Journal of Economic Behavior and Organization 240 (2025) 107293

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research paper

Nudging or nagging: The perils of persuasion☆

Andrea Guido [a,*], Denis Tverskoi [b], Sergey Gavrilets [c,b], Angel Sánchez [d,e], Giulia Andrighetto [f,g,h]

[a] Paris School of Business, France
[b] Center for the Dynamics of Social Complexity, National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996, USA
[c] Department of Ecology and Evolutionary Biology, Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA
[d] Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain
[e] Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, 50018, Zaragoza, Spain
[f] Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy
[g] Institute for Futures Studies, Stockholm, Sweden
[h] Institute for Analytical Sociology, Linkoping University, Sweden

ARTICLE INFO

Dataset link: https://researchbox.org/4590

JEL classification:
D7
D9

Keywords:
Persuasion
Nudges
Backfiring
Rule following
Psychological reactance

ABSTRACT

Over the past few years, policy-makers have enthusiastically resorted to mass persuasion to promote behavioral change. However, the one-size-fit-all approach adopted by most of policy-makers has often been criticized for leading to more heterogeneous and undesired behavior. A central unanswered question is whether the effect of persuasion depends on individuals' predisposition to follow externally-imposed rules. We provide experimental evidence on the heterogeneous effect of persuasion in a collective action problem. By studying the effect of appeals in an online experiment, we find that rule followers comply with the content of the appeal, while rule breakers react against it. Reactance to appeals among rule breakers emerges after some time and is robust even after controlling for social preferences, personal and social norms. Persuasive appeals have no overall effect on welfare, yet they introduce distributional disparities. Our findings raise awareness about the importance of individual heterogeneity when designing and evaluating behavioral interventions.

1. Introduction

Over the past few years, there has been an increasing interest around persuasive communication through public appeals to encourage socially-desirable behavior, especially in situations in which the individual and collective interests are not aligned (DellaVigna and Gentzkow, 2010; Matz et al., 2017; Gelfand et al., 2022). Public appeals are becoming integral part of policy makers'

☆ We are thankful to V. Capraro, E. Dimant, A. Festré, N. Gagnon, A. Gneezy, T. Jaber-Lopez, A. Koch, A. Malézieux, J. Nafziger, D. Nosenzo, R. Romaniuc, E. Spiegelman, A. Sutan, and A. Vostroknutov for comments. We also thank all participants to the IMEBESS conference in Lisbon, CIMEO Workshop in Rome, Aarhus University MOB Seminar Series, ESA European Conference in Bologna, the Friday meeting in Nice University, and the Economix Seminar Series in Paris Nanterre University. SG was supported by the U. S. Army Research Office grants W911NF-14-1-0637 and W911NF-18-1-0138, the Office of Naval Research grant W911NF-17-1-0150, the Air Force Office of Scientific Research grant FA9550-21-1-0217, and the John Templeton Foundation grant 62434. Replication material available at: https://researchbox.org/4590.

* Corresponding author.
E-mail address: a.guido@psbedu.paris (A. Guido).

https://doi.org/10.1016/j.jebo.2025.107293
Received 1 August 2024; Received in revised form 7 September 2025; Accepted 7 October 2025

toolbox to promote the adoption of behavior that are in line with their goals (Matz et al., 2017). Examples range from reducing tax evasion (Hallsworth et al., 2017), or promoting mask-wearing and vaccination (Van Bavel et al., 2020).[1]

While public appeals can be a cost-effective tool for behavioral change, their success is not always guaranteed. The experimental evidence of their effectiveness is mixed and, in some cases, highlights negative consequences. A striking pattern emerging from extant research is that appeals can lead to more polarized and heterogeneous behavior: some individuals react in compliance with the appeal, others backfire and go against it (Sunstein, 2017; My Boun and Ouvrard, 2019; Costa and Kahn, 2013). Some competing, yet not exhaustive, explanations have been put forward to explain backfiring behavior under the presence of appeals. A first explanation hinges upon social expectations, according to which individuals may believe that what motivates the intervention from an authority is a widespread lack of norm compliance (Sliwka, 2007; Nyborg and Rege, 2003). In this sense, appeals signal the absence of strong social norms. Recent work has also suggested that people display heterogeneous preferences for rule following (Kimbrough and Vostroknutov, 2015, 2016, 2018; Gächter et al., 2025) and that at the root of heterogeneous responses to appeals is the attempt to restore individuals' sense of autonomy to a perceived restriction on their behavioral freedom (Bryan et al., 2021; Sunstein, 2022; Reiff et al., 2021). This second conjecture finds theoretical support in the Psychological Reactance Theory (Brehm, 1966; Rosenberg and Siegel, 2018) which states that freedom of behavior is a central requirement in people's decision making. When threatened, people are motivated to restore their freedom by opposing to a rule or recommendation. Consequently, highly reactant people would react negatively to interventions that they perceive as more threatening to their behavioral freedom. Given that people display heterogeneous inclinations to follow rules, understanding how such personal inclination impacts the effectiveness of appeals is thus of crucial importance.

To study the effect of appeals and how they interact with individuals' dispositions to follow rules, we design a 36-day long online experiment implementing a collective-action game. Our experimental design is made of two main stages. In the first stage, we elicit a set of behavioral measures that account for subjects' predisposition towards following external rules, using both incentivized and non-incentivized tasks (Kimbrough and Vostroknutov, 2018; John et al., 1991). The second stage consists of an online Common Pool Resource game (Ostrom et al., 1992) lasting over 35 days. To measure the evolution and the effect of social norms, decisions in the Common Pool Resource game were preceded in each round by an incentive-compatible elicitation of personal normative beliefs, empirical and normative expectations (Bicchieri and Xiao, 2009). We randomly assigned participants to two conditions: a baseline condition with no messages (No-Message) in which participants had to make daily decisions in the Common Pool Resource game, and a message condition (Message) in which a daily message was displayed to promote a socially-beneficial behavior.[2]

Several aspects of our results are remarkable.[3] First, we report no overall effect of appeals on participants' extraction levels in the Common Pool Resource game. Yet, their effect is heterogeneous across individuals and increases behavioral variability: a fraction of participants in the experiment increase extraction levels, while others comply with the appeal content and reduce their extraction levels. Secondly, higher heterogeneity in behavior is explained by individual measures of rule compliance: rule followers comply more with the content of the appeal, while rule breakers go against it. Results hold even when controlling for social norms and pro-social preferences which strengthen the importance of individual predisposition to follow externally-given rules. When looking at the dynamics of extractions in the Common Pool Resource game, the gap between rule breakers and rule followers widens over time. Subjects who display lower dispositions for rule following react by increasing their extractions levels only after 10 days of receiving appeals. Lastly, we find that appeals have distributional effects in terms of welfare among individuals. While there is no difference in final earnings between rule breakers and rule followers in the absence of appeals, rule breakers end up earning substantially more than rule followers in the Message condition.

This study's contribution is three-fold. First, we add to the literature on persuasion (DellaVigna and Gentzkow, 2010) by providing some novel insights on the role of psychological reactance. Past experimental work has documented a higher behavioral variability under the presence of persuasive appeals, yet, to our knowledge, the role of psychological factors behind it has been overlooked (Croson and Marks, 2001; Dal Bó and Dal Bó, 2014). Second, we contribute to the more broad literature studying the heterogeneous effects of nudges (Sunstein, 2017, 2022). Past studies have suggested that nudge-based interventions have to align with individual preferences and predispositions in order to be effective (Bruns and Perino, 2021; Arad and Rubinstein, 2018). Lastly, our contribution is also methodological by overcoming the usual limitations of laboratory experiments in assessing repeated measures over time (Horton et al., 2011). Past work observing the dynamics of behavior and beliefs often involved subjects in a large number of decisions over a short frame of time. These decisions may become tedious and cognitively demanding tasks which in turn can affect the external validity of results. Our experimental timeline, spanning over several days, allows us to reproduce more closely persuasion interventions in the field, and overcome the difficulties standard laboratory experiments typically face.

The rest of paper is structured as follows. Section 2 presents our contribution to the related literature. Section 3 explains the experimental methods. Section 4 reports the results of the experiment, while Section 5 offers a global discussion. Lastly, Section 6 concludes.

[1] Persuasive communication is also used within organizations. As an example, effective leadership relies upon persuasive appeals to overcome coordination and cooperation failures (Brandts et al., 2015; Levy et al., 2011; Brandts and Cooper, 2007) or to encourage customers to adopt responsible conducts or soliciting customers' feedback to the company (Reiff et al., 2021).

[2] In the context of environmental and energy-preserving policies, these type of interventions have been previously defined as *green nudges* (Carlsson et al., 2021).

[3] Complementary findings from the same experiment are reported in Tverskoi et al. (2022) in which they provide an empirical test of Gavrilets (2021) utility model. Results show that, in the absence of appeals, the factors having the highest weight in agents' utility are personal norms and empirical expectations, while the presence appeals decreases the weight given to personal norms and simultaneously increases the weight given to peer conformity. Here we investigate another related aspect of the effect of appeals, which has to do with their heterogeneous effect based on individual predisposition to follow rules.

2. Related literature

Our work is related to the large body of economic research developing around the use of information provision and recommendations for behavioral change (Sunstein, 2022; DellaVigna and Gentzkow, 2010; Matz et al., 2017).

Past experiment have studied the effect of messages appealing to high levels of contributions in collective action problems. In most of these studies, appeals consist in messages exhorting a certain behavior in the game. The seminal work by Croson and Marks (2001) is one of the first experiments studying appeals in public goods games. Authors report no effect on average contributions.[4] The introduction of messages increases the variance in contributions rather than their mean level suggesting that while some subjects positively react to the intervention, others backfire to it. Similarly, Dale and Morgan (2010) find that appeals crowd out contributions in the public goods game. Asking to contribute the socially optimum level backfires relative to a condition with no messages.

Our work also relates to experiments using moral appeals — i.e., appeals to moral responsibility of one's actions to steer behavior. The seminal paper by Dal Bó and Dal Bó (2014) studies the effect of different types of informational nudges on contributions in a public goods game. Participants play repeatedly the public goods game in randomly shuffled groups for 10 rounds. After 10 rounds, a message is displayed to their screens. Authors design two types of moral appeals: *utilitarian* appeals – aiming at maximizing the group's welfare, and appeals based on the *golden-rule* – promoting reciprocal behavior. More specifically, utilitarian messages focus on the consequences of actions (for example, on others in one's group). While the golden rule principle abstracts from consequences and appeals to treat others in the way you wish others would treat you. Results show that both message types increase contributions and payoffs, in particular for the utilitarian one. Authors associate the positive effect of the messages to a temporary shift in subjects' preferences and expectations about others' cooperation levels. Yet, the effect of messages is not persistent. Contributions quickly decline to average levels of baseline periods, remarking a minor overall treatment effect over game repetitions. Similar results have been found also in other experimental work (Andrighetto et al., 2013).[5]

Our paper adds to this literature in several regards. First, we raise attention on the effect of individual heterogeneity in determining the success of nudges interventions, focusing in particular on rule following dispositions. Past related work has investigated the consequences of individuals heterogeneity in rule-following tendencies in social dilemmas (Kimbrough and Vostroknutov, 2015) showing that group composition inevitably affects the management of depletable resources. To the best of our knowledge, we are the first to explore the association between rule following dispositions and reaction to nudges. Such an issue becomes relevant in sight of future development of more targeted and personalized interventions. (Bryan et al., 2021).

Second, with our online set up, we overcome some flaws of laboratory experiments. Laboratory experiments have limited external validity in the sense of sample sizes, and decision time (Horton et al., 2011). Our experimental timeline spanning over several days allows us to observe the evolution of decisions, social norms, and other complementary measures that are difficult to observe in a short-term laboratory experiment, given the large quantity of questions and decisions asked to subjects at each round.

Past studies on the role of authorities are also related to our work. For example, Karakostas and Zizzo (2016) study the effect of authority in a destruction game. Similarly, Silverman et al. (2014) studied the effect of legitimacy and presumption of expert knowledge of authoritarian messages. What is worth noting from these studies, including ours, is that the experimenter is given the role of authority, and this could represent some forms of experimenter demand effect (Zizzo, 2010). However, providing the experimenter with such role as is a considerably more valid experimental test of the role of the authority than providing it to other experimental subjects.[6] Moreover, experimenter demand effect is less of an issue in our case, for several reasons. One has to do with the fact that sessions are conducted online, where the relational link between the experimenter and subjects becomes less relevant and direct. Second, the presence of possible experimenter demand effect is orthogonal to the research question, as we focus on the heterogeneity of reactions, rather than increasing general cooperation levels. Lastly, but most importantly, there is an external validity justification underlying the way we conceived our treatment, since we would expect that in the real world appeals likely come from authorities or from mediating third parties in a role of authority.

Finally, our work relates to the growing literature on the backfiring effects of nudges. Past work suggests that the success of a nudge depends on the alignment with individuals' predispositions to interventions itself (Sunstein, 2017; see also de Ridder et al., 2022 for a review). This implies that individuals may backfire to interventions because they hold conflicting preferences relative to the promoted behavior. In a laboratory public goods experiment, My Boun and Ouvrard (2019) find heterogeneous effects of recommendations on contributions. The design of the experiment allows to contrast the effect of recommendation messages vis-à-vis the introduction of a tax. Furthermore, participants in the experiment were classified according to their level of environmental concern, and sorted in homogeneous groups. While levying a tax on subjects increased contributions on average, the effect of messages is non trivial. Authors find that recommendations encourage higher cooperation only for a sub-sample of subjects (i.e., those highly concerned for the environment). At the same time, the implementation of recommendations seems to crowd out the level of contributions in comparison to the baseline for those less concerned about the environment. Further evidence comes from field experiments. Costa and Kahn (2013) found that energy conservation nudges depend on individuals' political ideologies. In particular, findings show that liberals, which are more likely to vote for environmentalist causes than conservatives in the U.S.,

[4] More precisely, authors find a null effect when valuations for the public good are homogeneous. Authors however do find an effect of appeals when returns from the public goods are heterogeneous.

[5] The weak persistence of moral appeals has also been documented in the field. In a randomize field experiment on energy conservation, Ito et al. (2018) find that subjects assigned to the moral suasion treatment induces short-run reductions in electricity usage. The effect of moral appeals vanishes over repeated interventions.

[6] Giving power to other experimental subjects would be a confounder in the identification of the authority effect relative to peer pressure effects.

are more responsive to information-based nudges than conservatives. Conservatives backfire to social information about energy consumption.[7]

Backfiring can emerge from individuals' reactions aiming at preserving freedom of choice. At the base of such behavior is the presence of psychological reactance (Brehm, 1966) according to which individuals display a state of motivation that leads individuals to regain lost freedom. For example, Fitzsimons and Lehmann (2004) find that unsolicited recommendations yields to backfiring and dissatisfaction in a consumer choice experiment. Similarly, Bruns and Perino (2021) find that recommendations are perceived as invasive, threatening personal freedom, and they increase anger among targeted individuals. These results are corroborated in a later study by the same authors (Bruns and Perino, 2023) showing that individuals' predisposition to the nudge (e.g., environmental concern) mediates the surge in psychological reactance. More generally, in a cross-country survey by Arad and Rubinstein (2018), authors report that a substantial proportion of subjects declare to act against paternalistic interventions, although they would have acted in line with the intervention in the absence of it. However, not all studies have reported evidence of psychological reactance in the presence of nudges. In a different, yet related, context, Cagala et al. (2024) study whether soft-commitments (a nudge-based intervention, Bryan et al., 2010) induce psychological reactance. Results show that the introduction of commitments does not increase the sense of threat to individual liberty.

All in all, these studies suggests that appeals may lead to counter-actions depending on personal predispositions. Our study contributes to this strand of literature. While we do not investigate in depth the psychological factors leading to backfiring (such as anger, or fear), to our knowledge we are the first to shed light on the moderating effect of individual predisposition to follow rules.

3. Methods and hypotheses

3.1. Experimental design

The study took over a total of 36 days, following the same methods as in Szekely et al. (2021). On the first day, we administered a series of pre-experimental online questionnaires. Participants completing all tasks of day 1 were invited to show up online on the following day to start a 35-day long Common Pool Resources game. Participants had 24 h to make decisions in each experimental day, starting from at 10am (CET). In what follows, we explain into details all the experimental stages.

3.1.1. Pre-experimental measures

On the first day of the experiment, participants went through several individual tasks (see Appendix D for further details). Not responding to any of these initial tasks resulted in the exclusion from the experiment.

Rule-Following Task. First, all subjects responded to the rule-following task (Kimbrough and Vostroknutov, 2018). The rule following task is a variant of the task used in Kimbrough and Vostroknutov (2015) and Kimbrough and Vostroknutov (2016), in which subjects' willingness to follow an experimenter stated rule at personal cost provides a measure of rule-following propensity. The task consists of a series of repeated individual decisions. Each individual is endowed with 20 balls to be put one-by-one either in a yellow or a blue bucket. Each ball in the blue bucket gives the participant 0.50 eurocents, while each ball in the yellow bucket gives 1 euro. The total earnings in this task is the sum of earnings from the buckets. As in the original implementation of Kimbrough and Vostroknutov (2018), instructions explicitly said that *"the rule is to put all the balls in the blue bucket"*. Hence, following the rule is per-se a costly action for the individual. Yet, each individual can freely choose how to allocate balls between buckets.

Big Five Inventory. Other measures used as complement to the rule-following task come from the Big Five personality questionnaire (John et al., 1991). The Big Five inventory includes 44 items. Participants rate each item on a 5-point scale ranging from 1 (disagree strongly) to 5 (agree strongly). Past research has found evidence showing association between personality traits and compliance with rules. In particular, agreeableness and conscientiousness are often found to account for rule-following propensity (Roberts et al., 2014; Bègue et al., 2015; Blagov, 2021).

Controls. Finally, we measure other individual traits that will be used as complementary measures and controls. First, we elicited prosociality levels using the Social Value Orientation task (Murphy et al., 2011). In the Social Value Orientation measure, subjects were asked to make 6 decisions in a series of incentivized dictator games that allocated money between themselves and a randomly assigned anonymous partner. These choices are then summarized to form a continuous measure of other-regarding preferences. Secondly, we measured risk aversion using the method of lotteries as in Dave et al. (2010). Lastly, we collected a set of demographic measures, such as age, gender, level of study and political orientation.

3.1.2. Common pool resource game

From the second day on, participants in the experiment played a Common Pool Resource (CPR) game for 35 days. CPR

[7] Other similar studies also found that political preferences play a role in the acceptance of behavioral interventions (see for example Tannenbaum et al., 2017).

environments are widely used to study consumption of depletable, rival resources (Ostrom et al., 1992) and represent an ideal artificial test bed to study the effect of nudge interventions. Examples span from attempts to reduce energy consumption, to pollution, and digital congestion. The procedure of the game followed that of previously implemented experiments (Ostrom et al., 1992; Cason and Gangadharan, 2015; Tverskoi et al., 2022). Groups of $n = 6$ participants were formed at random every day. Subjects in the experiment played anonymously, so their identity was hidden throughout rounds and among groups. Furthermore, to minimize in-group dynamics and ensure heterogeneity of individual types across groups, composition of groups was randomly determined every day.[8]

At the beginning of a day in the experiment, each group member received 30 tokens and decided how much to allocate them between a "common account" and a "personal account". Any token allocated to the common account yielded a payoff proportional to the ratio between one's allocation and to the group total. Allocations to the common account represents effort in extracting a common resource,[9] whose returns obtained by each individual depends on his/her individual extraction levels (x_i), as well as those of others in the same group (Σx_j). All tokens not allocated to the common account (the number of which is $30 - x_i$) were placed into the private account which did not give any extra return. The total monetary payoff in each round is therefore the sum of the payoffs from both the common and private account:

$$\pi_i = 30 - x_i + \frac{x_i}{\sum x_j}\left(a\sum x_j - b\left(\sum x_j\right)^2\right).$$

where, following the set-up of our experiment, $a = 15; b = \frac{1}{12}$.

Because groups are randomly reshuffled every day, and only some rounds are incentivized, we can assume that participants in the experiment will maximize their single-period utility function. Hence, the unique symmetric Nash equilibrium for individual i is $x_i^* = 24$ and the unique optimal Pareto symmetric solution is $x_i^O = 14$.

Notice that parameters of the payoff function have been chosen so that (i) predictions are in integer numbers, (ii) there is enough distance between the symmetric Nash equilibrium and the symmetric Pareto solution to facilitate statistical analyses. Furthermore, the non linearity of the payoff function brings solutions in the boundaries of the action space. Such feature makes appeals a focal point in the game, other than making it closer to reality as it is for social dilemmas in the field (Cason and Gangadharan, 2015; Ostrom et al., 1992; Anderies et al., 2011). To ease calculations when making allocation decisions, subjects were provided with a comprehensive table (in a pdf format, see experimental instructions in Appendix F) showing their personal return from the common account for each possible combination of their allocation and the total allocation of others in the same group. Moreover, subjects were given an online calculator displayed in all decision pages reproducing the content of the table.

At the beginning of each day, before making new decisions in the CPR game, all participant received information about i) the decisions of each subject in their own group (i.e., the single decisions taken by the other five group members), and (ii) their own individual payoffs from the previous day. No information was given about the decisions of participants in other groups in the experiment.

Every day, before each allocation decision, subjects were asked to answer three questions used to study their social expectations and personal norms (Bicchieri and Xiao, 2009). Questions allowed us to elicit subjects' beliefs about what is in their opinion the most appropriate behavior (personal normative beliefs) as well as what they expect others in their own group will do in a given round (empirical expectations), and what they believe is the most socially appropriate action (normative expectations). These latter two measures were incentivized. A full description of these measures can be found in the appendix (Appendix B).

3.2. Treatments

We randomized participants to two conditions. A baseline condition (No-message) in which no message was displayed during the CPR game. Participants in this condition, completed all the tasks on the first day, and then passed to the 35-day long CPR game. In the treatment condition (Message), an appeal message was displayed every day on participants' screens before making decisions in the CPR. The content of the message nudged participants to contribute the Pareto optimal solution. The exact wording was: "*Please, consider that the overall payoff of your group is maximized if each member contributes 14 tokens to the Common Account. This message is being displayed to all participants in the experiment*".. Such a message was displayed every day and subjects read it before making decisions in the CPR game.[10]

3.3. Conceptual framework and hypotheses

Our conceptual framework aims to provide guidance on how appeals shift behavior depending on rule-following predispositions.

[8] We implement a *stranger* matching protocol, hence, it may happen that previously encountered subjects are paired in the same groups more than once. Yet, given the large number of participants in our sessions, such a probability was minimized. Furthermore, subjects were not aware of the total number of participants in the whole sessions which makes impossible for them to estimate the probability of meeting same others.

[9] Following past works (Ostrom et al., 1992; Cason and Gangadharan, 2015), experimental instructions used neutral language, avoiding any reference to extraction of resources, and framing the game as an allocation decision. In what follows, however, we will refer to extraction levels when analyzing allocation decisions given the traditional implementation of CPR games to study extraction problems.

[10] The message was displayed in a separate screen preceding the decision page.

We depart from the fact that rules in society are either *top-down* (i.e., issued by authorities), or *bottom-up*, (i.e., emerging among peers). Adherence to either rule types can be heterogeneous, yet for different reasons. While individuals follow *bottom-up* rules out of their willingness to meet their peers' expectations (Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2015; Bicchieri, 2005), *top-down* rules are followed out of one's own predisposition to follow an authority's prescription (Gavrilets, 2021; Gavrilets et al., 2024; Rosokha et al., 2024; Gächter et al., 2025).[11] In particular, according to psychological theory of reactance, some may have a higher intrinsic motivation to follow a cue by an authority, while others may be less motivated or even act against it to restore a sense of freedom (Brehm, 1966; Rosenberg and Siegel, 2018).

We model this conceptual framework by assuming that individuals' utility function depends on their own monetary payoffs, expectations on how others believe one should behave (i.e., social norms), and appeals from an external authority. For the sake of simplicity, we assume these aspects can be modeled with a utility function in which deviations from norms and appeals by the authority generate either a utility cost or benefit:

$$U_i(x_i, x_{-i}) = \beta_i \cdot V_i(x_i, x_{-i}) - \gamma_i \cdot N(x_i - x^N) - \phi_i \cdot M(x_i - a) \tag{1}$$

where $N'(\cdot) > 0$, $N''(\cdot) < 0$, and $M'(\cdot) > 0$, $M''(\cdot) < 0$. We also assume that $N(0), M(0) = 0$.

The first component, $V_i(x_i, x_{-i})$, can be thought of as the monetary value of the action x for individual i, which also depends on the actions of others. The parameter β_i represents the weight that i gives to monetary payoffs. The second component captures individuals' tendency to comply with an existing norm x^N, which is the action that the individual or society considers appropriate or desirable in a given situation. The parameter γ_i captures the individual's sensitivity to norm compliance or deviation. Based on past literature, we assume it to be non-negative (Krupka and Weber, 2013), meaning that individuals generally like to adhere with prevailing norms and any deviation leads to disutility. However, it can be that an individual displays a negative γ as they like to go against what society prescribes. Notice that, in this simplified framework, we do not make any distinction between personal and social norms. However, the model can be extended to include both deviation from personal and social norms.[12] Additionally, we do not assume that preferences for norm following are belief dependent, i.e., they do not depend on i's expectations about norm compliance (e.g., descriptive norms) as highlighted in McBride and Ridinger (2021).

Similarly to Gavrilets (2021), the third utility component captures the degree of obedience to an appeal (a) from an authority on the individual's utility. We assume that the extent to which individuals follow the appeal is heterogeneous and captured by parameter ϕ_i. When $\phi > 0$, the individual has a preference to follow the appeal a and comply with the authority. On the opposite, when $\phi < 0$, the individual enjoys to deviate from the content of the appeal. This latter behavior can be the consequence of negative psychological reactance to external appeals due to a threat to or loss to their individual autonomy (Brehm, 1966; Rosenberg and Siegel, 2018). Lastly, under the absence of an authority and hence of an appeal a, we assume this component not to enter the utility function.

Under this simplified framework, an individual may deviate from a payoff-maximizing action for the sake of meeting bottom up and/or top down rules. However, their response can change depending on whether they like to adhere with either rule type. Some important aspects of our stylized framework need to be pointed out. First, we relate to the theoretical framework introduced by Kimbrough and Vostroknutov (2015) and Kimbrough and Vostroknutov (2018). In their framework, individuals have heterogeneous propensities to adhere with rules, which it then reflects into their levels prosociality. Our framework builds upon theirs by introducing the role of an external authority that can influence the behavior of individuals by sending a message that recommends or prescribe an action. In this sense, we add the role of psychological reactance and autonomy considered as the individual's sensitivity to the message from the authority. Secondly, one may assume that ϕ and γ are positively correlated, as an individual who has a preference to comply with social norms also likes to follow appeals. However, for simplicity reasons, we do not model such correlation, and assume them to be independent in our framework. Thirdly, we acknowledge that authority's messages can also impact social or personal norms in a subtle way, by, for example, providing a social reference of what one should do in a given situation. Our framework does not consider this possibility, yet we are able to measure and test whether norms change in the presence of appeals thanks to our experimental design.[13]

Following our conceptual framework, the effect of appeals may be heterogeneous depending on individuals propensities to adhere with authority (i.e., ϕ). We break this proposition down into two questions that can be addressed experimentally with our design. First, we look at overall behavioral variability. Under the presence of individual heterogeneity in the propensity to follow rules (i.e., heterogeneous distribution of ϕ), we expect that appeals lead to more behavioral variability than in their absence. This is also suggested by past literature, as discussed in Section 2, which found that appeals had no overall effect on average contribution levels, but they increased the variance of contributions (Croson and Marks, 2001; Dale and Morgan, 2010; My Boun and Ouvrard, 2019). For these reasons, we formulate our first hypothesis:

H1. We expect higher variability in extraction levels under the Message condition than in our No-Message condition.

[11] This is in line with the CRISP framework by Gächter et al. (2025) as well as the framework proposed by Gavrilets (2021), according to which individuals are mainly motivated by intrinsic motives to obey a rule and social forces. These frameworks also account for other reasons, such as avoiding punishment (extrinsic motivation) or social preferences. Given the absence of extrinsic motives in our study and the absence of social consequences in the rule-following task, we focus on intrinsic motivations and social expectations.

[12] For such a more comprehensive framework, see Gavrilets (2021), Gavrilets et al. (2024).

[13] We will discuss more thoroughly the possibility that messages can affect personal and social norms in the results section dedicated to appeals and norms.

We then study individual heterogeneous response to appeals based on their propensity to adhere with appeals. Following our conceptual framework, individuals with high ϕ are expected to adhere with the content of the authority's message. On the contrary, individuals with low ϕ are expected to negatively react. This conjecture is supported by previous research suggesting that nudge interventions are more effective if they in line with people's personal predisposition (Sunstein, 2017). For example, attempts to promote pro-environmental behavior are successful among subjects showing concern for the environment, while backfiring happens for those displaying low concern (My Boun and Ouvrard, 2019; Costa and Kahn, 2013). More generally, a good fraction of surveyed individuals act against paternalistic interventions, even if they would have behave according to what the intervention suggested under the absence of it (Arad and Rubinstein, 2018). Hence, as a result of our message intervention, we expect that:

H2. In our Message condition, individuals that display a higher (lower) predisposition to follow appeals are more likely to comply with the recommended behavior by reducing (increasing) their extraction levels.

3.4. Sample and procedures

We recruited 300 student subjects through the IBSEN[14] platform. Email invitations were sent in April 2021 and informed subjects that the experiments they were invited to participate would last several weeks. A full demographic description of our sample is included in the appendix (Table D1). Randomization was successful in all of our control variables and rule-following rates.[15]

Subjects were remunerated for some of the individual tasks on the first day (rule-following task, risk elicitation task, social value orientation), and from five randomly selected days of the CPR game (1 randomly selected day per week) in which they participated. In all selected days subjects could receive an extra payment based on the accuracy of their empirical and normative expectations (see Appendix B). Tokens were converted into Euros following a conversion rate of 30 Tokens = 1 Euro. To keep subjects engaged until the end of the experiment, 2 participants out 150 involved in each session were randomly selected to multiply by 20 their final payoff in the whole experiment. Participants were informed of this additional payment at the beginning of the experiment, and allowed us to obtain a low attrition rate (as mentioned above, only 17 out of 300 participants did not finish the experiment). Excluding the additional payment from the lottery, subjects earned an average of Euro 31.65.

Instructions used neutral language (see Appendix F). Subjects needed to pass a comprehension quiz in order to start the CPR game. A total of 17 subjects were excluded from the whole experiment (7 already in Day 2 because they missed the first tasks, 10 because they missed more than 5 decisions in the CPR game). On average, participants missed less than 1 choice over the whole experiment. As in previous work (Szekely et al., 2021), when missing a day of the experiment, subjects' decisions were replaced with that of a randomly chosen participant from the same experimental session and other group members were informed about the automatic decision.

We obtain informed consent from all subjects. The study received institutional ethical approval from the Ethics Committee of the Universidad Carlos III de Madrid. The online game was coded in oTree (Chen et al., 2016).

4. Results

In presenting the results, we first explore the general effect of appeals on extraction levels. We then focus on behavior variability across experimental conditions to explore whether messages have led to heterogeneous reactions. In the last part of this section, we explore the causes of higher behavior variability. In particular, we investigate the role of rule compliance measures elicited at the beginning of the experiment.[16]

4.1. Effects of appeals on extraction levels

We report no difference in terms of individual extraction levels in the CPR when comparing the No-Message condition to the Message condition. Average extraction is 20.38 in No-Message while it is 20.22 in the treatment (Student's t= 1.223, $p = 0.22$). Results from our model estimates report no significant difference between conditions (Table 1). Patterns of extractions are identical under both conditions following an increasing trend seen also in previous studies (Cason and Gangadharan, 2015) (Fig. 1; on average, 0.13 points each day, Table 1).

To investigate whether there was a differential effect of nudges on behavior, we look at the variability in extraction levels. A visual inspection of Fig. 2 provides suggestive evidence that extractions variability may differ across conditions. To explore these differences systematically, we perform a thorough analysis of individual choices in the experiment. Variability can be measured in two ways (Croson and Marks, 2001): (i) the variation of an individual's extractions over time (*within-individual variation*) and (ii) the variation of extractions at a given point in time among individuals (*between-individual variation*). For the former measure, we first calculate the individual's average extractions over the 35 periods ($\bar{x}_i$). Then, in each period we calculate the absolute difference between the individual's extractions in that period and the individual's average extractions (i.e., $|x_{it} - \bar{x}_i|$). This measure

[14] https://ibsen-h2020.eu/

[15] Pairwise comparison between conditions on age, gender, student status and rule-following rate were performed. All statistical differences are not significant at any conventional level (all p> 10%).

[16] We use mixed-effects models with random intercept at the individual level given the hierarchical nature of our data. Yet, all our results hold unchanged also when estimating linear models with errors clustered at the individual level. Full results are accessible at our study repository (https://researchbox.org/4590).

Table 1
Models of extraction levels and variability measures (between and within individuals). Mixed-effects models with random intercepts at the individual level. Demographics include age, gender and student dummy.

	Extraction x_i			Between Var. $\lvert x_{it} - \bar{x}_i \rvert$	Within Var. $\lvert x_{it} - \bar{x}_t \rvert$
	(1)	(2)	(3)	(4)	(5)
(Intercept)	20.32***	18.11***	19.43***	3.71***	4.27***
	(0.38)	(0.39)	(0.91)	(0.19)	(0.19)
Message	−0.16	−0.28	−0.27	0.73**	0.85**
	(0.54)	(0.56)	(0.56)	(0.27)	(0.27)
Day		0.13***	0.13***	−0.04***	0.04***
		(0.01)	(0.01)	(0.00)	(0.00)
Day * Message		0.00	0.00	−0.02**	−0.01*
		(0.01)	(0.01)	(0.01)	(0.01)
Demographics	No	No	Yes	Yes	Yes
Log Likelihood	−29714.51	−29349.97	−29351.03	−23825.54	−23947.72
Num. obs.	9800	9800	9800	9800	9800
Num. participant	293	293	293	293	293

Notes: robust standard errors in parentheses. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

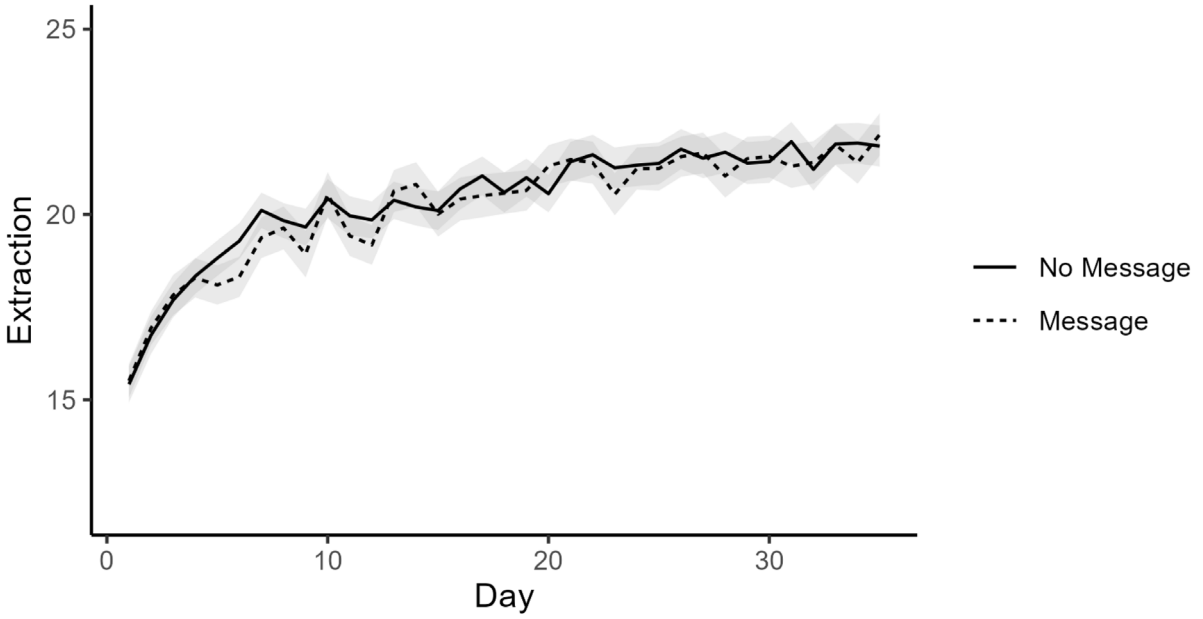


Fig. 1. Average extraction levels by experimental condition.

captures the extent to which individuals change their extractions over time. For the latter measure, we first calculate the average extractions of all participants in a treatment in each period ($\bar{x}_t$). Then, in each period, we calculate the absolute difference between a given individual's extractions and the session-level average extractions in that period (i.e., $|x_{it} - \bar{x}_t|$). This measure captures the extent to which individuals in the same treatment differ from each other in a given period. Results are reported in Table 1 (Models 4–5). While average levels are similar across treatments, extractions are more heterogeneous under Message conditions using both measures of variability. We obtain consistent results when performing a more conservative test on variance equality.[17] This is in line with previous work finding that moral appeals increase extraction variability in a voluntary contribution game (Croson and Marks, 2001).

Result 1. Extraction variability increases under the presence of appeals.

[17] A variance equality test shows a significantly higher variance in Message than in No-Message, $F = 0.8838$, $p < 0.001$.

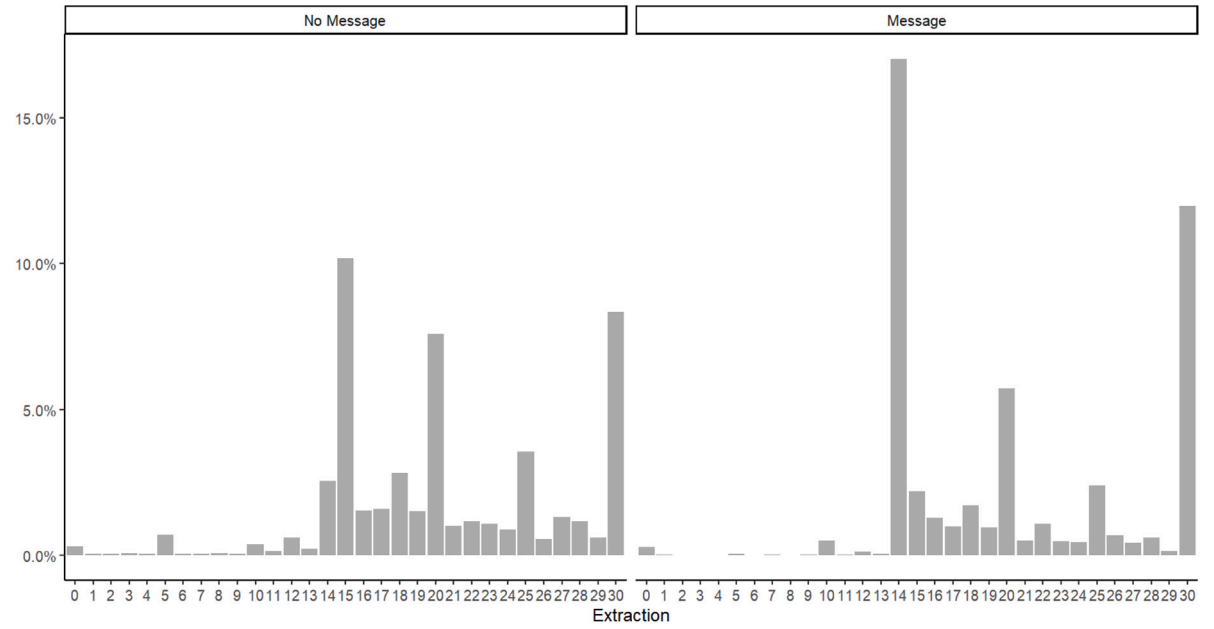


Fig. 2. Distribution of extractions by treatment.

4.2. Reactance and rule-following propensity

What leads to higher variability in extraction levels under the Message condition? Do appeals have a heterogeneous effect depending on individuals' propensity to follow rules? To shed light on these questions, we provide a twofold evidence. We first use measures of rule compliance in the Rule Following Task (RFT) to indirectly show that the higher variability under the treatment condition is explained by individuals' propensity to follow rules. Second, we offer more direct evidence of the heterogeneity in obedience to authority's appeals, and the consequences on extractions, by structurally estimating the utility function used in our theoretical framework.

Model-Free Results. An important step of our empirical identification of reactance towards authority is to report evidence in support of the heterogeneity in ϕ across individuals. As a first approach, we use the results from the RFT. It is important to acknowledge that this task has been used in past studies (Kimbrough and Vostroknutov, 2016, 2018) to capture the parameter γ, that is the agent's concern to follow social norms. We propose that the RFT can also capture a broader tendency to follow rules, including individuals' tendency to follow externally-imposed, top-down rules. From a conceptual point of view, this is possible because in the task itself, subjects are required to follow an arbitrary rule, imposed by a third-party (i.e., the experimenter) who plays the role of an authority. In support of this conjecture, Gächter et al. (2025) use a similar elicitation to disentangle motives behind rule following. Authors found that a large fraction of individuals follow rules out of their sense of duty or intrinsic motivation to follow what an authority (e.g., the experimenter) suggests, and not only their preferences to meet others' expectations.

We define the rule following rate as the ratio between actions in compliance with the rule (number of balls put in the blue bucket) over the total number of taken decisions (20 overall). Distributions of rule following rate are quite spread across the 0–1 interval, and statistically similar across conditions, showing that participant composition is comparable (Fig. 3, t=0.886, $p = 0.387$). About 35% of subjects in each condition has perfectly complied with the rule, while about 30% of them have never done so, similarly to previous studies (Kimbrough and Vostroknutov, 2018). Rule following rates do not correlate with our risk preference measures (Spearman $r = -0.08$, $p = 0.17$), nor gender (χ^2 test, $p = 0.77$), and age (Spearman $r = -0.04$, $p = 0.44$).

For convenience, in the summary statistics and figures that follow, we classify participants into two categories: Rule Followers, when an individual's rule following rate is higher than 50%, and Rule Breakers otherwise. Following this categorization, we classify 43% of our sample (N = 126) as Rule Breaker, and the remaining 57% (N = 167) as Rule Followers (see Table D1). Yet, in all analyses, we consider the rule following rate as a continuous variables.[18]

In Fig. 4a, we report extraction levels split between Rule Breakers and Rule Followers across experimental conditions. It is easy to notice that extraction levels are similar in the No-Message condition, while, under the presence of appeals, Rule Breakers

[18] Our results hold even when considering subjects with a rule following rate different than 1 as rule breakers as in Kimbrough and Vostroknutov (2018). Yet we keep the threshold of 50% as we believe that subjects close to 100% are more likely to be similar to rule followers than rule breakers. In all of our analyses, we privilege the continuous measure of rule following as main indicator.

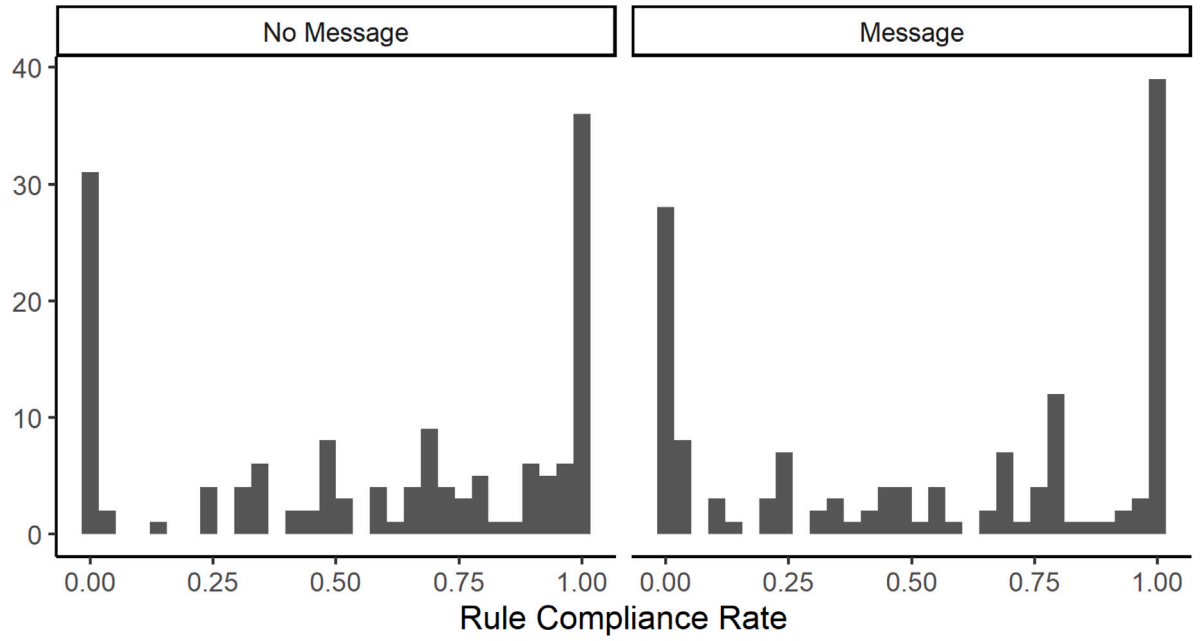


Fig. 3. Rule-following rates from RFT (percentage of balls in blue bucket).

display higher levels relative to Rule Followers. When looking at decisions by types, appeals seems to have an opposite effect. Rule Followers decrease their extractions under the Message condition, while Rule Breakers' increase them (Fig. 4b). In the last day of the experiment, rule breakers average extraction level is about 15% higher than rule followers (24.09 vs. 20.50, respectively). This is equivalent to a *Cohen's d* = 0.55, which is considered as a medium–large treatment effect size (Cohen, 1988). Estimates from regression models support this graphical evidence (Table 2). Model (1) and (2) analyze the behavior of rule breakers and rule followers separately. Predictor *Message* has a significantly different effect depending on types ($b = 1.85$ for rule breakers, $p = 0.023$; $b = -1.84$ for rule followers, $p = 0.009$). When considering types in the continuum using the measure of rule following, results hold unchanged. In the No-Message condition, rule following rate does not explain extraction behavior, while it has a significant effect under the Message condition (model 3, interaction term, $b = -4.44$, $p < 0.001$). Put differently, rule following rate plays a role only under the presence of appeals. These results hold unchanged even after controlling for social and personal expectations (see Appendix, Table D3).

Structural-Estimation Results. Results from the previous section hinge upon the assumption that rule following rates are a proxy of ϕ. However, they can also capture heterogeneity in γ_i as highlighted in past studies (Kimbrough and Vostroknutov, 2018). To offer a more direct evidence of the heterogeneity in ϕ and how such heterogeneity explains the difference in extraction variability, we estimate utility parameters. As a simplifying assumption, we assume linearity in $M(\cdot)$ and $N(\cdot)$, and plug empirical expectations into Eq. 1 to derive expected payoffs. In the Message condition, we estimate the full model including all utility components. In the No-Message condition, we estimate a restricted model by imposing $\phi = 0$, given the absence of appeals. However, as a sanity check, we also estimate the unrestricted model. We use two estimation techniques: first, we assume the existence of a representative agent and provide general estimates (i.e., by pooling data from all participants in the experiment); second, we report the results at the individual level (i.e., by estimating utility parameters for each participant by exploiting the repeated nature of our data). Further details on the procedures followed in the estimation are reported in Appendix A.

Results from the representative agent model indicate a general concern for social norms (i.e., $\gamma > 0$), and substantial heterogeneity with respect to ϕ (Table A1). This heterogeneity is systematically related to rule compliance: in the presence of appeals, Rule Breakers display lower values of ϕ than Rule Followers (Table A1, model 6). Individuals with the highest rule compliance rates exhibit a positive ϕ, whereas those with a rule compliance rate of zero display a negative ϕ. The estimated difference between these two extreme cases amounts to 0.091 ($p < 0.001$). As a robustness check, we find that conterfactual estimates of ϕ in the No-Message condition are statistically indistinguishable from zero and do not vary across rule-following types (Table A1, model 7), confirming that concern for authority and differences between types emerge only when appeals are present.

To provide a closer look at the distribution of parameters, and the relation with extraction in the CPR game, we report individual-level estimates (see Table A2 and Figure 6). We find that in the Message condition around 63% of participants have a positive ϕ, while the remaining 37% have a negative one. Again, this variation aligns with rule compliance: among Rule Followers, the median ϕ is 0.2439, which is significantly higher than the median for Rule Breakers ($\phi = -0.0214$; $p = 0.007$). Consistent with the representative agent results, the vast majority of participants display a positive γ. Importantly, we find no systematic differences in γ between rule-following types. The only exception is a decline in γ among Rule Breakers when moving from the No-Message to

Table 2
Regression models of extractions split by types. Mixed-effects models with random intercepts at the individual level.

Sample:	Extraction x_i				
	(1) Rule breaker	(2) Rule followers	(3) Both	(4) No Message	(5) Message
(Intercept)	18.34***	18.83***	18.05***	15.76***	21.15***
	(2.20)	(1.89)	(1.55)	(2.14)	(1.97)
Message	1.85*	−1.84**	2.29*		
	(0.81)	(0.70)	(0.93)		
Day	0.17***	0.10***	0.13***	0.13***	0.19***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
Rule following rate			1.43	1.68	−0.95
			(0.97)	(0.99)	(1.01)
Rule following rate * Message			−4.44**		
			(1.35)		
Rule following rate * Day				−0.01	−0.11***
				(0.02)	(0.02)
Demographics	Yes	Yes	Yes	Yes	Yes
Log Likelihood	−12593.29	−16723.34	−29338.93	−14504.72	−14799.68
Num. obs.	4199	5601	9800	4923	4877
Num. participant	126	167	293	148	145

Notes: standard errors in parentheses. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

the Message condition, which is accompanied by a relative increase in ϕ. Taken together, results from both estimation strategies underscore that rule compliance scores capture individuals' heterogeneous tendencies to respond to authority appeals.

When analyzing extraction levels, we find that individuals with $\phi < 0$ (i.e., those who react against authority) extract significantly more than those with $\phi > 0$ (Wilcoxon test on average individual contributions, $z = 8.585$, $p < 0.001$; see Figure 7). Regression analyses corroborate this finding (Table A3), indicating that participants with a negative ϕ consistently extract more than those with a positive ϕ.

Result 2. Individuals display heterogeneous preferences for authority obedience. Those with a higher (lower) ϕ steer behavior in compliance (against) the content of the appeal.

What are the consequences of appeals on the distribution of payoffs between rule breakers and rule followers? When pooling together subjects irrespective of their rule-propensity scores, we do not find any overall differences in final payoffs between conditions (No-Message and Message, respectively, average of individual payoffs: 3472.55 vs. 3515.01 points, $p = 0.52$). However, appeals seem to have a redistributive effect between types (Sunstein, 2022). While we find no difference in final payoffs in the No-Message condition between types (rule breakers and rule followers, respectively, earned on average, 3369.08 vs. 3543.09 points, Student's t-test, $p = 0.12$), we find higher payoffs among rule breakers than rule followers in the Message condition (3602.5 vs. 3441.9 points, Student's t-test, $p = 0.04$).

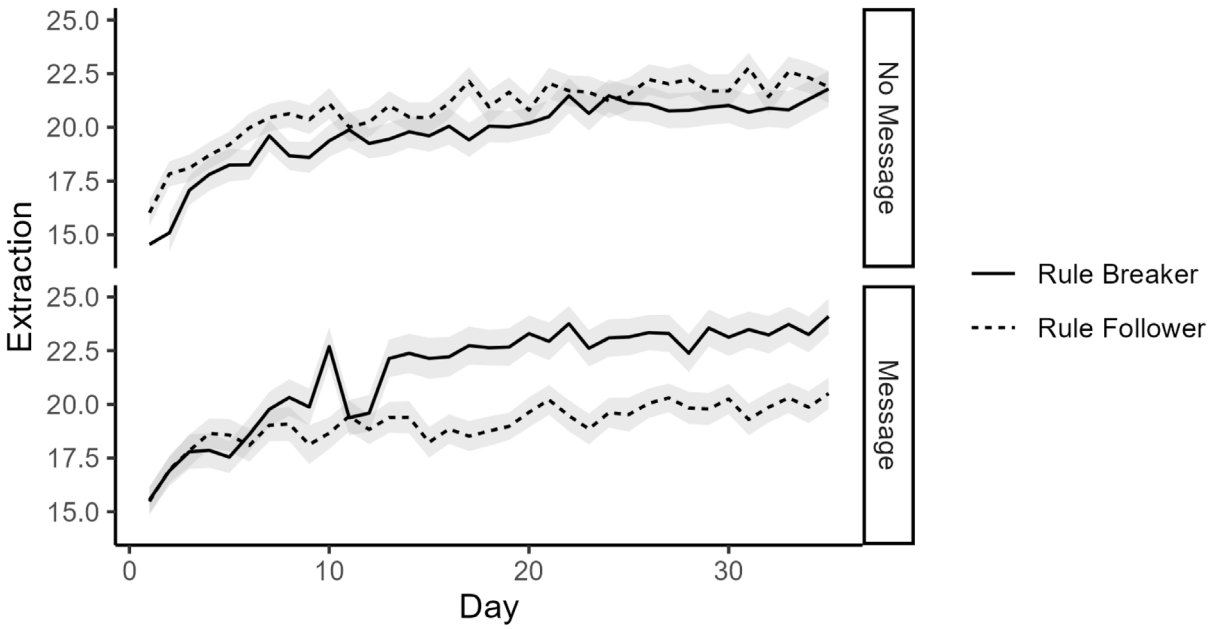
4.2.1. Emergence of backfiring over time

Previous analyses show that behavior of rule breakers is overall different from that of rule followers under the presence of nudges. In this section, we shed light on the dynamics over game repetitions. Although we do not make any conjecture on the dynamics of behavior over time, two scenarios are plausible. On the one hand, it is reasonable to expect that individuals may initially react to the message upon first receiving it, to then eventually reverting to the original state because of habituation over time. This is in line with the wearing off effect of nudges (Allcott and Rogers, 2014). On the other hand, another stream of work has also showed that recommendations and persuasive messages might not be immediately effective and behavioral change might happen over time (Fitzsimons and Lehmann, 2004; Allcott, 2011; Ito et al., 2018; Gravert and Kurz, 2021).

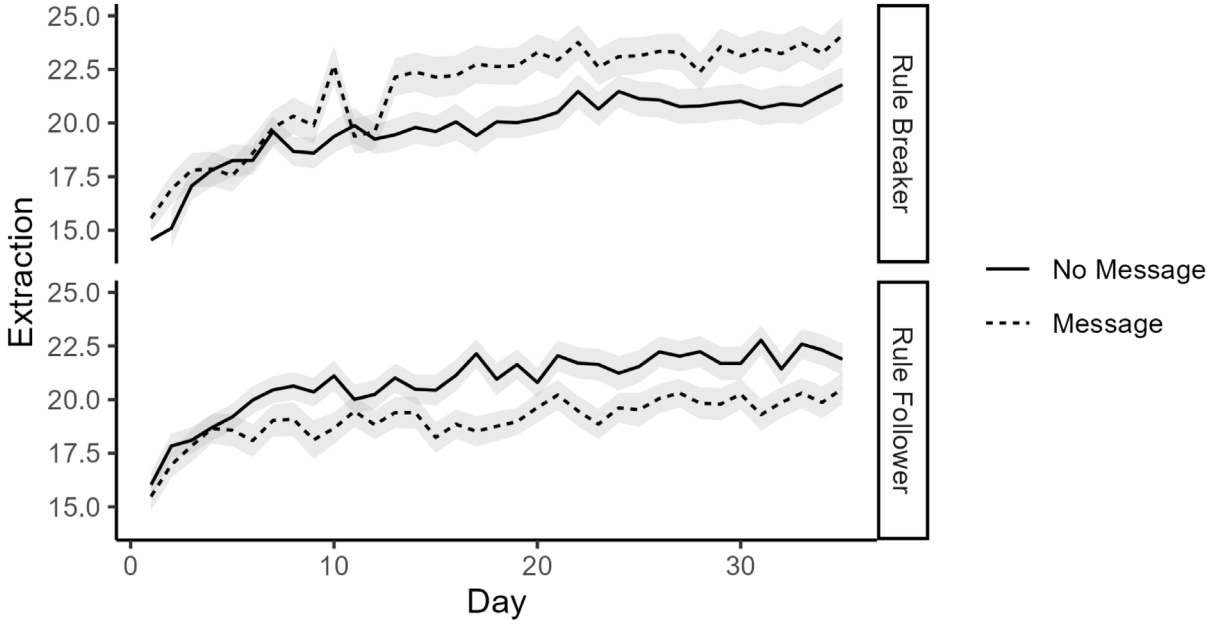
Our data suggest that differences between the rule breaker and rule followers widen over time. Fig. 4b shows that rule breakers increase their extraction levels compared to rule followers in the Message condition. Results from a regression model interacting *Day* with *Rule following Rate* confirms such intuition. We divide our sample by experimental conditions and run separate analyses for each subsample (Table 2, models 4 and 5). When analyzing data from the No-Message condition, we report no statistical difference between rule followers and breakers (model 4, $b = 1.68$, $p = 0.11$), nor any difference arise over time ($b = -0.01$, $p = 0.485$). We report a significant difference emerging through experimental days between rule followers and rule breakers extraction levels under our treatment condition (model 5, interaction term, *Rule following rate* and *Day*, $b = -0.11$, $p < 0.001$). Both norm followers and norm breakers had to receive the appeal several times before starting changing their behavior: rule breakers tend to increase their extractions levels, while rule followers to reduce them.

4.3. Appeals and norms

While our results report evidence of psychological reactance based on rule-following propensity, there is an alternative explanation of our results hinging upon social and personal norms. Appeals can produce a behavioral change by shaping norms



(a) Average extraction levels by experimental conditions for both types



(b) Average extraction levels by type under both experimental conditions

Fig. 4. Comparison of extraction levels by type and experimental condition. 95% confidence intervals reported in the shaded area.

in two plausible ways. On the one hand, an appeal can indirectly lead subjects to form normative expectations about what is the socially accepted behavior within our context, hence shaping the perception of what one should do in a given situation (McKenzie et al., 2006; Everett et al., 2015; Bicchieri, 2005; Moon and VanEpps, 2022).[19] On the other hand, they can also represent for some individuals a signal of general low compliance with a desired behavior, which motivates the need for an external intervention of an

[19] Tverskoi et al. (2022) propose a novel mathematical framework to model the dynamics and the co-evolution of behavior, beliefs and attitudes.

authority (Sliwka, 2007; Nyborg and Rege, 2003). For example, rule breakers may believe that appeals are motivated by a general low level of compliance with the prevailing norm resulting in higher levels of extractions. We test both conjectures by looking at social and personal norms elicited every experimental day using the method proposed by Bicchieri and Xiao (2009).

Fig. 5a reports personal normative beliefs, that is what one believes it is right to do, elicited across treatments and types. Form a visual inspection, personal norms slightly differ across conditions (left panel). Regression results show that personal normative beliefs are significantly lower in the Message treatment (Table D4, model 3, $b = -1.21$, $p = 0.002$). When analyzing by types (right panel), more insights about this change can be drawn. It is easy to notice that messages did not affect rule breakers' personal norms, while they slightly decrease those of rule followers (Table D4, models 1–2). Yet the effect is not significant when pooling all subjects (Table D4, model 4). In this case, when considering a continuous measure of rule following, regression estimates show that personal normative beliefs do not differ across types.

To investigate a possible change in empirical and normative expectations (Bicchieri and Xiao, 2009), Fig. 5b reports empirical expectations, that is what one believes others will do, and Fig. 5c depicts normative expectations, that is what one believes it is socially appropriate to do. For both variables, we report no significant difference between the No-Message and the Message condition. Irrespective of the type, appeals have no consistent impact on both normative and empirical expectations (see Appendix, Tables D6, D5). Table D6 in the Appendix reports evidence that appeals reduced rule followers' empirical expectations, yet the effect disappears when pooling considering all types.

Put together, these results show that appeals have not impacted the norms on extraction levels. We report some evidence of a change in rule followers' personal norms and empirical expectations. Yet, as shown in our regression estimates (Tables D6, D4), the effect of appeals is not consistent across models. Our results cast doubts about the validity of the two conjectures mentioned above: appeals do not systematically strengthen social and personal norms towards a desired behavior, nor they signal to some individuals a lack of compliance which motivates selfish behavior.

4.4. Robustness checks

To further support the idea that individual predisposition to follow rules moderates the effect of appeal messages, we back up our main results with measures that are related to rule following.

First are our elicited measures of personality from the Big Five questionnaire (John et al., 1991). Past work has found that measures of personality are generally correlated with individual willingness to comply with rules (Han, 2021; Otterbring and Festila, 2022; Bègue et al., 2015). Two measures among the five have been found to be significant predictors of one's willingness to follow rules, that is agreeableness and conscientiousness. Agreeableness is one of the five major dimensions of the Big Five inventory and describes individual differences in being likeable, pleasant, and harmonious in relations with others (Graziano and Tobin, 2009). Conscientiousness represents the propensity to follow socially prescribed rules, to be goal directed, to plan, and to be able to delay gratification (Roberts et al., 2009). People displaying higher level of agreeableness and conscientiousness were more likely to abide by public mass communications (Blagov, 2021; Nofal et al., 2020; Asselmann et al., 2020). Roberts et al. (2014) show that conscientiousness is fundamentally related with rule abiding tendency, self-control, and finally, morality and virtue.[20]

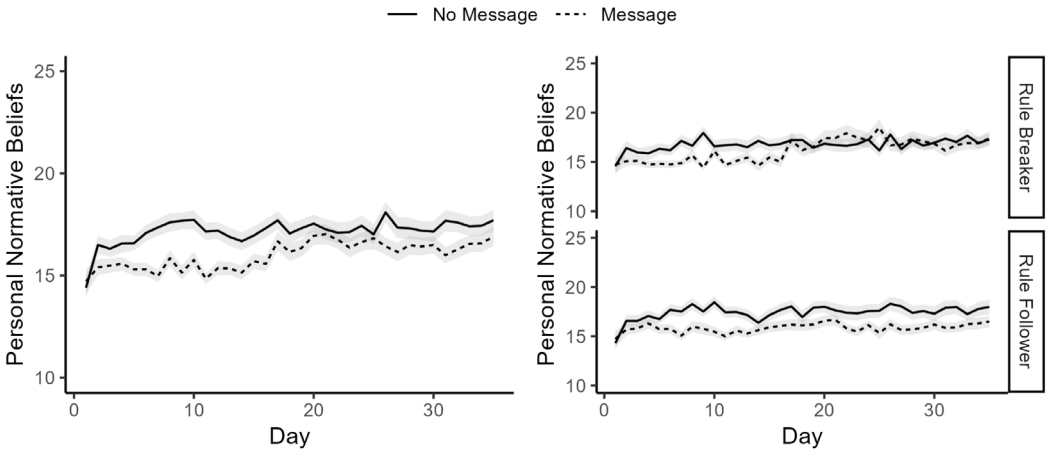
We find that agreeableness positively correlates with rule following rates (Spearman $r = 0.17$, $p = 0.004$). Conscientiousness positively correlates with rule following rates, yet the relation is not significant (Spearman, $r = 0.09$, $p = 0.14$). In Table D7, we regress all personality measures on extraction levels. None of these measures are significantly associated with extraction levels, with the exception of agreeableness which has a negative significant impact only in the Message condition ($b = -3.47$, $p = 0.003$). This evidence indicates that more agreeable participants, when receiving appeals messages, tend to significantly reduce extraction levels.

A second robustness check that we perform concerns individuals' level of prosociality measured via the Social Value Orientation task (Murphy et al., 2011). The effect of rule following rates on extraction levels may simply be the result of individuals' prosociality levels. Individuals complying with the rule in the rule-following task are not intrinsically rule followers, but simply less selfish than rule breakers. To disentangle between the effect of rule-following tendency and prosociality, we regress extraction levels on both rule following rates and SVO angle (Table D8). Results depicted in previous sections hold even under the addition of these covariates. SVO angle negatively correlates with extraction levels when regressed alone ($b = -0.10$, $p = 0.011$, Model 1). However, appeals still have a heterogeneous effect depending on subjects' rule following rates ($b = -3.04$, $p = 0.03$, Model 2). Rule followers reduce extraction levels under the Message condition, while rule breakers increase it. When including *Rule following rate*, levels of prosociality measured through SVO angle have no significant effect on extraction levels ($b = -0.08$, $p = 0.09$).

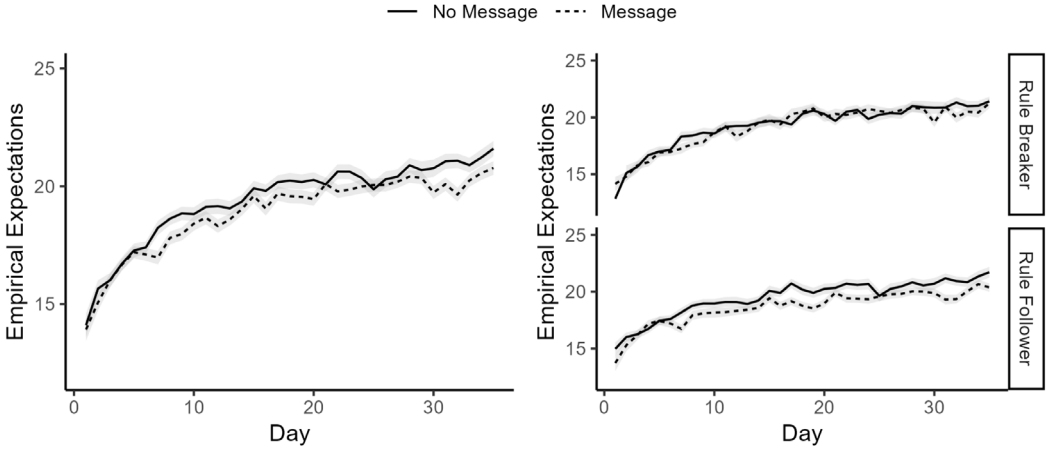
Lastly, we aim to disentangle an alternative explanations motivating Rule breakers' behavior. Their higher extraction levels may be motivated not by reactance to appeals, but because it is rational to do so if they strongly expect that others will decrease extraction levels. To shed light on this, we study empirical expectations (namely what one believes others will do) elicited during the CPR game. In the previous section, we have reported evidence showing that empirical expectations do not change across our two experimental conditions. As further evidence, Table D6 reports the regression estimates predicting empirical expectations among rule breakers and rule followers under both experimental conditions. By breaking down the sample in two, results show that appeals do not shift rule breakers' expectations, while they even decrease slightly those of rule followers. The effect of *Message* disappears when pooling together the two sub-samples.[21] From these estimates, it is clear that empirical expectations are not impacted by messages, which strengthens our conjecture that rule breakers react negatively to messages rather than simply acting opportunistically.

[20] As an interesting contrast, Tate et al. (2022) find no correlation between personality traits and rule following propensity in a sample of adolescents.

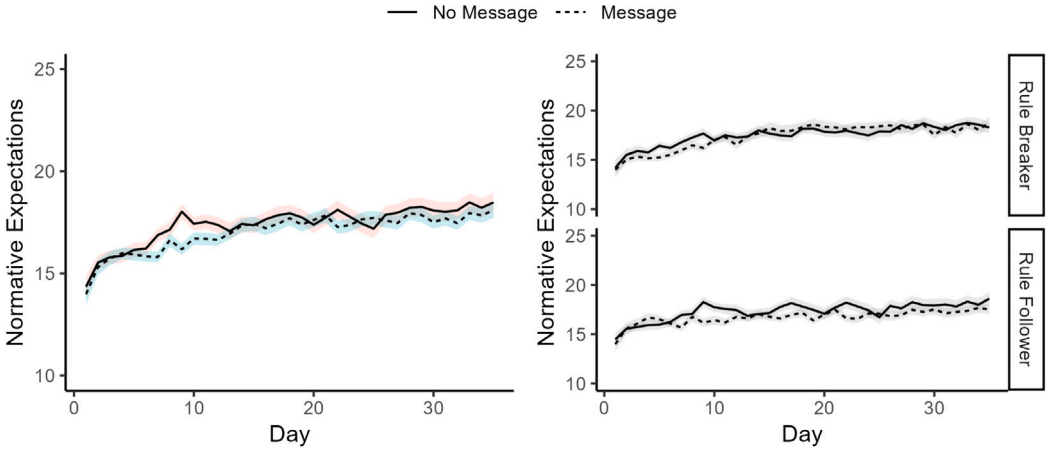
[21] Moreover, we find no difference in the variance of empirical expectations between types, and no evidence of rule breakers expecting more people in their group to choose an extraction level around 14 than rule followers do.



(a) Personal Normative Beliefs aggregated and by rule following types.



(b) Empirical Expectations aggregated and by rule following types.



(c) Normative Expectations aggregated and by rule following type

Fig. 5. Expectations over rounds. 95% confidence intervals reported in shaded areas.

5. Discussion

Our experimental evidence suggests two main findings: first, appeals increase the variability of extraction levels, leaving average levels unchanged; second, participants displaying higher level of rule compliance changed their behavior in line with the content of the appeal, while those displaying lower levels go against it.

Our first result is in line with past experimental literature (Croson and Marks, 2001; Dale and Morgan, 2010) documenting higher behavioral variability and heterogeneous reactions to appeals in the context of collective-action games. We contribute to these works by suggesting that rule-following propensity plays a key role in determining individuals' heterogeneous reaction. Our findings are in agreement with recent research on Psychological Reactance Theory (Brehm, 1966; Rosenberg and Siegel, 2018) and suggest that predisposition to a policy intervention, despite preserving freedom of choice, can be perceived as a threat to autonomy (Arad and Rubinstein, 2018; Sunstein, 2017; Bruns and Perino, 2021). To corroborate these results, we have also investigated whether personality traits can explain the difference in behavior observed between the two experimental conditions. Our results show that Big 5 personality traits do not correlate with rule following propensity (Tate et al., 2022), and do not explain the emergence of backfiring behavior, with the exception for agreeableness. More agreeable individuals steer their behavior under the presence of appeals, while those less agreeable go against it.

A novel aspect of our results is the fact that reactance to appeals does not emerge immediately, but over game repetitions. As reported in Fig. 4, the gap between rule breakers and followers widens as time unravels. This evidence is consistent with past research showing that recommendations and persuasive messages are not immediately effective (Fitzsimons and Lehmann, 2004; Allcott, 2011; Ito et al., 2018).

We also shed light on the role of alternative plausible channels. First, we investigate the role of social expectations (Bicchieri, 2005). Appeals can shape social norms, ultimately affecting behavior (McKenzie et al., 2006; Moon and VanEpps, 2022). Additionally, the intervention of an external authority may be considered as a signal of general lack of norm compliance (Sliwka, 2007; Nyborg and Rege, 2003), which would translate into a difference in individuals' empirical and normative expectations between experimental conditions. Yet, we do not report any significant difference in these measures across conditions. Second, we also rule out the role of individuals' level of prosociality, according to which, more prosocial individuals are more likely to adhere to appeals to enhance social welfare, while less prosocial individuals (e.g., those classified as competitive or individualistic) may even act to the detriment of others (Murphy et al., 2011). Results from our analyses show that prosociality, measured using the Social Value Orientation task, does not predict extraction levels, nor it changes the statistical significance of rule-following rates in our model.

These findings offer some insights for nudge-based policies (Carlsson et al., 2021; Sunstein, 2017, 2022). Nudge-bases policy campaigns should take into account individuals' heterogeneity in their propensity to follow rules. Nudge-based interventions have been criticized for their one-size-fits-all approach as it can lead to small effects or unintended consequences to some population subgroups (Bryan et al., 2021; Sunstein, 2022). Accounting for the extent to which people are susceptible to an intervention (recently defined as "nudgeability"; de Ridder et al., 2022) is a way to maximize the success of a policy, and our results show that the tendency to follow rules is an important aspect to consider.

However, nudge personalization opens the way to new debates. We believe that a major issue to be tackled in the future concerns the reaction of the public to personalization. While the recent development of information technologies provides an opportunity to design personalized persuasion via the collection of data (Matz et al., 2017; Mills, 2022), there is no consensus on how the public perceives personalization and the usage of personal data.[22]

6. Conclusion

Over the past years, policy-makers and organizations have enthusiastically resorted to public appeals for bringing about behavioral change. Yet, despite their low cost of implementation, there is no clear evidence on their effectiveness in promoting desirable behavior.

This paper shows that heterogeneity in rule following propensity is key to understanding the effect of appeals. Our work reports evidence from an experiment on the effect of socially-beneficial appeals in leading individuals to change their conduct towards the best solution for the whole group in a Common Pool Resource game. Results show no overall effect of appeals on subjects' extraction levels, but rather an increase in behavioral variability. Such heterogeneity is explained by measures of rule following, with rule followers complying more with the content of the message and rule breakers going against it. Our findings suggest that the effectiveness of informational nudges depends on individuals' disposition to follow externally-imposed rules. Thanks to our pre-experimental tasks, we rule out possible alternative mechanisms, such as other-regarding preferences. Finally, we find that the difference in behavior between types diverge over game rounds. This result calls for the importance of examining the effect of repeated messages on behavioral change. The effect on behavior might require time and occur only after people have listened to the message multiple times. However, the result may not be the desired one, as appeals might lead to backfiring effects.

Future research should keep furthering our understanding of heterogeneous responses to behavioral interventions. An interesting question that can be addressed experimentally could be whether leaving the opportunity to "switch off" appeals could reduce backfiring from rule breakers, or even solicit more cooperative behavior because their freedom of choice has been preserved. Similarly, our experimental results should be interpreted within the scope of the design, where the appeal was issued by the experimenter. Other forms of appeals, for instance those voiced by participants themselves, may produce different behavioral responses and remain an important avenue for future research. We leave this and related open questions for future research.

[22] Kozyreva et al. (2021) report mixed evidence on the acceptance of personalization across domains.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2025.107293.

Data availability

The data and scripts are available at the link indicated in the manuscript (https://researchbox.org/4590).

References

Allcott, Hunt, 2011. Social norms and energy conservation. J. Public Econ. 95 (9–10), 1082–1095.

Allcott, Hunt, Rogers, Todd, 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. Am. Econ. Rev. 104 (10), 3003–3037.

Anderies, John M., et al., 2011. The challenge of understanding decisions in experimental studies of common pool resource governance. Ecol. Econom. 70 (9), 1571–1579.

Andrighetto, Giulia, et al., 2013. Punish and voice: punishment enhances cooperation when combined with norm-signalling. PloS One 8 (6), e64941.

Arad, Ayala, Rubinstein, Ariel, 2018. The people's perspective on libertarian-paternalistic policies. J. Law Econ. 61 (2), 311–333.

Asselmann, Eva, et al., 2020. The role of personality in the thoughts, feelings, and behaviors of students in Germany during the first weeks of the COVID-19 pandemic. PloS One 15 (11), e0242904.

Bègue, Laurent, et al., 2015. Personality predicts obedience in a milgram paradigm. J. Pers. 83 (3), 299–306.

Bicchieri, Cristina, 2005. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press.

Bicchieri, Cristina, Xiao, Erte, 2009. Do the right thing: but only if others do so. J. Behav. Decis. Mak. 22 (2), 191–208.

Blagov, Pavel S., 2021. Adaptive and dark personality in the COVID-19 pandemic: Predicting health-behavior endorsement and the appeal of public-health messages. Soc. Psychol. Pers. Sci. 12 (5), 697–707.

Brandts, Jordi, Cooper, David J., 2007. It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. J. Eur. Econ. Assoc. 5 (6), 1223–1268.

Brandts, Jordi, Cooper, David J., Weber, Roberto A., 2015. Legitimacy, communication, and leadership in the turnaround game. Manag. Sci. 61 (11), 2627–2645.

Brehm, Jack W., 1966. A Theory of Psychological Reactance. Academic Press.

Bruns, Hendrik, Perino, Grischa, 2021. Point at, nudge, or push private provision of a public good? Econ. Inq. 59 (3), 996–1007.

Bruns, Hendrik, Perino, Grischa, 2023. The role of autonomy and reactance for nudging—Experimentally comparing defaults to recommendations and mandates. J. Behav. Exp. Econ. 106, 102047.

Bryan, Gharad, Karlan, Dean, Nelson, Scott, 2010. Commitment devices. Annu. Rev. Econ. 2 (1), 671–698.

Bryan, Christopher J., Tipton, Elizabeth, Yeager, David S., 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. Nat. Hum. Behav. 5 (8), 980–989.

Cagala, Tobias, et al., 2024. Commitment requests do not affect truth-telling in laboratory and online experiments. Games Econom. Behav. 143, 179–190.

Carlsson, Fredrik, et al., 2021. The use of green nudges as an environmental policy instrument. Rev. Environ. Econ. Policy 15 (2), 216–237.

Cason, Timothy N., Gangadharan, Lata, 2015. Promoting cooperation in nonlinear social dilemmas through peer punishment. Exp. Econ. 18 (1), 66–88.

Chen, Daniel L., Schonger, Martin, Wickens, Chris, 2016. Otree—An open-source platform for laboratory, online, and field experiments. J. Behav. Exp. Financ. 9, 88–97.

Cohen, Jacob, 1988. Statistical Power Analysis for the Behavioral Sciences. Academic Press.

Costa, Dora L., Kahn, Matthew E., 2013. Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. J. Eur. Econ. Assoc. 11 (3), 680–702.

Croson, Rachel, Marks, Melanie, 2001. The effect of recommended contributions in the voluntary provision of public goods. Econ. Inq. 39 (2), 238–249.

Dal Bó, Ernesto, Dal Bó, Pedro, 2014. "Do the right thing:" the effects of moral suasion on cooperation. J. Public Econ. 117, 28–38.

Dale, Donald J., Morgan, John, 2010. Silence is golden. Suggested donations in voluntary contribution games. Working Paper. University of California, Berkeley, USA.

Dave, Chetan, et al., 2010. Eliciting risk preferences: When is simple better? J. Risk Uncertain. 41 (3), 219–243.

de Ridder, Denise, Kroese, Floor, van Gestel, Laurens, 2022. Nudgeability: Mapping conditions of susceptibility to nudge influence. Perspect. Psychol. Sci. 17 (2), 346–359.

DellaVigna, Stefano, Gentzkow, Matthew, 2010. Persuasion: Empirical evidence. Annu. Rev. Econ. 2 (1), 643–669.

Everett, Jim A.C., et al., 2015. Doing good by doing nothing? The role of social norms in explaining default effects in altruistic contexts. Eur. J. Soc. Psychol. 45 (2), 230–241.

Fitzsimons, Gavan J., Lehmann, Donald R., 2004. Reactance to recommendations: When unsolicited advice yields contrary responses. Mark. Sci. 23 (1), 82–94.

Gächter, Simon, Molleman, Lucas, Nosenzo, Daniele, 2025. Why people follow rules. Nat. Hum. Behav. 1–13.

Gavrilets, Sergey, 2021. Coevolution of actions, personal norms and beliefs about others in social dilemmas. Evol. Hum. Sci. 3, e44.

Gavrilets, Sergey, Tverskoi, Denis, Sánchez, Angel, 2024. Modelling social norms: an integration of the norm-utility approach with beliefs dynamics. Philos. Trans. R. Soc. B 379 (1897), 20230027.

Gelfand, Michele, et al., 2022. Persuading republicans and democrats to comply with mask wearing: An intervention tournament. J. Exp. Soc. Psychol. 101, 104299.

Gravert, Christina, Kurz, Verena, 2021. Nudging à la carte: a field experiment on climate-friendly food choice. Behav. Public Policy 5 (3), 378–395.

Graziano, William G., Tobin, Renée M., 2009. Agreeableness. The Guilford Press.

Hallsworth, Michael, et al., 2017. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. J. Public Econ. 148, 14–31.

Han, Hyemin, 2021. Exploring the association between compliance with measures to prevent the spread of COVID-19 and big five traits with Bayesian generalized linear model. Pers. Individ. Differ. 176, 110787.

Horton, John J., Rand, David G., Zeckhauser, Richard J., 2011. The online laboratory: Conducting experiments in a real labor market. Exp. Econ. 14, 399–425.

Ito, Koichiro, Ida, Takanori, Tanaka, Makoto, 2018. Moral suasion and economic incentives: Field experimental evidence from energy demand. Am. Econ. J.: Econ. Policy 10 (1), 240–267.
John, Oliver P., Donahue, Eileen M., Kentle, Robert L., 1991. The Big Five Inventory—Versions 4a and 54. University of California, Berkeley, Institute of Personality, Berkeley, CA.
Karakostas, Alexandros, Zizzo, Daniel John, 2016. Compliance and the power of authority. J. Econ. Behav. Organ. 124, 67–80.
Kimbrough, Erik O., Vostroknutov, Alexander, 2015. The social and ecological determinants of common pool resource sustainability. J. Environ. Econ. Manag. 72, 38–53.
Kimbrough, Erik O., Vostroknutov, Alexander, 2016. Norms make preferences social. J. Eur. Econ. Assoc. 14 (3), 608–638.
Kimbrough, Erik O., Vostroknutov, Alexander, 2018. A portable method of eliciting respect for social norms. Econom. Lett. 168, 147–150.
Kozyreva, Anastasia, et al., 2021. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. Humanit. Soc. Sci. Commun. 8 (1), 1–11.
Krupka, Erin L., Weber, Roberto A., 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? J. Eur. Econ. Assoc. 11 (3), 495–524.
Levy, David M., Padgitt, Kail, Peart, Sandra J., Houser, Daniel, Xiao, Erte, 2011. Leadership, cheap talk and really cheap talk. J. Econ. Behav. Organ. 77 (1), 40–52.
Matz, Sandra C., et al., 2017. Psychological targeting as an effective approach to digital mass persuasion. Proc. Natl. Acad. Sci. 114 (48), 12714–12719.
McBride, Michael, Ridinger, Garret, 2021. Beliefs also make social-norm preferences social. J. Econ. Behav. Organ. 191, 765–784.
McKenzie, Craig R.M., Liersch, Michael J., Finkelstein, Stacey R., 2006. Recommendations implicit in policy defaults. Psychol. Sci. 17 (5), 414–420.
Mills, Stuart, 2022. Personalized nudging. Behav. Public Policy 6 (1), 150–159.
Moon, Alice, VanEpps, Eric, 2022. Giving suggestions: Using quantity requests to increase donations. J. Consum. Res. 10.
Murphy, Ryan O., Ackermann, Kurt A., Handgraaf, Michel, 2011. Measuring social value orientation. Judgm. Decis. Mak. 6 (8), 771–781.
My Boun, Kene, Ouvrard, Benjamin, 2019. Nudge and tax in an environmental public goods experiment: Does environmental sensitivity matter? Resour. Energy Econ. 55, 24–48.
Nofal, Ahmed Maged, Cacciotti, Gabriella, Lee, Nick, 2020. Who complies with COVID-19 transmission mitigation behavioral guidelines? PloS One 15 (10), e0240396.
Nyborg, Karine, Rege, Mari, 2003. Does public policy crowd out private contributions to public goods. Public Choice 115 (3–4), 397–418.
Ostrom, Elinor, Walker, James, Gardner, Roy, 1992. Covenants with and without a sword: Self-governance is possible. Am. Political Sci. Rev. 86 (2), 404–417.
Otterbring, Tobias, Festila, Alexandra, 2022. Pandemic prevention and personality psychology: Gender differences in preventive health behaviors during COVID-19 and the roles of agreeableness and conscientiousness. J. Saf. Sci. Resil. 3 (1), 87–91.
Reiff, Joseph, et al., 2021. When impact appeals backfire: Evidence from a multinational field experiment and the lab. Available at SSRN 3946685.
Roberts, Brent W., et al., 2009. Conscientiousness. The Guilford Press.
Roberts, Brent W., et al., 2014. What is conscientiousness and how can it be assessed? Dev. Psychol. 50 (5), 1315.
Rosenberg, Benjamin D., Siegel, Jason T., 2018. A 50-year review of psychological reactance theory: Do not read this article. Motiv. Sci. 4 (4), 281.
Rosokha, Yaroslav, et al., 2024. Evolution of cooperation in the indefinitely repeated collective action with a contest for power. Econom. Theory 1–31.
Silverman, Dan, Slemrod, Joel, Uler, Neslihan, 2014. Distinguishing the role of authority “in” and authority “to”. J. Public Econ. 113, 32–42.
Sliwka, Dirk, 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. Am. Econ. Rev. 97 (3), 999–1012.
Sunstein, Cass R., 2017. Nudges that fail. Behav. Public Policy 1 (1), 4–25.
Sunstein, Cass R., 2022. The distributional effects of nudges. Nat. Hum. Behav. 6 (1), 9–10.
Szekely, Aron, et al., 2021. Evidence from a long-term experiment that collective risks change social norms and promote cooperation. Nat. Commun. 12 (1), 1–7.
Tannenbaum, David, Fox, Craig R., Rogers, Todd, 2017. On the misplaced politics of behavioural policy interventions. Nat. Hum. Behav. 1 (7), 1–7.
Tate, Christopher, et al., 2022. The personality and cognitive traits associated with adolescents’ sensitivity to social norms. Sci. Rep. 12 (1), 15247.
Tverskoi, Denis, et al., 2022. Disentangling material, social, and cognitive determinants of human behavior and beliefs. SocArXiv preprint.
Van Bavel, Jay J., et al., 2020. Using social and behavioural science to support COVID-19 pandemic response. Nat. Hum. Behav. 4 (5), 460–471.
Zizzo, Daniel John, 2010. Experimenter demand effects in economic experiments. Exp. Econ. 13, 75–98.