.e. agree or disagree, on an event. He argues that (sic) *"as more and more individuals favor one position publicly, the reputational cost of favoring the alternative position rises, and vice versa. Thus, if p denotes the proportion of the population that publicly opposes the regime, each individual may be seen as having a threshold value of p above which he, too, will oppose the regime. If this occurs the "it is no mystery that seemingly small changes can wreak social havoc."* He illustrates this with an example (p-16, 1996) where a population is made up of ten individuals who have the following thresholds for speaking out against the regime:

Person	a b c d e f g h i j
Threshold	0 .2 .2 .3 .4 .5 .6 .7 .8 .9

One can see that individual *a* strongly opposes the regime and will speak out regardless. Next, in the strength of their opposition, *b* and *c*, who will speak out only if at least 20 percent of the population speaks out. Individual *j*, the person who most favors the regime, will speak

[4] Extrinsic rewards capture all kinds of rewards outside of the self, e.g., monetary, reputational or social approval. An example of this can be experimenter demand in experiments.

against it only if at least 90 percent of the population does so. He argues that the equilibrium outcome here is that only *a* speaks out implying that the regime has the support of 90% of the individuals. Now, consider that *b* has an unpleasant encounter with government officials. Due to this (experience) the threshold for *b* decreases to 0.1. This then starts a domino effect, with 20% of the population speaking out, *c*'s threshold is met, and so on, resulting in the toppling of the regime. The example above illustrates why preference reversal may have important consequences for social equilibria. That is, seemingly minor events, i.e. a bad experience, may tip the social equilibrium.

Preference falsification is related to other key phenomena in social psychology, particularly social norms, collective action and conformity (see Ross et al. 2023 for an overview). What these phenomena have in common is a focus on how individuals make decisions given strategic considerations in their interactions with others. Social norms influence individual actions to align with group standards according to expectations of acceptable behavior (Biccieri 2017). In collective action situations, individuals may face decisions about whether to contribute to a collective goal, considering the motivations and actions of other group members. Conformity may occur due to the fear of rejection or the desire to fit implicitly or explicitly with a reference group[5]. (e.g. Sherif, 1937; Asch, 1961, 1956; Moscovici, 1976; Cialdini and Goldstein, 2004).

Although all these concepts have overlaps, preference falsification is particularly focused on social dynamic effects in settings where norms are in the process of shifting or where there is uncertainty about which norms are emerging or gaining acceptance (Ross et al. 2023; Kuran 1995). Preference falsification is typically studied with reference to political and social events, using ethnographic case-studies (e.g. Kuran, 1991; Weeden 1999; Bhaumik 2002), and more recently, observational approaches (e.g. Jiang and Yang, 2016; Kalinin 2018; Shamaileh, 2019). Few studies have studied preference falsification using experiments (as typically used to examine conformity, social norms and collective action problems) where the experimenter has some control over extraneous factors. Given shifting or uncertain norms and different intrinsic motivations and preferences, online or laboratory experimental studies represent an important complement to typical approaches used to study preference falsification. For example, using experiments, intrinsic motivations can also be manipulated, by studying preference falsification with regards to different topics of varying importance to participants. External motivations (e.g. social pressures, monetary payoffs) can also be manipulated to increase or decrease the cost of

[5] Empirical studies suggest that the factors that motivate preference falsification include social pressures, fear of social or economic reprisal, desire for social approval or acceptance, and a perception that expressing true preferences or opinions may have negative consequences (Kuran, 1995).

not falsifying preferences. And norms (aka majority expressed preferences) can be experimentally manipulated so as to shift or change over time, allowing for elucidation of thresholds beyond which preference falsification occurs (or vice versa).

This study uses the experimental approach to do all three. First, we manipulate intrinsic motivations by examining how people respond vis a vis issues that may be perceived as moral versus neutral. Morality is important for individual decision making. Numerous studies on conformity have found that the (perceived) morality of an issue can increase resistance to it (Skitka, Bauman and Sargis, 2005; Hornsey et al., 2003, 2007; Aramovich, Lytle and Skitka, 2012; Skitka, Bauman and Mullen, 2008; Skitka, Washburn and Carsel, 2015) across a range of issues including torture (Aramovich, Lytle and Skitka, 2012), gay law reform, or atrocities against Aboriginal populations (Hornsey et al., 2007). We consider a neutral (i.e. baseline) and two moral (terrorism and Covid) issues. Importantly, we examine how *attitude strength* regarding the issue under study affects the likelihood that an individual will falsify her/his own private preference about the issue, providing us with additional insights about the interplay between intrinsic motivations and preference falsification. Secondly, we manipulate extrinsic motivations to falsify preferences using monetary incentives that increase over several rounds. Participants were offered monetary incentives to falsify their preferences and asked: what is the maximum monetary cost (i.e. 'willingness to forgo') an individual would be willing to bear in order to express his or her private preference for a moral and neutral issue? Simultaneously, we shifted the prevalent norm in the participants' experimental group (see below for details) to emulate shifting or uncertain norms (see Discussion later).

Two recent studies have examined conformism/preference falsification in one-shot interactions. Charness, Naef, and Sontuoso (2019) found that 'opportunistic conformism' may occur if it leads to a net increase in the individual's material payoff. They use the strategy method to implement a coordination game, where the team's payoff depended on the choices made by all team members. Meanwhile, Bursztyn et al. (2020) assessed Pakistani men's willingness to forgo money to preserve their anti-American identity. In an experiment, where they had to tick a box to accept a bonus from the U.S. government in exchange for the completion of a survey, they found that monetary incentives decreased the propensity of anti-American expression[6]. However, almost 25% of participants did forgo payments of up to R.s 100 ($0.35) to avoid misrepresenting their views via the checkbox.

[6] They also examined social influences by leading subjects to believe that their decision (ticking the box or not, the bonus acceptance) would be either private or public. When participants anticipated that their bonus will be publicly revealed to the other participants, the bonus rejection rate decreased significantly by 10%.

Our paper uses a novel approach developed in Duffy and Lafky (2018) where, depending on the attitude expressed by participants during Phase 1, each participant is assigned to a group of ten participants. Then, using the strategy method (Selten, 1967) they have to indicate whether they support, or oppose, the issue at hand (see the methods section for more details) for each possible combination of supporters and opposers in their group (ex., nine in favor - one opposed, eight in favor - two oppose etc.)[7]

Our design has several interesting features. We quantify and compare the average willingness to forgo to falsify preferences in neutral and moral frames by incrementally varying the number of individuals in that group with the same (or opposing) views. We use a moral frame, as it is an important element in decision making. Some philosophers have argued that morality is an essential part of one's identity (Parfit, 1984) and is an important trait as regards an individual's sense of identity (Strohminger and Nichols, 2014). Participants' preference falsification is thus elicited by decreasing the number of people holding the same (opposing) view[8]. We interpret this as participants facing increasing both social and payoff pressure. Increasing social pressure is presumably faced as the proportion of individuals in the group with the same viewpoint decreases, while payoff pressure occurs when the payoffs decrease as the participant maintains their view as the pressure to falsify increases.

We then assess how attitude strength and moral conviction influence preference falsification for both moral and neutral scenarios. This is important as morality can significantly influence individual attitudes thus resisting preference falsification. Note, as we obtain the joint effect of social and material payoff, any result obtained in our framework would be a joint effect of the two relative to previous studies that only focus on one of the two. Finally, our sample is more representative than laboratory experiments. Studying conformity in online environments with large samples allows for increased generalizability, enhanced ecological validity and the ability to achieve larger sample sizes (N=577), improving the statistical power and representativeness of findings.

While we find significant evidence of preference falsification for both moral and neutral frames, it was significantly reduced with moral issues. For moral issues, the maximum cost participants were willing to incur to not falsify their preference increases by almost 8.48 tokens ($0.85) with an increase in attitude strength of 1. Overall, the mean 'willingness to forgo' to express

[7] We had earlier planned ro run the study in the laboratory. However, we had to ran online experiments due to the Covid pandemic.

[8] To our knowledge we are the first to use this approach in this framework.

one's own (non-majority) preference is 57.83 tokens (n = 577), equivalent to $0.59 per person[9] and it increases with attitude strength, with the mean 'willingness to forgo' for the highest level of attitude strength (4/5) being 80% of the individual's endowment. Finally, framing matters, i.e. the "nature" of the moral issue matters. That is, relative to Covid-19, participants were significantly more willing to pay to express their private preference in the Torture treatment. The paper is structured as follows: we first present the hypotheses, then the methods and the results. We then discuss the results and conclude.

2. Hypotheses

Our first hypothesis[10] explores the extent to which the perceived morality of an issue affects the likelihood that an individual will falsify her/his own preferences? Non-incentivized results in moral and social psychology have shown that the moral framing of issues increases the likelihood that people resist majority influence (Hornsey et al. 2003; Hornsey et al. 2007, Aramovich, Lytle and Skitka, 2012) and distance themselves from those that are morally dissimilar (Skitka, Bauman and Sargis, 2005). Also, Skitka (2022) and Skitka et al. (2021) found that individuals who held strong moral convictions[11] about an issue were more likely to feel a sense of moral obligation to act on those convictions and stand up for their preferences, even in the face of social pressure or opposition. People with strong moral convictions are also less likely to seek approval from others for their true worldview. As a result, compliance on informational and normative grounds decreases (Skitka, Washburn and Carsel, 2015).

Given that moral framing results in individual's resisting majority influence and distancing from morally dissimilar individuals, individuals are more likely to act on these convictions and individuals with strong moral convictions are less likely to seek approval, we can conclude that behavior with moral issues will be different than behavior with neutral issues. This gives us hypothesis 1:

H1: Preference falsification will be higher in neutrally framed choice settings than in morally framed choice settings.

The second hypothesis deals with the relation between attitude strength and preference falsification. Individuals want to be consistent with their self-concept so that they maintain a positive self-view (Swann, 1987; Mazar et al., 2008). Hence, not being true to oneself may incur a

[9] Participants were willing to forgo more than 50% of their endowment to keep their private preference intact.

[10] Created the 29th of May, 2020 accessible https://aspredicted.org/XCW_YAG

[11] 'Moral conviction is a subjective assessment that one's attitude about a specific issue or situation is associated with one's core moral beliefs and fundamental sense of right and wrong' (Bauman and Skitka, 2009).

personal cost to the falsifier (Freud, 1923; Kuran, 1995; Duffy and Lafky, 2018; Bursztyn et al., 2020). Social psychologists have shown that the preference falsifier is burdened by the guilt or anger for not standing up for her or his preferences (tastes, preferences, personal standards, identity), something that has been identified as the intra-psychic need for consistency (Hornsey et al., 2007).

Attitude strength increases the stability and resistance of an attitude (Krosnick and Petty, 1995; Hornsey et al. 2003, 2007; Skitka, Bauman and Sargis, 2005; Skitka et al., 2021). Moreover, moral conviction, another measurement of attitude strength, predicts attitude over and above attitude strength (Hornsey et al., 2003; Skitka et al., 2005, Hornsey et al., 2007). Therefore, moral issues would induce a greater need for consistency for individuals, and a stronger protection of the sense of self. This gives our second hypothesis:

H2: *In morally framed choice settings, the higher the attitude strength of an individual towards an issue, the higher is the monetary cost they are willing to incur to avoid falsifying their preference.*

3. Methods

The study was pre-registered and took place from July to August 2020. Ethics approval was obtained from the Middlesex University. The study involved two phases: a pre-survey (Phase 1) and an experiment (Phase 2) conducted two weeks later. Both were conducted through Amazon Mechanical Turk[12].

Phase 1 -The pre-survey

We conducted a pre-survey (n = 1,484) which was used to identify the two topical (moral) issues with the greatest distribution of supporters and opposers potential subjects for the next phase, the topical (moral) issues to be used in our experimental study, and the level of attitude extremity (one component of attitude strength) that each participant attached to them. Participants answered four questions about their preferences on topical issues that could be classed as having a 'moral' element, as well as one 'neutral' issue (using a 7-point Likert-scale, from -3 strongly-oppose to +3 strongly-support).

[12] Replication studies (Rand, 2012) found MTurk data collection to be consistent with other methods whereas the population is more diverse than in Internet and American college samples (Buhrmester, Kwang and Golsing, 2011; Mason and Suri, 2011).

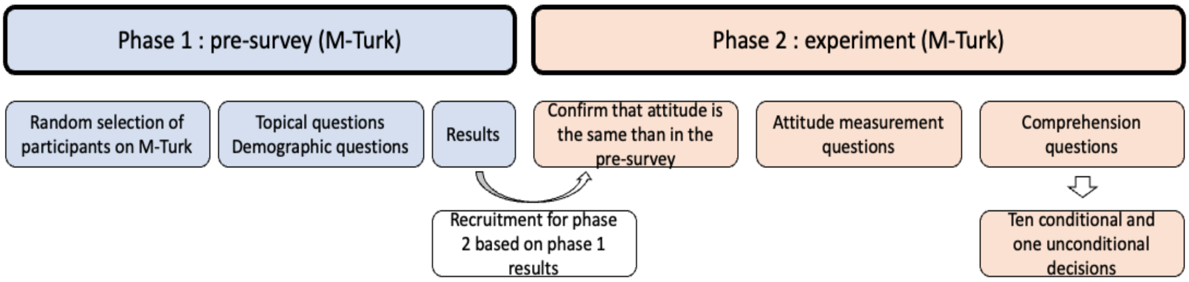
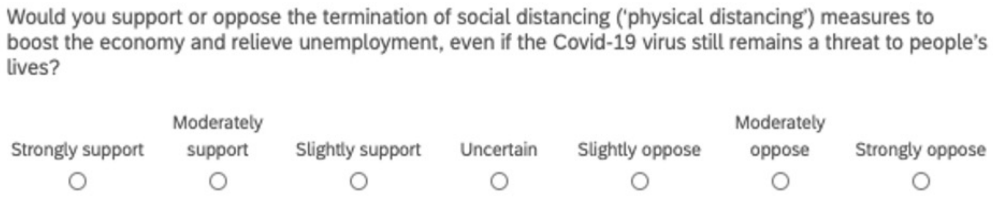

Phase 1 : pre-survey (M-Turk)
Phase 2 : experiment (M-Turk)
Random selection of participants on M-Turk
Topical questions Demographic questions
Results
Confirm that attitude is the same than in the pre-survey
Attitude measurement questions
Comprehension questions
Recruitment for phase 2 based on phase 1 results
Ten conditional and one unconditional decisions

Figure 1: Flow chart of the study.

The neutral (i.e. non-moral) issue is the preference of pasta over pizza and also serves as our control question. The 'moral' issues were selected from previous studies in social and moral psychology literature, as well as national surveys (Shwom, 2010 for gas emissions; Aramovich et al., 2012 for torture; Smith et al. GSS Survey, 2018 for gun permit and abortion). The question wording for most moral issues was obtained from previous studies. The question regarding Covid-19 is new. We ended up asking respondents about their views regarding torture, abortion, greenhouse gas emissions, gun permit, and Covid-19. An example of a question asked during the pre-survey is seen in Figure 2 below.


Would you support or oppose the termination of social distancing ('physical distancing') measures to boost the economy and relieve unemployment, even if the Covid-19 virus still remains a threat to people's lives?

Strongly support	Moderately support	Slightly support	Uncertain	Slightly oppose	Moderately oppose	Strongly oppose
○	○	○	○	○	○	○

Figure 2: Example of an attitude question in Phase 1 - Covid-19 topic.

The experiment: Phase 2

We used the pre-survey data to select the two topical (moral) issues with the greatest distribution of supporters and opposers (see Appendix B for more details about the results). This was done to have enough "supporters" and "opposers" for the second experimental phase which happened two weeks after the first experimental phase. This was important as later on we match participants to issues based on the revealed strength. The two issues with the greatest distribution of supporters and opposers were: Covid-19 and torture. The specific statements participants saw were:

- *Torture: "To what extent do you support or oppose the use of stress techniques when interrogating suspected terrorists, such as sleep deprivation, 'water boarding', long periods of hanging detainees by ropes in painful positions, etc.?"*

- *Covid-19: "Would you support or oppose the termination of social distancing ('physical distancing') measures to boost the economy and relieve unemployment, even if the Covid-19 virus still remains a threat to people's lives?"*

The average attitude strength (over a maximum score of 5) for Covid-19 was 3.69 and 3.39 for torture. As noted before, the neutral issue concerned preferences for pizza or pasta. See Appendix D for instructions.

As mentioned earlier participants were chosen based on the strength of their attitude in the pre-survey. Each participant was assigned to *only one* of the pre-survey issues (i.e., either torture, Covid-19 or food). We prioritized selecting strong supporters and opposers, over moderate/weak supporters and opposers as it would improve the possibility of getting statistically significant elicitations. For each moral issue, we sent one version to those who expressed opposition for the Covid-19 or torture question in the pre-survey and another version to those who expressed support for the Covid-19 or torture question (for the neutral issue, we distributed one 'prefer pizza' version and another 'prefer pasta' version). For instance, if a participant strongly opposed *the use of stress techniques (torture when interrogating suspected terrorists)* in the pre-survey, we sent him the *oppose version* of the torture links. The final experimental sample consists of n = 577 participants.

We also made sure that participants were consistent between the experiment and the pre-survey. We did this by asking them the same question in the pre-survey and the experiment. For instance, if they received the *oppose version* of the torture links, i.e. they had opposed *the use of stress techniques when interrogating suspected terrorists* in the pre-survey, then we asked them the same issue (*Would you support or oppose the use of stress techniques …*) but with only three choices : Support, neither support nor oppose, or oppose. Importantly, if their answer was not consistent with the one in the pre-survey, then they were withdrawn from the experiment.

Attitude strength measurement questions

We also asked those who were consistent between the experiment and the pre-survey to answer questions regarding attitude certainty on a scale of five (*How certain are you that of all the possible attitudes one might have toward this topic, your attitude reflects the right way to think and feel about the issue?*), attitude importance (*To what extent is your attitude about this topic personally important to you?*) and moral conviction about the issue (*To what extent does your attitude about this topic reflect your core moral values and convictions?*). These make up our attitude strength, which measures the extent to which the issue is important for the participant on a scale of five (Petty and Krosnick, 1995). It is computed by averaging attitude extremity, attitude importance, attitude certainty and moral conviction about the issue at hand. Attitude extremity refers to the extent to which the attitude deviates from neutrality for individuals and is measured by the extent to which they support or oppose an issue, on a scale of 7 from -3 “strongly support/oppose” to +3 “strongly oppose/support” (Abelson, 1995; Judd & Brauer, 1995). Attitude importance refers to the degree of psychological significance people attach to a given attitude. This is measured by asking how important or unimportant the issue was, on a scale of 5 from 1 “not at all important” to 5 “very important” (Boninger, Krosnick, Berent, & Fabriga, 1995). Attitude certainty refers to the degree that people feel sure about their position on a particular issue, on a scale of 5 from 1 “not at all certain” to 5 “very certain” (Gross, Holtz, & Miller, 1995). Finally, moral conviction refers to how an attitude is related to one’s core moral beliefs about right and wrong, on a scale of 5 from 1 “not at all reflect my core moral values and convictions” to 5 “very much reflects my core moral values and convictions” (Skitka, Bauman and Sargis, 2005; Aramovich et al., 2012).

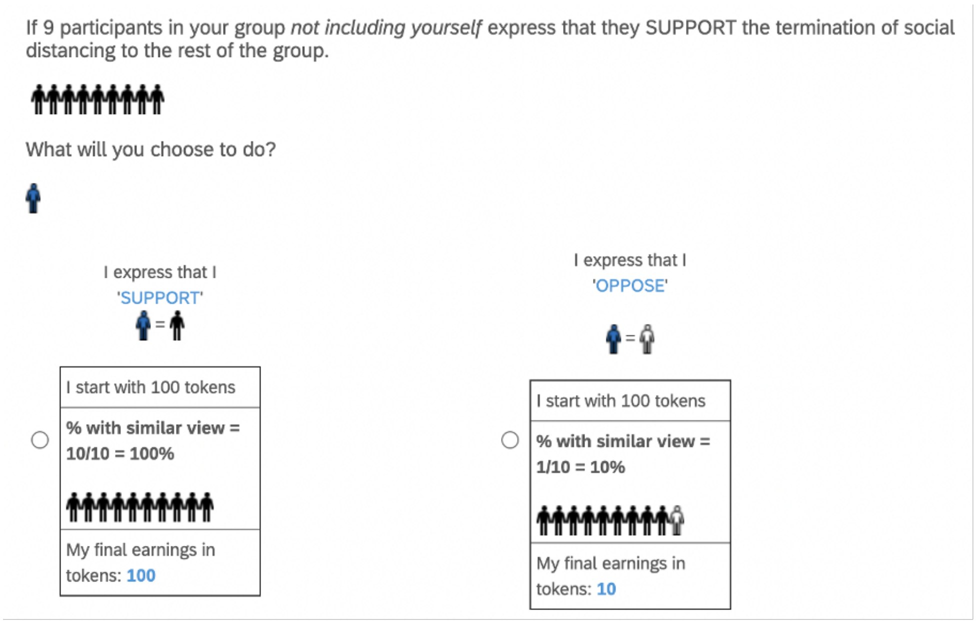
4. The experiment

We implemented an experiment (based on Duffy and Lafky, 2021) in which participants were asked to express their opinion towards issues taken from the pre-survey. Note that we use the strategy method to first elicit responses. Participants are then assigned to groups of ten players. The players were selected based on their answers in the pre-survey, so that each group was composed of group members with similar opinions over the issue. Each participant was paid $0.90 for participation, and allocated 100 tokens ($1.00) as their experimental endowment. They were then informed that their task was to indicate whether they opposed or supported the issue (food, torture or Covid-19). Participants were then explained how the final bonus payment was calculated: that the initial bonus of 100 tokens would be multiplied by the percentage of people in their group, during the experiment, with the same view as their own. For instance, if a participant chose to oppose torture in the experiment, then we multiplied 100 tokens by the number of participants in the group that also chose to oppose torture. If 5 out of 10 of the group members chose to oppose torture as well, that makes a final bonus of 100 * (5/10) = 50 tokens. Comprehension questions were asked and people who failed twice were withdrawn from the experiment.

Payoff calculation

Since the experiment was online and asynchronous, the payoff calculation could only be performed after all the participants completed the survey. As earlier mentioned, we implemented the strategy method (Selten, 1967) in order to elicit support or opposition in relation to the issue participants were assigned to. Participants indicated, for each composition of supporters and opposers in their group, if they supported or opposed the issue at hand to that same group. They had a total of ten decisions to make.

In the first decision, each participant had to express a preference in a homogeneous group in terms of pre-survey preferences (see Figure 3 below). In the second decision, the group composition changed so that the number of people with the same pre-survey preference decreased by one and the number of people with the opposite pre-survey preference increased by one, keeping the same total number of group members (10). The procedure was repeated from iteration 2-10. In this manner we could ask participants their choice for each possible group composition of supporters and opposers. By decision 10, each participant was in a group in which all remaining participants (nine) had opposing preferences about the issue at hand (based on the pre-survey answers). Importantly, note that participants could see their payoffs given the proportion with a similar view (figure-2, panels).



If 9 participants in your group *not including yourself* express that they SUPPORT the termination of social distancing to the rest of the group.

What will you choose to do?

○ I express that I 'SUPPORT'

I start with 100 tokens
% with similar view = 10/10 = 100%
My final earnings in tokens: **100**

○ I express that I 'OPPOSE'

I start with 100 tokens
% with similar view = 1/10 = 10%
My final earnings in tokens: **10**

Figure 3: Example of an experiment question in Phase 2 - Covid-19 supporters link, first conditional answer.

We put players with similar opinions at the beginning to capture the effect of resistance to preference falsification with different issues (neutral versus moral) and for different levels of attitude importance. Indeed, the way the experiment was presented to participants made them face increasing adversity: given that this was implemented using the strategy method, they responded to all group composition of supporters and opposers (the conditional decisions) in a sequential manner with no feedback[13].

To find out which conditional decision to apply for payoff calculation, we needed to know the exact composition of supporters and opposers. We asked participants to fill, prior to the ten conditional decisions, an unconditional decision: *"Considering that you are in a group of ten participants: Would you support or oppose the use of stress techniques when interrogating suspected terrorists, such as sleep deprivation, 'water boarding', long periods of hanging detainees by ropes in painful positions, etc.?".*

We then randomly assigned participants to groups of ten for the purpose of payoff calculation (after all the participants completed the experiment). Note, the pool of participants was composed of all the participants for each version of the questionnaire. We randomly assigned participants that opposed and supported the issue (in the above example, *the use of stress techniques when interrogating suspected terrorists*). Based on their unconditional decision, we computed how many supporters and opposers were there in each group. That matched necessarily one of the ten conditional decisions (e.g., 4 opposers and 6 supporters).

Eventually, we randomly determined which of their two decisions (unconditional or conditional) was relevant for the payoff calculation. One randomly selected participant had her or his conditional decision considered for payoffs calculation, while the 9 others had their unconditional decision implemented. The conditional decision was the one that corresponds to the actual number of unconditional supporters and opposers. This helped us compute the payoffs for each participant.

For example, for the issue "*the use of stress techniques when interrogating suspected terrorists"* we ended with 230 participants in total (supporters and opposers). They were randomly assigned to groups of ten. In each group, the last one was the participant for which the conditional decision was implemented. Based on the first nine participants' unconditional decisions, we computed that 3 participants supported and 6 participants opposed *the use of stress techniques when interrogating suspected terrorists*. We then looked, for the last participant of this group, for her or his conditional decision related to 3 supporters and 6 opposers. For this conditional decision, the player chose to oppose *the use of stress techniques when interrogating suspected terrorists*. That makes 3 supporters and 7 opposers. The participants who opposed were allocated with 70 tokens ($0.70), while the participants who supported were allocated with 30 tokens ($0.30).

[13] Participants spent, on average, 9.80 minutes to complete the questionnaire.

5. Results

Overall, we find that 57% of participants falsify their own preferences by conditional decision n°10. Preference falsification does not occur until decision 5, from decision 6 onwards, there is a steady increase in the proportion of falsifiers in each of the treatments. This was expected because the payoffs for falsifying preferences become greater than the payoffs for expressing one's private preferences only from decision 6. We discuss the results in detail below.

Moral versus neutral issues and preference falsification

Although preference falsification increases from decision 6 onwards for all treatments, we found this phenomenon to decrease with moral issues. The overall influence of the moral treatments (Torture and Covid-19) on preference falsification is significantly different from the neutral treatment (pizza versus pasta) in terms of the distributions of falsifiers and non-falsifiers (chi^2 test score = 57.958; p-value < 0.001%). Specifically, compared to the neutral treatment, for the Covid treatment the chi^2 score is 34.297 with p-value < 0.0001; for the Torture treatment the chi^2 score is 53.812 and p-value < 0.0001.

Moreover, in conditional decision number 10[14], while 80% of participants had falsified their preference in the Food/neutral treatment, only 50% (43%) had done so in the Covid-19 (Torture) treatments[15]. Importantly, the two moral treatments are not significantly different in terms of proportions of falsifiers (chi^2 test, score = 1.237; p-value 0.266). We show the proportion of falsifiers per treatment, as well as the 'expected' proportions of falsifiers in the case where only the monetary payoffs enter the utility function, in Figure 4 below. Our results thus support our first hypothesis: preference falsification is higher for neutral issues than for moral issues[16].

'Willingness to forgo' and preference falsification

We also estimate the mean 'willingness to forgo' by participants to express their own (non-majority) preferences. We first created a variable called 'opportunity cost' which is equal to the monetary loss (in tokens) someone would incur if she doesn't falsify her private preference, recorded as a continuous variable. Every time preference falsification occurred, we computed the

[14] When the participant making the decision is opposing (supporting) whereas the nine other group members chose to support (oppose) the issue at hand.

[15] Therefore 22% of participants in the pizza/pasta treatment did not change their mind whatever the material incentive to do so. It could suggest that they did not understand the experiment. The analysis of their comments in the discussion section suggests another explanation : they really care about their initial choice and want to be consistent with it.

[16] See Appendix C for panel data logistic regression results.

maximum cost an individual was willing to incur before falsifying. This cost equals 0 until decision 5 and is -10 in decision 6 and then decreases by 20 in each decision until decision ten where the opportunity cost reaches -90.

We find that the mean willingness to forgo is 57.83 tokens over all the participants (n = 577; s.d. = 40.56; median = 50) and is equivalent to $0.59 per person. In other words, participants are ready to forgo more than 50% of their endowment to keep their private preference intact in public. The same mean cost for participants with the highest level of attitude strength (equal to or above 4/5) was 79.5 tokens (n = 33; s.d. = 36.92; median = 100). This is equivalent to $0.80 per person or 80% of the endowment. We can conclude that the higher the level of attitude strength, the higher the monetary cost participants are willing to incur in order not to comply with the majority.

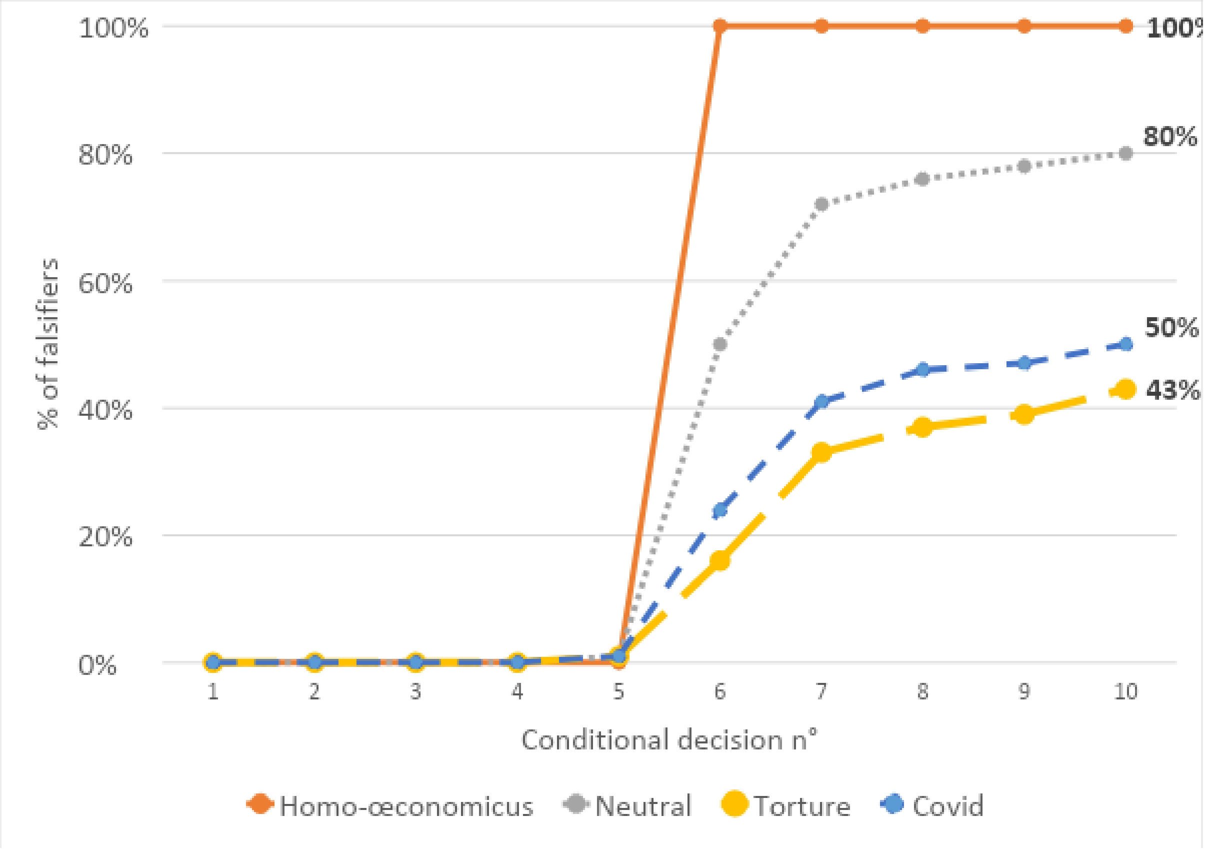


Figure 4: Preference falsification over conditional decisions per treatment. The conditional decisions are displayed on the horizontal axis while the percentage of falsifiers throughout the study is displayed on the vertical axis. The solid orange line (which ends up to 100%) refers to the 'expected' proportion of falsifiers if monetary payoffs are the only influence on choices, the dotted grey line (80%) refers to the neutral (Food) treatment, the dashed blue line (50%) refers to the Covid-19 treatment and the dashed yellow bordered line (43%) to the Torture treatment.

Participants were significantly more willing to pay to express their private preference about the issue in the Torture treatment as compared to the Covid-19. Participants in the Torture treatment with the highest level of attitude strength were willing to pay almost the entire endowment (89.55 tokens or $0.90 on average for attitude strength equal to or above 4/5) whereas the same kind of participants in the Covid-19 treatment were willing to pay 72.73 tokens ($0.73). In Table 5, below, we display the average amount that participants were willing to pay for each attitude strength level and each treatment.

The average willingness to forgo is lower in the Food treatment (39.62 tokens overall, s.d. = 37.1, median = 30) than in the Covid-19 (64.31 tokens, s.d. = 40.44, median = 100) and the Torture (71.15 tokens, s.d. = 37.32, median = 100) treatments. We hypothesized that participants in the neutral and moral treatments have the same distributions of ordinal outcomes (here the willingness to forgo). The hypothesis of equality of distributions was rejected at least at the 1% level (Wilcoxon-Mann-Whitney test,[17] z-score = 8.30, p-value = 0.000). In the second comparison, we hypothesized that participants in the Covid-19 and Torture treatments have the same distributions of ordinal outcomes. Here, the hypothesis of equality of distributions was not rejected (Wilcoxon-Mann-Whitney test, z-score = 1.56, p-value = 0.12). This is in line with our findings regarding the first hypothesis.

Table 1: Average maximum cost participants were willing to pay (per treatment) to avoid falsifying their preference for each attitude strength level. Amounts are in tokens.

	Treatments			
Level of attitude strength	**Covid**	**Food**	**Torture**	**All**
0		39.62		39.62
2	10.00		46.67	37.50
2.25	10.00		57.00	46.15
2.5	76.67		60.00	63.33
2.75	61.67		55.00	57.22
3	37.00		63.03	55.15
3.25	52.86		61.52	58.81
3.5	74.74		73.57	74.04
3.75	75.14		79.26	76.94
4	59.31		90.40	73.70
4.25	79.17		93.00	85.45
4.5	69.05		86.67	75.45
Average	**64.31**	**39.62**	**71.15**	**57.83**

[17] The distribution of our sample was not normal.

Influence of attitude strength on 'willingness to forgo' to avoid falsifying preferences

Our second hypothesis states that in morally framed choice settings, the higher the strength of subjects' attitudes to an issue, the higher the monetary cost they are willing to incur (i.e., willingness to forgo) to avoid falsifying their preference. Figure 5 below shows, for each level of attitude strength, the mean opportunity cost participants were willing to incur to not comply with the majority (and thus falsify their preference).

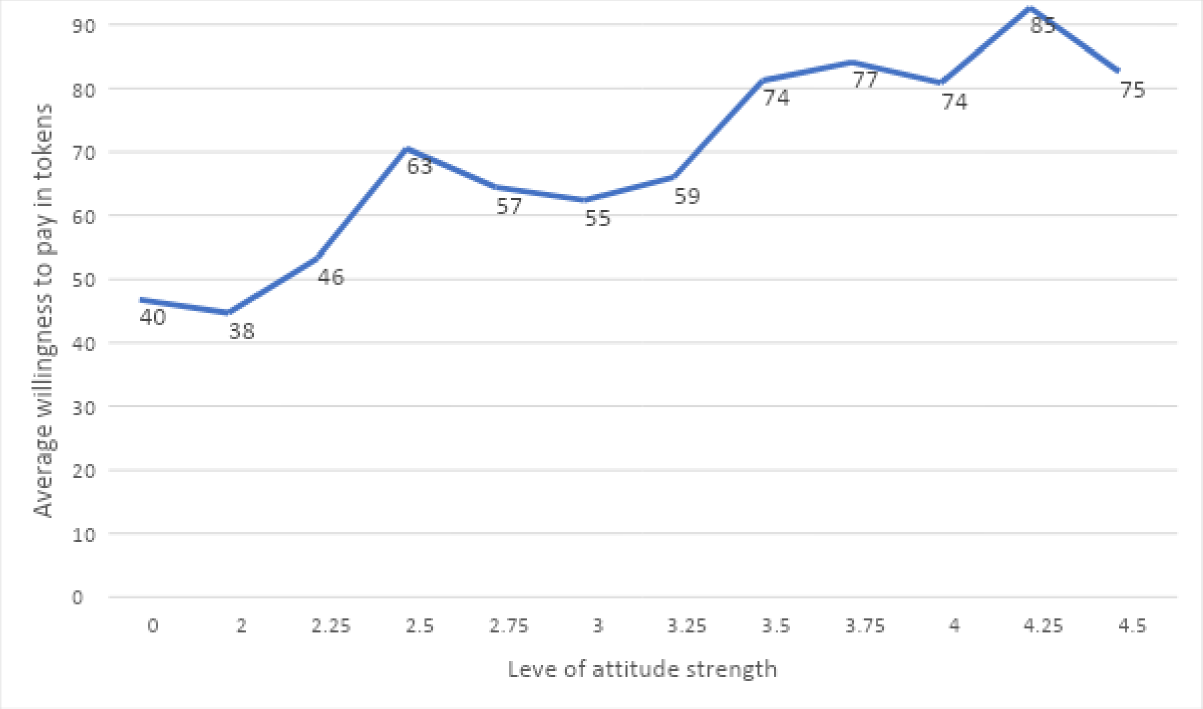


Figure 5: Average 'willingness to forgo' (in tokens) to avoid falsifying preferences for each level of attitude strength. On the horizontal axis are displayed the levels of attitude strength (computed as the average between attitude extremity, attitude certainty and attitude importance). On the vertical axis are displayed the mean maximum opportunity costs participants were willing to incur to not falsify their preference. One should read for an attitude strength level of 3 that the mean maximum opportunity cost participants were willing to incur to keep their own view is 55 tokens.

From figure 5 we can see that the maximum monetary cost incurred to avoid falsifying preferences increases as the level of attitude strength increases. We then assess the effect on an increase in one unit, in attitude strength, on the maximum cost an individual is willing to incur to express her or his private preference.

We performed regressions on cross sectional data using Tobit regression. In this model, the dependent variable was the opportunity cost, the independent variables were attitude strengths (computed as the average of attitude extremity, attitude certainty, attitude importance and moral conviction), age, sex, level of income, degree of liberalism, level of education and the degree to which one believes in God. The observations were censored to the left, because if an individual did not falsify her or his preference, we cannot know the precise amount she or he would have incurred. Therefore, we implemented an upper limit of 0. Table 6 below shows the Tobit regression results. For an increase of 1 in the level of attitude strength, the opportunity cost significantly increases by 8.48 tokens or $0.85 (t-score = -9.56; p-value = 0.000). The Age variable is weakly significant at the 5% level.

The second hypothesis is also supported by our results: with moral issues, the higher the strength of subjects' attitudes to an issue, the higher the monetary cost they are willing to incur to avoid falsifying their preference.

Table 2: Tobit regression results assessing the maximum cost (in tokens) an individual is willing to incur to express her or his private preference.

Dependent variable (opportunity cost in tokens)

Independent variables	**Coefficient (S.E.)**	
Intercept	26.93 (9.28)	***
Attitude strength	8.48 (0.89)	***
Age	0.24 (0.12)	**
Gender (1=female)	2.19 (3.09)	
Political orientation	-0.46 (0.94)	
Household Income ($)	-0.17 (0.52)	
Education	-1.47 (1.42)	
Belief in god	1.97 (1.08)	*

****p < 0.01 ; **p < 0.05 ; *p < 0.1*

+ The analysis is performed with attitude strength as the main independent variable (scale from 0 no attitude strength involved to 5 very high attitude strength). We controlled for Age (year of birth converted into the age as a continuous variable), Gender, Education (scale representing the highest level of education from 1 no degree to 9 PhD degree), belief in God (from 1 not at all to 5 very much), political orientation (scale from 1 strongly liberal to 7 strongly conservative) and Income (household income in dollars).

6. Discussion

While we don't study tipping points that can lead to changes in social equilibrium, we examine individuals' willingness to falsify their true preferences for a monetary reward when the proportion of individuals with like preferences varies. This phenomeon had previously been studied using non-incentivised survey methodology for one-shot interactions. To properly understand how preference falsification is related to social and material pressure one needs to vary the number of individuals holding similar views. Albeit, using the strategy method, this repeated games structure (i.e. varying proportions in a popultation), captures the marginal willingness to forgo that is otherwise not captured in a one shot framework. Additionally, we can then study how the expression of these preferences mapped into the willingness to pay/forego domain.

Given the strategy method our structure captures the weakest form of these forces that lead to preference falsification. That is, individuals in our experiment do not face any direct social pressure, and only respond to hypothetical scenarios. Any result thus obtained can only be stronger in environments were interaction is direct. We find that preference falsification occurs when the extrinsic rewards (including money or social approval) from expressing a preference contrary to one's views is greater than the intrinsic cost of not asserting one's views. This phenomenon has important consequences on the long-term dynamics of individual and collective, private and public preferences (Kuran, 1995). In the study of psychological costs and extrinsic rewards trade-offs, on the one hand, economists focus on why some people would depart from their (material) self-interest (Akerlof and Kranton, 2000; Bursztyn et al., 2020). On the other hand, psychologists focus on explaining individuals' motives to comply with the various pressures around them or on the motives to stand for their opinions (Jahoda, 1959; Hornsey and Jetten, 2004). We add to this literature by assessing individuals' willingness to forgo extrinsic rewards to avoid falsifying their own preferences regarding moral issues.

We compare preference falsification in neutral and moral issues and identify whether individuals with stronger attitudes about an issue were more resistant to PF than individuals with weaker attitudes. Finally, we assessed the maximum cost an individual was willing to sacrifice in order to keep her or his private preference and not comply with the majority. To our knowledge we are the first one's to study this with monetary incentives.

Our main contributions are the following. First, we introduce a novel design that allows us to explicitly calculate the willingness to forgo to express a private preference with real monetary stakes (unlike Bursztyn et al. (2020), who does not incrementally vary the number of opposers). We thus assessed how much would the private preference cost, on average, for each treatment of our study, and found that this cost increases with attitude strength and moral framing. More research is needed to disentangle social versus material effects (see limitations and further study below). Second, we developed a novel experimental design that can be used on MTurk using the strategy method (Selten, 1967) in online experiments, third we give new insights on current moral issues, such as Covid-19 (Bicchieri et al., 2021).

We found that, overall, preference falsification occurs once the majority holds the opposing view. More than half of the participants (57%) did falsify their preferences in an anonymous, online setting. This shows that, for these participants, the intrinsic, psychological cost of falsifying their private preference was not sufficiently high to motivate non-compliance with the opposite view. Importantly, between one third (in the Covid-19 the Torture treatments) and two thirds (in the Food treatment) did not want to incur any, or even a very small, cost and therefore chose to falsify their preference as soon as the cost became negative (conditional decision n°6) or very small (10 cents in conditional decision n°7). In the comments, participants who switched in those conditional decisions mentioned the maximisation of their payoff almost all the time. This suggests that material payoff is an important positive factor of preference falsification.

We also found that falsification was less frequent with moral issues than with neutral issues: while 80% of participants falsified their preference in the Food/neutral treatment, only 50% (43%) did so in the Covid-19 (Torture) treatments. Moreover, attitude strength about an issue makes it more resistant to social and/or material pressures: for an increase of 1 in the level of attitude strength, the opportunity cost increases by 8.48 tokens. This is expected, and confirms findings in the literature about moral mandates that when individuals stand up for a morally loaded private attitude, even if it is costly in terms of money or reputation (Skitka, Bauman and Sargis, 2005; Skitka and Wisniewski, 2011; van Zoomeren, Postmes and Spears, 2012; Skitka, Washburn and Carcel, 2015; Skitka and Morgan, 2021). These findings therefore suggest that the importance of material payoff is altered by morality.

Importantly, 58% (52%) of participants in the Torture (Covid-19) treatment chose not to falsify their preference whatever the costs. These participants scored higher than the falsifiers for every component of attitude strength (moral conviction, attitude extremity, attitude certainty and attitude importance). This result suggests that the psychological cost induced by preference falsification is greater than the material payoff for many of the participants in the moral treatments. However, results from Bursztyn (2020) show that those percentages would be lower with higher stakes and in public settings: it could be that anonymity and the online set up make the group not salient enough for participants so that they conform to the group behaviour (see meta study by Huang and Li, 2016). This is important for the dynamics of public preferences in the long term since the true preference is not completely overwhelmed by falsification. This result reinforces the idea that morality is an integral part of one's identity (Strohminger & Nichols, 2014) as a majority of the individuals chose not to falsify their (moral) beliefs.

Twenty two percent of participants in the neutral treatment did not falsify their preference at all. The analysis of their comments indicates clearly that participants understood the game and stuck to their preferences for two main reasons. First, they really prefer their initial choice over the switching one (“I would really like pizza for real.” ; “I really prefer pizza over pasta so I honestly chose my true selection each time.”). Second, they wanted to be consistent : meaning that neither payoff nor social influence would make them change their mind (“I prefer pasta and I also prefer not to lie about my preferences”; “I do not change my opinions based on others!”; “I went with what I would honestly like no matter the outcome, I didn't care too much about the bonus.”). Thus, this result confirms that self-assertion can be very important for some individuals (Freud, 1920; Kuran, 1995; Hornsey et al., 2007); Duffy and Lafky, 2018; Bursztyn et al., 2020). Future research could focus on those participants with the highest attitude strength scores and increase payoff to see if the result is robust.

The present study has some limitations. First, the number of participants per treatment (around 200) was limited because of the attrition rate in the second phase; this means that we could have lost representativeness or significance. Second, due to resource constraints, we limited the study to one neutral (food treatment) and only two moral issues (Torture and Covid-19). An obvious next step would be to replicate our study with more participants and with other moral and non-moral issues. Another limiting aspect was that we have used small stakes to induce social consensus. Note that in Bursztyn et al. (2020), increasing the stakes from 100 to 500 rupees decreased the number of participants who were willing to forgo the payment from 25% to 10%. Increasing the stakes, so that it is costlier to switch is a good way to check the robustness of the present results. Finally, we assumed that the monetary payoff is the main driving force because it is an online experiment in which people don't interact and will never do. However, even though the incentives were monetary, the way the experiment was designed was in terms of other people's preferences. Thus, in the present study we are not able to disentangle the effect of monetary payoff on preference falsification, from the social influence implied by our design. One possible development would be to use a design with no stakes at all – or negligible stakes to keeps the design consistent - to suppress (or minimize) the monetary effect and keep only the social influence effect, such as in Kundu and Cummins (2013) and Lisciandra et al. (2013) who both found that social influence triggered conformity for moral issues.

One of the main findings of the study is that preference falsification is more common when the stakes are lower and in a neutral moral issue, with the material payoff being an important factor in driving this behavior. In contrast, preference falsification was less common in a moral issue and among individuals with higher attitude strength. This suggests that the psychological cost of preference falsification may be greater when the issue is perceived as morally significant, and that some individuals may be more resistant to social and material pressures to conform to group norms. We also found that a small minority of participants did not falsify their preferences at all, and that this is driven by a combination of true preference and a desire for consistency. These findings contribute to our understanding of the factors that influence preference falsification and the ways in which individuals balance the costs and benefits of expressing their true preferences.

References

Abelson, R. P. (1995). Attitude extremity. Attitude strength: Antecedents and consequences, 4, 25-42.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. The quarterly journal of economics, 115(3), 715-753.

Aramovich, Nicholas P., Brad L. Lytle and Linda J. Skitka, 'Opposing torture: Moral conviction and resistance to majority influence' (2012) 7(1) Social Influence 21.

Aronson, E. (1969). A Theory of Cognitive Dissonance: A Current Perspective. Journal of Advances in Experimental Social Psychology, Vol. 4, Leonard Berkowitz, ed. New York: Academic Press, 1-34.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological monographs: General and applied, 70(9), 1.

Asch, S. E. (1961). Effects of group pressure upon the modification and distortion of judgments. In Documents of gestalt psychology (pp. 222-236). University of California Press.

Bauman, C. W., & Skitka, L. J. (2009). In the mind of the perceiver: Psychological implications of moral conviction. Psychology of learning and motivation, 50, 339-362.

Bicchieri, C., Fatas, E., Aldama, A., Casas, A., Deshpande, I., Lauro, M., ... & Wen, R. (2021). In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic. PloS one, 16(6), e0252892.

Boninger, D. S., Krosnick, J. A., Berent, M. K., & Fabrigar, L. R. (1995). The causes and consequences of attitude importance. Attitude strength: Antecedents and consequences, 4(7), 159-189.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?.

Bursztyn, L., Callen, M., Ferman, B., Gulzar, S., Hasanain, A., & Yuchtman, N. (2020). Political identity: Experimental evidence on anti-Americanism in Pakistan. Journal of the European Economic Association, 18(5), 2532-2560.

Charness, G., Naef, M., & Sontuoso, A. (2019). Opportunistic conformism. Journal of Economic Theory, 180, 100-134.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. Annu. Rev. Psychol., 55, 591-621.

Crutchfield, R. S. (1955). Conformity and character. American psychologist, 10(5), 191.

Davis, W. L. (2004). Preference falsification in the economics profession. Econ Journal Watch, 1(2), 359.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. The journal of abnormal and social psychology, 51(3), 629.

Duffy, J., & Lafky, J. (2021). Social conformity under evolving private preferences. Games and Economic Behavior, 128, 104-124.

Frank, Robert H., The Political Economy of Preference Falsification: Timur Kuran's Private Truths, Public Lies, Journal of Economic Literature , 1996, Vol. 34, No. 1, pp. 115-123.

Freud, S. The Ego and the Id, trans. James Strachey (New York: W. W. Norton, 1961; first German ed., 1923).

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105–2108. http://dx.doi.org/10.1126/science.1062872.

Gross, S. R., Holtz, R., & Miller, N. (1995). Attitude certainty. Attitude strength: Antecedents and consequences, 4, 215-245.

Hajnal, A., Vonk, J., & Zeigler-Hill, V. (2020). Peer influence on conformity and confidence in a perceptual judgment task. Psihologija, 53(1), 101-113.

Hornsey, M. J., & Jetten, J. (2004). The individual within the group: Balancing the need to belong with the need to be different. Personality and Social Psychology review, 8(3), 248-264.

Hornsey, M. J., Majkut, L., Terry, D. J., & McKimmie, B. M. (2003). On being loud and proud: Non-conformity and counter-conformity to group norms. British journal of social psychology, 42(3), 319-335.

Hornsey, M. J., Smith, J. R., & Begg, D. (2007). Effects of norms among those with moral conviction: Counter-conformity emerges on intentions but not behaviors. Social Influence, 2(4), 244-268.

Huang, G., & Li, K. (2016). The effect of anonymity on conformity to group norms in online contexts: A meta-analysis. International journal of communication, 10, 398-415.

Jahoda, M. (1959). Conformity and independence: A psychological analysis. Human Relations, 12(2), 99-120.

Jiang, J., & Yang, D. L. (2016). Lying or believing? Measuring preference falsification from a political purge in China. Comparative Political Studies, 49(5), 600-634.

Judd, C. M., & Brauer, M. (1995). Repetition and evaluative extremity. Attitude strength: Antecedents and consequences, 4, 43-72.

Kalinin, K. (2018). Linking Preference Falsification and Election Fraud in Electoral Autocracies: The Case of Russia. Political Studies, 66(1), 81-99.

Kelly, M., Ngo, L., Chituc, V., Huettel, S., & Sinnott-Armstrong, W. (2017). Moral conformity in online interactions: Rational justifications increase influence of peer preferences on moral judgments. Social Influence, 12(2-3), 57-68.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An over preference. Attitude strength: Antecedents and consequences, 1, 1-24.

Kundu, P., & Cummins, D. D. (2013). Morality and conformity: The Asch paradigm applied to moral decisions. Social Influence, 8(4), 268-279.

Kuran, Timur, 'Private truths, public lies: The social consequences of preference falsification.' (1995).

Lisciandra, C., Postma-Nilsenová, M., & Colombo, M. (2013). Conformorality. A study on group conditioning of normative judgment. Review of Philosophy and Psychology, 4(4), 751-764.

Martínez, D., Parilli, C., Scartascini, C., & Simpser, A. (2021). Let's (not) get together! The role of social norms on social distancing during COVID-19. PloS one, 16(3), e0247454.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior research methods, 44(1), 1-23.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research, 45(6), 633-644.

Moscovici, S. and Lage, E., (1976). Studies in social influence III: Majority versus minority influence in a group, 6(2). European Journal of Social Psychology 149.

Parfit, D. (1984). Reasons and persons. Oxford, England: Oxford University Press.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. Journal of theoretical biology, 299, 172-179.

Ross, D., Stirling, W. C., & Tummolini, L. (2023). Strategic theory of norms for empirical applications in political science and political economy. The Oxford Handbook of Philosophy of Political Science, 86.

Selten, Reinhard (1967): "Die Strategiemethode zur Erforschung des eingeschränkt rationalenVerhaltens im Rahmen eines Oligopolexperiments", Beiträge zur experimentellen Wirtschaftsforschung, Heinz Sauermann (ed.), Vol. I, Tübingen: J.C.B. Mohr (Siebeck), 136-168.

Sherif, M. (1937). An experimental approach to the study of attitudes. Sociometry, 1(1/2), 90-98.

Skitka, L. J. (2002). Do the means always justify the ends, or do the ends sometimes justify the means? A value protection model of justice reasoning. Personality and Social Psychology Bulletin, 28(5), 588-597.

Skitka, L. J., & Wisneski, D. C. (2011). Moral conviction and emotion. Emotion review, 3(3), 328-330.

Skitka, L. J., Bauman, C. W., & Mullen, E. (2008). Morality and justice: An expanded theoretical perspective and empirical review. Justice, 25, 1-27.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more?. Journal of personality and social psychology, 88(6), 895.

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. Annual Review of Psychology, 72, 347-366.

Skitka, L. J., Washburn, A. N., & Carsel, T. S. (2015). The psychological foundations and consequences of moral conviction. Current Opinion in Psychology, 6, 41-44.

Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen, General Social Surveys, 1972-2018 [machine-readable data file] /Principal Investigator, Smith, Tom W.; Co-Principal Investigators, Michael Davern, Jeremy Freese, and Stephen Morgan; Sponsored by National Science Foundation. --NORC ed.-- Chicago: NORC, 2018: NORC at the University of Chicago [producer and distributor]. Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org.

Strohminger, N., & Nichols, S. (2014). The essential moral self. Cognition, 131(1), 159-171.

Sunstein, Cass R. (2001). Republic.com. Princeton, N.J : Princeton University Press, http://www.loc.gov/catdir/toc/prin031/00045331.html

Swann, W. B., Griffin, J. J., Predmore, S. C., & Gaines, B. (1987). The cognitive–affective crossfire: When self-consistency confronts self-enhancement. Journal of personality and social psychology, 52(5), 881.

Tverskoi, D., Guido, A., Andrighetto, G., Sánchez, A., & Gavrilets, S. (2022). Disentangling material, social, and cognitive determinants of human behavior and beliefs.

Van Zomeren, M., Postmes, T., Spears, R., & Bettache, K. (2011). Can moral convictions motivate the advantaged to challenge social inequality? Extending the social identity model of collective action. Group Processes & Intergroup Relations, 14(5), 735-753.

Appendices

Appendix A – Moral issues used in the pre-survey and references

The table below displays question type (moral or neutral) and topics, the questions that have been asked to each participant of the pre-survey and the references from which we took the exact same wording for the question (except for Covid-19 and the neutral/Food issue that we have created on our own).

Table 7: Moral issues used in the pre-survey and references.

Question Type	**Question**	**Reference**
Moral – Gun permit	Would you support or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?	GSS Survey - Smith et al., 2018
Moral - Torture	To what extent do you support or oppose the use of stress techniques when interrogating suspected terrorists, such as sleep deprivation, 'water boarding', long periods of hanging detainees by ropes in painful positions, etc.?	Aramovich, 2012
Moral - Abortion	Do you support or oppose allowing abortion to remain a legal option in your country?	GSS Survey - Smith et al., 2018
Moral – Gaz emissions	Would you support or oppose a policy to reduce greenhouse gas emissions by taxing the use of carbon-based fuels such as coal, oil, and natural gas based on how much they contribute to climate change?	Shwom, 2010 (Global Environmental Change)
Moral – Covid-19	Would you support or oppose the termination of social distancing ('physical distancing') measures to boost the economy and relieve unemployment, even if the Covid-19 virus still remains a threat to people's lives?	Own question
Neutral - Food	Do you prefer pizza or pasta?	Own question

Appendix B – Results of the pre-survey: proportions of opposers and supporters of each moral issue

The figure below displays a bar chart graph for each issue (horizontal axis) the percentage of supporters, opposers and uncertain respondents in the pre-survey (vertical axis).

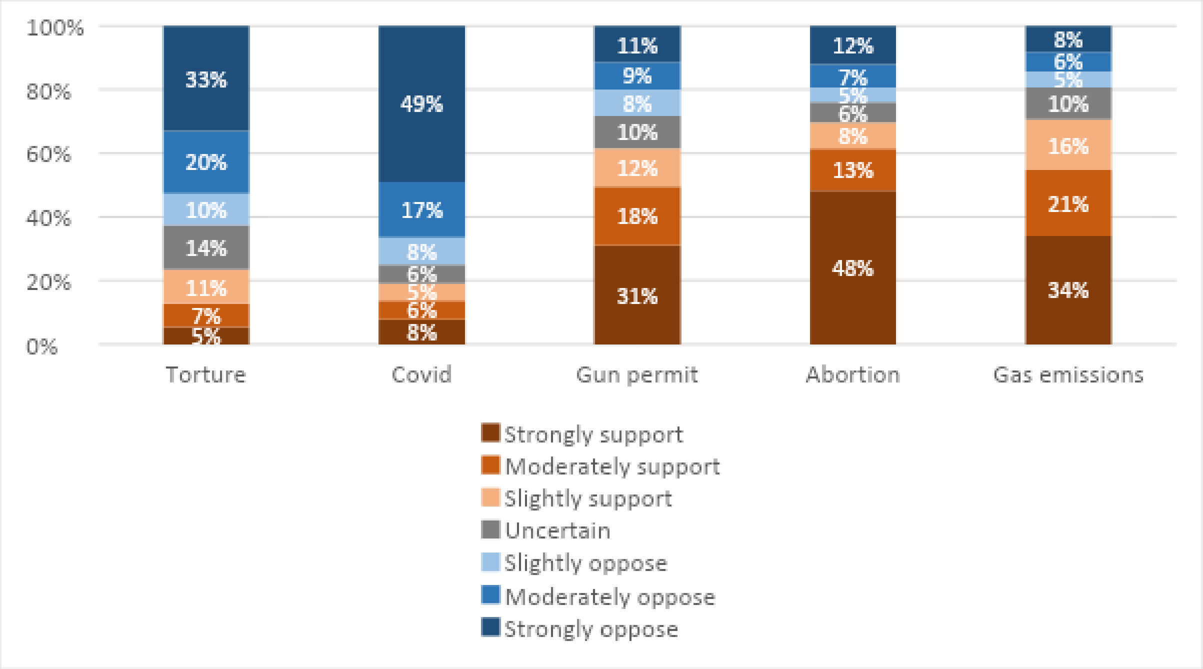


Figure 6: Proportions of opposers and supporters of each moral issue (n = 1,484).

Appendix C – Results of the panel data logistic regressions

In order to compare the relative impact of holding stronger attitudes about the issue at hand on the probability for individuals to falsify their preference, we use a logistic panel data regression with random effects. As we are dealing with ten repeated dichotomous choices for each individual, our dataset structure is suitable for panel data analysis. Panel data allow for individual heterogeneity and dynamic effect in individual behaviour (Greene, 2001).

To do this, we dropped the Food treatment from our previous dataset since we did not ask for the subject's attitude strength in this treatment. We set decisions as the time variable and whether or not an individual falsifies her or his preference (= 1 if they falsify their preference and 0 otherwise) as the dependent variable. The model predicts well the outcome (Wald chi2 with nine degrees of freedom > 50.00, p-value < 0.000).

The table below summarizes the odd ratios, standard deviations and p-values for attitude strength and control variables. It shows attitude strength is significant at least at the 1% level when predicting falsification.

Table 8: Panel data logistic regression results. Attitude strength is computed by averaging attitude extremity, attitude importance, attitude certainty and moral conviction about the issue at hand. We have performed the analysis both for attitude strength and moral conviction, and moral conviction provides stronger results. We controlled for decisions (the ten conditional answers), Age, Sex, 1.Education (equals 1 if the participant has obtained a 4-year college degree or more, equals 0 otherwise), God (from 1 not at all to 5 very much), and Income.

	Dependent variable (falsification)	
Independent variables	**Odds ratio (S.E.)**	
Intercept	0,03 (0,01)	***
Attitude strength	0,90 (0,02)	***
decisions	1,20 (0,03)	***
Age	0,99 (0,00)	
Sexe	0,99 (0,12)	
1.education_2	1,10 (0,14)	
God	0,97 (0,03)	
Income	1,00 (0,02)	

$***p < 0.01$; $**p < 0.05$; $*p < 0.1$

Controlling for decisions and sociodemographic variables, we found that attitude strength (computed as the mean of attitude importance, attitude certainty, attitude extremity and moral conviction) is significant at the 1% level (z-score = -4.17; p-value < 0.001). An increase of 1 unit in attitude strength decreases the likelihood of preference falsification by 10%. See table in Appendix 3 for the logistic panel data regression results.