An Informational Framework for Studying Social Norms: An Extended Version*

James Tremewan[†] [‡] **Alexander Vostroknutov**[§]

December 2020

The greatest achievements of humankind have resulted from our extraordinary ability to act together, to formulate and achieve common goals that benefit the community as a whole, in other words to cooperate. Comparisons between humans and other closely related species, such as apes and primates, unambiguously demonstrate that one crucial thing that distinguishes us from them is the presence of sophisticated culture—a collection of packages of social norms enhanced with high-fidelity information transmission mechanisms (e.g., language)—that allows us to sustain vital community-oriented behaviours across many generations (Henrich, 2017; Laland, 2018). Given the highly cooperative nature of the societies we live in, it is thus not surprising that social norms permeate all aspects of our daily lives. Consequently, concepts related to social norms have featured in most social sciences and philosophy at least since the time of Aristotle.

Mainstream economics has traditionally put very little weight on the role of social norms in economic behaviour. Nevertheless, its mathematical concepts—including preferences, utility, rationality, and game theory—combined with experimental methods have produced a set of valuable tools that have been used successfully to study incentivized social behaviour (Vostroknutov, 2020). Experimental economics was a late-comer to the topic of social norms. However, it can be a crucial addition to the methodological arsenal for studying them exactly because of its mathematically precise approach to formulating testable hypotheses.

*This is the extended version of the chapter in "Research Agenda in Experimental Economics" (2020, Cheltenham: Edward Elgar). We would like to thank Erik Kimbrough and Marie Claire Villeval for invaluable comments. All mistakes are our own.

[†]Department of Economics, University of Auckland, New Zealand. e-mail: james.tremewan@gmail.com.

[‡]Corresponding author.

[§]Department of Economics (MPE), Maastricht University, Tongersestraat 53, 6211LM Maastricht, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

Despite achieving various important insights, experimental economics' treatment of social norms remains sporadic, unstructured, and terminologically vague. We believe that, in order to make progress in the field, a general framework should be put forward that would allow researchers to speak the same language and to expand the applicability of experimental economics methods to broader sets of problems related to norm-related behaviour. In this paper, we propose such a framework based on the recent developments in experimental methodology and behavioural economic theory. The framework puts specific emphasis on the *transmission of information* relevant for normative reasoning, which is an integral part of understanding social learning, norm stability, norm uncertainty, among other topics, and their reflections in social behaviour. In essence, we construct an environment in which multi-modal information (private information, learned opinions, observed actions) is processed by norm-following agents in order to act in a world inhabited by other norm-following agents. We show how this framework can prove useful in interpreting the results of existing experiments, and in designing future experiments to carefully avoid confounds.

After developing a comprehensive framework for the study of social norms, we go on to use it to structure both a compendium of experimental methods that have been developed for studying this topic, and our views on the most important avenues for future research. The paper continues as follows: Section 1 gives a brief overview of terminology related to social norms and different definitions proposed by researchers from various social sciences in order to assist readers, new to the topic, in parsing the existing literature; in Section 2 we outline and discuss our proposed framework; in Section 3 we describe the tools that can be used for the experimental study of social norms, cite papers that illustrate these tools, and suggest how they might be used to further our knowledge of this fundamental aspect of human nature.

1 Definitions of Norms

An initial confusion to be cleared up in the definitions of norms is the dual use of the words *norm* and *normative.* For many economists, the first association with these words will be in relation to the distinction between positive and normative economics, the former being the study of "what is" and the latter the study of "what ought to be." That is, the word *normative* has a moral flavour. On the other hand, *normative* often refers in common usage to "what is typical," which is an *is* rather than an *ought* concept. Both uses of the words are ingrained in the literature on norms, and the words will be used in both ways in this paper. Often in the literature, people differentiate between *is* and *ought* using the terms *descriptive* and *injunctive*, simply defining descriptive norms as "what most people do" and injunctive norms as "what most people think is right." Many papers refer to either one or the other of these concepts as a social norm.

The only general agreement regarding the definition of the term *social norm* seems to be that there is no general agreement. In preparation for this paper, we ran a short survey of subscribers

to an experimental economics discussion group to gauge the personal views of researchers in the field on some key definitional issues, and indeed found considerable diversity.[1] For example, 46% of the 201 respondents regarded social norms and social preferences as completely distinct phenomena, while the other half viewed one as a subset of the other, or identified a partial overlap. While there were points of disagreement, there was also strong consensus on some issues that we will mention in this section.

Economic theorists have often viewed social norms as devices to enable coordination in the presence of multiple equilibria (e.g. Gintis, 2010). However 84% of respondents in our survey disagreed with the statement that coordination on one of multiple equilibria is a sufficient definition of the term *social norm*. A more widely accepted view of social norms is that they relate explicitly to beliefs about what others believe ought to be done. We follow the terminology of Bicchieri (2016), and call individual's beliefs about what ought to be done *personal norms* and beliefs about others' personal norms *normative expectations*.[2] In our survey, 85% of respondents agreed that normative expectations should be part of a definition of a social norm, much more than any other statement we suggested in this question. The idea that social norms are commonly held perceptions of what society believes is acceptable (i.e., shared normative expectations) is present in two of the most prominent definitions of social norms, those of Bicchieri (2016) and Elster (1989). Beyond this there is less agreement, and we now highlight some of the additional conditions some authors impose.

In some definitions, it must be the case that not only are there shared normative expectations, but that these expectations are correct. For example, the definition by Fehr and Schurtenberger (2018) states that social norms are "commonly known standards of behaviour that are *based on widely shared views* (our italics) of how individual group members ought to behave." In other words, personal norms and normative expectations generally coincide. Such a definition excludes an important class of phenomena which are regarded by others as social norms. Specifically, behaviour underpinned by *pluralistic ignorance* (Bicchieri, 2016). This is a situation where personal norms are different from shared normative expectations. For example, in the folktale "The Emperor's New Clothes" the observers understand that the Emperor is naked, however they keep pretending that he is not due to shared normative expectations (everyone believes that everyone believes that the Emperor is clothed). The opinions in our survey on whether personal norms should be included as a part of the definition of social norms are split: 40% of the respondents thought that "what one believes one ought to do" should be part of a definition of a social norm.

Elster (2009) includes in his definition that social norms are only adhered to due to fear of punishment ("the operation of social norms depends crucially on the agent being observed by

[1] The full survey and results can be found in Appendix A.

[2] Some authors use the term *moral norms* (e.g., Charness and Schram, 2013) or *private values* (d'Adda et al., 2020) instead of personal norms.

others"). An integral part of the definition of Bicchieri (2016) is that shared normative expectations only constitute a social norm when behaviour is *conditional* on those expectations as well as on *empirical expectations*, or beliefs about what others actually do. Bicchieri also requires conditionality in her definition of a descriptive norm, which is only defined as such if people are conforming to a behaviour because they believe others will also conform to it.

Overall, we see a large diversity of opinions on what norms are. The general framework we outline in the next section is rich enough to test the assumptions implicit in many of these definitions, while being relatively parsimonious in that the set of assumptions from which it itself is derived is rather small. Furthermore, there are hints that personal norms and empirical expectations may play an important role. Indeed, one of the important insights of our framework is that the effect of social norms, as represented by normative expectations, cannot be effectively studied without simultaneously accounting for these other types of "norm-related beliefs," which therefore play a major role in the remainder of the paper.

2 A Framework for Norms

In this section we develop a framework for investigating the relationship between norms and behaviour. In Section 2.1, we provide the basics of a general theory of "injunctive norms," showing how the normative acceptability or otherwise of an outcome can be derived from a game form. In Section 2.2, we follow the categorization of Bicchieri (2016) and give precise mathematical meaning to the four types of norm-related beliefs that she defines: factual beliefs, personal norms, normative expectations, and empirical expectations. With this at hand, we show how different kinds of informative signals, social or otherwise, lead to the updating of these norm-related beliefs. In Section 2.3 we incorporate all four types of beliefs into the utility function, through which they may influence behaviour.

Given the constraints of this paper, we only give simple examples to illustrate the mechanisms of the framework. We emphasise here that although many of the specific details of our framework as we outline it here may be debated, the main insights are likely to hold in any model with two general assumptions: 1) agents are influenced to some degree in their choices by normative concerns; and 2) that agents have some understanding of how others' normative views relate to factual beliefs, allowing information to be inferred from those normative views, and actions based upon them.

The main value of our framework is that it clearly separates the channels through which information, whether it be factual or related to social injunctive or descriptive norms, can affect behaviour: by shaping factual beliefs, personal norms, or normative or empirical expectations (Section 2.2); through a direct impact on utility caused by a desire to act in accordance with norms (Section 2.3); or by affecting perception of the likelihood of punishment (Section 2.3). We think that bearing this framework in mind when designing and interpreting experiments will

help researchers place their work in a broader context and avoid possible confounds.

2.1 A Theory of Injunctive Norms

In this section we present a general theory of injunctive norms following Kimbrough and Vostroknutov (2020b), further KV. While a general theory is not essential for the remainder of our framework (and can be easily replaced by another theory), it will give us a firm basis on which to build the example we focus on in the following section. Furthermore, we believe that *some* theory of injunctive norms is necessary to study social decision making because assuming that norms can vary freely with no structure allows everything to be explained in an ad hoc manner and is therefore unfalsifiable (Vostroknutov, 2020). Thus, KV's theory can be used as a starting point for future research on general theories of injunctive norms, which we will make use of in Section 3.1.

The theory builds on three premises: 1) that an individual feels (prospective) *dissatisfaction* about a certain outcome that can happen because of other materially better outcomes that could have happened instead; 2) that people are able to *empathise* with others (they understand others' dissatisfactions); 3) the process of gene-culture co-evolution favours social norms that *minimize aggregate dissatisfaction*, because individuals tend to bargain, argue, negotiate, etc. with one another, so the norm produces a balance between the interests of all parties involved.

To put this in mathematical terms, consider a finite set of players N and a set C of consequences that represent all possible outcomes of some interaction of the players.[3] Let us assume that there is a consumption utility function $u_i : C \to \mathbb{R}$ defined for each player i that associates with each consequence $c \in C$ a consumption utility that i receives in it. In this section we assume that all u_i's are common knowledge. Then

$$d_i(c_1, c) = \max\{u_i(c) - u_i(c_1), 0\}$$

denotes the *dissatisfaction* that player i feels about c_1 because of the possibility of c. Thus, player i is dissatisfied to the degree $u_i(c) - u_i(c_1)$, if c gives her higher consumption utility, and is not dissatisfied at all if c gives her lower consumption utility.[4] The *total dissatisfaction* that player i feels with respect to c_1 can then be defined as

$$D_i(c_1) = \int_{c \in C} d_i(c_1, c) dc. \tag{1}$$

This is just a summation of all dissatisfactions that are felt in c_1 because of all other consequences

[3] Typically, C represents the terminal nodes of a game. However, the game form (that defines players' strategies) does not need to be precisely specified, in which case C can be seen as a set of all possible outcomes of some unstructured bargaining process.

[4] This is the formulation proposed by KV. However, it has never been directly compared to other possible models in this framework. In Section 3.1 we discuss these alternatives in more detail.

in C (if there are finitely many consequences in C, then the integral is replaced by a sum). The *overall dissatisfaction* in c_1, which takes into account the dissatisfactions of all players, can be computed as

$$D(c_1) = \sum_{i \in N} D_i(c_1).$$

Here we use the assumption that players can empathise with each other, so they can compute the total dissatisfactions of other players. The knowledge of $D(c)$ for all consequences $c \in C$ gives each player an idea about how much dissatisfaction should be expected if c is achieved.

Next, KV postulate that the feeling of social appropriateness associated with a consequence in C—which is also called *normative valence*—is inversely proportional to the overall dissatisfaction in this consequence. To capture this idea, we define the *norm function* η as

$$\eta(c) = [-D(c)],$$

where $[\cdot]$ stands for the linear normalisation of the function $-D$ to the interval $[-1, 1]$ (this normalisation is necessary in order to have a common "normative space" for all games). So, for all consequences c that minimize the overall dissatisfaction D we have $\eta(c) = 1$, while for all consequences that maximize the overall dissatisfaction we have $\eta(c) = -1$.

To illustrate this concept, consider the Dictator game. Suppose that there is a pie of size 1 and that a dictator chooses to give $c \leq 1$ to a receiver and is left with $1 - c$. The set of consequences is given by all possible values of c and can be defined as $C = [0, 1]$. The consumption utilities are given by $u_D(c) = 1 - c$ for the dictator and $u_R(c) = c$ for the receiver. For any consequence $c \in C$ the total dissatisfaction of the dictator is $D_D(c) = \frac{c^2}{2}$, and the total dissatisfaction of the receiver is $D_R(c) = \frac{(1-c)^2}{2}$, both computed using definition (1). Thus, the overall dissatisfaction is given by

$$D(c) = D_D(c) + D_R(c) = \frac{c^2}{2} + \frac{(1-c)^2}{2}.$$

This is a parabola with a minimum at $c^* = \frac{1}{2}$. Thus, the norm function $\eta(c)$ is a downward sloping parabola with the equal split c^* being the most socially appropriate consequence, which we will also call the *injunctive norm*, defined as any consequence c in which $\eta(c)$ reaches its maximum. The consequences $c = 0$ and $c = 1$ are the least appropriate ones, with $\eta(0) = \eta(1) = -1$.

In order to use this theory to predict behaviour, KV assume that players maximize *norm-dependent utility* of the form

$$w_i(c) = u_i(c) + \phi_i \eta(c), \tag{2}$$

where $\phi_i \geq 0$ is an *individual propensity to follow norms*. This utility function represents the trade-off between increasing own consumption utility u_i and acting in accordance with the injunctive

norm η.[5] In the Dictator game, the optimal choice that maximizes $w_i(c)$ is $\frac{8\phi_i - 1}{16\phi_i}$ for $\phi_i \geq \frac{1}{8}$ and 0 for $\phi_i < \frac{1}{8}$. Thus, when the dictator cares little about following norms ($\phi_i < \frac{1}{8}$), she gives the receiver nothing. When she cares a lot about following norms ($\phi_i \geq \frac{1}{8}$) she chooses to give an amount proportional to ϕ_i that reaches the maximum amount of giving equal to $\frac{1}{2}$ as $\phi_i \to \infty$.

An important effect that this model can capture, which we will use in the next section, is that the norm changes depending on who needs money more, the dictator or the receiver. If we assume that the receiver has a different utility function, say $u_2(c) = \gamma c$ where $\gamma > 0$, then the injunctive norm becomes $c^* = \frac{\gamma}{1+\gamma}$. This makes sense: if the receiver needs money more than the dictator ($\gamma > 1$), then the norm is to give him more than half, giving everything in the limit ($\lim_{\gamma \to \infty} c^* = 1$); if the receiver needs money less than the dictator ($\gamma < 1$), then the norm is to give the receiver less than half, giving nothing in the limit ($\lim_{\gamma \to 0} c^* = 0$). The optimal choice, determined by the maximization of $w_i(c)$, follows the same pattern as above: for low ϕ_i the dictator will give nothing, and for high ϕ_i she will give increasing amounts that asymptotically reach $\frac{\gamma}{1+\gamma}$ as $\phi_i \to \infty$.

This example demonstrates how the theory predicts behaviour in *allocation games*, games with a single move by one player who chooses among arbitrary allocations. In the following sections we will use such essentially individual decision problems to showcase our framework. We do it in order to keep things simple, as well as because individual moral decision problems, akin to the Dictator game, constitute a large bulk of empirical data collected through surveys, thus making the framework presented here directly applicable to a wide range of empirical questions. For general games, special care should be taken when introducing the norm-dependent utility. For example, in games with sequential moves normative punishment should be explicitly modeled to account for norm violations as the game unfolds. We discuss punishment briefly in Section 2.3 and suggest that the readers who are interested in normative behaviour in generic games follow KV, who show how to deal with norm-dependent utility in general games with observable actions.

Another important factor that should be considered when using this model is *bounded rationality* (Simon, 1990). Indeed, the computation of the norm function η can be difficult: for each consequence each individual should compute the dissatisfactions of all players, which presumes knowledge of their utilities in all consequences and thus can be a daunting task in itself. To deal with this issue, Kimbrough and Vostroknutov (2020a) propose a methodology based on KV that allows the estimation of *moral rules* that can be used instead of η in some classes of games. Moral rules are simple but relatively good approximations of η that can be constructed for specific types of strategic interactions. For example, the rule "we should always share everything equally" might be widely used in Dictator-like situations, where a pie should be divided

[5] It does not always have to be a trade-off between own consumption utility and the norm function, they can be aligned in some contexts, for example when choosing between a Pareto-dominated and a Pareto-dominant allocation.

among some players. Higher social appropriateness of Pareto-dominant allocations is another example of a moral rule.[6] Overall, for some applications it might be reasonable to substitute the fully fledged computation of η with an approximation coming from a popular moral rule.

2.2 Norm-Related Beliefs

In this section we describe how various norm-related beliefs are formed in the presence of uncertainty, and how they can be affected by different types of new information. We take as our starting point the classification of beliefs in Bicchieri (2016), which appears to be gaining popularity in experimental economics, with many researchers already familiar with the terminology. Bicchieri (2016) distinguishes four types of beliefs that we present here together with the corresponding mathematical objects from our framework:

Factual beliefs: player i's beliefs about the utilities of consequences in C (beliefs about u_k for all $k \in N$);

Personal norm function: what player i believes is morally correct behaviour given her factual beliefs (represented by player i's own η_i);[7,8]

Normative expectations: player i's perceptions of personal norm functions of others given current information (i's beliefs about η_j for all $j \neq i$) – *normative information* that can come from, for example, credible statements by individuals about their moral views, or the outcome of a referendum on a moral issue;

Empirical expectations: player i's perceptions of how other players will behave given current information (model's predictions about the choices of players $j \neq i$ given normative expectations and beliefs about ϕ_k for all $k \in N$) – *empirical information* that may come from personal observation of others' behaviour, government statistics, and so on.

The main idea we wish to get across in this section is the interdependence of these four types of beliefs and the information that underlies them. Some of these relationships are obvious, but others less so. We will go through a series of examples to illustrate some key insights that arise from thinking about norms in this framework. To explain the general argumentation, we start

[6] As a matter of fact, it is a general result in the model of KV that a Pareto-dominant allocation always has higher normative valence than the allocation that it Pareto-dominates. Thus Pareto-dominance is an obvious candidate for a moral rule.

[7] In our framework, a *norm function* defines the beliefs about social appropriateness of all outcomes. By a *norm* we mean the argmax of a norm function. However, Bicchieri (2016) uses the term *personal norm* instead of our *personal norm function*. To avoid confusion with our own terminology, we will use *personal norm function* instead of Bicchieri's personal norm, and *personal norm* for the argmax of a personal norm function.

[8] When there is no uncertainty over the state of the world, η_i is known, and therefore it may seem more natural to economists to think of personal norms as preferences rather than beliefs. Here we are using the word "belief" in a more colloquial sense for convenience. However, when there is uncertainty, the personal norm will depend on beliefs about the state of the world, and is then truly a belief in the standard sense.

with an example of "flight-shaming." Then, we present a simple model with Bayesian agents, who update their beliefs in the Dictator game, that shows how such arguments can be modeled in our framework.

To give a concrete example of how uncertainty and asymmetric information influence norm-related beliefs, consider the social acceptability of flying to a holiday destination related to the modern phenomenon of "flight-shaming" (Gössling et al., 2020). Holiday-makers must decide how much convenience to sacrifice in order to benefit future generations by decreasing the potential costs resulting from anthropogenic climate change, but there is uncertainty over the degree to which emissions from the aviation industry will accelerate global warming.

How does factual information affect norm-related beliefs? First, imagine a situation where the latest and most reliable climate model is common knowledge. Because each individual's factual information is identical, *personal norms and normative expectations coincide*. Now suppose that you read an article, which suggests that climate change will proceed at a faster pace than previously thought (new factual information). You update your personal norm accordingly (flying becomes less personally acceptable), however you know that only some fraction of the population will have read the article, so you update your normative expectations to a lesser degree. Thus, *personal norms in a population can differ due to asymmetric information, but are likely to be correlated*. How much you update empirical expectations depends on your beliefs about how influenced other people are by normative concerns.

Now consider new normative information, for example a speech by Greta Thunberg emphasizing our moral duty to fly less. Clearly this will cause us to update our normative expectations, both because of direct information about Greta's personal norm and because this is likely to be a view shared by a broader swathe of the population. Empirical expectations will be updated to the degree to which we believe people are influenced in their choices by their normative beliefs. Furthermore, Greta's speech may influence our factual beliefs, and thereby our personal norms: knowing that Greta's personal norms are based on factual information, we can infer something about information she has that we may not be aware of. If she finds flying less acceptable than us, we can infer that her information suggests that flying is more harmful to future generations than we believed, and as far as we believe that her information sources are broader or more accurate than our own, we will update both our factual beliefs and personal norms.

New empirical information, for example a newspaper article showing that people are choosing to fly less and holiday more in their own country, will directly cause us to update our empirical expectations. But again, this may suggest that others have information we do not about the harms of flying, causing us to also update our factual beliefs, personal norms derived from those beliefs, and normative expectations.

This example demonstrates the intuition behind our framework and what it can offer to the researchers. To understand precisely how the four types of beliefs form and influence each other, we now consider a simple Dictator-game setup that only includes personal norm functions in

the utility, as in the previous section.[9] This is essentially a formalization of the verbal example we have just outlined, where the decision of the dictator is how frequently to fly, and the receiver represents future generations.

We start with a fixed set of consequences C (but not utilities) and a set of states of the world S that models uncertainty over all information relevant for a decision maker with the norm-dependent utility described by (2). Such information includes the utilities $u_i(c)$ for all $i \in N$ and all $c \in C$, and the individual propensities to follow norms ϕ_i for all $i \in N$. We assume that in each state of the world all these parameters are well-defined. Thus, if there is a common knowledge of the state of the world, then we are back to the problem without uncertainty as formulated in the previous section. It is important to emphasise at this point that any other specification of the norm function can be used instead of the dissatisfaction-based normative model of KV. Another utility specification will require a different definition of the state space.

Next, we assume that there is a time dimension with arbitrary number of periods. As time unfolds, players may receive informative signals about the state of the world that they use to update all four types of beliefs mentioned above. After that, they can take actions of two possible forms: 1) choices in an allocation game with outcomes C; and 2) credible communication of η_i, for example, a truthful public statement of opinion.[10] Actions of the first sort provide *empirical information* to any observers, whereas actions of the second sort provide *normative information* (see also the definitions of the norm-related beliefs above). New empirical and normative information, in conjunction with any new factual signals, can be used to update norm-related beliefs in the subsequent period.

To demonstrate the insights that can be gained from this very general setup, we provide simple numerical examples, defining a specific timeline of events chosen to illustrate some interesting implications. Consider the Dictator game with $C = [0, 1]$ as in the previous section where the receiver's utility function is $u_R(c) = \gamma c$. In our verbal example, γ represents the magnitude of welfare loss to future generations due to climate change. We consider an environment with only two dictators, 1 and 2, each of whom make choices in separate games with separate receivers, but who may sometimes observe some actions of the other.

In the first example, we illustrate how in our purely Bayesian framework, personal norm functions may differ from normative expectations, and how shocks to either empirical and normative expectations can result in changes not just in each of those expectations, but also in personal norm functions. To do this we introduce uncertainty over γ, which takes a value in $\{1, 6\}$ (the same for both receivers) depending on the true state of the world. For now we assume that $\phi_1 = \phi_2 = \phi$, and that this is common knowledge. Thus, we have two states of the world corresponding to different values of γ.

[9] The extension to more general norm-dependent utility, which includes normative and empirical expectations is straightforward (see Section 2.3).

[10] We assume for expositional purposes that all agents tell the truth and that this is common knowledge. This assumption can be easily relaxed in applications.

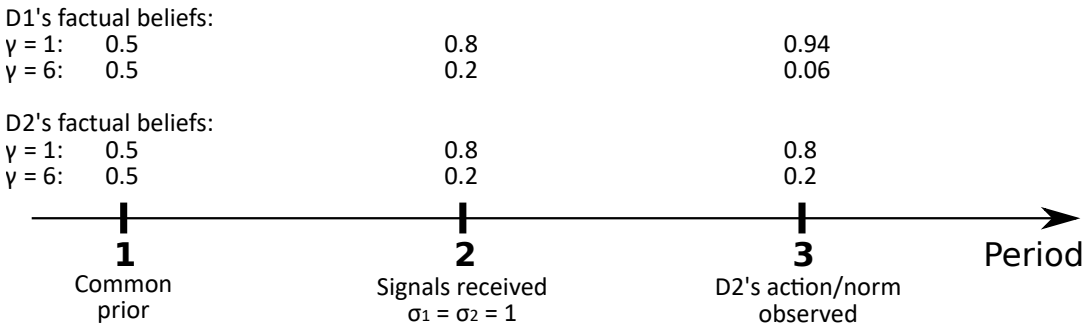

Figure 1: The sequence of events in the example.

In period 1, both dictators have the common uniform prior over γ that each state occurs with probability $\frac{1}{2}$, and norm-related beliefs and choices are determined on this basis (see Figure 1). In period 2, both dictators receive private noisy but informative signals that refine their information about the state of the world (*factual information*), allowing them to update their norm-related beliefs using Bayes' rule. Specifically, each Dictator i receives an independent signal $\sigma_i \in \{1, 6\}$, where $P(\gamma = \sigma_i | \sigma_i) = p > 0.5$ for $\gamma \in \{1, 6\}$. In period 3, Dictator 1 observes either the choice of Dictator 2 (*empirical information*) or her personal norm function (*normative information*), uses this to infer further information about the state of the world, and thereby updates her norm-related beliefs.

We track the changes in beliefs in a simplified manner by presenting them as expectations of some parameters instead of specifying full probability distributions. So, the expected values of γ will represent factual beliefs. To track the changes in (expected) personal norm functions and in beliefs about the personal norm function of the other dictator (normative expectations), we will report only the expected norms, or the consequences that maximize these expected norm functions. For empirical expectations we will similarly track the "expected" choice of the other dictator. In the examples below, we will call these expected values factual beliefs, personal norms, normative and empirical expectations, having in mind the respective beliefs associated with all these objects.

Figure 2 shows how Dictator 1's beliefs develop when $\phi_1 = \phi_2 = 1$, $p = 0.8$, and $\sigma_1 = \sigma_2 = 1$.[11] In period 1, the expected value of γ is 3.5 since $\gamma = 1$ or $\gamma = 6$ with equal probability (the left panel of Figure 2). Dictator 1's personal norm function and normative expectations coincide in period 1 and are equal to 0.78 (dashed lines on the right panel of Figure 2). This means that Dictator 1 believes that the most appropriate consequence is where the receiver gets 0.78; and also that Dictator 2 shares this view. These beliefs coincide because at this point both players have identical information about the state of the world (a common uniform prior). Dictator 1's optimal choice in period 1 will also coincide with her empirical expectations (the action that he believes Dictator 2 will choose; solid line on the right panel of Figure 2). This is because Dictator 1 believes that both players have the same personal norm functions and identical propensities to

[11] Calculations of each type of belief in each period for each combination of signals can be found in Appendix B.

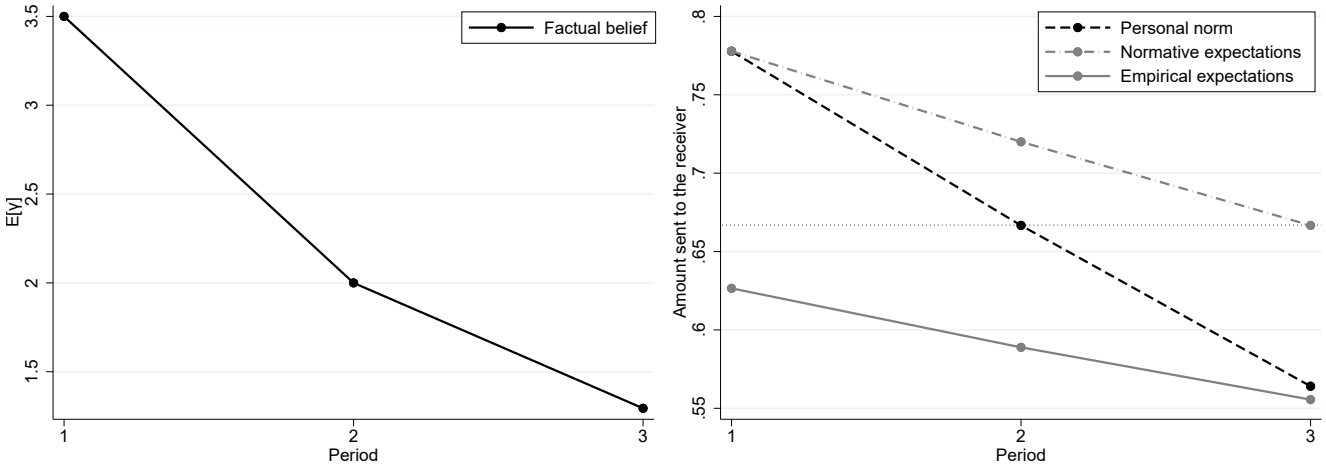

Figure 2: The evolution of the four types of beliefs of Dictator 1 when $\phi_1 = \phi_2 = 1$, $p = 0.8$, $\sigma_1 = 1, \sigma_2 = 1$.

follow them (ϕ_1 and ϕ_2 are equal). More generally, if all players have the same information, then they all have the same personal norm functions and normative expectations. However, own optimal choice and empirical expectations can in principle diverge if players have different ϕ_i's.

In period 2, Dictator 1 receives the signal $\sigma_1 = 1$ and updates her factual belief from 3.5 to $0.8 \cdot 1 + 0.2 \cdot 6 = 2$ (the left panel on Figure 2). Dictator 1's personal norm function gets updated so that the most appropriate (expected) consequence is to give the receiver less than in period 1, since there is a higher chance now that he has low γ. Dictator 1's normative expectations also decrease, but to a lesser degree: the signal $\sigma_1 = 1$ implies an increased probability that $\sigma_2 = 1$, but there still remains a chance that $\sigma_2 = 6$. This leads to the divergence of personal norm functions and normative expectations even in this case when both dictators know that they receive messages from the same noisy signal structure. The decrease in normative expectations naturally leads to a decrease in empirical expectations, informed by Dictator 1's new knowledge of Dictator 2's personal norm function.

Given that ϕ_2 is common knowledge, the observation of either Dictator 2's personal norm function or action in the game will perfectly reveal his signal $\sigma_2 = 1$ since he chooses different actions when receiving different signals. Therefore, in period 3, Dictator 1 after observing the choice of Dictator 2 or being informed of his personal norm function, should update as if she received a second message $\sigma_1 = 1$. Her factual beliefs drop to 1.29, and as a result her personal norm falls as well (the left panel on Figure 2). Knowing that Dictator 2's information is solely the single signal $\sigma_2 = 1$, Dictator 1's normative expectations are now the same as her own personal norm function when she had only one signal $\sigma_1 = 1$ in period 2 (the dotted horizontal line on the right panel of Figure 2). For this same reason, her empirical expectations are now the same as her optimal action in period 2. Overall, this example demonstrates how new factual information deduced from own signal and from observing another person in the same role percolates through the norm-related beliefs.

In our second example we show how observing normative or empirical information can have

different effects on Dictator 1's norm-related beliefs when she has uncertainty over ϕ_2 in addition to the uncertainty over γ. Assume now that, in period 1, Dictator 1 believes that $\phi_2 \in \left\{\frac{5}{8}, 1\right\}$, each with probability $\frac{1}{2}$, and that this is private information of Dictator 2. In this case we have four states of the world corresponding to all possible combinations of different γ and ϕ_2. With these parameters, in period 3, Dictator 2 can choose a low amount (if $\phi_2 = \frac{5}{8}, \sigma_2 = 1$), an intermediate amount (if $\phi_2 = \frac{5}{8}, \sigma_2 = 6$ or $\phi_2 = 1, \sigma_2 = 1$), or a high amount (if $\phi_2 = 1, \sigma_2 = 6$). Thus, Dictator 2's choice in period 3 may not perfectly reveal his signal about γ.

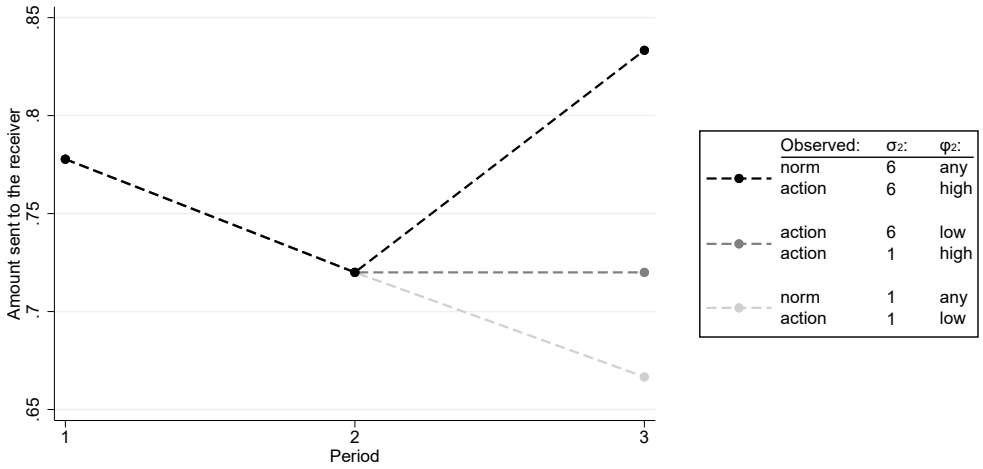

Figure 3: The evolution of Dictator 1's personal norm function with additional uncertainty over ϕ_2.

Suppose as before that $\sigma_1 = 1$. For this case, Figure 3 shows the evolution of Dictator 1's personal norm function for all combinations of Dictator 2's signals and the information about these signals revealed to Dictator 1 in period 3 (she can observe the norm function of Dictator 2 or his action). In periods 1 and 2, before Dictator 2's signals are revealed, the personal norm function of Dictator 1 is not influenced by the information about ϕ_2, so it remains the same as in the previous example. If, in period 3, Dictator 1 observes the personal norm function of Dictator 2 (the paths with "norm observed" on Figure 3), she always updates her own personal norm function, because σ_2 can be directly inferred (black and very light grey lines on Figure 3). However, empirical information (the paths with "action observed" on Figure 3) is only revealing if either Dictator 2's observed choice was low, when it must be that $\phi_2 = \frac{5}{8}$ and $\sigma_2 = 1$, or the choice was high, which happens only when $\phi_2 = 1$ and $\sigma_2 = 6$. In these cases, the observed choice perfectly reveals σ_2, and Dictator 1 updates her personal norm function same way as when directly observing the personal norm function of Dictator 2. When Dictator 2's choice is intermediate, it cannot be determined whether he is a strong rule-follower but believes γ is most likely small ($\phi_2 = 1$ and $\sigma_2 = 1$), or he is less influenced by norms but believes γ is most likely large ($\phi_2 = \frac{5}{8}$ and $\sigma_2 = 6$). In this case, Dictator 1 learns nothing new about the true state of the world and so does not update her personal norm function (middle grey line on Figure 3). The same logic applies to normative expectations, which follow a similar pattern.

In this example, the situation is quite different for empirical expectations, which are shown in Figure 4. If, in period 3, Dictator 1 observes the action of Dictator 2 (low, intermediate, or high amount), this observation becomes the empirical expectation because Dictator 1 knows that Dictator 2 has received no new information besides σ_2 and ϕ_2 (the darkest, the lightest, and the middle lines on Figure 4). However, if she receives only normative information, she never learns anything about ϕ_2, and her empirical expectations are then represented by the distributions over actions given uncertain ϕ_2 (the other two lines on Figure 4).

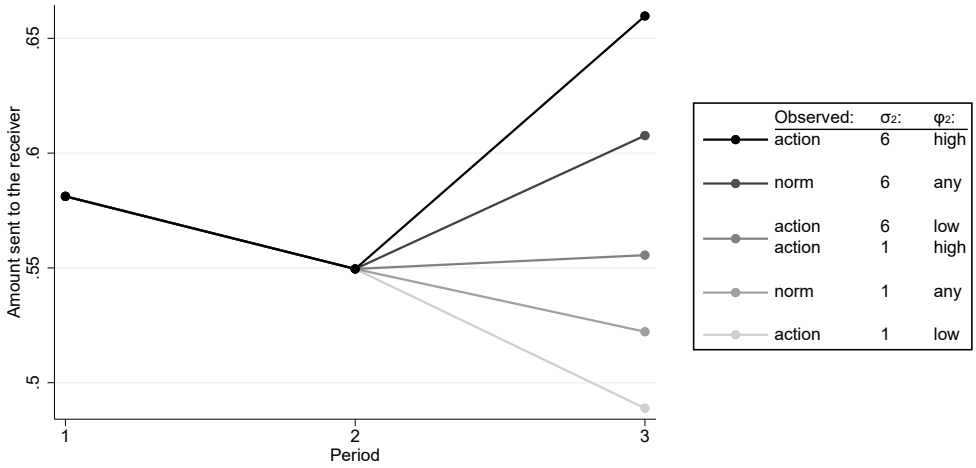

Figure 4: The evolution of Dictator 1's empirical expectations with additional uncertainty over ϕ_2.

Four important intuitions are illustrated by this model. First, it is always the case that, with symmetric information about utilities and payoffs, personal norm functions coincide with normative expectations. If there is no reason to believe others have different information about the consumption values of consequences, then there is no reason to believe their personal norm functions will be different. Thus, differences in observed behaviour can come only from different propensities to follow norms ϕ_i. This suggests that with symmetric information people should tend to interpret deviations from the norm as a signal of selfishness or disregard for norms.

Second, personal norm functions and normative expectations *can differ* when one is aware that others have different information, *but will be correlated* if information about the true state of the world received by different players is correlated, as is the case when players receive messages from the same signal structure. Therefore, if personal norms and normative expectations both independently affect utility, failure to control for both will mean that the effect of changes in one will be wrongly attributed to the other. This suggests that both personal norm functions and normative expectations should be estimated in situations when there is reason to believe that people have correlated signals.

Third, empirical information about choices of others can change behaviour *even if one does not directly care about conforming with others' choices*, because it reveals information about the state of the world and causes people to update their personal norm functions. In this case, it is possible that the new optimal actions will be closer to the observed action, but for purely informational

reasons that have nothing to do with conformism per se. This does not imply though that conformism does not exist (see next section).

Fourth, normative and empirical information (e.g., opinions and actions) can have asymmetric effects when there is uncertainty over the degree to which others follow norms. Specifically, we should expect that, in the world where there is little uncertainty over ϕ_i's, normative and empirical signals should be similar in their ability to reveal new information. Though, empirical information will not be as revealing as normative information because in general some types of players will choose to pool on some actions. However, when there is a lot of uncertainty about ϕ_i's, normative information can influence norm-related beliefs less than empirical information. This is simply due to the fact that in this case knowing someone's personal norm function does not inform players about these individual's propensity to follow them.

We now point out a couple of ways in which our assumptions can be relaxed. The possibility that different groups of people may have fundamentally different normative viewpoints can be accommodated by allowing agents to update their beliefs conditionally on their knowledge of how another agent they are observing forms his own injunctive norms. In other words, as long as you believe you understand the mapping between the states of the world and someone's injunctive norm functions, you will still update your norm-related beliefs based on their actions and normative statements. To give an example, when Donald Trump announces that compelling the public to wear face masks is wrong, his detractors—knowing how he reacts to factual information about the increasing death toll—might conclude that the situation with Covid-19 is getting worse, thus extracting useful information from the normative statement of someone who does not share their normative viewpoint.

In addition, as discussed at the end of the previous section, we would like to emphasize that belief updating can also be substantially affected by bounded rationality. Numerous systematic biases in updating beliefs have been identified, including motivated reasoning, confirmation bias, and so on and so forth. These biases can create distortions in updating of all the norm-related beliefs discussed in this paper, but can be taken into account by explicitly introducing models of bounded rationality into the framework.

2.3 Generalised Norm-Dependent Utility

In the previous sections we discussed how norm-related beliefs can be formed and reshaped by new information. In this section we are looking at how these beliefs might affect decision making through a generalized norm-dependent utility function that includes all four types of norm-related beliefs. This though can be problematic since, as we have seen above, personal norm functions, normative and empirical expectations are interconnected. In what follows, we do not explicitly refer to their relationships, but consider them only *after all updating has taken place*. In other words, from this point we treat these beliefs as separate functions without explicitly

specifying the connections between them. This is not a problem as long as it is understood that the interconnectedness is taken care of beforehand.

It is commonly believed that people have an intrinsic desire to follow norms. In our survey, 92% of respondents stated that people's choices are influenced by personal norms, as we assumed in the model in the previous section, and 96% agreed that people follow "social norms" even when anonymous.[12,13] Norm-dependent utility functions have already been proposed, for example Krupka and Weber (2013) consider a utility function with only material payoffs and normative expectations. Bašić and Verrina (2020) extend this utility function to include personal norms. We simply propose the further inclusion of empirical expectations, giving us the following generalized norm-dependent utility function:

$$W_i(c) = u_i(c) + \phi_i \eta_i(c) + \phi_i^N \eta_i^N(c) + \phi_i^E \delta_i(c). \tag{3}$$

Here ϕ_i, ϕ_i^N, and ϕ_i^E are *intrinsic* propensities to follow personal norms, normative expectations, and empirical expectations respectively. We view these parameters as individual constants that are *context-independent* (see more on this in "Situational Factors and Norm-Following" in Section 3.3). The function $\eta_i^N : C \to [-1, 1]$ represents normative expectations as an "aggregated" norm function that depends on beliefs about others' personal norm functions $(\eta_k)_{k \in N}$, for example averaging across all individuals. Similarly, $\delta_i : C \to [-1, 1]$ is a function that depends on beliefs about the distribution of others' actions, for example the distribution of actions normalised to $[-1, 1]$. These two functions capture the idea that utility is increasing in the degree to which one conforms to what one believes others think is right and what one believes others will do. However, it is not obvious what these functions should look like when the decision maker believes there is heterogeneity in personal norm functions and actions, particularly if there are multiple modes. We return to this point in Section 3.2.

We are not claiming here that these parameters are necessarily all non-zero, which is an empirical question that has yet to be fully addressed. Rather, given the likely correlation in factual beliefs, normative and empirical expectations highlighted by the model in the previous section, failure to begin with such a general utility function runs the risk of biasing estimates of remaining parameters. Finally, some current theories can be seen as imposing constraints on this general empirical model. For example, Elster's definition of norms implies that $\phi_i^N = 0$ under anonymity. The "conditionality" in Bicchieri's definition of a social norm implies that both ϕ_i^N and ϕ_i^E must be strictly positive.

[12] In the question about social norms we did not specify whether we were referring to injunctive or descriptive norms, so it is unclear whether they believed that either normative or empirical expectations, or both were important.

[13] For an explanation of why such preferences may have evolved, see Henrich (2017).

Punishment

It is generally recognised that punishment of norm violators is a crucial mechanism for maintaining social norms, and one that is common in all human societies (Henrich, 2017). Punishment can include present and future financial losses, physical harm, or the psychological cost of experiencing social disapproval. Fear of punishment is a fundamentally different motivation for following social norms than the internalised rule-following tendencies described earlier: it is extrinsic rather than intrinsic. When we move away from the allocation games considered previously (complete impunity), punishment must be modelled explicitly to fully untangle how new information affects norm-related beliefs and behaviour.

It seems natural that the level of punishment one might expect for a given action is increasing in the degree of the norm violation, as expressed in such ancient ideas of justice as "an eye for an eye" and "let the punishment fit the crime" (see Fehr and Fischbacher (2004a) for experimental evidence). KV model this explicitly, considering punishment as a norm in itself, where the most appropriate thing to do is to punish norm violators. They derive the level of punishment owed to a norm violator from the game form, which allows punishment to be included in our framework without any additional complications. Expected punishment is therefore a function of normative expectations, which determines beliefs about level of punishment others see as due for a given action, and beliefs about the rule-following propensity of potential punishers (i.e., the likelihood that others will be willing to pay for conforming to the punishment norm).

An understanding of the connection between norm violation and punishment can have consequences for the updating of norm-related beliefs. For example, on observing someone being punished unexpectedly for an action, one should see that action as less socially appropriate. Also, the possibility of punishment augments the impact of empirical information on normative expectations: when we observe someone violating our perception of an injunctive social norm, this can inform us that the violator does not expect to be punished, which signals that our perception may be wrong or overstated. On the other extreme, the existence of severe punishments creates a situation in which even selfish individuals start to follow norms (rationally preferring it to being punished), which can make empirical information useless for updating normative expectations since everyone is choosing the same action.

Overall, the possibility of punishment can change the way all kinds of norm-related beliefs are updated. So, if in some empirical application there is a reason to expect a tangible effect of punishment or reputation damage, then its influence should be explicitly taken into account in order to cleanly identify the parameters in (3).

2.4 Summary and Discussion

The framework we have outlined flows naturally from essentially only a few basic assumptions: injunctive norms would be agreed upon given symmetric information; people's behaviour is in-

fluenced by norms; there is uncertainty about the world; people can learn from others' behaviour and normative statements. It highlights three channels through which shocks to norm-related beliefs can influence behaviour: via the effect on other norm-related beliefs (Section 2.2); through a direct effect on utility (Section 2.3); and finally, through punishment (Section 2.3). Separately identifying the existence or importance of these channels will require careful measurement and clever experimental designs, and may be challenging in many settings (especially field experiments). These challenges will be addressed in the following section. However, even when one cannot separately identify each factor, bearing in mind our framework will help to interpret norm-related behaviour.

3 Experimental Methods and Research Directions

The purpose of this section is twofold. In addition to proposing various directions of research, informed by the framework described in Section 2, we also aim to provide a reference for readers who are new to the field by explaining various experimental methods that have been developed for the study of norms. This paper is not intended as a review of the literature, and the papers we cite are chosen to illustrate methods and approaches rather than provide an overview of the current state of knowledge. The focus is not on what is known, but on what is unknown.

The structure mimics that of Section 2. In Section 3.1 we discuss building and testing theories of (personal) injunctive norms. In Section 3.2 we focus on norm-related beliefs, first describing methods for measuring such beliefs, then how one might explore their formation and transmission. Section 3.3 relates to the connection between beliefs and actions: how norm-related information can change behaviour, and how to measure individuals' tendencies to follow norms. In this section we also discuss the challenges of eliciting beliefs and actions from the same individual (order effects), and experiments related to punishment.

3.1 Injunctive Norms

There are many potential lines of research that are necessary to further our understanding of injunctive norms. KV show that the model described in Section 2.1 can account for the behaviour in a wide variety of experimental games, however the evidence is based on previous experiments that were not specifically designed to test it. Therefore, there is a need for direct experimental evidence to support the theory or to determine strategic contexts in which it does not work. Some tests along these lines can be found in Panizza et al. (2020a) and Merguei et al. (2020). In addition, many assumptions of the theory need to be directly verified. For example, KV assume that injunctive norms are constructed solely from the dissatisfactions due to unachieved outcomes with *higher* utility. However, it is not obvious that joy from not achieving outcomes that bring *lower* utility does not play a role. In principle, dissatisfaction, joy, or some other re-

actions to counterfactual payoffs can enter the calculation of the norm function η. Evolutionary models in the style of Gavrilets and Richerson (2017) can be used to determine what kinds of norm functions have better survival chances and thus have higher probability of being internalised in the process of evolution. In addition, the theory of KV contains a variety of "cultural" parameters that should be empirically estimated *before* the model can be applied to a specific situation. These include the parameter that defines the importance of punishment (σ in KV) and individual norm-aggregation weights that determine social relationships that can influence the injunctive norms like social status, ingroup/outgroup, kin and ownership claims (see KV). Some of these parameters can be measured by means of third-party Dictator games (see e.g., Chen and Li, 2009), however currently there is no settled experimental methodology to estimate them.

As we mentioned before, Kimbrough and Vostroknutov (2020a) argue that the computation of η can be complex due to the need to aggregate a lot of information about others (their dissatisfactions). The inability to perform these computations can lead to the emergence of moral rules that replace η in some contexts whenever the costs of using a moral rule are lower than the benefits of decreased computational complexity. At this point we do not know in which contexts and to which degree people use moral rules instead of fully-blown injunctive norms η. Kimbrough and Vostroknutov (2020a) describe a methodology that can help to answer such questions. In particular, they show how to determine whether some given moral rule is likely to emerge in a specific class of games. However, a separate theory that would predict which moral rules should emerge in which classes of games can be invaluable for empirical applications, where it is unreasonable to assume that people can compute η, for example because the environment is too complex.

3.2 Norm-Related Beliefs

In this section we first address the question of how to measure the four types of beliefs that provide the foundation of our framework, then consider how to investigate belief formation and transmission.

Measuring Beliefs

There are numerous well-established methods for incentivized elicitation of beliefs about actual outcomes or choices of others under the condition that these events can be later observed and compared to subjects' responses (for a survey, see Schlag et al., 2015). These methods allow us to directly measure a variety of factual beliefs, as well as empirical expectations, and expected punishment. For some factual beliefs, such as the future impact of climate change, the objective truth is unknowable, so cannot be used to incentivize elicitation. Prelec (2004) provides an incentivized mechanism for such cases when one's own beliefs are correlated with those of others.

Measuring personal norms is more problematic, as it is difficult to see how they can be elicited in an incentivized fashion. If it is true that normative and empirical expectations do not influence anonymous behaviour (as suggested by Elster, 2009), then anonymous third party allocation and punishment games unconflated with self-interest (e.g., Fehr and Fischbacher, 2004b) would be incentive-compatible for elicitation of personal norms. However, the role of normative and empirical expectations under anonymity is something that remains to be established (see Sections 2.3 and 3.3).

The most straightforward way to measure personal norms is simply to ask subjects how appropriate is each choice from their perspective (to measure norm functions η_i), or which choice is the most appropriate (to measure norms c_i^*). Traditionally, economists have been suspicious of any beliefs elicited from people who have no explicit incentive to tell the truth. This attitude appears to be softening a little in some quarters, with increasing evidence that unincentivized responses can provide useful information.[14] However, there is good reason to believe that responses to questions about personal norms may in some circumstances be systematically biased: "social desirability bias" would lead subjects to misreport in the direction of their normative expectations, especially for sensitive topics (Krumpal, 2013). Possible partial remedies to ameliorate this concern are credibly ensuring anonymity (Hoffman et al., 1994) and having subjects sign oaths to tell the truth (Jacquemet et al., 2013).

One might think that it is necessary to incentivize elicitation of personal norms in order to achieve a fully incentivized elicitation of the higher-order normative expectations. However, the method introduced in Krupka and Weber (2013), further KW, cleverly sidesteps this problem. Subjects are asked to state how socially appropriate each possible action is, and are paid if they guess the modal response of other subjects. This incentivizes subjects to state not the degree of appropriateness that they themselves believe is correct, but to match others' responses. Formally this is a coordination game with multiple equilibria, but the authors persuasively argue that the truth-revealing equilibrium will be the most likely due to its focality. This mechanism essentially allow us to directly measure η_i^N from the model in Section 2.2.[15] This method has proved popular and has been used in many studies (e.g., Gächter et al., 2017; Barr et al., 2017; Hoeft et al., 2020; Chang et al., 2019; Kassas and Palma, 2019).

A possible limitation of the KW task is that this method only elicits a central tendency of the belief distribution, while other features may also be important (see Section 3.3).[16] d'Adda et al. (2020) introduce a method for eliciting beliefs about the full distribution of personal norms, however it relies on the preliminary unincentivized elicitation of personal norms. First, each

[14]For risk, time, and social preference elicitation see Falk et al. (2016); for a review of experimental comparisons of unincentivized and incentivized belief elicitation see Charness et al. (2020).

[15]In a situation where information is common knowledge (and the subject pool homogeneous), the method of KW also elicits personal norms, which in this context will be identical to normative expectations (see Section 2.2).

[16]In KW, subjects were incentivized to guess the modal appropriateness level. However, other payment schemes can be used to elicit beliefs about the mean or median. See Schlag et al. (2015) for a review of possible methods.

subject is asked to state which action they find personally most appropriate, and is then asked to guess the distribution of other subjects' responses to the initial question; finally, subjects are paid according to how close their guesses are to the true distribution. This allows the computation of not only a central tendency, but also perceptions of the level of consensus on which action is most appropriate (what the authors call the "partiality" of the norm). Note, however, that this method elicits a distribution of the norm c^* rather than the full norm function η.

We believe that there is a room for improvement of the elicitation methods for norm-related beliefs. As discussed in Section 2.3, it is far from clear what aspects of normative expectations are important for behaviour, especially in the cases when there is a perception that personal norm functions are widely dispersed or have multiple modes. Optimal behaviour may depend on the values of η_i and η_i^N in all consequences C as well as on the uncertainty related to these objects. Or perhaps it is simply a central tendency that largely determines behaviour. In order to investigate this question, we need more precise estimates of norm functions. Following this direction, Merguei et al. (2020) propose a modified KW task where appropriateness levels are elicited on a continuous instead of a 4-item Likert scale as in KW. This allows for more accurate estimates that can be used for within-subjects analysis. However, this method still relies on eliciting only the mode of the belief distribution. There is still a need for new methods that can estimate not only the whole norm functions, but also the uncertainty related to them.

Belief Formation

Few studies testing the impact of norm-related information measure its effect on beliefs, and instead look directly at behaviour (see Section 3.3 for some examples). However, an experiment in d'Adda et al. (2020) shows how the impact of normative information on different types of beliefs can be investigated in the lab. Subjects are shown one of three distributions of personal norms regarding the most appropriate decision in a modified dictator game that were elicited in a previous session of a similar experiment: a baseline; one with a lower average but similar variance; and one with a similar average but greater variance. This is followed by elicitation of personal norms, normative expectations, and empirical expectations regarding decisions in the game. They find that being exposed to the distribution with the lower average reduces all three types of beliefs; the high variance distribution has minimal directional impact, but increases the dispersion of responses in the elicitation of normative and empirical expectations.

The impact of empirical information can be studied in much the same way, by informing subjects of the distribution of actual choices of some previous participants. For example, Bicchieri and Xiao (2009) compare the impact of shocks to both empirical and normative expectations on themselves and each other. In two (out of six) treatments, information was provided about what percentage of subjects in an earlier session chose a fair division in a dictator game. In the other two treatments, the information about how many others said one should divide money fairly was provided instead. In two further treatments, combinations of the two types of informa-

tion was given. Both types of information were found to shift both types of beliefs, confirming that empirical expectations can affect normative expectations, and vice versa, as our framework predicts.

Illustrating how explicitly measuring beliefs can aid our understanding of the mechanism underlying behavioural change, Panizza et al. (2020b) study shocks to empirical expectations in mini-allocation games by asking subjects to predict the choices of someone else, who participated in the experiment before. Their experiment compares several competing hypotheses regarding the nature of the shift in behaviour due to these shocks including hypotheses related to norm following. By measuring normative expectations using the KW task, the authors show that the change in behaviour caused by the empirical information are driven by an intermediate impact on these expectations.

Merguei et al. (2020) is one of the few experiments that considers the aggregation of normative expectations η_i^N. Before subjects play the Dictator game with second-party punishment (à la Fehr and Fischbacher, 2004b), each receiver is shown the normative expectations of her dictator, elicited with the KW task. The goal of the study was to estimate how receivers aggregate their own norm-related beliefs with those of the dictator for the purpose of punishing him. The authors find that, out of two available norm functions, receivers choose the cheapest, or the one which prescribes less costly punishment due for the dictator. This can be seen as an opportunistic choice of a norm in case of normative uncertainty and sheds some light on how η_i^N is aggregated from personal norms.

Much work needs to be done to better understand how norm-related information changes beliefs. There are few studies on the degree to which people use Bayesian updating when incorporating new normative information, or suffer from, for example, self-serving bias or confirmation bias. We know little about how uncertainty in the four types of beliefs affects the processing of new information. An interesting topic which has received a lot of attention in the fields of psychology and human evolutionary biology, is how information from different sources is given different weight in the updating of beliefs, for example selective attention to ingroups or those with higher status/prestige (e.g., Rendell et al., 2011; Henrich, 2017). Experimental economics methodology using incentivized games and a formal framework such as the one proposed in this paper, may have a lot to contribute in answering these questions.

3.3 Norm-Related Behaviour

In this section we discuss the connection between norm-related beliefs and behaviour, essentially the nature of the ϕ's in utility function (2). We begin with experiments that shock beliefs and look for a behavioural response that implicitly tests the hypothesis that at least one of ϕ_i, ϕ_i^N, or ϕ_i^E is not equal to zero. As shown in Section 2.2, all kinds of information can affect all kinds of beliefs, so behavioural change resulting from a shock to empirical expectations, for example, does not

imply that $\phi_i^E \neq 0$, as the change might result from an intermediate effect on personal norms or normative expectations.[17] We then move on to experiments that estimate the average value of ϕ_i^N across a population, and finally describe a method for measuring ϕ_i^N at the individual level. We finish with a discussion of experiments on punishment and situational factors on anti-normative behaviour.

Testing the Impact of Information on Behaviour

When looking at the influence of norms and norm-related information on behaviour, it is, of course, not necessary to measure propensities to follow norms, or identify intermediate affects through beliefs. While it may be useful to understand the precise channels through which behavioural change can occur—and is definitely of scientific interest—for many practical purposes it is only the ultimate effect on actions that are of concern. This is particularly the case for field experiments, where the objective is simply to find a solution to a specific problem. Also, in natural field experiments, gathering auxiliary data on beliefs may not be possible.

Numerous lab experiments present subjects with normative or empirical information from subjects in earlier sessions and examine the effect on behaviour, in much the same way as the studies cited in Section 3.2 (all those papers looked at the impact on actions as well as beliefs). An alternative approach is to inform subjects of another subject's choice within the same session (Gächter et al., 2013). Such information may be seen as more relevant or credible than experimenter-supplied information from another session. Such distrust would not be unreasonable, given that typically experimenters choose to give information from extreme sessions, in order to maximize possible treatment effects.

There is a substantial literature of field experiments investigating the potential of empirical or normative information to enact behavioural change in many different domains, with varying success. Empirical information about peer behaviour has been disseminated in a variety of different ways, for example: posters and flyers on university campuses aimed at reducing alcohol consumption (Wechsler et al., 2003); reports on energy usage from an energy company to reduce consumption (Allcott, 2011); and letters from an employer to increase savings (Beshears et al., 2015). Normative information can be transmitted in a number of ways: emoticons signalling social approval or disapproval have been used in household energy consumption reports (Allcott, 2011) and on electronic displays near hand-rub dispensers in hospitals (Gaube et al., 2018); Bursztyn et al. (2018) informed Saudi husbands of the results of a survey of their peers on the acceptability of female labour force participation.

Another valuable tool that researchers in social norms should be aware of is "vignette studies" (Chapter 2, Bicchieri, 2016). A vignette presents subjects with a description of a scenario,

[17] If conditionality is assumed to be part of the definition of a social norm (Bicchieri, 2016), one could also argue that failing to reject $\phi_i^N = 0$ doesn't imply that people do not respond to social norms, but rather that there is no social norm associated with the behaviour under study.

which can be used as the basis for eliciting norm-related beliefs. These are particularly useful when the behaviour under study is not easily replicated in the lab (e.g., child marriage), and when information from the experimenter is unlikely to be effective in shocking beliefs in the field (e.g., because the subject pool have firm beliefs based on long experience regarding the acceptability or prevalence of a behaviour in their community). In Brouwer et al. (2019), vignettes are used to refine understanding of treatment effects in a field experiment, illustrating how the advantages of a field experiment (external validity) can be combined with the advantages of a lab experiment (control and precision of measurement).

Measuring Norm-Following Propensities

The first attempt to measure the propensity to follow norms was undertaken in Krupka and Weber (2013). They used dictator game choices combined with the average normative expectations of a *separate* set of subjects in conditional logit regressions to estimate the population average of ϕ_i^N. They found that it is positive and significant as expected, with subjects willing to sacrifice around \$5-6 to comply with the social norm.

Whether or not it is sufficient to use average normative expectations from subjects who are not making the actual decisions to estimate propensity to follow norms is debatable. If there is an unambiguous social norm, in the sense that normative expectations in the population are reasonably well-aligned, then it is immaterial from whom the normative expectations are elicited. However, theoretically (as shown in Section 2.2) and empirically (Merguei et al., 2020), normative expectations can vary widely, and the measurement error resulting from using an average will bias estimates of ϕ_i^N.

Another reason to elicit actions and normative expectations from separate subjects is to avoid possible order effects. The concern is that subjects may select an action to be consistent with their previously stated norms, or state norms to justify an earlier taken action, inflating the correlation between norms and behaviour. One study (d'Adda et al., 2016) did not find any order effects for the KW task, and at least two papers have analysed the effect of the within-subjects norm elicitation on behaviour (Thomsson and Vostroknutov, 2017; Panizza et al., 2020a). However, it remains to be seen how robust the finding of no order effects is across games and subject pools, so it seems prudent to take precautions when eliciting both actions and normative expectations from the same subjects.[18] Bašić and Verrina (2020) guard against order effects by eliciting normative beliefs online, four weeks before a lab experiment where actions are elicited. If such a time delay is impractical, beliefs can be elicited "behind the veil of ignorance," for example before subjects know whether they will take the role of dictator or receiver in an allocation game (d'Adda et al., 2020).

KW, and the subsequent studies using their method, estimate a utility function which only

[18]Similar issues have been studied with respect to elicitation of empirical expectations, and evidence for the existence, and even direction, of order effects has been mixed. See Schlag et al. (2015) for a survey.

contains utility from money and concern for normative expectations, i.e. they assume $\phi_i = \phi_i^E = 0$. If, as argued in 2.2, personal norm functions and normative expectations are likely to be positively correlated, there will be an omitted variable bias inflating estimates of ϕ_i^N. Bašić and Verrina (2020) follow similar methodology to KW, but elicit both normative expectations and personal norm functions from the same subjects, who make decisions in an assortment of games. They find that while normative expectations and personal norms are highly correlated, there is sufficient variation to estimate the separate impact of each on decision making.

While many questions can be addressed with estimates of population averages, it is also important to investigate heterogeneity in individual propensity to follow norms. This is necessary, for example, in order to better understand the role of individual behaviour in repeated interactions. Measuring individual propensity to follow norms was the goal of Kimbrough and Vostroknutov (2016), who proposed an individual rule-following task to measure ϕ_i directly. In this task and its later modification (Kimbrough and Vostroknutov, 2018), subjects were offered a trade-off between having more money and following an artificial costly rule devised by the experimenters. "Rule-following" subjects chose to forgo significant amounts of money in order to stick to this rule, whereas "rule-breaking" subjects kept most of the money and broke the rule. The important part of the experiment was to see how these individual measures of rule following correlate with behaviour in various social dilemmas. Kimbrough and Vostroknutov (2016) found that rule-followers give more money in the Dictator game, have higher rejection threshold in the Ultimatum game, give higher percentage back in the Trust game, and when grouped with other rule-followers sustain cooperation in the repeated Public Goods game. These findings validated both the measure of ϕ_i obtained from the rule-following task and the norm-dependent utility specification (2).[19]

The advantage of the rule-following tasks in measuring ϕ_i is that they are "assumption-free," in the sense that we do not need to assume anything specific about the norm-dependent utility to obtain these measures. A more assumption-dependent approach is to use "heavy" within-subjects designs. For example, in Panizza et al. (2020a) each subject makes choices in over a 100 different mini-allocation games with two outcomes and two players. These binary choices are then plugged into an individual regression that assumes random utility of the form (2). Individual ϕ_i then comes out as a coefficient on the normative term in the utility. Panizza et al. (2020a) also use the rule-following task within the same subjects and show that the estimates of ϕ_i obtained from the regression correlate with the estimates from the rule-following task, thus showing that the regression estimates are reliable. Notice however that in the regression method the estimates of ϕ_i depend on the assumptions about the shape of the norm function η_i. It is thus possible to measure ϕ_i in a variety of different ways and most likely there are other methods that no one has yet thought about.

[19] Kimbrough and Vostroknutov (2018) also tested the robustness of the Dictator game result in five countries and found even stronger correlation of dictator choices with the proxies for ϕ_i measured in the rule-following task.

One of the most interesting and urgent directions of research suggested by our framework is to carefully establish which of personal norm functions, normative expectations, and empirical expectations influence behaviour the most. As far as we are aware, there are as yet no studies that carefully account for all these beliefs simultaneously, and therefore, given their likely correlations, no convincing evidence for any one of them. Clearly this will be easier in the laboratory, where anonymity can easily remove confounds due to possible changes in expected punishment. However, the relative importance of different types of beliefs are likely to be highly context dependent, so field experiments are crucial, perhaps supported by auxiliary information from vignette studies. One of such field experiments is reported in Rössler et al. (2019).

Another fascinating question is the degree to which a tendency towards rule-following is genetically or culturally determined. Rejection in ultimatum game was found to be heritable (Wallace et al., 2007), but further evidence is required from tasks focussed specifically on norm-following, such as that in Kimbrough and Vostroknutov (2018). To answer this and other related questions, such as whether ϕ_i changes with age, we will need evidence from neuroeconomics and cross-cultural, developmental, and twin studies. Some initial evidence is provided in House et al. (2020) who find that children across different societies start following norms and responding to novel norms at a similar age, suggesting a universal psychology for norm-following.

Punishment

Behaviour consistent with punishment of anti-normative behaviour can be observed in many experimental games, such as rejection of unfair offers in the Ultimatum Game. However, to narrow down the alternative explanations for such behaviour, separate punishment mechanisms must be introduced: after a game is played, subjects can be allowed to choose to pay for the destruction of the earnings of other participants in the game (Fehr and Gachter, 2000). Paying a cost to enforce norms is termed "costly" or "altruistic" punishment. Such punishment can be viewed as altruistic in the sense that that one is sacrificing one's own welfare to support social norms, which benefits society as a whole.

Such "second-party punishment" still admits the possibility that punishment is driven by personal revenge rather than upholding social norms. To eliminate this explanation, "third-party punishment" can be introduced. In such games, subsequent to decisions being made in a game, a subject who is otherwise unaffected by the other players' choices may engage in costly punishment (Fehr and Fischbacher, 2004a). Both second- and third-party punishment was frequently observed in the lab, and was typically effective in supporting cooperative behaviour. However, in the early experiments, no retaliation by the punished party was possible. Allowing retaliation makes subjects less willing to punish (Nikiforakis, 2008), reducing it's effectiveness.

There are few field experiments on norm enforcement. The most widely used paradigm was introduced in Balafoutas and Nikiforakis (2012) to study violations of an anti-littering norm. Here, an associate of the experimenter throws rubbish on the ground in front of others, while a

second researcher records responses and characteristics of potential responders. Overt punishment (e.g., verbal reprimand) of littering is rarely observed in the field (4% in the original experiment), largely due to fear of retaliation. Higher levels of norm-enforcement can be observed using an experimental design that allow for *indirect punishment*, which is unlikely to result in counter-punishment: here the litterer subsequently drops the contents of a bag, and onlookers can punish the norm violation by neglecting to assist (Balafoutas et al., 2014). Brouwer et al. (2019) uses this paradigm to examine whether parents use punishment of third parties as a tool to teach social norms to their children. This study also elicits normative expectations from a separate subject pool using a vignette study to further investigate the cause of observed treatment effects, illustrating how the advantages of a field experiment (external validity) can be combined with the advantages of a lab experiment (control and precision of measurement).

A promising space for future field experiments on norm enforcement, which has not, to the best of our knowledge, been thus far exploited, is the internet. Online experiments have some advantages over offline: they are relatively cheap, and people are notoriously happy to criticize perceived norm violators on message boards and social media sites, ensuring sufficient norm enforcement to observe treatment effects. It is also a socially important area to study, as online shaming can have severe consequences (Ronson, 2017).

Situational Factors and Norm Following

There are situational factors, besides informational conditions, that may affect people's tendency to follow norms. These may provide useful mechanisms to alter norm-related behaviour. The most frequently studied is anonymity. Using KW style estimation, both Krysowski and Tremewan (2020) and Bašić and Verrina (2020) find evidence that anonymity decreases willingness-to-pay to adhere to social norms. It may be tempting to conclude from the parameter estimates that the intrinsic desire to follow norms is reduced by anonymity. However, neither study controls for expected punishment, which is likely to be higher without anonymity. Given that the expected level of punishment is most likely positively correlated with the social unacceptability of an action, failing to control for it explicitly will inflate estimates of ϕ_i^N in the treatments without anonymity, where punishment is possible. Further studies are required to fully understand the mechanisms through which anonymity impacts norm following (Krysowski and Tremewan (2020) find some evidence that norms themselves may differ depending on the degree of anonymity).

The "broken windows effect" is another example of a situational factor that may influence norm following. The name comes from Wilson and Kelling (1982) who suggested that more serious crime can be reduced by cracking down on more minor but highly visible offences such as vandalism and littering. This type of effect can be studied using field experiments where the experimenter manipulates the untidiness of the environment, such as Ramos and Torgler (2012), who find that academics are more likely to litter in an already disorderly department

lounge. This effect could be caused either by the disorder serving as empirical information that influences norm-related beliefs (Section 2.2), or as evidence that norm violations are unlikely to be punished (Section 2.3). In a similar vein, Berger and Hevenstone (2016) show that people are less likely to punish littering in a train station when there is already rubbish around. Here the empirical information implied by the rubbish could shift beliefs about the strength of either an anti-littering norm or a norm of punishing littering.

Note that the effects of both anonymity and "broken windows" can be explained within our framework, without complicating the model by assuming that the *intrinsic* desire to follow norms can be situation dependent. We think that this also is likely to be true of other situational factors and, for reasons of parsimony, recommend carefully considering the channels described in Section 2 before adding further parameters.

4 Conclusion

In this paper we have outlined a framework that describes how rational norm-following agents make decisions and process various types of information to update their norm-related beliefs. It shows how uncertainty of various kinds affects moral judgment that in its turn influences behaviour. On the theoretical side, we demonstrate with simple examples that norm-related beliefs—personal norms, normative and empirical expectations—are interconnected, and that a change in any of them can result in updating of the rest. This suggests that on the empirical side care should be taken when estimating the effects of various beliefs on behaviour. Specifically, given that all types of norm-related beliefs are correlated, statistical misinterpretation of their effects is possible when only some are taken into account. This implies that a proper analysis of social behaviour should take into account much more features of the context than is typical in experiments today. This paper points out what these features are and how to measure some of them. Nevertheless, the framework is new and many details are still untested. We believe that the current experimental methodology for studying norms is already reasonably well-developed and allows any researcher to apply our framework in the lab or field. However, for many types of questions, further development and testing of experimental methodology is still necessary.

The framework suggests a set of parameters that are relevant for the understanding of normative decision making and that can be measured experimentally. These include the parameters that were introduced by KV: the propensity to follow norms ϕ_i and various "social" parameters that enter the calculation of personal injunctive norm function. Apart from this, it also suggests some new variables that can have dramatic influence on information transmission and interpretation. One of these new variables is information quality pertaining to a specific agent: if it is believed that this agent reliably provides good information (she is known to receive high quality signals or is more capable of interpreting them), then her influence on the beliefs and the behaviour of others can be much stronger than that of the agent who is believed to possess

only low quality information or is incapable of interpreting it (Morgan et al., 2012; Vostroknutov et al., 2018).

On a final note, we do not claim that our framework is the ultimately right one to study social norms. As we have shown in the second half of the paper, the theoretical foundation of the framework is new and is not yet thoroughly tested, which can lead to revisions of some "modules" of the theory. However, we believe that even without understanding all the details perfectly, this framework has the capacity to structure our thinking about normative decision making and makes much clearer all the intricate connections between facts, beliefs, and behaviour.

References

Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics 95*(9-10), 1082–1095.

Balafoutas, L. and N. Nikiforakis (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review 56*(8), 1773–1785.

Balafoutas, L., N. Nikiforakis, and B. Rockenbach (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences 111*(45), 15924–15927.

Barr, A., T. Lane, and D. Nosenzo (2017). On the social appropriateness of discrimination. Technical report, CeDEx Discussion Paper Series.

Bašić, Z. and E. Verrina (2020). Social norms, personal norms and image concerns. working paper.

Berger, J. and D. Hevenstone (2016). Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows. *Rationality and society 28*(3), 299–319.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015). The effect of providing peer information on retirement savings decisions. *The Journal of finance 70*(3), 1161–1201.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C. and E. Xiao (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making 22*(2), 191–208.

Brouwer, T., F. Galeotti, M. C. Villeval, et al. (2019). Teaching norms in the streets: An experimental study. Technical report.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2018). Misperceived social norms: Female labor force participation in saudi arabia. Technical report, National Bureau of Economic Research.

Chang, D., R. Chen, and E. Krupka (2019). Rhetoric matters: a social norms explanation for the anomaly of framing. *Games and Economic Behavior*.

Charness, G., U. Gneezy, and V. Rasocha (2020). Experimental methods: Eliciting beliefs. Technical report, mimeo.

Charness, G. and A. Schram (2013). Social and moral norms in allocation choices in the laboratory. mimeo, University of California at Santa Barbara and University of Amsterdam.

Chen, Y. and S. X. Li (2009). Group identity and social preferences. *American Economic Review 99*(1), 431–57.

d'Adda, G., M. Drouvelis, and D. Nosenzo (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics 62*, 1–7.

d'Adda, G., M. Dufwenberg, F. Passarelli, and G. Tabellini (2020). Social norms and private values.

Elster, J. (1989). Social norms and economic theory. *Journal of economic perspectives 3*(4), 99–117.

Elster, J. (2009). Norms. In *The Oxford handbook of analytical sociology*.

Falk, A., A. Becker, T. J. Dohmen, D. Huffman, and U. Sunde (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences.

Fehr, E. and U. Fischbacher (2004a). Third-party punishment and social norms. *Evolution and human behavior 25*(2), 63–87.

Fehr, E. and U. Fischbacher (2004b). Third-party punishment and social norms. *Evolution and human behavior 25*(2), 63–87.

Fehr, E. and S. Gachter (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90*(4), 980–994.

Fehr, E. and I. Schurtenberger (2018). Normative foundations of human cooperation. *Nature Human Behaviour 2*(7), 458–468.

Gächter, S., L. Gerhards, and D. Nosenzo (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review 97*, 72–86.

Gächter, S., D. Nosenzo, and M. Sefton (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association 11*(3), 548–573.

Gaube, S., D. Tsivrikos, D. Dollinger, and E. Lermer (2018). How a smiley protects health: A pilot intervention to improve hand hygiene in hospitals by activating injunctive norms through emoticons. *PloS one 13*(5), e0197465.

Gavrilets, S. and P. J. Richerson (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences 114*(23), 6068–6073.

Gintis, H. (2010). Social norms as choreography. *politics, philosophy & economics 9*(3), 251–264.

Gössling, S., A. Humpe, and T. Bausch (2020). Does 'flight shame'affect social norms? changing perspectives on the desirability of air travel in germany. *Journal of Cleaner Production*, 122015.

Henrich, J. (2017). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hoeft, L., W. Mill, and A. Vostroknutov (2020). Normative acceptance of power abuse. mimeo, University of Mannheim, MPI Bonn, Maastricht University.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith (1994). Preferences, property rights, and anonymity in bargaining games. *7*, 346–380.

House, B. R., P. Kanngiesser, H. C. Barrett, T. Broesch, S. Cebioglu, A. N. Crittenden, A. Erut, S. Lew-Levy, C. Sebastian-Enesco, A. M. Smith, et al. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour 4*(1), 36–44.

Jacquemet, N., R.-V. Joule, S. Luchini, and J. F. Shogren (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management 65*(1), 110–132.

Kassas, B. and M. A. Palma (2019). Self-serving biases in social norm compliance. *Journal of Economic Behavior & Organization 159*, 388–408.

Kimbrough, E. and A. Vostroknutov (2016). Norms make preferences social. *14*(3), 608–638.

Kimbrough, E. and A. Vostroknutov (2018). A portable method of eliciting respect for social norms. *Economics Letters 168*, 147–150.

Kimbrough, E. and A. Vostroknutov (2020a). Injunctive norms and moral rules. mimeo, Chapman University and Maastricht University.

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

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity 47*(4), 2025–2047.

Krupka, E. L. and R. A. Weber (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association 11*(3), 495–524.

Krysowski, E. and J. Tremewan (2020). Why does anonymity make us misbehave: different norms or less compliance? *Economic Inquiry*. forthcoming, https://doi.org/10.1111/ecin.12955.

Laland, K. N. (2018). *Darwin's unfinished symphony: how culture made the human mind*. Princeton University Press.

Merguei, N., M. Strobel, and A. Vostroknutov (2020). Moral opportunism and excess in punishment decisions. mimeo, Maastricht University.

Morgan, T. J. H., L. E. Rendell, M. Ehn, W. Hoppitt, and K. N. Laland (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society of London B: Biological Sciences 279*(1729), 653–662.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics 92*(1-2), 91–112.

Panizza, F., A. Vostroknutov, and G. Coricelli (2020a). Meta-context in social decisions. mimeo, Maastricht University, University of Trento, and University of Southern California.

Panizza, F., A. Vostroknutov, and G. Coricelli (2020b). Norm conformity leads to extreme social behavior. mimeo, University of Trento, Maastricht University, University of Southern California.

Prelec, D. (2004). A bayesian truth serum for subjective data. *science 306*(5695), 462–466.

Ramos, J. and B. Torgler (2012). Are academics messy? testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics 8*(3), 563–577.

Rendell, L., L. Fogarty, W. J. Hoppitt, T. J. Morgan, M. M. Webster, and K. N. Laland (2011). Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences 15*(2), 68–76.

Ronson, J. (2017). So you've been publicly shamed, 2015. *London: Picador 217*.

Rössler, C., H. Rusch, and T. Friehe (2019). Do norms make preferences social? supporting evidence from the field. *Economics Letters 183*, 108569.

Schlag, K. H., J. Tremewan, and J. J. Van der Weele (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics 18*(3), 457–490.

Simon, H. (1990). A mechanism for social selection and successful altruism. *250*(4988), 1665–1668.

Thomsson, K. and A. Vostroknutov (2017). Small-world conservatives and rigid liberals: Attitudes towards sharing in self-proclaimed left and right. *135*, 181–192.

Vostroknutov, A. (2020). Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik 42*(1), 3–39.

Vostroknutov, A., L. Polonio, and G. Coricelli (2018). The role of intelligence in social learning. *Scientific reports 8*(1), 6896.

Wallace, B., D. Cesarini, P. Lichtenstein, and M. Johannesson (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences 104*(40), 15631–15634.

Wechsler, H., T. E. Nelson, J. E. Lee, M. Seibring, C. Lewis, and R. P. Keeling (2003). Perception and reality: a national evaluation of social norms marketing interventions to reduce college students' heavy alcohol use. *Journal of studies on alcohol 64*(4), 484–494.

Wilson, J. Q. and G. L. Kelling (1982). Broken windows. *Atlantic monthly 249*(3), 29–38.

A Survey and Results

This survey was advertised to members of the Economic Science Association, and received 201 responses.

In the following questions, please choose the option that is the closest to your personal views.

1. Which of the following statements do you most agree with?:
   - Social preferences are a subset of social norms **(21%)**
   - Social norms are a subset of social preferences **(20%)**
   - Social preferences and social norms describe completely distinct phenomena **(46%)**
   - None of the above **(12%)**

2. Can an individual's behaviour be influenced by social preferences when they know they are not observed (i.e. acting under complete anonymity)?
   - Yes **(97%)**
   - No **(3%)**

3. Can an individual's behaviour be influenced by social norms when they know they are not observed (i.e. acting under complete anonymity)?
   - Yes **(96%)**
   - No **(4%)**

4. Do you believe that people have personal norms that can influence their behaviour, regardless of what other people think or do?
   - Yes **(92%)**
   - No **(8%)**

5. Do you think coordination on one of multiple equilibria is a sufficient definition of the term "social norm"?
   - Yes **(16%)**
   - No **(84%)**

6. Which of the following define a "social norm"? (you can select none, or more than one response)
   (a) a commonly observed behaviour **(49%)**
   (b) a belief about what others believe ought to be done **(85%)**
   (c) what one believes one ought to do **(40%)**
   (d) Other (please specify) **(19%)**

   Combinations of answers, excluding respondents who chose "other":

- none **(0%)**
- (a) only **(4%)**
- (b) only **(22%)**
- (c) only **(1%)**
- (a) and (b) only **(26%)**
- (a) and (c) only **(2%)**
- (b) and (c) only **(18%)**
- (a), (b), and (c) **(21%)**

7. People are following a social norm when (you can select none or more than one)

   (a) they conform to what they think other people do **(66%)**
   (b) they follow what they believe they ought to do **(46%)**
   (c) they follow what they believe others believe ought to be done **(88%)**
   (d) Other (please specify) **(7%)**

   Combinations of answers, excluding respondents who chose "other":

   - none **(0%)**
   - (a) only **(3%)**
   - (b) only **(3%)**
   - (c) only **(17%)**
   - (a) and (b) only **(3%)**
   - (a) and (c) only **(32%)**
   - (b) and (c) only **(11%)**
   - (a), (b), and (c) **(32%)**

8. Social norms are about

   - actions **(54%)**
   - outcomes **(1%)**
   - both **(44%)**

9. Social preferences are about

   - actions **(15%)**
   - outcomes **(23%)**
   - both **(62%)**

10. People may think that something ought to be done, just because they often observe others doing it

    - true **(90%)**
    - false **(10%)**

B Derivations

B.1 Asymmetric information on payoffs and perfect information on norm-following

Let $\lambda \in \{1, \bar{\lambda}\}$, each with equal probability (in the main text, $\bar{\lambda} = 6$). Assume for now that players only care directly about personal norm, and place weight $\phi_1 = \phi_2 = \phi$ on normative valence (i.e. $u(c) = 1 - c + \phi\eta_P(c)$). Assume also that ϕ is sufficiently large that dictators always give strictly positive amount.

Useful equations:

- Let q be the probability with which a dictator believes $\gamma = 1$.
- Expected dissatisfaction:

$$D(c) = q\left[\frac{c^2}{2} + \frac{(1-c)^2}{2}\right] + (1-q)\left[\frac{c^2}{2} + \frac{\gamma(1-c)^2}{2}\right] = \frac{c^2}{2} + \frac{E(\gamma)(1-c)^2}{2}$$

- $\eta_{E(\gamma)} = \frac{2(1+E(\gamma))c}{E(\gamma)^2}\left[2E(\gamma) - (1+E(\gamma))c\right] - 1$
- Norm: $c_i^*(E(\gamma)) = \frac{E(\gamma)}{1+E(\gamma)}$
- Utility maximizing choice: $\hat{c}(E(\gamma)) = \frac{4\phi E(\gamma) + (4\phi-1)E(\gamma)^2}{4\phi(1+E(\gamma))^2}$

Let γ_P be the expected value of the receiver's multiplier given all information (i.e. signal, observation of others actions, etc.). Let γ_N be belief about the *other dictator's* belief about the expected value of the receiver's multiplier.

Period 1

- $\gamma_P = \gamma_N = \frac{1+\bar{\gamma}}{2}$
- Personal norm and normative expectations (same in first period because info symmetric): $c_P^* = c_N^* = \frac{\gamma_P}{1+\gamma_P} = \frac{1+\bar{\gamma}}{3+\bar{\gamma}}$
- Empirical expectations and action (same in first period because info symmetric): $c_E^* = \frac{4\phi\gamma_P + (4\phi-1)\gamma_P^2}{4\phi(1+\gamma_P)^2}$

Period 2

- $P(\gamma = 1 | \sigma_1 = 1) = p$ (by definition of p)
- If $\sigma_1 = 1$, $\gamma_P = \gamma_P^1 = p + (1-p)\bar{\gamma}$, else $\gamma_P = \gamma_P^\gamma = 1 - p + p\bar{\gamma}$

- $$\begin{aligned} P(\sigma_2 = 1|\sigma_1 = 1) &= P(\sigma_2 = 1|\gamma = 1)P(\gamma = 1|\sigma_1 = 1) + P(\sigma_2 = 1|\gamma = \bar{\gamma})P(\gamma = \bar{\gamma}|\sigma_1 = 1) \\ &= \frac{P(\sigma_2=1)P(\gamma=1|\sigma_2=1)}{P(\gamma=1)}p + \frac{P(\sigma_2=1)P(\gamma=\bar{\gamma}|\sigma_2=1)}{P(\gamma=\bar{\gamma})}(1-p) \\ &= \frac{\frac{1}{2}p}{\frac{1}{2}}p + \frac{\frac{1}{2}(1-p)}{\frac{1}{2}}(1-p) \\ &= 2p^2 - 2p + 1 \end{aligned}$$
- Personal norm after $\sigma_1 = 1$: $c_P^*(\gamma_P^1) = \frac{p+(1-p)\bar{\gamma}}{1+p+(1-p)\bar{\gamma}}$
- Personal norm after $\sigma_1 = \gamma$: $c_P^*(\gamma_P^\gamma) = \frac{1-p+p\bar{\gamma}}{2-p+p\bar{\gamma}}$
- Normative expectations after $\sigma_1 = 1$:

$$\begin{aligned} c_N^* &= P(\sigma_2 = 1|\sigma_1 = 1)c_P^*(\gamma_P^1) + P(\sigma_2 = \gamma|\sigma_1 = 1)c_P^*(\gamma_P^\gamma) \\ &= (2p^2 - 2p + 1)c_P^*(\gamma_P^1) + (2p - 2p^2)c_P^*(\gamma_P^\gamma) \end{aligned}$$

- **Personal norm is different from normative expectations.**
- Empirical expectations if $\sigma_1 = 1$:

$$c_E^* = P(\sigma_2 = 1|\sigma_1 = 1)\hat{c}(\gamma_N^1) + P(\sigma_2 = \gamma|\sigma_1 = 1)\hat{c}(\gamma_N^\gamma)$$

- Choice if $\sigma_1 = 1$: $\hat{c}(\gamma_P^1)$

Period 3

- Both personal norms and actions differ depending on signal, so observing other's personal norm or action perfectly reveals the signal they received. Dictator 2 receives no extra info, so choice and personal norm won't change and Dictator 1's empirical expectations in Period 3 are Dictator 2's choice in Period 2 (either directly observed, or inferred from info on personal norm).
- If $\sigma_1 = 1$ and $\sigma_2 = \bar{\gamma}$:
  - $\gamma_{P,2} = \frac{1+\bar{\gamma}}{2}$ (signals contradict so uninformative)
  - Personal norm: $c_P^* = \frac{1+\bar{\gamma}}{3+\bar{\gamma}}$
  - Normative expectations: $c_N^* = \frac{1-p+p\bar{\gamma}}{2-p+p\bar{\gamma}}$
  - Empirical expectations: $\hat{c}(1 - p + p\bar{\gamma})$
- If $\sigma_1 = 1$ and $\sigma_2 = 1$:
  - $\gamma_{P,2} = \frac{p^2}{p^2+(1-p)^2} + \frac{(1-p)^2}{p^2+(1-p)^2}\bar{\gamma}$
  - Personal norm: $c_P^* = \frac{\gamma_{P,2}}{1+\gamma_{P,2}}$
  - Normative expectations: $c_N^* = \frac{p+(1-p)\bar{\gamma}}{1+p+(1-p)\bar{\gamma}}$
  - Empirical expectations: $\hat{c}(p + (1-p)\bar{\gamma})$

B.2 Asymmetric information on both payoffs and norm-following

Set up same as before but rule-following parameter of Dictator 2 (ϕ_2) is random and private information:

- $\phi_2 \in \{\phi_L, \phi_H\}$, each w.p. $\frac{1}{2}$.
- Suppose ϕ_L is such that a Dictator's choice is the same for $\sigma_2 = 1, \phi_2 = \phi_H$ and $\sigma_2 = \bar{\gamma}, \phi_2 = \phi_L$.

Choice, personal norms, and normative expectations in Periods 1 and 2 are same as in previous section, because they are independent of ϕ_2.

Period 1

- Empirical expectations: $c_E^* = \frac{1}{2}\frac{4\phi_L\gamma_P+(4\phi_L-1)\gamma_P^2}{(1+\gamma_P)^2} + \frac{1}{2}\frac{4\phi_H\gamma_P+(4\phi_H-1)\gamma_P^2}{(1+\gamma_P)^2}$

Period 2

- Empirical expectations if $\sigma_1 = 1$: average of equivalent expressions in previous section over $\phi \in \{\phi_L, \phi_H\}$

Period 3

We now need to treat observation of Dictator 2's personal norm and action separately.

Dictator 1 observes Dictator 2's Period 2 personal norm - shock to normative expectations

- Personal norm perfectly reveals Dictator 2's signal, so personal norms and normative expectations same as in previous section.
- Empirical norms are averages of equivalent expression in previous section over $\phi \in \{\phi_L, \phi_H\}$.

Dictator 1 observes Dictator 2's Period 2 choice ($c_{2,2}$) - shock to empirical expectations

- If $c_{2,2} = \hat{c}(\phi_L, \gamma_P^1)$ or $c_{2,2} = \hat{c}(\phi_H, \gamma_P^\gamma)$, then *both* signal and ϕ_2 perfectly revealed, and we are back to the equivalent expressions in previous section. **Empirical expectations are different from the case where norms are observed, because we know ϕ_2 (personal norms and normative expectations are the same).**
- If $c_{2,2} = \hat{c}(\phi_L, \gamma_P^\gamma) = \hat{c}(\phi_H, \gamma_P^1) = c_M$:
  - Personal norm and normative expectations same as Period 2, because Dictator's action uninformative about the true state (and therefore about σ_2).
  - Empirical expectation: $c_E^* = c_{2,2}$ (Dictator 2 gets no new info)