Journal of Economic Behavior and Organization 197 (2022) 624–642

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

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

Moral opportunism as a consequence of decision making under uncertainty☆

Nitzan Merguei, Martin Strobel, Alexander Vostroknutov*

Department of Economics (MPE), Maastricht University, Tongersestraat 53, Maastricht 6211LM, the Netherlands

ARTICLE INFO

Article history:
Received 9 March 2021
Revised 17 March 2022
Accepted 20 March 2022

JEL classification:
C91
C92

Keywords:
Moral opportunism
Normative punishment
Excessive punishment
Social norms
Norm elicitation

ABSTRACT

When people with different normative beliefs interact, moral opportunism—or the tendency to follow the norm that brings the highest material benefit—can arise. Our conjecture is that this behavior is a consequence of expected norm-dependent utility maximization under uncertainty. Using a novel theoretical framework for studying social norms, we experimentally test this idea in the Dictator game with second-party punishment. The theory links normative beliefs and punishment strategies, which allows us to study what determines recipients' punishment choices after they are shown normative beliefs of their dictators. We corroborate the theory and find that many recipients indeed act in accordance with the maximization of expected norm-dependent utility that has a negative flavor of moral opportunism. We also find that some of the recipients are excessive punishers: they punish a lot, but not according to their normative beliefs. Excessive punishment is exhibited by recipients from Southern Europe, but not by recipients from the rest of Western Europe.

1. Introduction

Asking for directions in India can be an unusual experience. You will be pleasantly surprised that everyone you ask gladly explains to you how to get to your destination, just to realize afterwards that the directions are dead wrong. Later, while boiling over such episode, you learn that in India it is considered impolite to not provide directions, and that you need to look for subtle body-language signs to tell whether what you are hearing is a "real thing" or a polite nonsense. The realization that this is a local tradition calms you down since people who gave you directions were not deliberately violating a social norm as you see it, but instead followed another one. You decide not to reprimand, or otherwise punish, them even though they deliberately lied to you.

This example shows how people react when they encounter contradictory norms in a multicultural setting. In general, however, it is not very clear what reactions the multiplicity of norms can (or should) produce. When you face a situation in which you know that others have different opinions about the social norm that applies, should you insist that your beliefs

☆ We would like to thank the participants of the MU-CEN seminars at Maastricht University for invaluable comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All mistakes are our own.

* Corresponding author.
E-mail addresses: nitzan.merguei@gmail.com (N. Merguei), m.strobel@maastrichtuniversity.nl (M. Strobel), a.vostroknutov@maastrichtuniversity.nl (A. Vostroknutov).

https://doi.org/10.1016/j.jebo.2022.03.020

are “more correct” than theirs? Should you just stick to what they think or take a “convex combination”? Or perhaps you should choose the norm that is “cheaper”? In this paper, we attempt to shed some light on these issues using a novel theoretical framework (Kimbrough and Vostroknutov, 2020, further KV) that allows to analyze moral decisions within the utilitarian tradition (Kessler and Leider, 2012; Krupka and Weber, 2013) and conduct an experiment that allows us to tackle these questions in a quantifiable way.

Ideas about how people approach situations where multiple possible rules of conduct apply have been floating around for quite some time (Hoffman and Spitzer, 1985; Frey and Bohnet, 1995; Kagel et al., 1996; Straub and Murnighan, 1995; Bicchieri, 2008; Bicchieri and Chavez, 2010; Fershtman et al., 2012). Studies scattered across social sciences seem to discern one pattern that emerges in a variety of different contexts. The pattern is that people—when facing a multiplicity of different social norms or rules pertaining to the same situation—choose to follow the one that brings them the highest material benefit.[1] For example, Kassas and Palma (2019) find that in a Dictator game with ambiguous ownership claims subjects exhibit a “self-serving bias” by interpreting the situation in a way that favors them materially. Other studies (e.g., Eftedal et al., 2020) talk about “principled” versus “opportunistic” motives when reacting to injustice. van Baar et al. (2019) report that some subjects engage in *moral opportunism* defined as a strategy of switching between different moral principles depending on which one is less costly.

In this paper, we follow the steps of the authors cited above, however, our question is defined differently. While most studies provide somewhat circumstantial evidence of moral opportunism under various, often hard-to-verify, assumptions about the perception of the multiplicity of norms, we focus on an interaction between two players where one *knows exactly* the normative beliefs of the other. Thus, our question is how a player—who knows that she is interacting with someone who holds different opinion about social appropriateness of outcomes—incorporates this knowledge into her decisions. This goes beyond previous literature that is concerned with how a player chooses to follow one of the many presumably possible norms regardless of what others believe. We are interested in *how* and *why* behavior changes when (different) normative beliefs of others are known.

To investigate these questions, we make a simple observation—general to Bayesian decision making under uncertainty with norm-dependent utility specifications (Tremewan and Vostroknutov, 2020)—that any uncertainty about prevalent norms leads to expected-utility-maximizing behavior that takes into account all these possible norms (through the probabilistic beliefs about their occurrence). This is an instance of standard expected utility maximization that, by taking into account all possibilities, arrives at an optimal decision that might be “wrong” from the individual perspective of any of possible norms considered separately (or in situations where there is no uncertainty). Given that utility maximization is involved and the (norm-dependent) utility incorporates consumption as well as normative components, it is not surprising that each agent when maximizing *expected* utility of such kind will end up choosing something that from *some* normative perspective will be considered opportunistic. We conjecture, therefore, that “moral opportunism” is a natural consequence of norm-dependent utility maximization under uncertainty.

To connect this idea to the Indian example above, ask yourself: What would *you* do (as a person who considers lying inappropriate) when someone asks you for directions in India? In case you do not know the directions, would you lie or just admit that you do not know where the destination is? If you do the former you violate the “truth-telling” norm, if you do the latter you violate the local custom. The logic of optimization under uncertainty suggests that you will take into account all consequences of these alternative choices, in terms of both consumption and morality, and make the choice that maximizes your overall norm-dependent utility. So, you may optimally choose to lie about directions even though this behavior might look “opportunistic” from the perspective of those who have never been to India.

In order to study moral opportunism while taking into account normative uncertainty, we have chosen the context of normative punishment (Fehr and Fischbacher, 2004). The reasons for this are severalfold. First, punishment, and specifically costly punishment, is a purely normative phenomenon related to norm violations (Mackie, 1982; Fehr et al., 2002; Fehr and Fischbacher, 2004; Falk et al., 2005; Masclet and Villeval, 2008; Elster, 2009), which makes it an ideal candidate to test hypotheses about how divergent norms are aggregated. Second, costly punishment is an example of the ultimately unselfish behavior in one-shot interactions, which makes it unconfounded by material incentives (Fehr and Gächter, 2002). The only choice that punishers have is to lose less or more money. Thus, changes in punishment strategy can reveal how exactly normative beliefs of others get incorporated into the decisions. Third, individual normative beliefs are complex objects that assign social appropriateness to all outcomes in a game (Kessler and Leider, 2012; Krupka and Weber, 2013), so it is not easy, if possible at all, to uncover how own and others’ normative beliefs are aggregated from a single choice in, say, a Dictator game. We need to know the reactions of subjects in all possible contingencies in order to have a glimpse at this aggregation process. Second-party punishment with strategy method provides us with such an opportunity. Finally, a theory of injunctive norms by KV, which includes a model of normative punishment, presents us with the means to predict the punishment strategy from elicited beliefs about social norms. With this model, it becomes possible not only to study the prevalence of normative punishment and specific deviations from it, but also—and more importantly—to test hypotheses about the nature of the norm-aggregation process and moral opportunism in particular. If the model is supported by the

[1] Notice that we focus on multiplicity of *social norms* (shared by most members of a community) rather than personal norms, which, in essence, are just individual unconstrained beliefs (see, e.g. Bašić and Verrina, 2021). The reason is that multiplicity of personal norms does not present a problem of choosing between them, since people are not exactly supposed to follow someone else’s personal norm if it is different from the social norm. This is the same with different individual preferences: if someone has different preferences, it does not mean that others should follow them.

evidence from our experiment, then we can imagine that the predictions of the model will hold in other settings as well. Given that the framework of KV is general and applies to all games, this paves the way to a theory of moral opportunism for various strategic interactions.

The experiment can be summarized as follows. In the first stage, we use a novel continuous norm elicitation task (CNE task) to elicit subjects' beliefs about social appropriateness of all actions in the Dictator game. This is a version of the well-known task proposed by Krupka and Weber (2013), with the only difference that subjects choose appropriateness levels on a continuum instead of the discrete 4-item Likert-scale. This modification was necessary since we needed the recipients to gain a good understanding of the beliefs of the dictators, which cannot be done with only four possible levels of appropriateness. In the next stage, paired subjects play the Dictator game with second-party punishment as in Fehr and Fischbacher (2004). While dictators decide on their offers, recipients choose how much they would like to punish them for each possible choice (strategy method). The treatment manipulation involves showing the recipients the beliefs of their dictators that just got elicited in the CNE task. Thus, recipients, when formulating their punishment strategy, know precisely what the dictators believe is appropriate. They can also compare their own and dictator's beliefs on a graph that presents this information in a nice, user-friendly manner. The control is the same except that the beliefs of the dictator are not shown. Finally, in the last stage of the experiment, subjects participate in another, also incentivized, CNE task where they are asked to correct their normative beliefs in case they find it necessary. This second CNE task is used to estimate the perceived normative uncertainty in individual subjects by measuring how much they have adjusted their beliefs.

We find that the behavior of our subjects is generally consistent with moral opportunism as a consequence of decision making under uncertainty. Subjects act "opportunistically" and choose to punish according to their own or dictator's normative beliefs, depending on which entails less punishment costs. We also find that recipients' individual degree of moral opportunism correlates with how normatively uncertain they are, which supports our main idea that moral opportunism and normative uncertainty are related. Notice that unlike the studies cited above that see moral opportunism as something that only "people without moral values" would engage in, our theory and experiment suggest that *anyone*, even highly norm-abiding individuals, can be morally opportunistic simply because under uncertainty any expected utility maximization can be considered wrong from some normative perspective.

Having a model of normative punishment also allowed us to get a better look at specific properties of punishment strategies that *did not follow* the model's predictions. The model predicts, unsurprisingly, that agents, who have low propensity to follow norms in general (close to standard selfishness), should not punish others at all due to costs related to implementing the punishment. However, such selfish behavior is one thing that we *do not* observe in our experiment at all (with very few exceptions). What we do observe is exactly the opposite: the model of normative punishment does not fit in certain cases because some subjects punish *too much*. Specifically, we observe what we call *excessive punishment*, or a strategy to subtract some fixed amount of money from the dictator regardless of her actions. The model of normative punishment cannot accommodate such behavior simply because it assumes that punishment ensues from dictator's breaking a norm, which can never produce uniform punishment of all available actions. We find that around half of our subjects use excessive punishment to some degree. More importantly, we find that normative punishment and excessive punishment are in a sense mutually exclusive: the more a recipient follows one strategy, the less she follows the other. Interestingly, the average punishment costs are twice higher in the group of excessive punishers than in the group of normative punishers.

We did not design our experiment to test any specific demographics-related hypotheses. Nevertheless, we find a strong effect of nationality on being normative or excessive punisher. In particular, we find much more excessive punishment among recipients from Southern Europe than among recipients from other Western European countries, who are predominantly normative punishers. Together with the observation that the average normative beliefs in these groups of recipients are virtually the same, this suggests a specific cultural component related to punishment that drives this difference. We do not have any specific hypotheses to account for this, however, it is reminiscent of anti-social punishment detected by Herrmann et al. (2008), who also found that it is prominent in Southern and Eastern Europe, as well as in Arabic countries, but not in other Western European countries. In addition, this difference might be related to the tightness or looseness of societies (Gelfand, 2019), a trait that defines how strictly the norms are followed. Within this framework, the results of Dimant et al. (2022) regarding uncertainty about the behavior of others are coherent with related findings in this study. We believe that our results can inspire more experiments that can focus on fleshing out the cultural difference that we have discovered.

2. Framework

2.1. Moral opportunism and normative uncertainty

The view that we promote in this study—namely, that the phenomenon of "moral opportunism" is the consequence of utility maximization under normative uncertainty—is a direct implication of the models with norm-dependent utility specification advocated in the social norms literature (e.g., Kessler and Leider, 2012; Krupka and Weber, 2013). Indeed, suppose that elements of some set C (consequences) describe the allocations in a Dictator game;[2] suppose also that the dictator has

[2] In this section we use Dictator games to illustrate how our argument works. The analysis applies equally well to any other game (see KV).

some norm-dependent utility $U(c) = u(c) + \phi\eta(c)$, where $u(c)$ is the consumption utility; $\eta : C \to [-1, 1]$ is a *norm function* that specifies social appropriateness of consequences in C, and $\phi \geq 0$ is a fixed, individual norm-following parameter. In this literature, it is assumed that the decision maker *maximizes* this utility function. Thus, already without normative uncertainty, the agent faces a *trade-off* between personal consumption and following norms. So, if ϕ is low, the agent will not choose the normatively best option or *the norm* (the consequence c^* with the highest $\eta(c^*)$), but rather something that brings her higher consumption utility than in c^*.

In principle, we can already call this type of behavior "moral opportunism," because the socially appropriate behavior coded in η is opportunistically traded-off for personal gain. However, this is not how this term is usually used in the literature cited in the Introduction, where it is assumed that there are "multiple norms." Our view is that the same maximization of norm-dependent utility as above is responsible for this specific brand of moral opportunism as well. Suppose that the setting is as above, but with a difference that now the agent thinks that there are two norm functions η_1 and η_2 that realize in two states of the world with probabilities p and $1 - p$. In this case, the agent maximizes $u(c) + \phi(p\eta_1(c) + (1 - p)\eta_2(c))$, which will produce some optimal solution depending on the inputs $(u, \eta_1, \eta_2, p, \phi)$. This behavior might be "morally opportunistic" for two reasons: 1) because ϕ is low (selfish behavior, as above) and/or 2) because of the interaction of η_1, η_2, and p. For example, suppose that we have a very rule-following agent with arbitrarily high ϕ, but with $\eta_2(c) = 1$ for all $c \in C$ (all consequences are equally appropriate, which implies that η_2 does not contain any normative information). The introduction of this possibility with any probability $1 - p$ will already shift the optimal choice towards selfishness, because the convex combination of any η_1 with a constant makes η_1 flatter, decreasing marginal utilities and shifting optimal choice. Moreover, as $p \to 0$, we will observe the dictator choosing more and more selfish options *not because she is selfish*, but rather because she is following the flat norm function that says that any consequence is equally appropriate and thus that in this case only consumption utility matters. This is closer to the typical description of moral opportunism in the literature, and also the type of opportunism we are interested in in this paper. An important point we try to make here is that moral opportunism is not some "special" kind of behavior, but that it is simply the consequence of expected norm-dependent utility maximization under uncertainty, which we formulate in a definition.

Definition. *Suppose that in the Dictator game with consequences C an agent, who believes that the applicable norm function is η_1, chooses some action c_1. Now, suppose that new information arrives suggesting that another norm function, η_2, is also possible. Then, we call agent's behavior* ***morally opportunistic*** *if after receiving this information her choice changes to $c_2 \neq c_1$.*

This definition summarizes the idea described above. It says that behavior is morally opportunistic when the agent is *sensitive* to new information about other norms or, in other words, information about other norms makes her change her behavior. Such change in behavior is *moral*, because it is driven exclusively by the new information about norms in the Dictator game. Such change is also *opportunistic from the perspective of* η_1 because information about the new norm makes agent choose differently than when she only took η_1 into account. For the sake of expositional clarity, we do not discuss this definition further, however in Appendix A we provide more involved argumentation.

With this definition in mind, we can think about situations or games that we can use to test this idea. If we tried to use subjects' behavior in a Dictator game with two possible norm functions η_1 and η_2, it would not give us a good test of the norm-dependent utility maximization simply because the maximization problem will be sensitive to all parameters $u, \eta_1, \eta_2, p, \phi$, and it would not be easy or even possible to separate their influences. We could introduce second-party punishment (by the recipient). This would focus dictator's behavior more on following norms, but again it would not resolve the identification issue. However, if in the Dictator game with second-party punishment, we provide *the recipient* with the information on the norm function of the dictator, then we can observe which norm function the recipient uses for punishment (her own, the dictator's, or some mixture of both). In this case, individual norm-following propensity of the recipient (her ϕ) is not very important as long as it is high enough that the recipient chooses to punish at all. The belief p and consumption utility u also become secondary, since the recipient chooses the amount of punishment that should be related to the appropriateness of the *behavior* of the dictator. This leaves us with only functions η_1 and η_2 that can influence the punishment choice, and this is the setup we chose to analyze experimentally. In the following subsection we lay down the theoretical underpinnings of this setting in more detail.

2.2. Model of normative punishment

In order to estimate how recipients formulate their punishment strategies using some possible norm function(s), we use a model of *normative punishment* that maps norm functions into punishment choices. Such a model was developed by KV as a part of their general theory of injunctive norms. Given some norm function, this theory postulates that players feel resentment when they observe that an action of a previous mover is inconsistent with reaching the most socially appropriate outcome prescribed by the norm function. This resentment motivates them to punish norm violators.[3] Thus, KV propose a model of normative punishment that takes an arbitrary norm function defined on the outcomes of some game and produces a *punishment norm function*, or a separate injunctive norm that prescribes how players, who maximize norm-dependent utility, should punish others for norm violations. In this section, we apply the model of normative punishment

[3] As is explained below, the amount of resentment felt by a player is proportional to the difference in social appropriateness of the most appropriate outcome and the one that has been reached.

to the Dictator game in order to understand punishment behavior of recipients and the trade-off between the two norm functions that they make.

We start with defining some notation consistent with the experiment we describe below. Let $C = \{0, 10, \ldots, 100\}$ be the set of amounts that a dictator can keep for herself, which corresponds to the set of allocations in the Dictator game. Let $\eta_r : C \to [-1, 1]$ be the *recipient's norm function* normalized to the interval $[-1, 1]$.[4] Suppose a dictator chooses the outcome $c^* \in C$ such that $c^* = \arg\max_{s \in C} \eta_r(s)$, or the outcome that the recipient deems most socially appropriate.[5] In this case, the recipient does not punish the dictator. If however the outcome c, chosen by the dictator, is not the most socially appropriate according to η_r, then the punishment mechanism is activated. To compute how much the dictator must be punished, we need to find 1) the interval of dictator's possible "after-punishment" payoffs and 2) recipient's *resentment* of outcome c that determines how a payoff from this interval is chosen.

To determine the interval of after-punishment payoffs we proceed as follows. The lowest bound of the interval is taken to be the lowest payoff that the dictator can achieve in the game, which corresponds to the harshest punishment possible. This payoff is equal to 0. The highest bound of the interval is determined as follows. In the typical cases when the dictator decides to choose an outcome that gives her higher payoff than c^*, or $c > c^*$, the highest bound is determined by the *Deterrence principle* (KV), which states that punishment after the dictator chose such c should deter her from deviating from the most socially appropriate outcome. Thus, the dictator's after-punishment payoff should never exceed the payoff that she would have gotten should she have chosen the most appropriate outcome c^*. In some non-typical cases, when the dictator chooses an outcome $c < c^*$ (usually when the dictator chooses to give more than half of the pie to the recipient), the highest bound of the interval is equal to c. In other words, if the dictator has chosen the amount for herself that is less than what she could have obtained by choosing the most socially appropriate outcome, she is still violating the norm and should be punished, so her payoff should be less than what she has. Overall, the highest bound of the interval of the after-punishment payoffs can be expressed as $m_c = \min\{c, c^*\}$. Thus, for any choice c of the dictator, the amount of money that the dictator ought to be left with after punishment should lie in the interval $[0, m_c]$. Resentment of the action c pins down the exact choice of punishment in this interval.

We define *resentment* that the recipient feels after the dictator has chosen some $c \in C$ as the size of norm violation equal to $r_c = \eta_r(c^*) - \eta_r(c)$. In other words, resentment is the highest when the action with the lowest appropriateness value was chosen and is zero for the action with the highest appropriateness value (c^*). KV postulate an *Eye-for-an-Eye principle*, which states that punishment should be proportional to the resentment r_c.

For each $c \in C$, we have computed the interval $[0, m_c]$ in which the dictator's after-punishment payoff should lie. We have also determined the amount of resentment r_c that the recipient feels. KV suggest a simple formula for the amount of payoff that the dictator *should be left with* after optimal punishment by a norm-following recipient: $p_c = (1 - \frac{r_c}{2})m_c$. The recipient's maximization problem that generates this outcome is solved in Appendix B. Notice that $p_c = 0$ when resentment is the highest ($r_c = 2$), so the dictator should be left with nothing. When the resentment is the lowest ($r_c \to 0$) the dictator should end up with the payoff of m_c.[6]

According to this model the amount that the recipient with the norm function η_r should subtract from the dictator who chooses outcome c is $y_{rc} = c - p_c$ (as long as ϕ is high enough for optimal punishment to be positive). Here c is the payoff that the dictator has chosen for herself, and p_c is the payoff that she should be left with after punishment. This calculation allows us to formulate the general decision problem of the recipient in the presence of two norm functions η_r and η_d (dictator's norm function). It would be natural to assume that the recipient will use only one or the other norm function for punishment. However, given that there is no clarity about what punishment should be used, we consider a more general problem of the recipient. Suppose that the recipient *chooses* a punishment strategy that is a mixture of punishments prescribed by η_r and η_d, namely $y_c(s) = s y_{rc} + (1 - s) y_{dc}$ for some $s \in [0, 1]$ and where y_{dc} is the punishment amount prescribed by η_d when outcome c was chosen. Assume as well that the recipient needs to pre-commit to a punishment strategy before observing the move of the dictator (strategy method in the experiment) and thus believes that the actions of the dictator come from some distribution F over C. Then, the problem of the recipient can be formulated as follows:

$$\max_{s \in [0,1]} E_F[100 - c + \phi(p\eta_r(c) + (1 - p)\eta_d(c)) - y_c(s)/3].$$

Here $100 - c$ is the consumption utility of the recipient from offer $100 - c$; the norm-dependent term is as in Section 2.1; and $y_c(s)/3$ stands for the cost of punishment: as in the experiment, it is a third of the amount subtracted from the dictator, $y_c(s)$, which in its turn depends on punishment strategies prescribed by η_r and η_d. The problem above has the same solution as this one:

$$\min_{s \in [0,1]} E_F[y_c(s)] \quad \text{or} \quad \min_{s \in [0,1]} s E_F[y_{rc}] + (1 - s) E_F[y_{dc}].$$

[4] The appropriateness levels in the experiment take values in the interval [0,10]. The normalization is performed by dividing each appropriateness level by 5 and subtracting 1. This is not strictly necessary, but is done to be consistent with KV.

[5] In principle, a norm function can have multiple maxima. The theory can be easily extended to deal with this case. However, in our experiment 92% of all norm functions elicited from subjects have a single maximum, so for this reason as well as the ease of exposition we restrict our discussion to the norm functions with a single maximum.

[6] We will not deny that this model is completely ad hoc. However, no other model of punishment exists in the literature, so we needed to start somewhere. Fig. 8 of KV shows that this model fits rather well with the average punishment strategies in Fehr and Fischbacher (2004).

This problem has a solution $s^* = 0$ or $s^* = 1$ depending on whether the cost $E_F[y_{rc}]$ or $E_F[y_{dc}]$ is smaller. This is what we call moral opportunism in our experiment: while subjects still choose to punish, which shows that they take norms into account in their decisions, they nonetheless do so in a way that minimizes their costs given that under normative uncertainty it is not obvious what exactly should be punished.

To understand whether y_{rc} or y_{dc} will be used as a punishment strategy (or whether $E_F[y_{rc}] < E_F[y_{dc}]$ holds), we can split these terms into two parts corresponding to the Deterrence and to the Eye-for-an-Eye principles. Using the formula for y_{rc} we get:

$$E_F[y_{rc}] = E_F[c - m_c] + E_F[m_c r_c/2],$$

and similarly for y_{dc}. The first term $E_F[c - m_c]$ is the cost of following the Deterrence principle: it is the expected cost of deterring the dictator from choosing something inappropriate. The second term $E_F[m_c r_c/2]$ corresponds to the Eye-for-an-eye principle as it is the expected cost produced by resentment r_c.

For any arbitrary norm functions η_r and η_d that differ in both the most appropriate action and the amount of possible resentment, we cannot tell what s^* will be since it will be influenced by the two terms within y_{rc} and y_{dc} in non-trivial way. However, if we focus on the norm functions η_r and η_d that are different on only one dimension, then we can formulate some common heuristics to determine s^*. Suppose that $\eta_d(c) = b\eta_r(c)$, where $b \in (0, 1)$, so the two norm functions peak at the same place, then the first term is the same for both of them and s^* is determined solely by the second terms. From the definition of r_c it is clear that the resentment from η_d is just b times the resentment from η_r. Therefore the optimal choice will be $s^* = 0$, since η_d is cheaper to punish. So "on average," the norm function with the *smaller spread* defined by the difference between its maximum and minimum values will be cheaper (assuming same peaks), and this is what we are testing in our experiment: whether people indeed choose the norm function with the smallest spread (because it is indicative of cheaper punishment).

Another possibility is to imagine two norm functions with the same spread, but different peaks (e.g., if η_d is obtained from η_r by "moving" its graph to the left or to the right). In this case the second term is going to be (roughly) constant, and the first term should determine s^*. Here the optimal choice is that of η_r or η_d ($s^* = 1$ or $s^* = 0$) depending on which one has the peak closer to the selfish action of the dictator (the cheapest norm function in this sense is the one that peaks at $c = 100$). So, the general rule in this case is to punish according to η_r or η_d depending on which peak is closer to $c = 100$. We test this possibility in our experiment as well. In the most general case, when η_r and η_d differ on both dimensions, these two "guidelines" (choose the norm function with smaller spread, or choose the norm function with the rightmost peak) will still make for a good decision. Thus, our interest is in whether subjects use any of them.

3. Experimental design

The experiment consisted of three tasks. The first task was a continuous version of the norm elicitation task proposed by Krupka and Weber (2013). The main difference between the original norm elicitation and our task (CNE task) is that in the CNE task subjects could choose the evaluations of appropriateness on a continuous scale and not on a 4-Likert scale. We decided to use this new task because it allowed subjects to express their normative beliefs in a much more precise way than the original norm elicitation task allows. Since the *norm functions* elicited from dictators were later presented to recipients, we needed them to be as informative as possible, so that recipients could properly learn the beliefs of the dictators.[7] In the second task, participants played the Dictator game with second-party punishment as in Fehr and Fischbacher (2004). This task was used to collect data on the punishment strategies of the recipients. The third task was the same as the first one. Here participants were given the opportunity to adjust their norm function reported in the first CNE task. This was done in order to control for the possibility that some recipients may have very uncertain beliefs in the beginning of the experiment, which could influence their behavior, and to check if the observed norm function of the dictators influenced their beliefs. After the main part of the experiment, demographic information was collected. Subjects knew that the experiment consists of three parts, but only learned about what they are before each part.

The experiment had two treatments that differed with respect to the information shown to the recipients. In the Main treatment, recipients were shown their own norm function elicited in the CNE task as well as the norm function of the dictator with whom they were paired.[8] In the Control treatment recipients were not presented with the norm function of the dictator, and could only observe their own norm function. The Control treatment was used as a benchmark for measuring punishment decisions based exclusively on recipients' own norm functions.

The two treatments were run simultaneously in all sessions. Overall, 206 subjects participated in the experiment (58% female, average age 21 years old): 138 in the Main treatment and 78 in the Control treatment.[9] All sessions were run in May

[7] The problem with the 4-Likert scale is the following. Suppose that we need subjects to evaluate the appropriateness of 11 actions in the Dictator game. If they can choose only four levels of appropriateness, then many actions are forced to be assigned the same appropriateness level. This would not be too helpful if recipients wanted to learn the normative beliefs of their dictators. At the same time, having a discrete but finer Likert scale, say, an 11-Likert scale is also not a very good idea, because then subjects in a session with 32 people can spread around these 11 categories making the one chosen by the majority much more random than with the 4-Likert scale. This is not good for eliciting correct beliefs from subjects who now have to also worry about which exact category will attract this small majority. The continuous scale solves both of these problems.

[8] In the first CNE task, dictators were informed that their elicited beliefs may or may not be used in the later parts of the experiment.

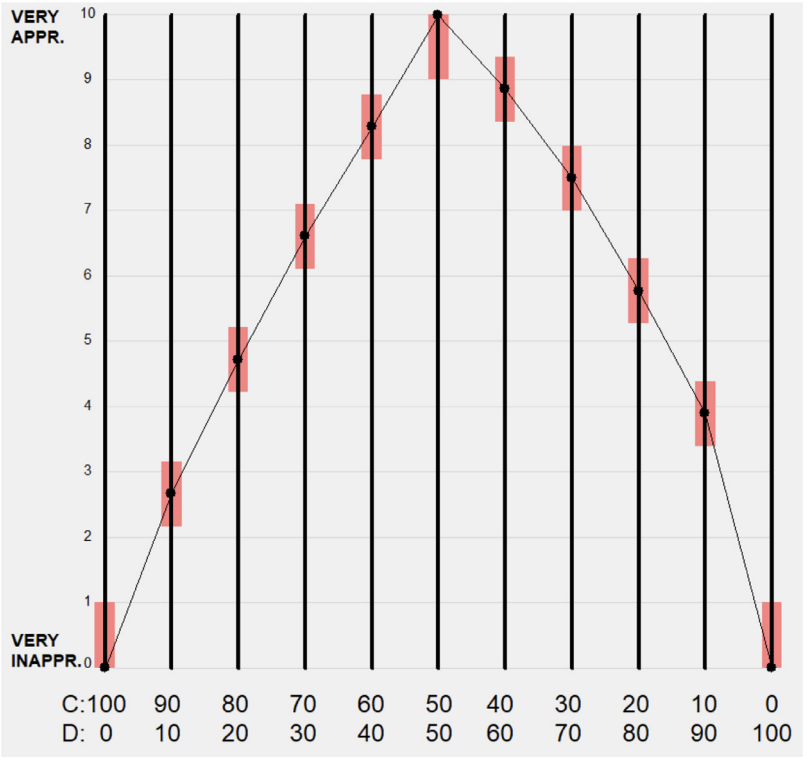

Fig. 1. Continuous norm elicitation task. The *x*-axis displays different allocations of points to Individuals C and D. The *y*-axis corresponds to the appropriateness evaluations from very inappropriate to very appropriate on an arbitrary scale from 0 to 10.

2019 at Maastricht University. There were no pilots, and no data were discarded. The software was programmed in z-Tree (Fischbacher, 2007). Instructions can be found in Appendix G.

3.1. Continuous norm elicitation task

In the CNE task each subject was asked to evaluate social appropriateness of actions in the Dictator game. Subjects were presented with a hypothetical situation in which Individual C has 100 points that he or she can keep or share with Individual D in increments of 10 points.[10] Thus, Individual C can choose one of the 11 available actions: give 0, 10, 20 ... 100 points to Individual D. Subjects were asked to evaluate social appropriateness of each of these actions.

Figure 1 shows the screen on which subjects were choosing their evaluations. For each action of Individual C on the *x*-axis subjects could drag the black circle with the mouse and place it anywhere on the corresponding vertical line. Before the choice, all 11 circles were at the lowest position corresponding to very inappropriate evaluation.[11]

The main difficulty with the continuous norm elicitation as compared to the discrete one of Krupka and Weber (2013) is the correct incentivization. The idea of eliciting normative beliefs in general is that subjects should choose evaluations that they believe are shared by the majority of other subjects in the session. For the Likert-scale task of Krupka and Weber (2013) this is straightforward: subjects are paid for an evaluation if they choose one of the four appropriateness levels that the majority of others have chosen as well. With the continuous scale this is no longer possible. We devised the following method of paying subjects. As the black circle corresponding to some action was moving on the screen, so did the red rectangle (see Fig. 1), which was centered at the location of the black circle. Subjects were told that they are paid for the *percentage* of other subjects whose evaluations fell inside the red rectangle. For example, if all other subjects choose their evaluations inside the red rectangle, then the payment is 3 Euro. If proportion $X \in [0, 1]$ of others put their evalu-

[9] The number of subjects in the Control treatment was designed to be half of that in the Main treatment. The reasons is that in the Main treatment we have two general conditions: when the dictator's norm function predicts *more* punishment than recipient's and when it predicts *less* punishment. Thus, we aimed at collecting the same number of observations in the two conditions of the Main treatment and in the Control treatment.

[10] Subjects were told that each point is "hypothetically" exchanged for 5 Euro cents, same rate as in the actual Dictator game in the next task.

[11] Notice that the decisions for all 11 actions were made on the same screen. This was done for two reasons. First, this helped subjects to choose continuous values that reflected their beliefs about the appropriateness of all actions *relative* to each other. Second, this same screen was used to present the beliefs of the dictators to recipients in the Dictator game that followed. The task of learning dictator's beliefs was made simpler since the beliefs were presented in a familiar way.

ations inside the rectangle, then the payment is 3*X* Euro. This procedure ensured that subjects were incentivized to put their evaluations at the spot where they believed most others are evaluating the action. The reason for using percentages of others instead of the absolute number is to make sure that the procedure does not depend on the number of subjects in the session. If we were to pay for the number of other subjects who put their evaluations in the red rectangle, then the payment would depend on the session size, which is not a desirable feature, as incentives would then depend on it.[12]

The final complication with this payment procedure is the theoretical possibility of "unraveling" that can take place if the payments at the values close to the border are not appropriately taken care of. Suppose that the red rectangle is always centered at the location of the black circle. In this case, when the black circle is placed close to the maximal or minimal possible valuation, some part of the red rectangle will be outside the range of possible values. This will decrease the number of others who can be counted for the payment, thus creating an incentive to place the black circle away from the border. In fact, given some *atomless* belief about the evaluations of others, it is a strictly dominant strategy to put the black circle further away from the border, so that a larger part of red rectangle covers the possible range of values. Like in the Beauty Contest game (Nagel, 1995), repeated elimination of dominated strategies will lead to unraveling or to a unique undominated choice in the middle of the interval.[13] To fix this problem, we have restricted the movement of the red rectangle close to the borders. When the black circle gets closer than half of the length of the rectangle to the border, the rectangle gets "stuck" and stops moving beyond the possible range of values. On Fig. 1 this can be seen for actions 0/100, 50/50 and 100/0.[14] This way the unraveling process does not start and any choice of valuation can be a best response to some belief about the choices of others.[15]

3.2. Dictator game with second-party punishment

When subjects were making their choices in the CNE task they knew only that the experiment consists of three parts, but without being aware of what exactly the other two parts are. It was important to keep it this way, so that dictators—when choosing their evaluations in the CNE task—would not strategically alter their evaluations. For the Dictator game with punishment all subjects read the same on-screen instructions about the procedure that closely followed Fehr and Fischbacher (2004). Each subject was endowed with 50 points (1 point = 0.05 Euro) and then assigned the role of a dictator or recipient.[16] Dictators additionally received 100 points that they could choose to share with the matched recipient in 10-points increments. While dictators were choosing how many points to share, recipients decided by how much they would like to punish their dictator by subtracting points from them.

In order to understand which norm function(s) guided the punishment decisions of recipients, we used the strategy method. Each recipient indicated by how much they would like to punish the dictator for every possible transfer, before knowing the actual allocation chosen by the dictator. Recipients could punish the dictators using the 50 points they were endowed with. They had to forgo 1 point in order to deduct 3 points from the dictator. Thus, the 50 points endowment allowed recipients to deduct all points from the dictators, even if they had received nothing from them. It was not possible however to leave the dictators with negative points, which was explained in the instructions (see Appendix G). To make it clear how many points are paid for punishment and how many points are subtracted from the dictator, participants made their choices on a user-friendly screen that displayed this information for every action (see Fig. 2).[17]

Before making their punishment decisions, recipients in the Main treatment were presented with information about the normative beliefs of the dictator from the previous CNE task. In order to measure if punishments made by recipients were to some extent based on the normative beliefs of the dictators, it was important to match subjects with different norm functions. Therefore, they were paired based on their norm functions elicited in the CNE task. For every subject the software computed a *value* associated with her norm function. Specifically, it subtracted the appropriateness level for the action where the dictator gives nothing to the recipient (100/0) from the appropriateness level for the action leading to equal split (50/50). Subjects with different values thus obtained were matched into pairs.[18]

[12] One may think that this procedure is too complicated to describe in the instructions. However, the instructions are actually rather simple (see Appendix G). In no session of the experiment did we encounter any questions from subjects or expression of confusion related to this procedure.

[13] Notice that this argument critically depends on the atomless nature of the beliefs. If everyone believes that some number X will be chosen, then it is optimal to choose X as well. However, notice that this is not an atomless belief.

[14] Subjects learn this special feature of the rectangle movement on a separate screen where they can practice moving the black circle on a single empty slider. They are told in the instructions to move the circle to the border to make sure that they understand that the rectangle gets stuck when the circle gets too close to the border. It is explained to them that this is done in order to keep their chances of winning the same as when the circle is away from the border (see Appendix G).

[15] This also creates the situation in which the expected utility is the same for any choice of valuation in the interval equal to the half-length of the rectangle next to the border. However, it is a relatively small price to pay as compared to the possibility of unraveling.

[16] The actual assignment procedure was not random and depended on subjects' norm functions (see below). Subjects were told that they were "matched" with another subject in the session without using the word "random."

[17] In order not to confuse the dictator from the hypothetical situation presented in the CNE task with the real dictators that recipients were matched with, in the instructions real dictators were called "Participant A" and recipients "Participant B" as compared to Individuals C and D in the hypothetical situation.

[18] Notice that this value is different from the theoretical spread defined in Section 2, which was the difference between the highest and the lowest values of the norm function. We chose to match subjects according to this other measure of spread since ex ante we believed that most subjects will indicate 50/50 split as the most appropriate. This would have made the two measures the same. In reality, a significant proportion of subjects indicated other

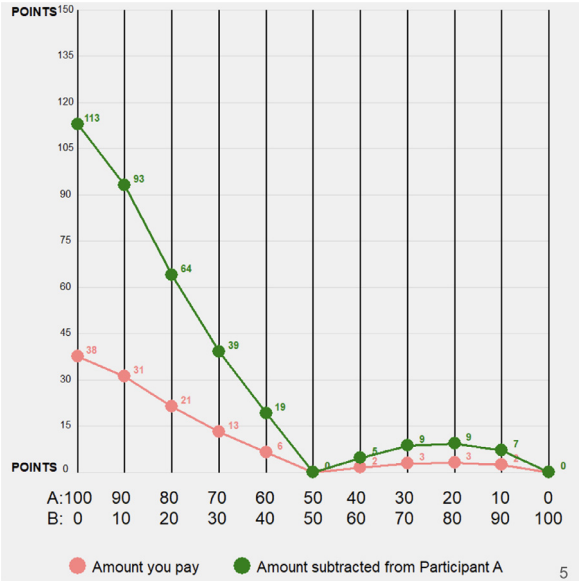

Fig. 2. The interface used by recipients to choose their punishment strategy.

To test the implication of the theory in Section 2, our goal was to maximize the number of pairs with sufficiently different spreads. Thus, the software ranked all subjects in a session according to their measures of spread. The ranking was then divided halfway and alternating roles of dictators and recipients were assigned to the subjects in the top half of the ranking. Notice that these subjects were *not* matched with each other. Rather, each of them was matched with an equidistant subject in the bottom half of the ranking. Figure 10 in Appendix C demonstrates graphically how the matching was done.[19]

Figure 3 shows the screen on which the norm functions in the Main treatment were shown to each recipient before they could make their punishment choices. Here recipients could see their own norm function and the norm function of the matched dictator. In the Control treatment recipients saw the same screen only without the norm function of the dictator. Dictators were not presented with this screen at all and thus could not see the norm function of the matched recipient. Dictators also did not possess an *explicit* knowledge that their norm function is shown to the recipients. To avoid deception, in the instructions for the CNE task all subjects were informed that the information about their choices in the CNE task may or may not be used in the other parts of the experiment. Finally, since we were concerned that recipients may not be able to remember the two norm functions when making their punishment choices, we gave them a possibility to refer back to this screen by pressing a button while making their punishment decisions. The screen then reappeared next to the graph on which recipients were choosing punishments (see Fig. 2).

3.3. Second CNE task

To account for recipients with uncertain normative beliefs who were unsure about how norm function should look like in the first CNE task, we allowed all subjects to update their norm functions in the repetition of CNE task after the Dictator game. In this task all subjects saw their elicited norm function from the first CNE task and were given a chance to update their beliefs (as before by moving the black circles). The incentives were exactly the same as in the first CNE task. This task allowed us to test hypotheses related to different punishment behavior of recipients with uncertain beliefs as compared to those with certain beliefs who did not update their norm functions even after observing the beliefs of their dictators.

choices as the most appropriate (see Section 5.1). The fact that the matching was not done exactly in accordance with our definition of the spread does not influence any results in any way, because the goal of the matching was to make sure that the norm functions in each pair of subjects are different enough.

[19] Subjects were not informed about this matching procedure. Thus, they could have thought that the other norm function that they see is over-represented in the population and thus could have over-reacted as compared to the case when they knew that the frequency of this other norm function is low. This is indeed true, and—in strict accordance with our definition of moral opportunism—we expect that subjects would have not taken such norm function into account (because, for a rare norm function, uncertainty is low). Nevertheless, we did not inform subjects about the frequency of the other norm function in the population, because our purpose was not to see how people react to different frequencies of other norm functions, but rather to test the hypotheses about the *existence* of moral opportunism. The exaggerated feeling of uncertainty artificially created in our experiment helps us to test our hypotheses better than with the alternative design.

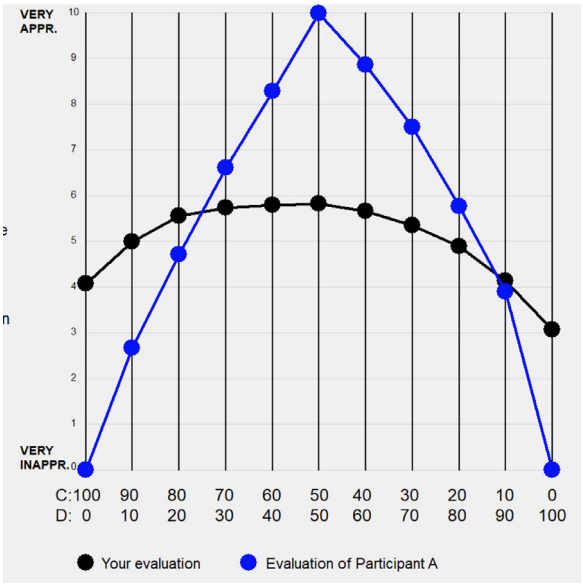

Fig. 3. The screen on which the norm functions of a matched pair were presented to the recipients in the Main treatment.

3.4. Payment

At the end of the experiment subject received a show up fee of 3 Euro and were paid for all three tasks according to their decisions and the decisions of others. No feedback about their earnings was provided to them after each part. They saw their payments only at the very end of the experiment. Subjects were paid according to the following procedure. In the first CNE task, one random action was drawn to determine the payment. Subjects received money based on the percentage of other subjects who reported a similar appropriateness level for that particular action as described above. In case 100% of others chose within subject's red rectangle, the maximum amount of 3 Euro was earned. In the Dictator game, dictators earned: 50 points + 100 points – transfer made to the recipient – points deducted by the recipient. Recipients earned: 50 points + transfer made by the dictator – points paid to punish the dictator for the action that she has actually chosen. Each point was worth 0.05 Euro. The second CNE task was incentivized in the same way as the first one.

4. Variables and hypotheses

Before we get to the results of the experiment, we describe how we analyze the data and lay down some hypotheses. The theory presented in Section 2 allows us to test to which degree subjects' punishment strategies are defined by their normative views. In the Control treatment, where only own norm function is observed, this can be done by means of the panel-data regression with 11 data points for each recipient i:

$$Punishment_{ci} = \chi + \phi y_{rci} + \varepsilon_{ci}, \tag{1}$$

Here $Punishment_{ci}$ is the amount of punishment that recipient i specified in the Dictator game for the outcome c and y_{rci} is the predicted punishment strategy from her own norm function in the first CNE task. The value of the regression coefficient ϕ defines how much the optimal punishment influences i's choices, and the regression coefficient χ defines the "excess" punishment that should be zero from the model's perspective, but might be different in reality.

The next step is to understand how to connect recipients' punishment strategies with two norm functions that they observe in the Main treatment, namely their own norm function and that of the dictator. We assume that the punishment strategy of a recipient is a convex combination of the two punishment strategies generated by the two norm functions that the recipient observes (like in Section 2.2). For all recipients in the Main treatment, we estimate the regression similar to (1):

$$Punishment_{ci} = \chi + \beta_r y_{rci} + \beta_d y_{dci} + \varepsilon_{ci}.$$

Here y_{dci} is the predicted punishment strategy computed from the dictator's norm function that is observed before (during) the Dictator game. From the estimates of β_r and β_d we can tell which norm function is used by recipients. To make these coefficients easier to interpret and consistent with (1), we reformulate the above regression model as follows:

$$Punishment_{ci} = \chi + \phi(\gamma y_{rci} + (1-\gamma) y_{dci}) + \varepsilon_{ci}, \tag{2}$$

where $\phi = \beta_r + \beta_d$ and $\gamma = \beta_r/(\beta_r + \beta_d)$. Here $\gamma \in [0, 1]$ is the weight that determines reliance on the own punishment norm function versus the dictator's. In what follows we will formulate our hypotheses and results in terms of ϕ and γ.

We propose several hypotheses related to the punishment choices in the Main treatment that we can test with our data. The simplest and most straightforward hypothesis is that the observation of the dictator's norm function does not change the recipients' punishment behavior in any way.

Hypothesis 1. *($\gamma = 1$)* The punishment strategy of the recipients in the Main and Control treatments is based only on the recipients' own norm functions, even if they know the dictators' norm functions. The behavior in both treatments should be the same.

Of course, the purpose of this study is to determine how exactly recipients react to the dictator's norm function, so we deem it likely that recipients, who are confronted with a different norm function, would take it into account. One possibility—that we cannot ex ante discount—is that they might decide to punish dictators based solely on dictators' norm functions. Such recipients might be more concerned that dictators follow the behavior they themselves believe to be appropriate, even if it is different from what recipients think. Note, however, that only recipients who receive the information about the dictators norm function can use this approach. Recipients who do not have this information can only punish based on what they believe to be the norm function.

Hypothesis 2. *($\gamma = 0$)* The punishment strategy of the recipients in the Main treatment is based exclusively on the dictators' norm functions.

Another possibility, as was suggested in Section 2.2, is that recipients, who are presented with the norm function of the dictator, punish based on some convex combination of punishment strategies coming from their own and dictator's norm functions. We do not include this as a separate hypothesis since the estimates of the regression (2) will always give us some value of γ consistent with this idea.

The scenario that our experiment was designed to test though, is when a recipient bases her punishment on one norm function out of the two that brings the lowest cost of punishment.

Hypothesis 3. ***(Moral Opportunism)*** The recipients in the Main treatment, who choose to punish at a cost, should do so based on the norm function that prescribes less punishment and allows them to keep more money.

Finally, our design allows us to test another hypothesis related to moral opportunism, namely that people engage in it due to normative uncertainty. By comparing the norm functions elicited from recipients in the first and the second CNE tasks, we can determine how much they change their beliefs after observing dictators' norm functions. If subjects are very certain about their appropriateness estimates in the first CNE task, which implies no normative uncertainty from their perspective, then we should not see any changes in the second CNE task. This also should imply that recipients should not be morally opportunistic since they know which of the two norm functions is the "correct" one – their own norm function. However, if there is a considerable normative uncertainty from the recipients' perspective, then they should adjust their beliefs in the second CNE task incorporating the information obtained from the observation of the norm functions of their dictators (Tremewan and Vostroknutov, 2020). In this case, recipients should update their norm functions in the direction of the dictators' norm functions.

Hypothesis 4. ***(Normative Uncertainty)*** Moral opportunism should be present only when there is a considerable normative uncertainty that is revealed by recipients' updating their beliefs towards the observed norm functions of the dictators.

5. Results

5.1. Norm functions

We start our analysis by looking at the norm functions elicited in the first CNE task. Since at this point in the experiment all subjects in both treatments faced the same task, we work with all data. Figure 4 displays individual norm functions. We can observe a large heterogeneity in norm-related beliefs, which demonstrates that even in the simplest situations like the Dictator game subjects have very different opinions about what others believe is appropriate. Importantly, the beliefs differ in their overall "shape": many of them have a peak at 50/50 split, but there are also other types, for example, monotonically increasing and decreasing beliefs. The objective heterogeneity in beliefs does not immediately imply that the subjects feel uncertain about them. However, given the multiplicity of different shapes, it is likely that they are. We come back to this question in Section 5.5.

To compare our belief elicitation method to that of Krupka and Weber (2013), we compute the average norm function shown in Figure 11 in Appendix D. It looks very similar to the original norm function in the Dictator game obtained by Krupka and Weber (2013) and replicated by many other studies, for example Kimbrough and Vostroknutov (2016, 2018). Notice that in these papers the norm functions were elicited with the Likert-scale version of the task. The fact that the continuous and the discrete scales generate the same average results is encouraging, it shows that the type of the scale is not particularly important for the aggregate results. However, our continuous method has an advantage when individual measures of the norm function should be used in the statistical analysis.

Result 1. There is a large heterogeneity in normative expectations. The average norm function looks very similar to previous studies, which validates the CNE task.

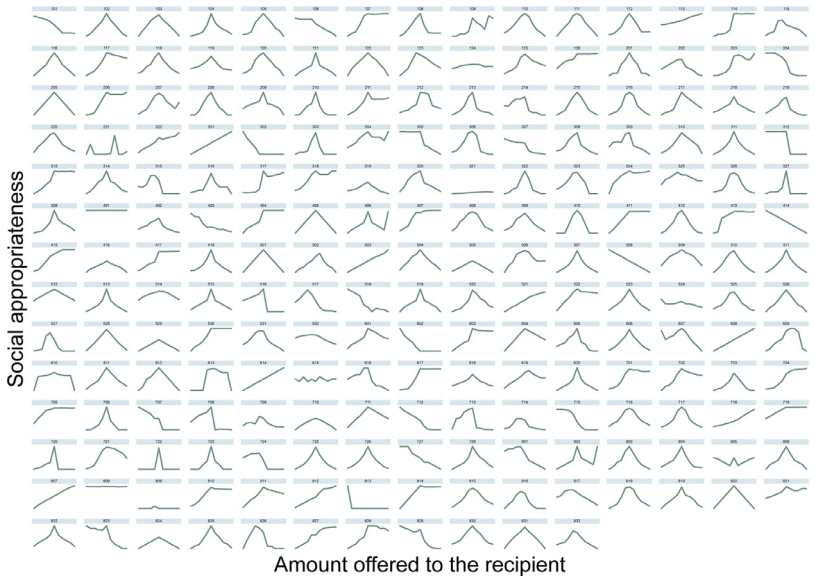

Fig. 4. Individual norm functions elicited in the first CNE task for all subjects in both treatments.

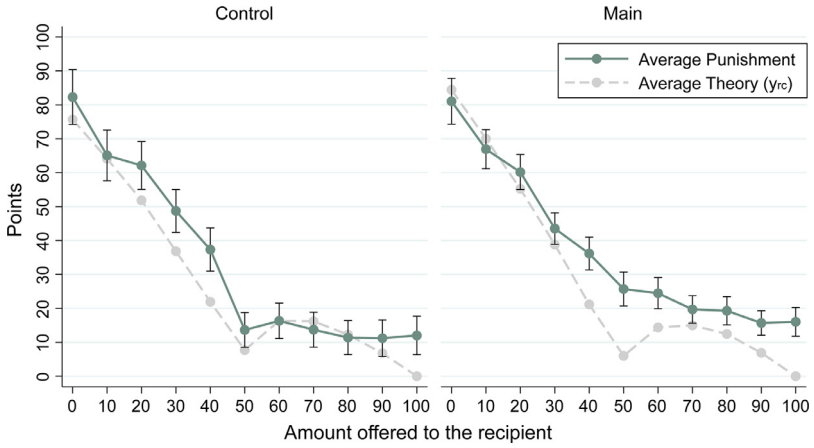

Fig. 5. Left panel. Average punishment and theoretical predictions in the Control treatment. **Right panel.** Average punishment and theoretical predictions (only own norm functions of the receivers are used) in the Main treatment. The error bars are ±1SE.

5.2. Average punishment

Average punishment across all actions meted out by recipients was 34 points in the Control treatment and 37 points in the Main treatment (the difference is not significant). Figure 5 shows the average points recipients were willing to subtract from their dictators for every action. Since we are dealing with heterogeneous norms (see Section 5.1), it is not surprising that punishment is imposed for every level of transfer. For instance, 45% of recipients were willing to punish their dictator for an equal offer. However, only 68 recipients (61%) believed that the most socially appropriate offer is half of the endowment. From this perspective, it is not strange to observe some punishment of 50/50 split. These features are also present in the Fig. 5 of Fehr and Fischbacher (2004), which are qualitatively very similar to our Control treatment.

Figure 5 also shows the average predictions of the model of normative punishment for receivers' own norms (y_{rc}) in both treatments.[20] In the Control treatment, where subjects do not have information about the norm function of the dictator, we see a decent fit of the expected theoretical predictions to the actual average punishments (the left panel of Fig. 5). We believe that this is an important result that shows that 1) a simple model of normative punishment can in principle account for punishment behavior and 2) since the model only uses norm functions extracted in the first CNE task, this result shows that subjects' norm-related beliefs dictate their punishment strategy. In the Main treatment (the right panel of Fig. 5), we

[20] For each subject and each offer we compute y_{rc} as described in Section 2.2. Figure 5 reports averages of these y_{rc} over all subjects in a treatment.

Table 1
The Baseline model. Random-effects panel regressions of punishment choices on the model's predictions for the Control and Main treatments. Errors are clustered by subject and robust. Standard errors in parentheses.

Treatment:	Control	Main
ϕ	0.676***	0.734***
	(0.113)	(0.089)
γ		0.730***
		(0.162)
const	14.930**	15.204***
	(5.176)	(4.429)
N observations	374	759
N independent	34	69

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2
Model fit comparisons of the five models. "df" stands for the degrees of freedom. Δ is the difference in BIC (AIC) from the Spread Model. The models are sorted by the value of BIC.

	df	AIC		BIC	
		value	Δ	value	Δ
Spread model	11	7331		7382	
Spread/Peak model	17	7316	-15	7394	12
Baseline model	5	7377	46	7400	18
Peak model	11	7367	36	7418	36

see that the fit of the model with own norm function only is not as good. The reason, of course, is that in this treatment recipients observe dictators' norm functions as well, which changes their punishment choices.

Result 2. The model of normative punishment seems to fit the data in the Control treatment relatively well.

5.3. Moral opportunism

In this section we look at the determinants of the punishment decisions made by the recipients and test whether our subjects are using any of the two guidelines coming from the maximization of the norm-dependent utility (Section 2). The guidelines are that norm-following agents should choose to punish according to the norm function with the smallest spread or the one that has the peak closest to the selfish choice of the dictator ($c = 100$).

We start with testing the general models presented in Section 4. Table 1 shows the estimates of ϕ and γ in the two treatments obtained from fitting regression models (1) and (2). As described in Section 4, in the Main treatment we run the standard linear regression (see Table 6 in Appendix E.1) and then compute non-linear transformations of the coefficients to obtain γ (and ϕ with linear transformations). We call this model the Baseline model (for the comparisons with other models).

The regression for the Control treatment shows that ϕ is significant, which means that the model of punishment has some predictive power (overall R^2 is 19%). It is important to notice that the constant term is high and significant as well. This demonstrates that the recipients have a tendency to subtract money from the dictators "uniformly" for all actions. This does not fit the model of punishment, but demonstrates that subjects subtract *more* than the model predicts. We will come back to this finding in Section 5.5. The regression for the Main treatment (overall R^2 is 18%) has a similar significant estimate of ϕ and a significant estimate of γ that tells us that in this regression specification subjects seem to rely more on their own norm function (with the weight around 0.7) than on the dictator's norm function. These findings reject Hypotheses 1 and 2 presented in Section 4.

Next, we consider two models related to the two guidelines that test moral opportunism. In the Spread model we introduce categorical variables that partition observations in the Main treatment into three groups: recipients whose norm function had the same/larger/smaller spread than the dictator's.[21] In the Peak model we partition observations depending on whether recipient's most appropriate outcome (the peak) is at the same/higher/lower level of c than that of the dictator. We also consider Spread/Peak model with both sets of categorical variables fully interacted (9 groups).

Table 2 shows the BIC and AIC for these models. The Spread model performs the best in terms of BIC. Interestingly, the Spread/Peak model—which of course has lower AIC because the additional degrees of freedom are unaccounted for—does *not* perform better than the Spread model in terms of BIC. This suggests that our subjects are sensitive to spreads of the norm functions, but not the peaks. This is also supported by the finding that the Peak model is the worst in terms of both AIC and BIC.

[21] As in Section 2, the spread of a norm function is defined here as the difference between its maximum and minimum values.

Table 3
The Spread model. Transformed coefficients from the random-effects panel regression in Table 7 in Appendix E.1. Errors are clustered by subject and robust. Standard errors in parentheses.

Spread:	$D = R$	$R < D$	$D < R$
ϕ	1.105***	0.693***	0.657***
	(0.151)	(0.150)	(0.117)
γ	0.661**	1.110***	0.271
	(0.253)	(0.283)	(0.151)
const	0.937	18.236**	18.825*
	(6.384)	(6.962)	(7.669)
% observations	20%	45%	35%

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4
Model fit comparisons of the eight models. "df" stands for the degrees of freedom. Δ is the difference in BIC (AIC) from the Spread Model. The models are sorted by the value of BIC.

	df	AIC		BIC	
		value	Δ	value	Δ
Spread model	11	7331		7382	
Min/Max model with spreads	11	7342	11	7393	11
Baseline model	5	7377	46	7400	18
Min/Max model	5	7384	53	7407	25
Shape model	8	7378	47	7415	33
Peak model	11	7367	36	7418	36
Resentment model	5	7689	358	7712	330
Resentment model with spreads	11	7661	330	7712	330

In the rest of this section, we look more closely at the winning Spread model. We hypothesize, accordingly, that the reaction to the dictator's norm function depends on its spread relatively to the spread of the recipient's norm function. Specifically, the arguments in Section 2 suggest that recipients might use only one norm function with the smaller spread, or that γ is 0 or 1 depending on which spread is larger. To test this idea, we analyze the choices of the recipients in the Main treatment. We split them into three groups: 1) with equal spread of the two norm functions (group $D = R$, 20% of recipients); 2) with the recipient's norm function having smaller spread (group $R < D$, 45% of recipients); and 3) with the dictator's norm function having smaller spread (group $D < R$, 35% of recipients).

To estimate ϕ and γ in the three groups we define the categorical variables, which represent them, and run the regression shown in Table 7 in Appendix E.1. Table 3 shows the estimates of ϕ and γ for the three groups obtained by transforming the coefficients in that regression. First, notice that the estimates of ϕ are very significant in all three groups. This suggests as before that norm functions are good predictors of punishment strategies (overall R^2 is 22%). In addition, we see that when subjects observe two norm functions with equal spread (group $D = R$) they put the weight 0.661 on their own norm function (γ coefficient). This means that recipients mix the two punishment strategies in approximately equal proportion. However, most importantly, we can see that in the group $R < D$, where the spreads of the recipients' norm functions are smaller than the dictators', recipients choose to follow their own punishment strategy (γ is close to 1 and significant).[22] In the group $D < R$, where dictators have a smaller spread of the norm function, recipients put little weight on their own punishment strategy ($\gamma = 0.271$, not significant) and instead follow the punishment strategy emerging from the dictator's norm function. To test if the difference between these γ coefficients is significant we compute their difference $\gamma_{R<D} - \gamma_{D<R}$ and find that it is equal to 0.839 with $p = 0.009$. This demonstrates that recipients in the Main treatment follow the norm function with the smaller spread, which supports Hypothesis 3.

Result 3. Recipients in the Main treatment, who observe two norm functions, follow the punishment strategy resulting from the norm function with the smallest spread. When the spreads are the same, they mix the punishments strategies, defined by the two norm functions, in roughly equal proportion.

5.4. Comparison with other models

A natural question at this point is whether there are other characteristics of the norm functions that might be influencing the way they are aggregated for the punishment decisions in the Main treatment. We have analyzed several additional models summarized in Table 4.

In the Shape model, we classify all norm functions into four categories: peaky shape (weakly increasing on the left of 50/50 split and weakly decreasing on the right), weakly increasing, weakly decreasing, and miscellaneous. We categorize the

[22] The coefficient γ is in interval [0,1] as long as coefficients β_r and β_d are positive (as suggested by the theory). One of the estimates of γ is bigger than 1 due to the presence of some negative insignificant coefficients in the regression in Table 7 in Appendix E.1.

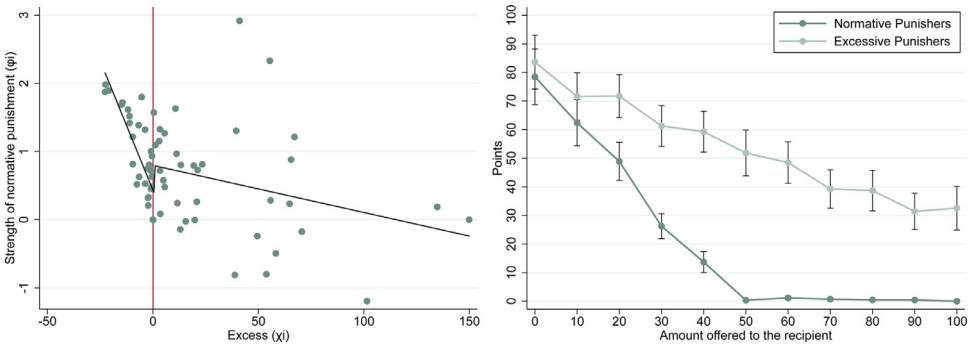

Fig. 6. Left panel. Individual coefficients χ_i and ϕ_i for the recipients in the Main treatment. The black line represents the fitted OLS regression with two regimes: χ_i lower and higher than the median (the red line). 69 observations. **Right panel.** Average punishment strategies for the recipients below the median χ_i (normative punishers) and above the median (excessive punishers). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

recipients in the Main treatment according to whether they observe the two norm functions of the same shape or different shapes (two categories). This defines the categorical variable that are used in the regression in place of the one that defined spreads before. In terms of AIC and BIC, the Shape model is not doing very well in comparison to the Spread model.

The Resentment model uses different independent variables than the Spread model. Here, instead of y_{rc} and y_{dc} we use resentments r_c computed for the recipients' and dictators' norm functions (See Section 2.2). Thus, in the Resentment model we assume that subjects react only to the size of norm violation and disregard the amount of money that dictator has (which is used to transform resentments into y_{rc} and y_{dc}). We consider two specifications of the Resentment model: with partitioning recipients into the Spread model categories (with spreads in Table 4) and without such partitioning. Both specifications fare much worse than the Spread model in terms of BIC and AIC.

The last specification is the Min/Max model where the independent variables are $\min\{y_{rci}, y_{dci}\}$ and $\max\{y_{rci}, y_{dci}\}$ instead of y_{rc} and y_{dc}. The idea here is that subjects, who face two norm functions and strive to minimize costs, may choose to punish each action following the norm function that prescribes the cheapest punishment *for that action*. This is another way that moral opportunism can be defined. Table 4 shows that the fit of the two specifications of this model, with and without partitioning recipients into spread categories, are worse than the Spread model.

Result 4. The additional models that we considered do not fit our data better than the Spread model.

5.5. Excessive punishment

An attentive reader could have noticed that the constants in regressions in Tables 1 and 3 are very high and significant, which means that some subjects tend to punish their dictators equally for any outcome they choose. This is not consistent with our model of normative punishment. So, to understand where this "excessive punishment" comes from, we perform a more detailed analysis of individual punishment choices. To do this, we estimate individual regressions for each recipient in the Main treatment using the 11 data points that define their punishment strategies:

$$Punishment_{ci} = \chi_i + \phi_i(\gamma_i y_{rci} + (1-\gamma_i)y_{dci}) + \varepsilon_{ci}. \tag{3}$$

This is the same regression as in (2) only run for individual subjects.[23] Here χ_i is the subject-dependent constant that estimates how much excessive punishment the recipient is applying; ϕ_i represents the individual measure of adherence to normative punishment; and γ_i controls for the (potentially morally opportunistic) choice of a punishment strategy. If our subjects try to spend as little money as possible without being perceived as norm violators, then they should show little excessive punishment (small χ_i) and high propensity to stick to normative punishment, be it opportunistic or not (high ϕ_i).

The left panel of Fig. 6 shows the coefficients χ_i and ϕ_i on a scatter plot. There is a significant negative relationship between them (Spearman's $\rho = -0.46$, $p = 0.0001$). This already demonstrates that higher ϕ_i is associated with lower χ_i, which is in line with the idea that subjects who choose to follow normative punishment, as defined in Section 2, do not use excessive punishment. We fit an OLS regression, shown in Fig. 6 as a black line, that estimates the relationship between χ_i and ϕ_i separately for the subjects with above- and below-median χ_i (the median is shown as the red line). Both coefficients on χ_i are significant ($p < 0.016$). We can also see that the recipients below median have high average ϕ_i and negative χ_i (0.86 and −5.63 respectively), whereas recipients above median have lower ϕ_i and high χ_i (0.54 and 36.94). Thus, this analysis suggests that our subjects are located on a continuum between exclusively following normative punishment without excess (normative punishers) and not following normative punishment with large excess (excessive punishers).

[23] Actually, we run one regression with dummies for individual subjects interacted with each independent variable.

Even though we define normative and excessive punishers as the two extremes of a continuum, we will abuse these definitions for expositional purposes. We will define two *groups* of subjects called "normative punishers" and "excessive punishers" divided by the median of χ_i. Most of our subjects do not fit either of the strict definitions of normative or excessive punishers, however we still classify them as such in order to emphasize the differences between these two groups.[24] The average punishment strategies of the recipients in these two groups are shown on the right panel of Fig. 6. It is rather clear that normative punishers follow the strategy that resembles our theoretical predictions (see Fig. 5), whereas excessive punishers are influenced to some extent by the normative punishment strategy—the curve is downwards sloping—but mostly punish all outcomes at the same high level. As expected from the figure, the average punishment among excessive punishers is 53.62, which is much higher than 21.15 for normative punishers (ranksum test, $p < 0.0001$, 69 observations).

Result 5. Recipients in the Main treatment can be divided into normative and excessive punishers. The former follow the model of normative punishment without excess, whereas the latter do not follow the model of normative punishment, but instead apply large excessive punishment.

From the analysis above it is not clear whether or not the presence of normative and excessive punishers somehow depends on the two norm functions observed in the Main treatment. We conduct the same analysis as above for the Control treatment and the pooled data from both treatments. The results are presented in Appendix E.2.[25] Both analyses show qualitatively same results as above with all coefficients and tests close in values and significant, which suggests that being a normative or excessive punisher is independent of observing multiple norm functions and can be seen as some kind of an individual characteristic.

Result 6. Recipients in the Control treatment also behave like normative and excessive punishers. So this trait does not depend on the presence of two norm functions in the Main treatment.

5.6. Normative uncertainty

In Section 5.3, we have established that recipients in the Main treatment behave in a way consistent with the idea of moral opportunism. In this section, we connect moral opportunism with perceived normative uncertainty and test our Hypothesis 4. To do that we need to determine how much recipients in the Main treatment adjust their norm functions in the second CNE task. For each recipient i we compute an individual measure of average adjustment as

$$\pi_i = \frac{1}{|C|}\sum_{c\in C}\frac{a_{ci}-b_{ci}}{d_{ci}-b_{ci}},$$

where b_{ci} is the appropriateness level chosen by recipient i for outcome c in the first CNE task (before the Dictator game); a_{ci} is the appropriateness level chosen after the Dictator game; and d_{ci} is the appropriateness level chosen for outcome c by i's dictator. The idea here is that for each outcome c the fraction above is equal to 0 if there is no adjustment ($a_{ci} = b_{ci}$) and to 1 if the normative valence is adjusted exactly to the dictator's level. So, $\pi_i = 0$ when there is no adjustment in any outcome, and $\pi_i = 1$ when the whole norm function is adjusted to the dictator's. The value of π_i is positive whenever the adjustment happens in the direction of the dictator's norm function and is negative when the adjustment goes in the opposite direction. It can also be thought of as the average measure of how much the appropriateness levels are adjusted towards the dictator's norm function in percentage terms.

The left and right panels of Fig. 7 show the histograms of π_i in the Control and Main treatments. In the Control treatment we make the exact same calculation of π_i with only difference that the recipients do not observe their dictators' norm functions. This serves as a benchmark to which we can compare the adjustments in the Main treatment. The left panel of Fig. 7 illustrates that in the Control treatment around 65% of recipients do not adjust their norm functions, while the rest do it in a random way: the average π_i in the Control treatment is -0.00725, not significantly different from zero. In the Main treatment the picture is very different. We have 40% of recipients who do not adjust their norm functions. However, the majority of those who do change them in the direction of their dictator's norm functions, which is attested by the fact that most adjusting recipients have positive π_i between 0 and 1 (see Fig. 7). The average π_i in the Main treatment is 0.22, which is significantly different from zero (t-test, $p = 0.045$). The distributions of π_i in the Main and Control treatments are also significantly different (ranksum test, $p = 0.0054$).[26]

[24] We chose to split the sample by the median of χ_i because this median is essentially zero, which allows us to cleanly separate subjects with positive χ_i and others. We could have divided the sample by the median of ϕ_i, or could have used some other method of grouping subjects. However, all we want with this analysis is to demonstrate that the subjects with high ϕ_i and subjects with high χ_i exhibit different punishment strategies. For this purpose, we believe, the division by χ_i does the job well enough.

[25] Notice that we can estimate individual coefficients χ_i and ϕ_i in the Control treatment in the same way we did it in the Main treatment taking into account only own norm function.

[26] The average π_i of the recipients in the Main treatment who do adjust (excluding 40% of those who do not change their norm function) is 0.34 (different from zero, t-test, $p = 0.045$) and different from the distribution of π_i for adjusting recipients in the Control treatment (ranksum test, $p = 0.019$) where the average is -0.016 (not significantly different from zero).

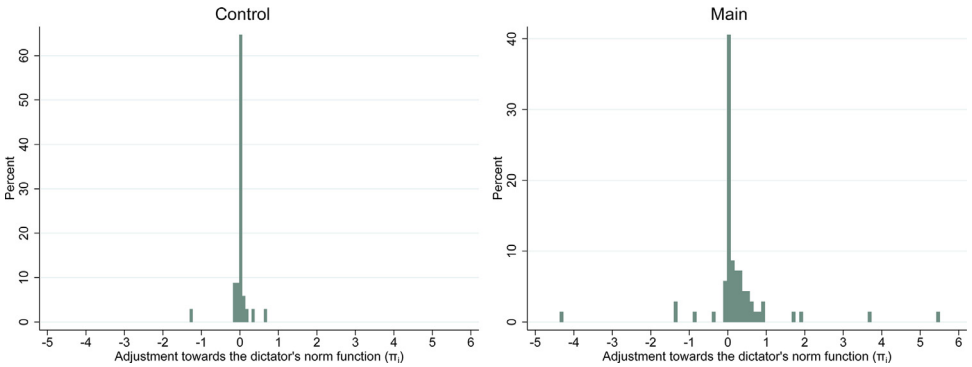

Fig. 7. Left panel. Histogram of π_i in the Control treatment. **Right panel.** Histogram of π_i in the Main treatment.

Table 5
Robust OLS regressions of $\bar{\gamma}_i$ on π_i in the Main treatment. Observations with $R = D$ are dropped.

Punishers:	All	Normative	Excessive
π_i	0.504^{-}	0.123*	0.470
	(0.283)	(0.056)	(0.556)
const	-1.134	0.320	-2.412
	(1.428)	(0.226)	(2.763)
N independent	56	28	28

$^{-}$ $p < 0.1$, * $p < 0.05$.

Result 7. 60% of the recipients in the Main treatment adjust their norm functions and do so in the direction of the dictators' norm functions on average. This suggests the presence of considerable normative uncertainty, which is consistent with the hypothesized origins of moral opportunism.

This result shows that, on average, recipients in the Main treatment adjust their norm functions in the direction of the dictator's norm function. Our task now is to connect the rates of individual adjustments π_i with some measure of moral opportunism to check if subjects with higher rate of adjustment π_i are also those who are more morally opportunistic. To do that we make use of individual coefficients γ_i computed in Section 5.5. These coefficients specify to which degree subjects use their own norm function in the presence of another one coming from the dictator. Notice that γ_i is *not* a measure of moral opportunism, because if the dictator's norm function is cheaper then morally opportunistic subjects will follow it and their γ_i will be close to zero. To obtain a meaningful measure of moral opportunism we construct a new variable $\bar{\gamma}_i$ that is equal to γ_i for subjects whose norm function has a smaller spread than the dictator's and $1 - \gamma_i$ for subjects whose norm function has a larger spread than the dictator's. These correspond exactly to the conditions $R < D$ and $D < R$ in Section 5.3. The intuition behind this definition of $\bar{\gamma}_i$ is the following. Suppose we are in the environment $R < D$ where recipient's norm function should be chosen by a morally opportunistic agent. Then, the size of γ_i determines the extent of moral opportunism: the closer γ_i is to 1, the more morally opportunistic the agent is. In the condition $D < R$ the situation is the opposite: the closer $1 - \gamma_i$ is to 1, the more morally opportunistic the agent is. For the analysis below we drop the subjects from the condition $R = D$ as it is unclear which norm function they use.

To provide evidence for Hypothesis 4, we regress $\bar{\gamma}_i$ on π_i in Table 5. In the leftmost regression with all punishers (recipients), the resulting coefficient on π_i is weakly significant with $p = 0.081$. This suggests that the higher the adjustment of a subject, the higher weight she puts on the norm function that is cheaper. To reformulate, the more normatively uncertain the subject is, the more morally opportunistic her behavior becomes. This result is not particularly strong (10% significance), which might be due to the presence of excessive punishers who do not follow the normative punishment strategy. Thus, we divide the sample into normative and excessive punishers (the middle and right columns in Table 5). Here we see that for normative punishers the coefficients on π_i is now significant with $p = 0.036$ and for excessive punishers it is not. Moreover, the range of predicted values of $\bar{\gamma}_i$ is within [0,1] for normative punishers (as it should be) and is negative for excessive punishers (which does not make sense). This provides additional evidence that excessive punishers do not follow normative punishment whereas normative punishers do. The result for normative punishers provides tangible support for Hypothesis 4.

Result 8. Normative punishers with higher rate of adjustment π_i tend to put more weight on the cheaper norm function (higher $\bar{\gamma}_i$). In other words, more uncertain normative punishers are more morally opportunistic. The same is not true for excessive punishers.

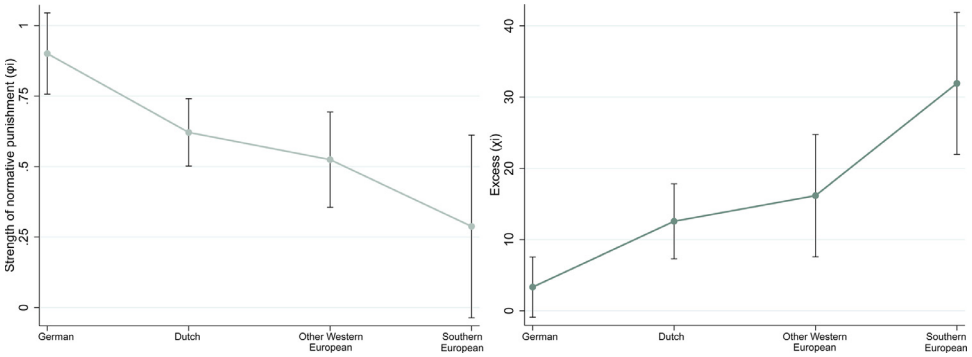

Fig. 8. Left panel. The average ϕ_i across four nationality categories. **Right panel.** The average χ_i across four nationality categories. The error bars are ±1SE.

6. Discussion

In this final section we present the unhypothesized but rather intriguing relationship between being a normative/excessive punisher and nationality. In our sample, there were 40 different nationalities thanks to an international mixture of students in Maastricht University, where the data were collected. All demographic details can be found in Tables 9, 10, and 11 in Appendix F. We analyze the distributions of ϕ_i and χ_i by nationality. Given that our experiment was not explicitly designed to test for specific demographic effects, for this analysis we pool the data from both treatments. We believe that this is legitimate since the analysis in Appendix E.2 shows that the relationship between ϕ_i and χ_i does not depend on the treatment.

Figure 8 shows the averages of ϕ_i and χ_i for four subsets of recipients from both treatments: German (26 subjects); Dutch (20 subjects); Other Wester European (18 subjects); Southern European (19 subjects).[27] We see that subjects from Germany can be squarely classified as those who follow normative punishment and do not punish in excess. They have the highest average ϕ_i and the lowest excessive punishment (not significantly different from zero). Subjects from Southern Europe are on the other extreme: they do not follow normative punishment (insignificant average ϕ_i) but have the highest excessive punishment. The Dutch and other Western European subjects lie somewhere in between. The regressions of ϕ_i and χ_i on nationality dummies, gender, and age presented in Table 8, Appendix E.3 support these results. For ϕ_i, we have a 10% significant effect of Southern Europeans (baseline is Germans). We also see a 5% significant effect of Southern Europeans for χ_i. Gender and age are not significant in both specifications.

These results suggest that Souther Europeans have very different approach to punishment then other Western Europeans. Specifically, they seem to punish their dictators excessively regardless of their own as well as others' normative beliefs. This becomes even more evident when we look at the average norm functions of these subjects presented in Figure 12 in Appendix D. The average norm functions across the four nationality categories are almost identical, which makes it clear that the differences in punishment decisions are not directly associated with the perceived social norms, but are rather determined by some other factors that are most likely cultural. Interestingly, Panizza et al. (2020) also find similar excessive punishment in an experiment run in Italy. All this evidence suggests that we have stumbled upon an important cultural difference in attitudes towards punishment that is very large, detectable in small samples, and seems to appear in other studies as well.

The behavior of excessive punishers (or Southern Europeans) does not fit our model of normative punishment. Indeed, the model stipulates that punishment happens after a norm was violated, and we do find that many subjects follow this general guideline when choosing punishment strategies. Excessive punishers do not follow normative punishment, but instead punish the dictator uniformly for any action. How should we think theoretically about this deviation from the model? One possibility is that excessive punishers punish others because they have different *identity* (Akerlof and Kranton, 2000; Kimbrough and Vostroknutov, 2022). Indeed, examples of violence towards people of different identity abound, and it is clear that this violence is not triggered by some actions of the victims, but rather by who they are. To test this idea, experiments can be conducted that, for example, modulate the perceived identity of the dictator (via, say, minimal group paradigm, Tajfel et al., 1971; Chen and Li, 2009). Another possibility is that *emotions* play a role in choosing punishment strategies. It is not inconceivable that subjects who are upset tend to punish others uniformly, regardless of their actions (for similar ideas, see e.g. Klimecki et al., 2016). It can be that subjects get upset because they expected to be dictators but were assigned the roles of recipients.[28] Finally, the difference between normative and excessive punishers can be related to cultural *tightness* or *looseness* (Gelfand, 2019). Indeed, according to this view, people in tight cultures (e.g., Germany) follow norms in a more

[27] Other Western European category includes UK, France, Austria, Switzerland, Belgium, Ireland, and Luxembourg. Southern European category includes Greece, Italy, Spain, and Portugal.

[28] It is not very clear though why Southern Europeans should be more upset than Northern Europeans (in Maastricht).

precise and thought-through fashion than people in loose cultures (e.g., Southern Europe), see also Dimant et al. (2022). In terms of punishment, this results in carefully meted normative punishment among the participants from tight cultures, and more arbitrary punishment in loose cultures, which is similar to what we observe.

7. Conclusion

In this study we analyze the phenomenon termed in the literature *moral opportunism*. When people face normative uncertainty, or there are several possible norms in a given situation, their behavior becomes opportunistic in the sense that they choose to follow the norm that minimizes their costs. We hypothesize that this behavior is a consequence of expected norm-dependent utility maximization. We use a novel theoretical framework that allows to model moral opportunism, test it in an experiment, and show that the behavior of subjects is indeed in line with the theoretical predictions.

We contribute to the existing literature on moral opportunism by looking at the environment where subjects have different normative beliefs that are directly observed by them and the experimenters, which allows for direct test of hypotheses related to moral opportunism. In the previous literature, the exact nature of normative disagreement was not directly observed, but only hypothesized. This is also one of the first studies that demonstrates how new theoretical framework for studying social norms (Kimbrough and Vostroknutov, 2020; Tremewan and Vostroknutov, 2020) can be used in order to predict and experimentally detect novel behavioral phenomena. We hope that our findings and theoretical methodology can be useful for other researchers who study normative decision making.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2022.03.020.

References

Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. Q. J. Econ. 115 (3), 715–753.
van Baar, J.M., Chang, L.J., Sanfey, A.G., 2019. The computational and neural substrates of moral strategies in social decision-making. Nat. Commun. 10 (1), 1–14.
Bašić, Z., Verrina, E., 2021. Personal Norms- and Not Only Social Norms-Shape Economic Behavior. Discussion Paper 2020/25. MPI Collective Goods.
Bicchieri, C., 2008. The fragility of fairness: an experimental investigation on the conditional status of pro-social norms. Philos. Issues 18, 229–248.
Bicchieri, C., Chavez, A., 2010. Behaving as expected: public information and fairness norms. J. Behav. Decis. Mak. 23 (2), 161–178.
Chen, Y., Li, S.X., 2009. Group identity and social preferences. Am. Econ. Rev. 99 (1), 431–457.
Dimant, E., Gelfand, M.J., Hochleitner, A., Sonderegger, S., 2022. Strategic Behavior with Tight, Loose, and Polarized Norms. SSRN 4004123.
Eftedal, N.H., Kleppesto, T.H., Czajkowski, N.O., Sheehy-Skeffington, J., Roysamb, E., Vassend, O., Ystrom, E., Thomsen, L., 2020. Disentangling principled and opportunistic motives for reacting to injustice: a genetically-informed exploration of justice sensitivity. bioRxiv.
Elster, J., 2009. Norms. The Oxford Handbook of Analytical Sociology.
Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73 (6), 2017–2030.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25 (2), 63–87.
Fehr, E., Fischbacher, U., Gächter, S., 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Hum. Nat. 13 (1), 1–25.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137.
Fershtman, C., Gneezy, U., List, J.A., 2012. Equity aversion: social norms and the desire to be ahead. Am. Econ. J. 4 (4), 131–144.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Frey, B.S., Bohnet, I., 1995. Institutions affect fairness: experimental investigations. J. Inst. Theor. Econ.(JITE)/Zeitschrift für die gesamte Staatswissenschaft 286–303.
Gelfand, M., 2019. Rule Makers, Rule Breakers: Tight and Loose Cultures and the Secret Signals That Direct our Lives. Scribner.
Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.
Hoffman, E., Spitzer, M.L., 1985. Entitlements, rights, and fairness: an experimental examination of subjects' concepts of distributive justice. J. Legal Stud. 14, 259–298.
Kagel, J.H., Kim, C., Moser, D., 1996. Fairness in ultimatum games with asymmetric information and asymmetric payoffs. Games Econ. Behav. 13 (1), 100–110.
Kassas, B., Palma, M.A., 2019. Self-serving biases in social norm compliance. J. Econ. Behav. Organ. 159, 388–408.
Kessler, J.B., Leider, S., 2012. Norms and contracting. Manage. Sci. 58 (1), 62–77.
Kimbrough, E.O., Vostroknutov, A., 2016. Norms make preferences social. J. Eur. Econ. Assoc. 14 (3), 608–638.
Kimbrough, E.O., Vostroknutov, A., 2018. A portable method of eliciting respect for social norms. Econ. Lett. 168, 147–150.
Kimbrough, E.O., Vostroknutov, A., 2020. A Theory of Injunctive Norms. Mimeo, Chapman University and Maastricht University.
Kimbrough, E.O., Vostroknutov, A., 2022. Affective Decision-Making and Moral Sentiments. Mimeo, Chapman University and Maastricht University.
Klimecki, O.M., Vuilleumier, P., Sander, D., 2016. The impact of emotions and empathy-related traits on punishment behavior: introduction and validation of the inequality game. PLoS ONE 11 (3), e0151028.
Krupka, E.L., Weber, R.A., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11 (3), 495–524.
Mackie, J.L., 1982. Morality and the retributive emotions. Crim. Justice Ethics 1 (1), 3–10.
Masclet, D., Villeval, M.-C., 2008. Punishment, inequality, and welfare: a public good experiment. Soc. Choice Welfare 31 (3), 475–502.
Nagel, R., 1995. Unraveling in guessing games: an experimental study. Am. Econ. Rev. 85 (5), 1313–1326.
Panizza, F., Vostroknutov, A., Coricelli, G., 2020. How Conformity Can Lead to Extreme Social Behaviour. Mimeo, University of Trento, Maastricht University, University of Southern California.
Straub, P.G., Murnighan, J.K., 1995. An experimental investigation of ultimatum games: information, fairness, expectations, and lowest acceptable offers. J. Econ. Behav. Organ. 27 (3), 345–364.
Tajfel, H., Billig, M.G., Bundy, R.P., Flament, C., 1971. Social categorization and intergroup behaviour. Eur. J. Soc. Psychol. 1 (2), 149–178.
Tremewan, J., Vostroknutov, A., 2020. A Research Agenda in Experimental Economics. Edward Elgar Publishers. An Informational Framework for Studying Social Norms

Moral Opportunism as a Consequence of Decision Making under Uncertainty

Nitzan Merguei, Martin Strobel, Alexander Vostroknutov

Appendix (for online publication)

A Definition of Moral Opportunism

In this appendix we discuss in more detail the definition of moral opportunism given in Section 2.1. As was mentioned in the Introduction, various studies of moral opportunism define it differently and often only for specific contexts. We try to bring all these definitions to some common denominator that can be mathematically formalized. We start with observing that moral opportunism is typically defined as a *characteristic of behavior*. Some choices made by an agent can be deemed morally opportunistic or not. Therefore, we also will stick to this general idea and define moral opportunism as a *moral judgement* of an action by others.

The assumption that moral opportunism is a *moral* judgement immediately leads to the following observation. At least in the framework of KV, a moral judgement can only be made *in comparison* with some ideal behavior prescribed by some normative guideline. Thus, moral opportunism is in principle a relative concept that only makes sense from the perspective of some moral system. This implies that instead of talking about moral opportunism as such, we need to talk about *moral opportunism from the perspective of some normative guideline*.

To make this argument more concrete, we consider *allocation games* (Tremewan and Vostroknutov, 2020) with some pre-specified norm function. An allocation game is any game where only one player acts in a single decision node with actions leading to consequences in some set C. Each consequence $c \in C$ maps to an allocation of resources among N players through (vector) utility function $u : C \to \mathbb{R}^N$ with components u_i representing consumption utilities for all $i \in N$. For example, the Dictator game is an allocation game that satisfies this definition. As a prerequisite, we also consider an arbitrary norm function $\eta : C \to [-1, 1]$ that defines social appropriateness of consequences in C.[1] Let us talk about a tuple $\langle C, u, \eta\rangle$ as an allocation game with a norm function.

Given these definitions, let us now think about what norm-following agents can choose in $\langle C, u, \eta\rangle$ and whether their choices can be called morally opportunistic. As is common in social norms literature, suppose that an agent i is the decision maker in $\langle C, u, \eta\rangle$ who maximizes norm-dependent utility

$$U_i(c) = u_i(c) + \phi_i\eta(c).$$

Depending on ϕ_i, the optimal choice of i will be some $c_i^*(\phi_i, \eta)$ that maximizes U_i. This means that agents with different ϕ_i's will choose differently in $\langle C, u, \eta\rangle$. However, none of these different choices can be called morally opportunistic, which according to other researchers should involve competing moral principles. Thus, when agents with various ϕ_i choose in $\langle C, u, \eta\rangle$ with a single fixed norm function η, their behavior is characterized by different degrees of following η rather than by moral opportunism. This is an important observation that will be used later in our definition of moral opportunism.

This argument leads to the following conclusion. Suppose that we have two agents with one choosing in $\langle C, u, \eta_1\rangle$ and another in $\langle C, u, \eta_2\rangle$. The differences in their environments come only in the form of different social appropriateness of the actions defined by η_1 and η_2 (the rest is the same). Both agents think that there is only one norm function defining social appropriateness: η_1 and η_2 respectively. In this situation, the *behavior of neither agent is morally opportunistic*, because both believe that only one norm function defines social appropriateness of the actions (and moral opportunism involves at least two possible norm function for each agent).

In order to define moral opportunism, we need to consider a situation where agent i believes that there are at least two norm functions η_1 and η_2 that possibly describe social appropriateness in an allocation game. Let us denote such environment by $\langle C, u, \eta_1, \eta_2\rangle$. What choice should i make in $\langle C, u, \eta_1, \eta_2\rangle$ for it to be considered

[1] In general, KV define social appropriateness over *consequences* of a game. In case of allocation games the set of consequences is in one-to-one relationship with the actions of the decision maker. Thus, we can also talk about social appropriateness of actions in allocation games.

morally opportunistic? As we mentioned above, morally opportunistic choice can only be detected in comparison to some moral guideline. Therefore, we need to choose a moral position from which to judge i's actions. Suppose that a group of people thinks that only η_1 applies to the allocation game defined by C and u and that agent i's propensity to follow norms is ϕ_i. Then, this group expects that i will choose $c_i^*(\phi_i, \eta_1)$ in the environment $\langle C, u, \eta_1 \rangle$, which to them is a benchmark against which they judge i's behavior. Now, suppose that the group also knows that agent i is unsure about which norm function, η_1 or η_2 applies (but is sure about C and u). So, the only difference between $\langle C, u, \eta_1 \rangle$ and $\langle C, u, \eta_1, \eta_2 \rangle$ is the knowledge of i that η_2 is possible. In this case, we argue that *any choice* by i that is not equal to $c_i^*(\phi_i, \eta_1)$ should be considered morally opportunistic. The reason is the following. If agent i changes her behavior from $c_i^*(\phi_i, \eta_1)$ when she gets aware of the possibility of η_2, this means that η_2 is somehow important for i, for otherwise she would not change her choice. Given that i cares about η_2 in some way, this is morally opportunistic from the perspective of η_1, because, from this perspective, η_1 is the only correct moral view of the situation and if someone does not agree with it, then this person is not fully respecting the morality of η_1. Thus, any choice of i that disagrees with $c_i^*(\phi_i, \eta_1)$ is morally opportunistic from the perspective of η_1. We summarize this with the general definition of moral opportunism.

Definition. *Suppose agent i is choosing in some allocation game $\langle C, u, (\eta_k)_{k=1..K} \rangle$ where she believes there are K possible norm functions η_k for $k = 1..K$. Then, for each given k, i's choice $c \in C$ is **η_k-opportunistic** if $c \neq c_i^*(\phi_i, \eta_k)$, which is the optimal choice of i in the environment $\langle C, u, \eta_k \rangle$.*

With this definition at hand we can now think about the conditions under which agents who maximize norm-dependent utility are morally opportunistic. One observation that we make in this regard is that the structure of C can determine how much opportunistic behavior we should expect to see. For example, suppose that C contains only two elements, so the allocation game has two actions. In this case half of the conceivable norm functions will be prescribing the choice of one action and half of the other. In such conditions, it is likely that even if agent i believes that there are many norm functions, her choice might not look opportunistic from the perspective of any of them simply because all these norm functions prescribe the same choice as most socially appropriate (suppose i has very high ϕ_i so her consumption utility does not matter). This suggests that the amount of moral opportunism that we can observe in some environment depends on how many choices there are and how many different norm functions prescribe different alternatives as the most appropriate.

Keeping in mind this fact about discrete sets C, let us define a class of allocation games in which such problems do not arise and players have ample opportunities to change their choices. We do this in order to understand general implications of *expected* norm-dependent utility maximization for morally opportunistic behavior. Suppose that C is a closed interval on the real line and suppose that all u_i and η_k are twice continuously differentiable. Suppose as well that agent i considers two norm functions η_1 and η_2 and believes that the former occurs with probability p and the latter with probability $1 - p$. Then i maximizes the expected norm-dependent utility

$$U_i(c) = u_i(c) + \phi_i(p\eta_1(c) + (1-p)\eta_2(c)).$$

For the interior solution (assume one exists) the first order conditions define the optimal choice c^* as satisfying

$$\frac{u_i'(c^*)}{p\eta_1'(c^*) + (1-p)\eta_2'(c^*)} = -\phi_i.$$

Notice that in this formulation c^* depends on both η_1 and η_2 in a way that takes into account the whole functions. For general functions η_1 and η_2 that do not peak at the same place, the solution c^* above will be both η_1-opportunistic and η_2-opportunistic, because c^* will in general be different from the solutions c_1^* and c_2^* to either

$$\frac{u_i'(c_1^*)}{\eta_1'(c_1^*)} = -\phi_i \quad \text{or} \quad \frac{u_i'(c_2^*)}{\eta_2'(c_2^*)} = -\phi_i.$$

The same logic can be applied to any number of norm functions.[2]

This observation allows us to formulate a result that relates moral opportunism and maximization of expected norm-dependent utility.

Result. *When agent i maximizes expected norm-dependent utility in continuous allocation games with twice continuously differentiable utilities and norm functions, she will be in general η_k-opportunistic for all functions η_k that i believes can occur with some non-zero probability.*

This result essentially says that *anyone* who maximizes expected norm-dependent utility is morally opportunistic from any perspective by the nature of the expected utility maximization that is sensitive to all possible norm functions. As we mentioned in the Introduction, this conclusion is what makes our approach different from the previous research where moral opportunism was considered as something only exhibited by people "without moral values." The analysis in this appendix suggests that moral opportunism can be a general phenomenon that any expected norm-dependent utility maximizer can be susceptible to.

[2]Notice that if we have a group of agents such that all of them believe that η_1 happens with probability p and η_2 with probability $1-p$, then the choice of c^* is not morally opportunistic from their $(p \circ \eta_1, (1-p) \circ \eta_2)$-perspective.

B Computation of Optimal Punishment

In this section we explain how we came up with the optimal punishment strategy used to construct independent variables y_{rc} and y_{dc}. We apply the theory described in KV. According to this theory, a norm violation creates an additional "punishment norm function" that is followed in the same way other norms are. For the case of the Dictator game in which the dictator chooses action $c > c^*$, the punishment norm function $\mu_c(x)$ is shown in Figure 9. When $c \leq c^*$ then point c on the graph merges with the point $m_c = \min\{c, c^*\}$.

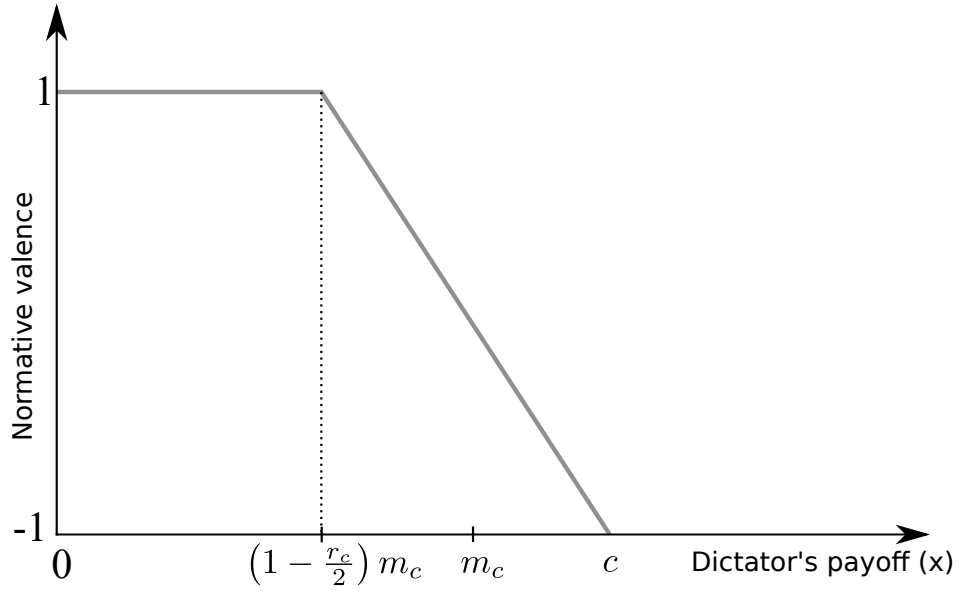

Figure 9: The punishment norm function $\mu_c(x)$ for action c in the Dictator game as defined by KV.

To understand what is the optimal punishment choice after the dictator chose c, we assume that the recipient solves the following maximization problem that determines the optimal payoff x left to the dictator after punishment:

$$\max_{x\in[0,c]} a(1-\sigma)\mu_c(x) - (c-x)/3.$$

Here $a \geq 0$ is the recipient's propensity to follow norms; $\sigma \in (0,1)$ is the parameter that determines the importance of punishment (see Kimbrough and Vostroknutov, 2020); and $(c-x)/3$ is the recipient's cost of leaving the dictator with x. Let $y_{rc} = c - \left(1-\frac{r_c}{2}\right)m_c$. Then the optimal solution is $x_1^* = \left(1-\frac{r_c}{2}\right)m_c$ whenever $y_{rc} \leq 6a(1-\sigma)$ and $x_2^* = c$ otherwise. This is simple to see. The maximand is a piece-wise linear function that has a peak at x_1^* if the decreasing part of $\mu_c(x)$ is steeper than $-15°$ (this is $45°$ times $1/3$, the punishment multiplier), and is monotonically increasing otherwise. Thus, for low values of $a(1-\sigma)$ the optimal solution is to not punish at all and leave the dictator with his earnings (recipient's cost is zero), and for high $a(1-\sigma)$ the solution is x_1^*, which generates optimal amount of punishment equal to $y_{rc} = c - \left(1-\frac{r_c}{2}\right)m_c$.

C Matching Procedure

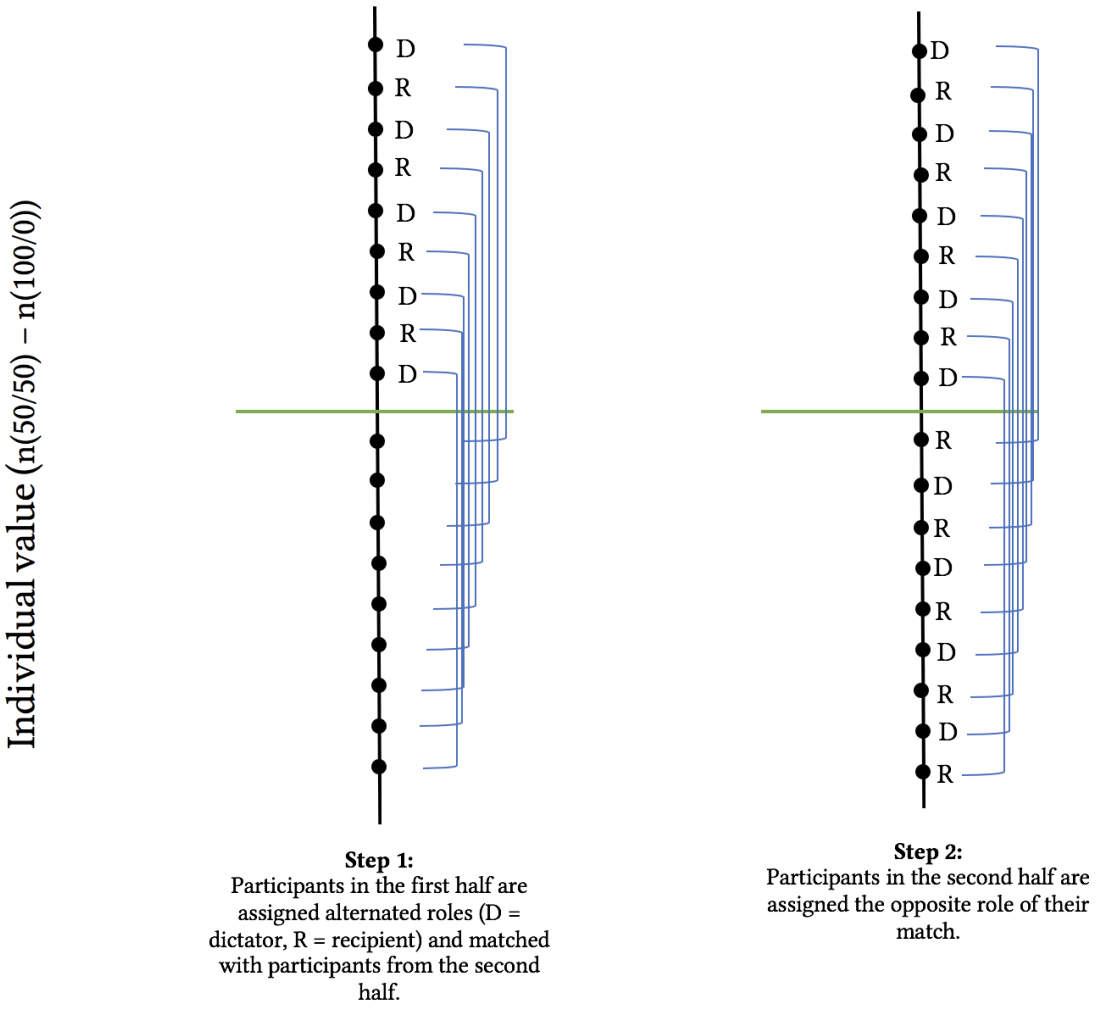

Figure 10: A graphical representation of the matching procedure.

D Additional Graphs

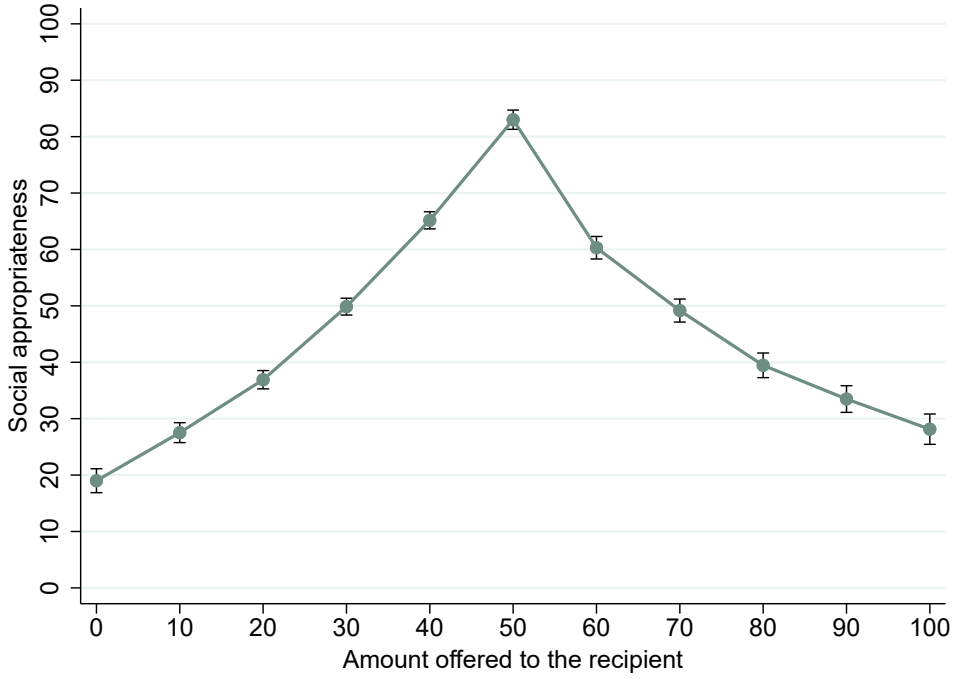

Figure 11: Average norm function elicited in the first CNE task (all subjects). The error bars are ±1SE.

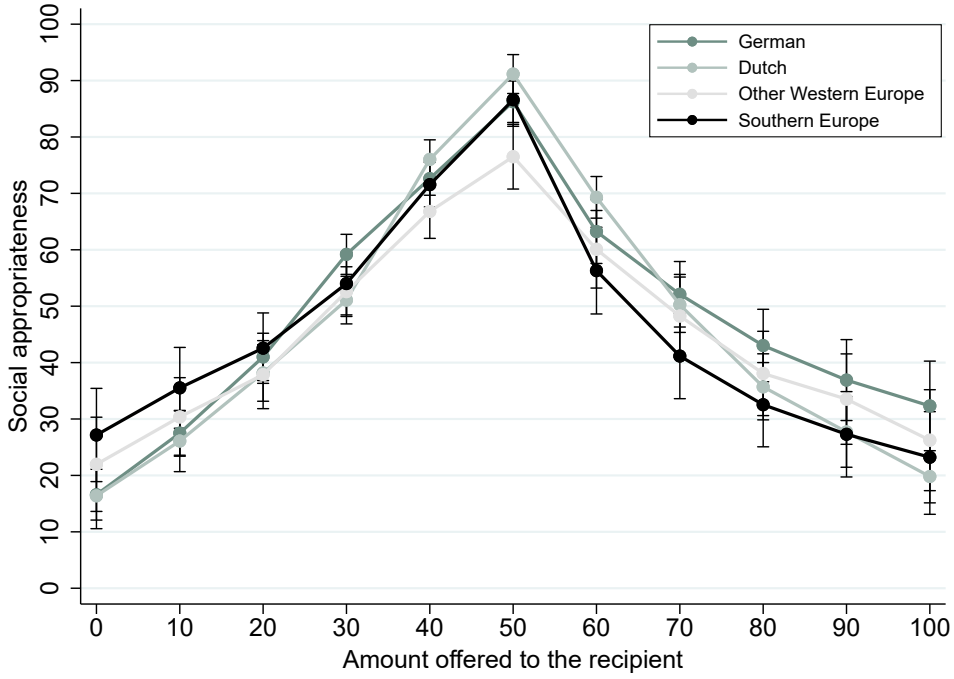

Figure 12: Average norm functions by nationality (recipients only). The error bars are ±1SE.

E Additional Results

E.1 Section 5.3

Treatment:	Control	Main
y_r	0.676***	0.536***
	(0.113)	(0.123)
y_d		0.198
		(0.126)
const	14.930**	15.203***
	(5.176)	(4.429)
N observations	374	759
N independent	34	69

Table 6: Random-effects panel regressions of the punishment decisions on the predictions of the model in the Control and Main treatments. Errors are clustered by subject and robust. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$R<D$	17.299
	(9.446)
$D<R$	17.888
	(9.978)
y_r	0.731**
	(0.268)
y_d	0.375
	(0.300)
$R<D \times y_r$	0.038
	(0.345)
$D<R \times y_r$	-0.553
	(0.290)
$R<D \times y_d$	-0.451
	(0.356)
$D<R \times y_d$	0.104
	(0.321)
const	0.938
	(6.384)
N observations	759
N independent	69

Table 7: Random-effects panel regression of the punishment decisions on the predictions of the model in the Main treatment. The baseline is the group $D = R$ with equal spread. $R<D$ and $D<R$ represent the dummies for the other two groups. Errors are clustered by subject and robust. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

E.2 Section 5.5

Figure 13 shows the coefficients ϕ_i and χ_i in the Control treatment on a scatter plot. The Spearman's $\rho = -0.64$, $p < 0.0001$. In the fitted OLS regression (black line), only the coefficient on below-median χ_i is significant (coefficient -0.092, $p < 0.001$). The recipients below median have high average ϕ_i and negative χ_i (0.82 and -3.54 respectively), whereas recipients above median have low ϕ_i and high χ_i (0.06 and 39.04). The differences between both coefficients are significant between these groups (ranksum tests, $p < 0.0439$). The average punishment among excessive punishers is 47.57 and 20.38 among normative punishers (ranksum test, $p = 0.0004$).

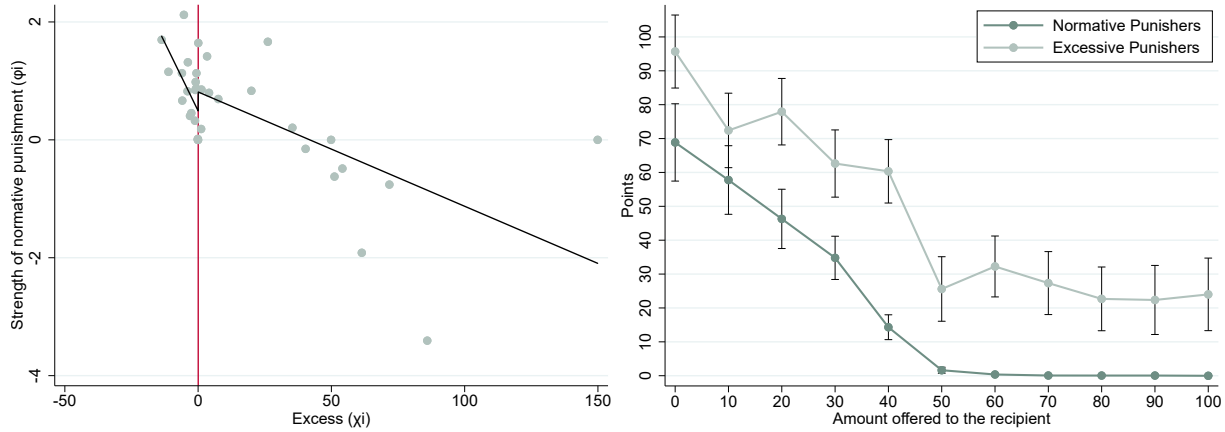

Figure 13: **Left panel.** Individual coefficients χ_i and ϕ_i for the recipients in the Control treatment. The black line represents the fitted OLS regression with two regimes: χ_i lower and higher than the median (the red line). 34 observations. **Right panel.** Average punishment strategies for the recipients below the median χ_i (normative punishers) and above the median (excessive punishers) in the Control treatment.

Figure 14 shows the coefficients ϕ_i and χ_i in both treatments. The Spearman's $\rho = -0.51$, $p < 0.0001$. In the fitted OLS regression (black line), both coefficients on χ_i (below and above median) are significant ($p < 0.003$). The recipients below median have high average ϕ_i and negative χ_i (0.85 and -4.95 respectively), whereas recipients above median have low ϕ_i and high χ_i (0.37 and 37.65). The differences between both coefficients are significant between these groups (ranksum tests, $p < 0.008$). The average punishment among excessive punishers is 51.69 and 20.81 among normative punishers (ranksum test, $p < 0.0001$).

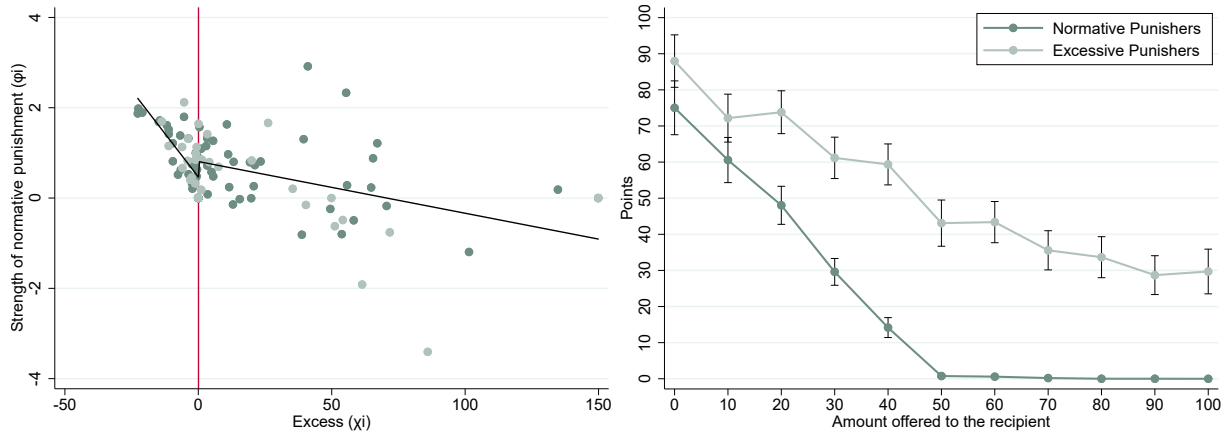

Figure 14: **Left panel.** Individual coefficients χ_i and ϕ_i for the recipients in both treatments. The black line represents the fitted OLS regression with two regimes: χ_i lower and higher than the median (the red line). 103 observations. **Right panel.** Average punishment strategies for the recipients below the median χ_i (normative punishers) and above the median (excessive punishers) in both treatments.

E.3 Section 6

	ϕ_i	χ_i
Dutch	-0.305	7.195
	(0.193)	(7.184)
Other Western European	-0.392	12.030
	(0.238)	(8.587)
Southern European	-0.588^{-}	29.267*
	(0.349)	(11.674)
Male	0.141	5.026
	(0.203)	(7.877)
Age	0.044	3.251
	(0.040)	(2.937)
const	-0.072	-67.187
	(0.887)	(61.385)
N independent	83	83

Table 8: Robust OLS regressions of individual coefficients ϕ_i and χ_i on nationality, sex, and age. $^{-}$ $p < 0.1$, * $p < 0.05$.

F Demographics

Variable	Obs.	Mean	Std. Dev.	Min	Max
Male	206	.4126	.492	0	1
Age	206	21.35	1.979	19	30
Years of study	206	2.126	1.275	0	7

Table 9: Demographics.

Field of study	Freq.	Percent	Cum.
Business	97	47.09	47.09
Economics, Business Economics	24	11.65	58.74
Fiscal Economics	23	11.17	69.90
Law	23	11.17	81.07
Finance	1	0.49	81.55
European Studies	8	3.88	85.44
Business intelligence Master	1	0.49	85.92
Arts and Culture, humanities, life/social sciences	12	5.83	91.75
Health studies and Psychology	5	2.43	94.17
Political studies	4	1.94	96.12
Engineering, Mathematics	2	0.97	97.09
Econometrics, Operations Research	6	2.91	100.00
Total	206	100.00	

Table 10: Field of study.

Nationality	Freq.	Percent	Cum.
German	53	25.73	25.73
Dutch	33	16.02	41.75
Other Western European	40	19.42	61.17
Scandinavian	3	1.46	62.62
Southern European	37	17.96	80.58
Eastern European	12	5.83	86.41
Anglo	5	2.43	88.83
South American	4	1.94	90.78
Chinese	3	1.46	92.23
Other Asian	9	4.37	96.602
Miscellaneous	7	3.40	100.00
Total	206	100.00	

Table 11: Nationality.

G Instructions

G.1 General Instructions

You are now participating in a decision making experiment. If you read the following instructions carefully you will be able to earn money in addition to your show-up fee of 3 Euro. Your earnings will depend on your decisions and the decisions of other participants. You will never get to know the identity of other participants you were matched with, nor they will get to know yours. Your earnings will be paid to you in CASH at the end of the experiment. The payment at the end of the experiment is also anonymous, that is, no other participant will know how much you earned in this experiment.

During the experiment you are not allowed to communicate with anybody. In case of questions, please raise your hand. Then we will come to you and answer your questions privately. Any violation of these rules excludes you immediately from the experiment and you will not be able to earn money.

The experiment consists of three parts. You will receive instructions for each part before it starts.

G.2 Part I

The information you provide in the following task may or may not be used later on in the experiment. In the task you will be asked to evaluate a hypothetical situation and decide whether taking certain actions would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior". By socially appropriate we mean behavior that most people agree is the "right" thing to do.

In the following task, you will read a description of a scenario. The description corresponds to a situation in which a person, "Individual C", must make a decision. In the scenario Individual C has 11 different possible actions he/she could take. The actions impact the payoff of Individual C as well as some other person, "Individual D". In the task, you are only asked to evaluate the decisions Individual C could make. After you read the description of the situation, you will be asked to evaluate all the different possible actions available to Individual C. For each of these actions you will decide the level of social appropriateness or inappropriateness of that action. In each of your responses, we would like you to answer as truthfully as possible, based on your opinions of what constitutes socially appropriate or socially inappropriate behavior.

After you have finished your evaluations, the computer will randomly select one of the possible actions Individual C could take. Each action is equally likely to be selected. The computer will calculate the proportion of participants whose evaluations are similar to yours for that randomly selected action. You will get paid according to this proportion. On the right you see an example of a slider for one possible action. The black circle shows the appropriateness level of an action that ranges from very inappropriate (VERY INAPPR.) to very appropriate (VERY APPR.). You will be able to move the circle by dragging it along the slider (please try to move it now). The more participants put the black circle anywhere in the vicinity of your choice (indicated by the red rectangle) the more you will get paid. To give you an example: if 100% of other participants in this room choose the appropriateness levels that fall within the red rectangle you will receive 3 Euro, if 50% of the participants choose within the red rectangle you will receive 1.5 Euro, and if no one chooses the level similar to yours (the choices of all other participants are outside the red rectangle), you will receive nothing. In general, if proportion X of participants chooses within the boundaries of the red rectangle you will receive 3*X Euro.

Please try to move the black circle around. Notice that when you reach the boundary, the red rectangle does not diminish in size. This means that your chances to earn money are not decreased when you choose to put the black circle close to the boundary.

G.3 Hypothetical Situation

Imagine that Individual C has been invited to the experiment and paired with another anonymous Individual D so that neither individual will ever know the identity of the other individual with whom she/he is paired.

Individual C reads the following instructions: *You will receive 100 points **(10 point = 0.5 Euro)** that you can keep or share with Individual D. You can give between 0 and 100 points to Individual D in increments of 10.*

Now look at the graph on the right side of the screen and consider the 11 possible actions available to Individual C. The actions (how the individual could distribute the points between him/herself and Individual D) are written on the horizontal axis (e.g., C:30; D:70 means that individual C decided to keep 30 points and to give 70 points to Individual D). By dragging the black circle up and down the slider you can indicate the level of social appropriateness for every possible action. The most appropriate level is at the top of the graph while the least appropriate level is at the bottom.

Remember: you will get paid based on the proportion of other participants who choose a similar appropriateness level (any choice that falls within the borders of the red rectangle) for a randomly chosen action. Specifically, you will receive 3*X Euro if proportion X of other participants in this room chooses within the boundaries of the red rectangle. Your payment DOES NOT depend on the actual decision made by Individual C.

Please make a choice for each action of Individual C. Press OK when you are done. If for some reason the black circle is outside the boundaries of the red rectangle please raise your hand and we will fix that.

G.4 Part II

You will now be matched with another participant and assigned a role of either "Participant A" or "Participant B". First, both Participant A and Participant B are given 50 points each to be used in the game as described below (**10 point = 0.5 Euro**).

In the game Participant A **additionally** receives 100 points. Participant A can decide how to share (or not share) these 100 points with Participant B. Participant A can give any amount X of points between 0 and 100 to Participant B. Participant A can only give in increments of 10 points.

At the same time Participant B decides if and how many points to subtract from Participant A depending on A's actions. Specifically, Participant B chooses how many points to subtract from Participant A for each amount that A can choose to give to B: 0, 10, 20, ..., 100 points. For each of these cases, for Y points that Participant B decides to subtract from A, B will pay one third of that amount (or Y/3 points). Participant B can subtract between 0 and 150 points (and pay between 0 and 50 points). In case more points are subtracted than Participant A has left, Participant A will receive 0 points.

After both participants make their choices the amount of points subtracted from Participant A and the subtraction payment incurred by Participant B are determined by the action that Participant A has chosen (to share X points with B).

The payoffs of Participants A and B are calculated as follows. Participant A receives 50 points and additional 100 points to share with B; A shares X points with B and gets Y points subtracted by B. Thus, A's payoff is 50 + 100 - X - Y. Participant B receives 50 points; then B receives X points from A and pays Y/3 points for subtracting points from A. Thus, B's payoff is 50 + X - Y/3.

Please click OK if you are ready to go on. If you have any questions, please raise your hand and wait for help.

G.5 Instructions for Dictators

The computer has assigned you the role of **Participant A**.

Both you and Participant B have received 50 points (10 points = 0.5 Euro). You received additionally 100 more points that you can keep to yourself or share with Participant B.

If you would like to share any points with Participant B please indicate the amount on the right of the screen. Enter multiples of 10: 0, 10, 20, ..., 100 points.

Please click OK if you are ready to go on.

G.6 Instructions for Recipients in Main Treatment

The computer has assigned you the role of **Participant B**.

On the right side of the screen you can see the evaluations that you and Participant A, with whom you are paired, made in the previous part of the experiment. Specifically, the black graph shows how **you** have evaluated the social appropriateness of hypothetical actions of Individual C. The blue graph shows the same evaluations of **Participant A** with whom you are paired in this part of the experiment.

You do not have to make any choice on this screen. It is for your information only. You will be able to open this graph again when you make your decisions.

Please press OK button when you are ready to go on.

G.7 Instructions for Recipients in Control Treatment

The computer has assigned you the role of **Participant B**.

On the right side of the screen you can see the evaluations that you made in the previous part of the experiment. Specifically, the black graph shows how you have evaluated the social appropriateness of hypothetical actions of Individual C.

You do not have to make any choice on this screen. It is for your information only. You will be able to open this graph again when you make your decisions.

Please press OK button when you are ready to go on.

G.8 Subtraction Decisions for Recipients in Main Treatment

As Participant B you can choose how many points to subtract from Participant A depending on his/her choice. Remember, Participant A chooses how to divide 100 points between him/her and you (10 point = 0.5 Euro). The graph on the right of the screen shows 11 possible choices of Participant A. For example, A:30 B:70 means that Participant A has chosen to give you 70 points and keep 30 points. Your task is to determine how many points you would like to subtract from Participant A conditional on his/her choice.

To make your choice please drag the green circles on each of the 11 sliders. The green number next to the green circle indicates how many points you want to subtract from Participant A if he/she chooses a specific allocation of points. The red circles (and red numbers) indicate how many points you would need to pay for this subtraction: it is the number of points you choose to subtract divided by three.

When Participant A has made his/her choice, it will determine which one of the 11 possibilities will be used to calculate the amount of points subtracted from A and the payment that you will need to make for the subtraction.

For your convenience, you can see the evaluations that you and Participant A made in the previous part of the experiment by pressing the ”SHOW THE EVALUATIONS” button below these instructions (to hide the graph press ”HIDE THE EVALUATIONS” button).

Please make sure that you make your choice for all 11 possibilities. Press OK when you are ready to go on.

G.9 Subtraction Decisions for Recipients in Control Treatment

As Participant B you can choose how many points to subtract from Participant A depending on his/her choice. Remember, Participant A chooses how to divide 100 points between him/her and you (10 points = 0.5 Euro). The graph on the right of the screen shows 11 possible choices of Participant A. For example, A:30 B:70 means that Participant A has chosen to give you 70 points and keep 30 points. Your task is to determine how many points you would like to subtract from Participant A conditional on his/her choice.

To make your choice please drag the green circles on each of the 11 sliders. The green number next to the green circle indicates how many points you want to subtract from Participant A if he/she chooses a specific allocation of points. The red circles (and red numbers) indicate how many points you would need to pay for this subtraction: it is the number of points you choose to subtract divided by three.

When Participant A has made his/her choice, it will determine which one of the 11 possibilities will be used to calculate the amount of points subtracted from A and the payment that you will need to make for the subtraction.

For your convenience, you can see the evaluations that you made in the previous part of the experiment by pressing the ”SHOW THE EVALUATIONS” button below these instructions (to hide the graph press ”HIDE THE EVALUATIONS” button).

Please make sure that you make your choice for all 11 possibilities. Press OK when you are ready to go on.

G.10 Part III

On the right of the screen you see the evaluations you made in PART I of the experiment. Remember, you were choosing the appropriateness levels of the hypothetical actions of Individual C, who was dividing 100 points between him/herself and Individual D.

In this part you can choose to maintain or revise your evaluation. After you make your new evaluations (or keep the old ones), the computer will randomly select one of the possible actions Individual C could take. Each action is equally likely to be selected. For that randomly selected action you will get paid according to the proportion of other participants who gave an answer similar to yours. To give you an example: if 100% of other participants in this room choose the appropriateness levels that fall within the red rectangle you will receive 3 Euro, if 50% of the participants choose within the red rectangle you will receive 1.5 Euro, and if no one chooses the level similar to yours (the choices of all other participants are outside the red rectangle), you will receive nothing. In general, if proportion X of participants chooses within the boundaries of the red rectangle you will receive 3*X Euro.

Your payment only depends on the proportion of people who choose a similar appropriateness level to yours, and DOES NOT depend on the decisions made before by you or the participant you were paired with. You can now revise or maintain the social appropriateness/social inappropriateness of every action.

Please click OK if you are ready to go on.

References

Kimbrough, E. O. and Vostroknutov, A. (2020). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Tremewan, J. and Vostroknutov, A. (2020). *A Research Agenda in Experimental Economics*, chapter An Informational Framework for Studying Social Norms. Edward Elgar Publishers.