CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2019-02

Cristina Bicchieri and
Eugen Dimant

Nudging with Care: The Risks and Benefits of Social Information

April 2019

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

The most recent version of the working paper can always be downloaded following this link.

Nudging with Care: The Risks and Benefits of Social Information

Cristina Bicchieri[1,2,**], Eugen Dimant[1,3]

[1] *University of Pennsylvania, Behavioral Ethics Lab*
[2] *Wharton School of Business*
[3] *Identity and Conflict Lab*

Abstract

Norms and nudges are both popular types of interventions. Recent years have seen the rise of 'norm-nudges' - nudges whose mechanism of action relies on social norms, eliciting or changing social expectations. Norm-nudges can be powerful interventions, but they can easily fail to be effective and can even backfire unless they are designed with care. We highlight important considerations when designing norm-nudges and discuss a general model of social behavior based on expectations and conditional preferences. We present the results of several experiments where norm-nudging can backfire, and ways to avoid these negative outcomes.

Keywords: Norm-Nudges, Nudge, Social Information, Social Norms
JEL: B41, D01, D9

1. Introduction

In recent years, public policy has been paying serious attention to what prevents individuals or groups from adopting beneficial practices or abandoning harmful ones. Nudges are a popular behavioral approach, built on the assumption that individuals often make sub-optimal decisions and that mild nudges can lead them to behave in beneficial ways. Behavioral nudges draw on cognitive biases that are difficult to avoid, and aim to fill the intention-behavior gap. For example, they have been employed to help people stop smoking, take their medications, or save more in their 401(k) plans. Nudges work in a simple

*We would like to thank Syon Bhanot, Jon Jachimowicz and two anonymous referees for helpful comments and suggestions.
**Corresponding author
Email addresses: cb36@sas.upenn.edu (Cristina Bicchieri), edimant@sas.upenn.edu (Eugen Dimant)

This version: April 19, 2019

and economical way by re-framing the choice architecture to redirect behavior, without forbidding any option or significantly changing economic incentives (Thaler and Sunstein, 2008). Nudges are not restricted to situations where they make choices easier or exploit inertia and procrastination. They may also involve giving people social information, as when taxpayers are alerted that a majority pay taxes on time (Hallsworth et al., 2016), or reminding them that some behaviors are inappropriate, as when drivers in Bogota were publicly given green "thumbs up" or red "thumbs down" depending on their driving behavior (The Guardian, 2013). In these cases, people are nudged towards behaviors that are socially beneficial, as opposed to behaviors that only benefit the individual. What needs to be changed are collective behaviors, especially those that can produce negative externalities for the community. To achieve such changes, nudging with social information about what others do or approve/disapprove of (in the same context) is the tool used to induce behavior change (e.g., Allcott, 2011).

Such nudges implicitly rely on the assumption that the target behaviors are *interdependent*, in the sense that the preference for paying taxes on time or refraining from bad driving can be influenced by the expectation that most people pay their taxes on time (*empirical expectation*) or, as in the case of driving, that bad driving is publicly disapproved of (*normative expectation*).[1] We define a *norm-nudge* as a nudge whose *mechanism* of action relies on *eliciting social expectations* with the intent of *inducing desirable behavior*, under the assumption that individual preferences for performing the target behavior are *conditional on social expectations*. Norm-nudging may provide information about what 'most people' in the same situation do, or what 'most people' in the same situation approve or disapprove of. In the first case, we aim to induce or even change individuals' empirical expectations about how others behave, in the second we aim to induce (or change) their normative expectations about what others believe is the right thing to do.

Norm nudging, to be effective, must correctly identify the mechanisms through which different types of information affect behavior and understand the specific context in which the target behavior occurs (see discussion in Gino et al., 2019). An operational definition of norms is a first step in this direction, which we provide in Section 2 using simple and measurable concepts such as expectations and preferences (Bicchieri, 2006). This is a step forward with respect to the common psychological definition in terms of perceptions of how

[1] Normative expectations are 2nd-order beliefs about the normative beliefs of other people (Bicchieri, 2006).

common or desirable a behavior is (Cialdini and Trost, 1998). Indeed, the simple perception that a behavior is widespread or approved may not induce congruent behavior, unless individual preferences are conditional on such perceptions. Conditionality of preferences, in turn, has to be measured, to assess if social expectations can have an influence on choice.

How to change such (often wrong) beliefs is a topic of great interest in the literature (Bicchieri and Mercier, 2014), a topic we discuss in Section 4. Often we see a combination of non-social beliefs and social expectations, and we need to disentangle their respective influence on the target behavior. As we discuss in Section 2, it is most important to carefully establish the nature of the target behavior, what motivates it, and only subsequently test which intervention may be most effective at changing it. Because of these complexities, norm nudging should *not* be thrown into the same category as all other nudges. Compared to other traditional nudges, norm-nudges require a meaningfully different and nuanced approach in order to be effective, and failing to do so will doom the intervention to failure.

Since norm nudging mainly involves giving information with the aim of changing expectations and thus behavior, in this paper we explore some pitfalls in reporting information about group behavior and beliefs, with the goal of improving norm-nudging practices. Even assuming that conditional preferences do exist, and thus norm-nudging is appropriate, there are several problems that may make norm-nudging ineffective. Here we look in particular to cases in which information is negative or uncertain, as well as the more general problem of the asymmetry between empirical and normative information in the inferences individuals tend to draw from each of them.

For example, pointing out how many others engage in socially harmful behavior may have the unintended consequence of making the behavior natural and permissible, and end up encouraging it. Campaigns against corruption often fall into this trap (e.g., Bicchieri and Ganegonda, 2016; Dimant and Schulte, 2016). Especially when pluralistic ignorance is present, as when there is a divergence between private beliefs and prevailing norms, letting people know how common a bad norm is will only reinforce it, even if it contradicts their personal beliefs or self-interest. Here, it may be effective to convey truthful information about what a majority really thinks. When a norm-nudge is introduced to shift people away from a behavior they enjoy, they will do what they can to reinterpret the behavior in such a way that allows them to keep behaving in the same way without reservations. Uncertain information invites self-serving 'reinterpretations' (Dana et al., 2007). A common uncertainty involves *reference networks*. Reference networks are the strongest influence on behavior: what people in one's ethnic group, gender, religious or political community do

and think exert a much greater influence than people who are perceived as dissimilar (Hogg and Turner, 1987). When disseminating information about what others do or approve of, it is essential to single out the relevant reference network. In our work on sanitation in India, we were surprised to realize that neighbors, unlike family or close friends, were not a reference network for making decisions about sanitation practices (Shpenev et al., 2019). Disseminating information about neighbors' behavior in this case would have little effect.

Asymmetry in what people infer from different kinds of information is another reason why providing social information may be ineffective. Publicizing any behavior, whether good or bad, may lead the receiver to conclude that the behavior is also approved of. In this case empirical information leads to parallel normative conclusions (Eriksson et al., 2015; Lindström et al., 2018; Bicchieri et al., 2019b). Conversely, when getting normative information about common approval of good behavior, one may not infer that most people behave in the appropriate way. Words are not deeds and normative information does not always support parallel empirical conclusions, which is why normative appeals often do not get the desired result. Finally, when normative and empirical information are not congruent, we frequently see that the empirical information exerts a greater pull than the normative one (Bicchieri and Xiao, 2009). A prominent sign prohibiting littering may be disregarded in a place full of garbage. This is not surprising: social norms are obeyed because normative expectations tell us that transgression will prompt negative consequences. Expecting or observing others to misbehave – especially in large numbers – can render punishment ineffective or irrelevant, weakening normative expectations. Whether norm-nudging aims to create a norm to curb anti-social behavior, or to abandon a harmful or inefficient norm, changing social expectations is in order. Such interventions will be more effective if they take into consideration how easy it is for social information to backfire.

We discuss some of the existing literature on the topic of norm nudges in Section 3. In Section 4, we report on experiments we conducted to explore (and provide solutions to) problems created by uncertain and negative social information, as well as by the presence of asymmetries in what we infer from empirical and normative information. We present our conclusions in Section 5.

2. Understanding Social Behavior

An important task any intervention aimed at changing behavior should consider is to diagnose the nature of the target behavior. The first feature of a target behavior we

should understand is whether it is socially *independent* or *interdependent*. When behaviors are independent, the motivation to undertake the behavior is *unconditional* on a person's social expectations. For example, our collective custom of using umbrellas when it rains is motivated by a (shared) need to protect ourselves from water, not by our expectations that other people around us use umbrellas. Similarly, our (typically common) refusal to harm an innocent person is motivated by our belief about what is right, not by our expectation about what those around us approve or do (Bicchieri, 2016, p. 31). In cases of customs and moral norms, we may observe people to behave in similar ways, and even expect others to do likewise, but this does not mean that their social expectations are a motivating factor. Norm-nudging in these cases will be ineffective, as it presupposes that the relevant preferences are conditional on social expectations.

When behaviors are interdependent, the motivation to undertake such behaviors is *conditional* on a person's belief about what is commonly done and/or what is commonly approved within that person's reference network. Think of fashions or fads. In these cases, the desire to imitate the trendy, the successful, or possibly to be correct (as in 'social proof'), motivates conformity to others' behavior. Deutsch and Gerard (1955) were the first to show that, in all these cases, informational and normative components, either alone or together, may motivate conformity. *Normative* here simply means that sometimes conformity is due to social pressure, as when a boy will get a tattoo to be liked and feel accepted in a valued group. Fashions, fads, social proof, and getting a tattoo are all examples of behavior driven by unilateral expectations. Someone following a fashion cares about what "fashion trendsetters" wear, but the reverse is not true. Similarly those who read reports consumers have written on Amazon about refrigerator models care about others' opinions, but the opposite is not true either. In all these cases, being informed about what others in similar situations do may be effective, as the desire to imitate (as in fashions and fads), to be right (as in social proof) or be accepted (as with boys' tattoos) is a powerful motivator.

Another example of interdependent behavior in which expectations are mutual, or multilateral, is coordination. Coordination on language rules, dress codes, etiquette, and all sorts of mutual signals stems from the desire to harmonize our actions with others who are instrumental in helping us reach specific goals. Here, again, expectations about what others do in a similar situation motivate choice via conditional preferences. All cases of interdependent behaviors where preferences are conditional on empirical expectations only are what Bicchieri (2006) calls *descriptive norms*. Note that the term 'descriptive norm' is widely used in the psychological literature to mean the perception of what is commonly

done, what is usual and customary (Schultz et al., 2007). This definition overlooks a most important point: some such behaviors are socially unconditional, but others are not. Collective habits, customs, fashions, and conventions are all "descriptive norms" in the psychologist sense, but it does not consider the (un)conditionality of such. If our goal is to change some of these regularities, we have to be sharper in distinguishing one from the other. For example, since a custom is a pattern of behavior such that individuals (unconditionally) prefer to conform to because it meets their needs, changing it may just call for providing people with better means to satisfy their needs. On the contrary, a common signaling system influences action via the joint force of expectations and the preference for coordinating with others. To change behavior, expectations have to change, and the main challenge here is to induce different (and mutual) empirical expectations.

To summarize, a descriptive norm is a *behavioral pattern* such that individuals prefer to conform to it on condition that they believe that most people in their reference network conform to it (empirical expectation) (Bicchieri, 2016). As we mentioned, the conditional preference makes the difference.

Perhaps the best examples of interdependent behavior are social norms. In a social norm, behaviors are interdependent, as preferences for conformity are conditional on *both* empirical and normative expectations Bicchieri (2006, 2016). Since social norms tell us how we ought (or ought not) to act, they are often confused with injunctive norms (Cialdini et al., 1991; Rivis and Sheeran, 2003). As is the case with descriptive norms, the psychological definition of injunctive norm as what people perceive as desirable or approved of overlooks a most important characteristic: the presence or absence of social conditionality. What we collectively believe ought to be done, what is socially approved or disapproved of, could be a shared moral or religious norm, or a social norm proper (Bicchieri, 2016, p. 31). Moral norms lack social conditionality: we comply with moral norms out conviction about what is right, and do not condition our choice on what others do or believe, whereas conforming to a social norm is *motivated* by social expectations. Changing moral behavior is much harder than changing conformity to a social norm, since we can not rely on just changing social expectations.[2]

[2]This does not mean that what is a moral or religious norm to some may not be a social norm to others. Wearing a veil is a case in point. For many Muslim women, it is a valued sign of religious identity that they wear proudly, while others living in a strict Muslim country may wear it only because sanctions could be severe if they do not.

To summarize: A *social norm* is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectation) (Bicchieri, 2016). Here, too the conditional preference makes the difference. Also note that social norms include both a descriptive component and a normative one, whereas the common definition of injunctive norm only includes a normative component. Social norms, to exist, need *both* components. As we shall see in the Bicchieri, Dimant and Sonderegger (2019b) experiment that we report in Section 4, a social norm that was merely injunctive, telling people what "most others approve of", would be insufficient to induce pro-social behavior. Since a main function of social norms is to curb selfish behavior in favor of collective welfare, it is important that the normative 'should' be supported by evidence of congruent behavior.

We characterize the distinction between the theory we endorse and the usual descriptive/injunctive separation in the following Figure 1:

Four Types of Behavior

	Independent Behavior (Unconditional Preferences)	**Interdependent Behavior (Conditional Preferences)**
Descriptive	*Custom* You **prefer** to do X because **you believe X meets your needs**. Your choice does not depend on others doing X or thinking that you should do X.	*Descriptive Norm* You prefer to do X because **you expect others to do X.** Your choice depends on your **empirical expectations** of others' behavior.
Injunctive	*Moral Rule* You **prefer** to do X because **you believe X is the right thing to do**. Your choice does not depend on others doing X or thinking that you should do X.	*Social Norm* You **prefer** to do X because **you expect others to do X and you believe that others think that you should do X**. Your choice depends on both empirical and normative expectations.

Figure 1: Independent vs. Interdependent behaviors

The horizontal rows represents what are commonly called descriptive and injunctive norms, respectively: The behaviors in the first row refer to the perceived prevalence of a behavior, behaviors in the second row refer to the perceived degree of social approval/disapproval of the behavior (Cialdini and Trost, 1998). The problem with this 'hor-

izontal' definition is that it does not differentiate between conditional and unconditional preferences for the specific behavior. Customs and descriptive norms proper are grouped together, as are moral and social norms. The vertical columns represents a more useful way to differentiate among behaviors. Customs and moral norms are grouped together "vertically" as behaviors that are socially unconditional. Descriptive and social norms are grouped together because both are conditional on social expectations. Norm-nudging will be ineffective with customs and moral norms, effective with descriptive and social norms proper. A first, necessary step to design interventions to effect behavior change is to assess behavior 'vertically'. Not targeting the appropriate behavior, for example by confusing a custom with a descriptive norm proper, will prevent us from achieving the expected results.

Take as an example the many unsuccessful attempts to eliminate the still frequent practice of child marriage (Bicchieri et al., 2014). Any field intervention must first assess the preferences individuals have, the options they have to choose from, and the beliefs they have about these options. Collective practices like child marriage can be sustained by two kinds of preferences, namely (socially) unconditional and conditional preferences, and two kinds of beliefs, namely nonsocial beliefs and social expectations. Child marriage may be a moral rule, a descriptive norm, a social norm proper, or just be a simple, shared traditional custom. Mapping the full range of preferences and beliefs makes it possible to determine what type of practice child marriage is, and design effective interventions. An intervention in this case may take different forms. If we establish that child marriage is a social norm, an intervention aimed at changing both empirical and normative expectations is in order; if it is just a descriptive norm, changing empirical expectations will be sufficient, but if we determine that child marriage is a moral norm, no intervention aiming at changing social expectations will be successful, as informing individuals that others are changing behavior, or that waiting to marry until later is now approved of will not change what is felt to be a moral imperative and an unconditional preference for obeying it. Nor will such an intervention work if people have beliefs about the risks that an unmarried girl will incur (such as rape and unwanted pregnancy) and the advantages of early marriage (better adjustment to the husband and in-laws, more pregnancies, and so on). In this case, these beliefs must change in order for behavior to change.

In sum, norm-nudging, to be effective, requires understanding what motivates individuals to choose particular actions. A key advantage of defining behavior in terms of conditional or unconditional preferences and beliefs (expectations) is that we can independently measure and quantify these primitive constructs (and hence norms). Belief-elicitation pro-

tocols can be used to measure whether individuals hold sufficiently high empirical and normative expectations, and hence to determine whether a *consensus* exists that a norm applies to a given situation. Mutual consistency of second-order beliefs (normative expectations) measures the degree of consensus that a norm applies to a specific situation. Independent and direct norm elicitation (e.g., Bicchieri and Chavez, 2010) can be combined with indirect derivation from subjects' choices (e.g., Krupka and Weber, 2013). Experimental games provide an ideal way to measure compliance, or how much individuals adhere to the norm. For example, even if there is agreement among subjects about a norm of equal sharing in a typical Ultimatum game, an acceptance threshold below the norm and a low offer show low compliance. Consensus about an equal sharing norm may not translate into overall compliance, since many other factors also influence norm compliance. We know that privacy reduces conformity (Allen, 1965; List et al., 2004; Bolton et al., 2019), and factors such as punishment, peer pressure, and social proximity enhances it (Fehr and Gächter, 2000; Mas and Moretti, 2009; Bicchieri et al., 2018b; Dimant, 2019). Preference-elicitation mechanisms can be employed to measure the extent to which preferences for compliance are conditional on these beliefs (e.g., Gächter et al., 2018; Bicchieri et al., 2019a).

Though norm-nudging is often conducted in the field, designing experiments that test the underlying mechanism of the behavior are useful in clarifying if indeed the hypothesized mechanism drives the behavior (Gino et al., 2019).[3] Measuring expectations and conditional preferences can be successfully done in the field, as Bicchieri's work on sanitation in India testifies (Bicchieri et al., 2018a). In Section 4, we present a set of experiments that directly assess whether a norm exists, and under which conditions it will be followed.

3. Norm-Nudges

There has been increasing interest in public policy to explore norm-nudges. For example, Mols et al. (2015) suggest that, since individuals are members of social groups, new norms must be created to successfully change behavior. Reijula et al. (2018) also point out that policy-makers must understand the limitations of nudging, mostly because nudging focuses on *individual behavior change*, whereas we often need *collective change*. As discussed,

[3]By 'mechanisms' we mean what motivates people to choose particular behaviors and both descriptive and social norms rely on different motives. Arguably, individuals are not always aware or consciously processing the information available to them and may obey norms as default rules, without much thinking. Whether norm-nudging involves system 1 or system 2 is an open discussion (Löfgren and Nordblom, 2019).

norm-nudging is clearly useful when behaviors are interdependent. Interdependencies may already exist, as is the case with the provision of public goods, where people have to collectively contribute (e.g., taxation). Alternatively, they may be created, as when individuals who want to lose weight participate in weight loss groups such as Weight Watchers. Even in an interdependent context, it is important to differentiate between descriptive and social norms, as they rely on different kinds of expectations: empirical only in the former, empirical and normative in the latter. Following a descriptive norm is always aligned with an individual's self-interest, but not so conforming to a social norm, since social norms exist precisely to alleviate the tension between individual and collective welfare.

Norm-nudging is typically performed by relaying information about what other people do. In this case, we may want to induce the perception that there exists a descriptive norm by letting people know what others in a similar situation do, as when we inform individuals of what their neighbors are doing. Social influence may work via the desire to imitate (those more influential or more knowledgeable), coordinate with, or be accepted by a valuable group. The hope of intervention designers is that individuals' preference for conformity will be conditional on the information they are given.

Comparing a person to his peers, neighbors, or friends has often proven to be an effective way to change behavior. Allcott (2011) report that American households who got mailers comparing their own electricity consumption to that of their neighbors reduced their consumption as much as they would have if the cost of power had risen by 11-20%. Bhanot (2018) discusses a natural field experiment looking at the water consumption of 40,000 California households. Participants were assigned to a control group and three conditions: No Drop, Drop, and Injunctive Drop. In all three conditions participants were given information about their water usage and how it compared to both the average and the most efficient (top 20%) households. The Injunctive Droplet condition contained a water droplet visual which was either a smiley, neutral, or a frowny face depending on the household's water usage compared to similar households. The results show that the Injunctive Drop condition effectively reduced water consumption significantly. Similarly, Ferraro et al. (2011) report large effects for water consumption where a comparison to neighbors was much more effective than either information provision or normative messages asserting the evils of water overconsumption. Another study by Goldstein et al. (2008) shows that telling hotel guests that a majority of other guests reuse their unwashed towels prompted a large number of guests to do the same. In comparison, making an environmental (normative) appeal to save the water consumed by washing used towels did not have any effect. Though

these studies make very clear the power of empirical information (what others do), we also know that simply being informed about what other people do may not be effective. For example, Schultz et al. (2007) showed that informing homeowners about how much electricity they are using in relation to other homeowners in the area led those who were above average in electricity use to curb their usage and those who were below average to do the opposite. Everyone appeared to trend towards their updated empirical expectations.

There are many reasons why empirical information may be ineffective, or even backfire. An obvious one is misunderstanding the relevant reference network. Norms are properties of groups, not individuals, so it is important to clearly identify the reference network of norm-followers. Simply stated, a reference network for a specific behavior are the people whose actions or beliefs one takes into account when deciding what to do. Uncertainty about the relevant reference network may lower norm compliance. For example, informing individuals that "most people save energy by reducing use of air conditioners at peak times" may lead to several interpretations of who those individuals are. They may be neighbors, or instead people who live in other, different and cooler areas, and in this case, a self-serving interpretation may lead one to think that, in this particular environment, keeping air conditioning at full power is fine. As discussed in Bicchieri et al. (2018b) in Section 4, uncertainty about the relevant reference network led players to discount information about the (high) percentage of players behaving pro-socially. If a specific behavior is common in another group, why should one think it is also common in one's group? Specifying the relevant reference network helps avoid self-serving interpretations. A good example is Hallsworth et al. (2016) study of how informing overprescribing General Practitioners (in the UK) that they prescribed antibiotics more than 80% of other GPs in their area significantly reduced antibiotics prescriptions. Since all GPs are part of the National Health Service and there was no reason to believe any specific difference in their area might exist, the message was effective.

Another reason why empirical information may backfire is when the messenger is not trusted. Stibe and Cugelman (2016) point to credibility and the suspicion of hidden intentions as reasons why the message may not be effective, and the extensive literature on the failure of legal interventions may be useful in understanding why issues of credibility and trust mar otherwise positive messages (Stuntz, 2000). Credibility is also particularly important in cases of pluralistic ignorance (Miller and McFarland, 1987) where a descriptive norm is misperceived. When individuals engage in social comparison and from often limited observation infer common behavior, but cannot transparently communicate their

true preferences, public revelations of real participation rates (if lower than they appear) can have a major impact. Berkowitz and Perkins (1987) have touted the effectiveness of such *belief shocks* on college drinking rates, provided the source of the message is trusted.

Finally, empirical messages may be ineffective because people tend to reject information that is inconsistent with their beliefs. Bicchieri and Mercier (2014) point out how the more central and entrenched a belief is, the more difficult it is to dislodge it. Being informed that most people in one's neighborhood recycle may have no effect on someone who is convinced that recycling pollutes the environment. It has also been shown that past behavior moderates the effect of empirical information, especially when past behavior was well established (Frey and Meier, 2004), probably because the associated beliefs are well entrenched. A possible way to moderate this negative effect is to introduce observability. Yoeli et al. (2013) show that when registration to curb air-conditioning during peak demand periods was made public, participation increased. Observability is often tied to reputation, and reputation effects may dampen the influence of previous beliefs, provided they are not too central or related to one's self-identity. However, recent experimental evidence about observability with respect to a pro-social action (giving to a charity) show that being observed by a stranger in absence of consequences may backfire (Bolton et al., 2019).

Norm-nudging can also appeal to social norms proper. This is typically done by presenting normative messages, either in the form "it is good to do..." or more explicitly stating that "most people approve/disapprove of...". The jury is still out about the effectiveness of such messages. Brent et al. (2017) demonstrated an effect of normative appeal messages in nudging for water conservation. A classic example is Cialdini et al. (1990)'s manipulation of the salience of normative massages. When the environment pointed to a negative descriptive norm (littering) but subjects were exposed to a positive normative message, the latter had a greater effect on behavior. In the already mentioned experiment by Schultz et al. (2007), signaling normative information in addition to the empirical information (by adding a smiling emoticon next to a homeowner's electricity usage if it was below average or a frowning one if it was above average), those who consumed above average continued to reduce their consumption, while those who consumed below average maintained their originally low energy usage (see also Bhanot, 2018).

However, Bicchieri and Xiao (2009) showed that when empirical and normative messages are incongruent, adding the normative message ("most people believe one should...") could be ineffective. In that case, the (negative) empirical information wins. This result makes sense because social norms usually prescribe behavior that may be at odds with

other, narrowly self-interested motives. When we realize that anti-social behavior is frequent, and pro-social behavior has a cost, we do not feel obliged to obey the pro-social norm. The very existence of conditional preferences guarantees it. Note that conditionality implies that, at any given time, a social norm may exist without being followed.[4] If empirical or normative expectations are not met, a norm will not be obeyed. Corruption is an example. Think of two communities that hold similar normative expectations about the inappropriateness of bribing: in both cases, individuals believe that bribing is disapproved, i.e. a no-bribing norm exists. In one community, this general social disapproval for bribing is accompanied by evidence that bribing is infrequent, so that empirical and normative information available to individuals in that community are congruent. In the other community, however, the disapproval for bribing is accompanied by widespread evidence that bribing is common, as empirical and normative information are incongruent. If individuals in the latter community observe a sufficient number of transgressions, they may transgress too, since their compliance is conditional upon what others do. Consequently, compliance (i.e. norm following) may be lower in the latter than in the former community, even if members of both communities hold similar beliefs about what is socially appropriate. As we shall discuss in Section 4, incongruity of normative and empirical expectations is a crucial factor in norm-transgression (Bicchieri et al., 2019a).

Moreover, normative messages alone, especially when cast into a "should" injunction, may provide the wrong type of information: if I need to be told what is the right thing to do, it means many people behave badly. In this case, the message may almost give permission to misbehave. Indeed, Chaudhury et al. (2006) report that, while 25% of Indian teachers and 19% of Bangladeshi teachers are missing from school each day, between 75% and 81% show up. Where teachers have reason to believe that teachers' show up rates are low, signaling the comparatively high attendance rates should be much more powerful than a simple message telling teachers they *should* improve their attendance. This normative message may give the wrong impression that many teachers are indeed absentees (if we have to tell teachers to improve their performance, then it would be reasonable for them to infer that many teachers are not behaving as they "should"), causing teachers to reinforce their negative expectations.

[4]This is not the case with descriptive norms: if a descriptive norm is not followed, it ceases to exist (think of September 3, 1967, when the traffic in Sweden switched from driving on the left-hand side of the road to the right. A coordinated change in expectations immediately induced different behavior).

4. The Critical Role of Norm Information

How norm-nudging information is presented is critical to behavioral results. Often we have (and provide) only one type of information: empirical (what others do) or normative (what others approve/disapprove of). What do people infer from each type of message? Are there asymmetries in the interpretation of these different kinds of information? Can information backfire? In all cases in which we want to nudge collective pro-social behavior, and we provide some normative or empirical information, we must be be alert to the role the decision architecture plays in shaping outcomes, and be careful to avoid the possibility of biased interpretations and belief manipulation. In what follows we show the results from a few recent experiments that highlight the potential pitfalls but also the benefits of norm-nudging.[5]

4.1. The Effects of Explicit Norm Information.

In a recent experiment, Bicchieri et al. (2018b) we examine the effect of punishment on norm conformity, especially when punishment leads to suboptimal outcomes. In particular, we explore whether this negative effect is due to a lack of perceived legitimacy of rule enforcement, as is commonly assumed, or is instead due to the uncertainty of the message, enabling agents to choose how to interpret it and thus justify selfish behavior. In our case, the uncertainty is about the appropriate *reference network*, or the group of players that players can meaningfully compare with. There is a difference, however, between a normative message and an empirical one, as the normative message, even in the presence of the former uncertainty, is much less manipulable than the empirical one. Participants play a standard Trust Game in which we vary the message that trustees receive and examine their effect on the amounts returned to the investor. We vary:

- The presence of weak punishment for non-compliance. If imposed, the punishment is less than the benefit from non-compliance.
- What type of information was provided (no info, empirical info, normative info). The information provided was based on a previous survey in the same trust game context and was of the following form:

[5]For brevity, we only present a shortened version of the experiments and their key results.

1. Empirical info: "*In a previous survey, most participants in the role of a trustee returned at least half of the transferred amount.*"
2. Normative info: "*In a previous survey, most participants said that the trustee should return at least half of the transferred amount.*"

Since non-compliance, which is subject to punishment, is defined as returning less than 50% of the tripled amount sent by the investor, we dub the case in which the investor sends half (all) of her endowment the 'Low Compliance Cost' ('High Compliance Cost') condition. An interesting result in Figure 2 is that when compliance is more expensive, the combination of punishment and empirical information about others' conformity has detrimental effects on reciprocity.

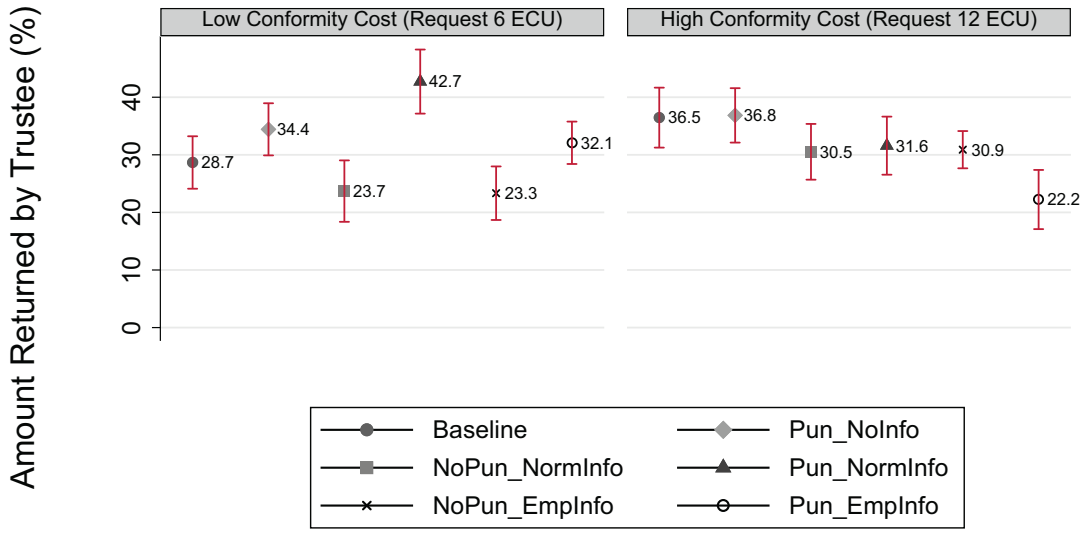

Figure 2: Amounts per Low Compliance Cost vs. High Compliance Cost returned by trustees as percentages of amount received from investors. Baseline: no punishment or norm information; Pun_NoInfo: punishment without norm information; NoPun_NormInfo: no punishment with normative information; Pun_NormInfo: punishment and normative information; NoPun_EmpInfo: no punishment with empirical information; Pun_EmpInfo: punishment and empirical information. Figure appeared originally in Bicchieri et al. (2018b).

An explanation that is in line with recent experimental literature is that individuals attempt to exploit the wiggle-room of norm-based interventions to avoid compliance (Konow,

2000; Dana et al., 2007; Spiekermann and Weiss, 2016). Our findings indicate that individuals tend to choose self-serving beliefs (and behavior) more often when faced with empirical information because it is easier to 'interpret' what other people do, whereas they have a harder time distorting information about what is normatively appropriate. The negative effect of combining empirical information with punishment in the high compliance cost condition seems to be due to the *uncertainty* of the empirical information. Participants in the high cost condition who do not wish to conform may exploit their uncertainty about the *reference network*: the empirical information may refer to the low cost group, the high cost group, or both. Exploiting this uncertainty means forming the belief that the empirical information refers to the low cost group only, as for them it is cheap to comply. For the low cost condition players, this rationalization does not work, since if participants complied in the high cost condition then they surely complied in the low cost condition as well. This manipulation does not occur with the normative message. In our culture, a norm of reciprocity is widely shared, and it would be odd to claim that it only applies to specific groups. Normative manipulation can of course occur, but this usually only happens when different norms apply to the same situation, which is not the case here.[6] One can conclude that being specific as to the relevant reference network may prevent self-serving interpretations, especially when information about what others have done is provided.

4.2. Self-Serving Belief Distortion: How We Process Norm-Uncertainty

In another experiment, Bicchieri et al. (2019b) we explore the relationship between norm-uncertainty and lying. Our goal was to understand how individuals use information (empirical or normative) to form beliefs about whether a norm applies and how belief formation is affected by the source of the uncertainty and awareness of the possibility of subsequent lying. This paper explicitly deals with two problems related to norm-nudging. On the one hand, we show that providing information about an uncertain state of the world may lead, under the right circumstances, to belief distortion. Such distortion is particularly worrisome when it *justifies* anti-social behavior. On the other hand, belief distortion is facilitated by an inherent asymmetry between normative and empirical information.

The experimental design is a modification of the Fischbacher and Föllmi-Heusi (2013) die-under-the-cup paradigm. Participants are engaged in two ways:

[6] For comprehensive analysis of cases in which normative information can be manipulated, see Bicchieri and Chavez (2013).

1. They have to provide their beliefs about an uncertain state of the world (deciding which of two alternative empirical messages such as "*the majority lies/does not lie*" is true or which of two alternative normative messages such as "*the majority approves/does not approve of lying*" is true).

2. Roll a die anonymously and report the outcome. Whether or not the report is truthful is up to the participant.

We examine belief distortion by varying whether participants know about the upcoming choice of rolling the die when they are asked about the true state of the world. Holding the type of uncertain state of the world constant (i.e., empirical or normative information) and only varying whether individuals know that they will have the opportunity to roll the die (and lie), any differences in reported beliefs can be attributed to belief distortion.

An important prediction of our model is that belief distortion of the empirical information will be common because non-compliance in this condition is more costly than in the normative condition. The reason for this is the asymmetry between what we infer from empirical versus normative information, which we verify experimentally. For example, when we get information that "most people do not lie" in our specific context, we typically infer that most people disapprove of lying. Thus believing that most people do not lie and then choosing to lie invites (potential) social disapproval. If instead we are presented with information that "most people disapprove of lying", we do not necessarily infer overall honest behavior to the same degree. Words and deeds are often at odds, and in general individuals are less likely to unambiguously infer good behavior from a positive normative attitude. To justify lying in our context, it is not necessary to believe that most people disapprove of lying. It is instead enough to 'twist' our empirical expectation and convince ourselves that most people lie, because in this case lying must *not* be disapproved of. So if one must choose a state of the world, and there is an advantage to lying, we predict that belief distortion will only occur with empirical messages. Indeed, we find compelling evidence that individuals engage in self-serving belief distortion in order to lie, but we observe belief distortion *only* in the context of uncertainty about what others do (empirical uncertainty), not in the context of uncertainty about what others approve of (normative uncertainty). Figure 3 illustrates the main results. In particular, we find that, in the Empirical information condition, only 47.2% of participants who were not aware of an upcoming cheating opportunity (CPU) believe that a majority of previous participants lied in the same situation, whereas 62.8% of participants who were aware of an upcoming

cheating opportunity (CPK) believe that a majority of previous participants lied in the same situation. This is compelling evidence for belief distortion which, in turn, leads to a significant increase in lying behavior (28.8% vs. 38.7%). In the Normative information condition, we do not find any belief distortion when the true normative state of the world is uncertain, and in consequence lying behavior is unaffected (36.4% vs. 38.5%).

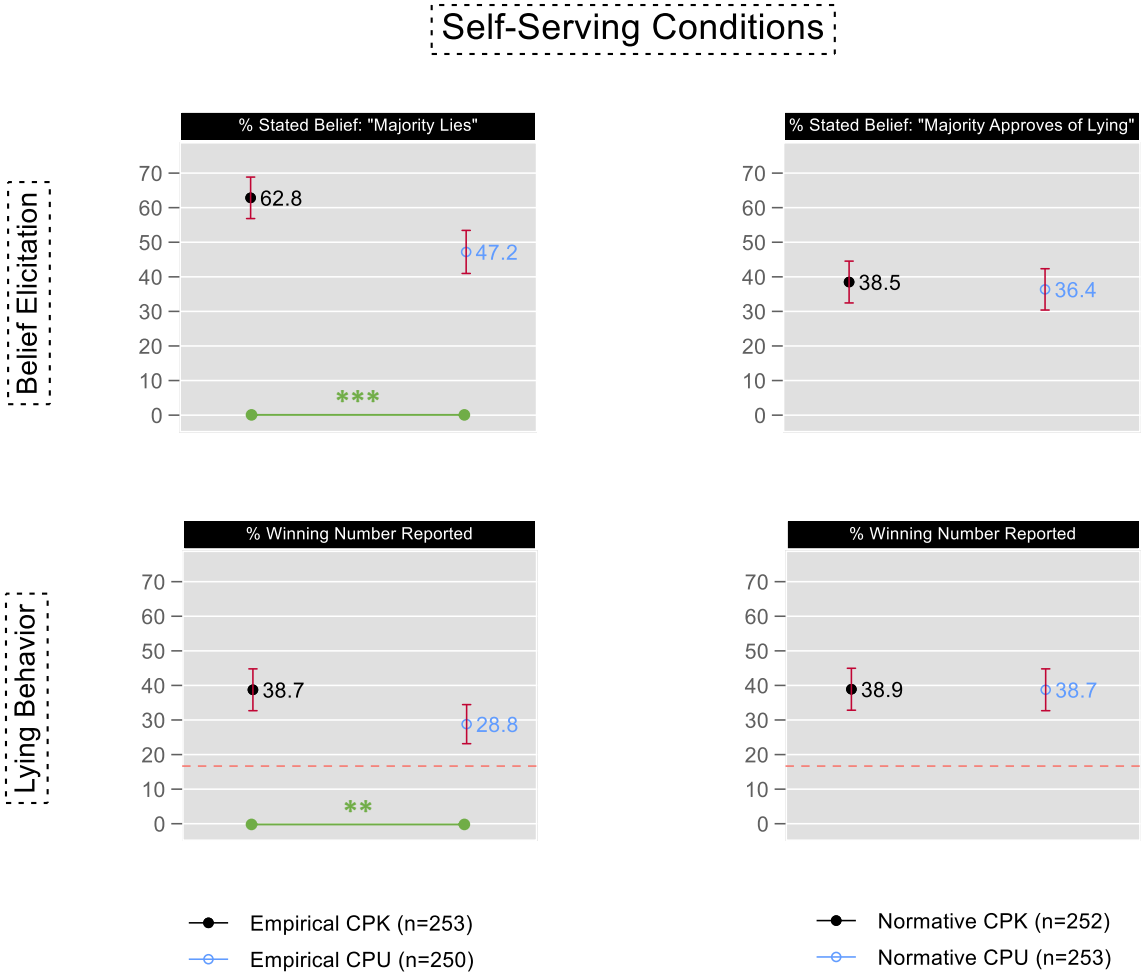

Figure 3: Stated beliefs and lying behavior across both empirical and normative conditions when cheating opportunity was known (CPK) or unknown (CPU) at the time of belief elicitation. Differences in beliefs within the same norm information domain (empirical or normative) indicate self-serving belief distortion. Stars indicate significant differences at the conventional levels of *p<0.1, **p<0.05, and ***p<0.01. Figure appeared originally in Bicchieri et al. (2019b).

To test our assumption about the asymmetry of what we infer from empirical or normative information, we administered a simple follow-up survey that included two variations:

1. Giving empirical information and then eliciting normative beliefs
2. Giving normative information and then eliciting empirical belief

- "The majority of participants *did not lie* for their own benefit. How many participants *approved* of lying?"
- "The majority of participants *did not approve* of lying. How many participants *lied* for their own benefit?"[7]

As argued above, variation in information in this context produces variation in what one infers about the normative appropriateness or frequency of the behavior. When participants are told that the majority of participants *did not lie*, they infer that the majority (77.48%) disapprove of lying (Figure 4, left panel). We do not observe the reverse to the same degree (Figure 4, right panel): When participants are told that the *majority disapproves of lying*, they infer that only 47.65% will be honest.

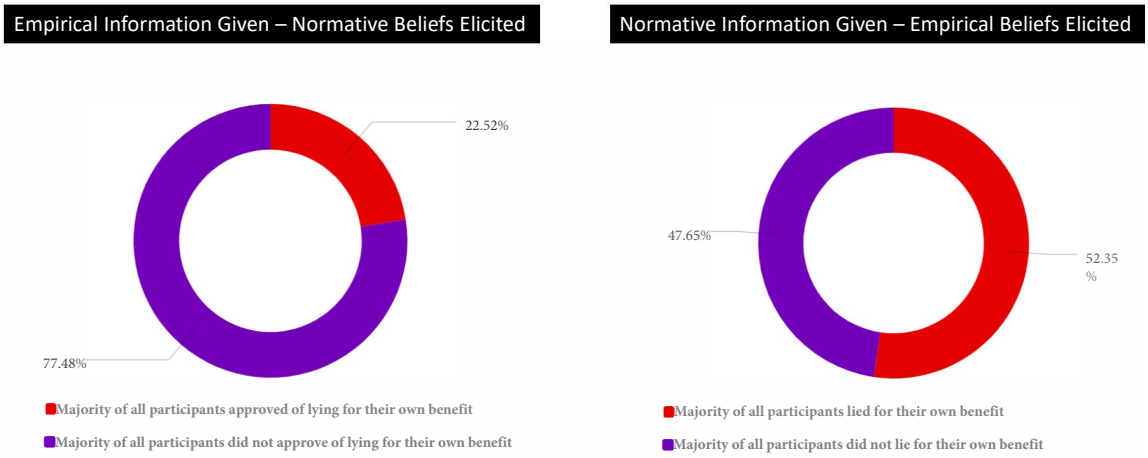

Figure 4: Distribution of normative (empirical) information inferred from being provided empirical (normative) information. Figure appeared originally in Bicchieri et al. (2019b)

[7] In both conditions on average participants think that "majority" means 71.44% and 71.84%, respectively.

An important conclusion to be made is that we have a tendency to infer the normative from the empirical, much less vice versa.[8] This has major consequences for norm-nudging. Providing empirical information will often make people conclude that common, widespread behavior (good or bad) is acceptable, at a minimum, and often even approved of. Our result is in line with recent work showing that commonness of observed behavior influences its moral status (Lindström et al., 2018). For example, it has been shown that when tax evasion is perceived as more common, it will also be seen as more acceptable, which in turn influences its prevalence (Eriksson et al., 2015). This conclusion seems to condemn us to a cascade of bad behavior whenever people realize how common the behavior is. When bribing, corruption, cheating and a host of other anti-social or plainly harmful behaviors (such as wife beating or female genital cutting) are perceived as common and thus 'normal' and approved of, it is hard but not impossible to reverse them. In the next experiment, we have explored ways to mitigate the effects of getting information about anti-social behavior.

4.3. How to Contain Norm Erosion

In a recent paper, Bicchieri et al. (2019a) we study the dynamics and erosion of norm compliance among peers. In a repeated and non-strategic setup, individuals can actively comply with or violate a pro-social norm of giving to a charity. We isolate and study the erosion of norm compliance by varying the observability of peer behavior and the social proximity among peers across treatments. Our design uses a variant of the extended dictator game as originally proposed by List (2007) and Bardsley (2008). In this variant, called take-or-give (ToG) donation game, each subject makes a donation decision towards a charity. The game starts with the subject and the charity both provisionally endowed with the same amount of money and each participant can decide whether to:

- Give some or all of her money to the charity
- Leave the equal split as is
- Take some or all of the money from the charity

Participants make the first choice in isolation and are then randomly grouped with other participants where each participant continues to make the same decision towards a

[8] A case of inferring normative from empirical would be a situation where we believe that "most people misbehave". Here we may have an interest to infer that they also approve the bad behavior.

charity for the next 19 rounds. Across three experimental conditions, we systematically vary how much each participant knows about her group members. The variations include:

1. *NoInfo*: no information is ever revealed about any group member

2. *Observation*: full transparency exists with respect to the behavior of all group members in the sense that participants observe a history table indicating the behavior of one's group members (and only of those) over all periods.

3. *ObservationSP*: same as Observation but with one additional piece of information. Participants now also know their degree of social proximity with each group member. Proximity is introduced by asking a question about the success of a Philadelphia sports team and participants observe what other participants replied to the question and what they did with respect to the charity.

The novelty of this approach is that it relies on social identity. Social identity theories assume that individuals categorize themselves and others into groups (Tajfel, 1982). According to social identity theory, identifying with a group induces conformity to what is perceived as appropriate group behavior and aversion to what is perceived as inappropriate (Turner, 1985; Akerlof and Kranton, 2000). We would thus expect less frequent self-serving use of empirical information when group identity is activated. Note that conformity to group behavior presupposes the existence of group norms, and it is therefore important to know if a norm in fact exists. The results of our norm-elicitation protocol following Bicchieri and Chavez (2010) show that giving to a charity, as well as not taking from a charity, are generally expected and approved behaviors. In an in-group setting, perception of appropriate behavior is especially influenced by what other group members do, so we hypothesize that there is *heightened conditionality* of preferences with respect to empirical information in settings where individuals interact with socially proximate others. Provided group identification is triggered, individuals are expected to respond more strongly to the behavior of in-group members than to the behavior of anonymous others.

Overall, we find that exposure to peers drives the erosion of norms by facilitating the spread of norm violations in that individuals react to anti-social behavior (taking) but not to pro-social behavior (giving). In the presence of social proximity, however, individuals are influenced by observing *both* examples of norm violations and norm compliance. Both negative and positive behaviors are contagious, spread within groups,and end up stabilizing the donation norm roughly at its initial level. In line with group identity theory, this result

holds only among cohesive groups (where all group members shared the relevant dimension of identity) and is absent among non-cohesive groups. Results are illustrated in Figure 5.

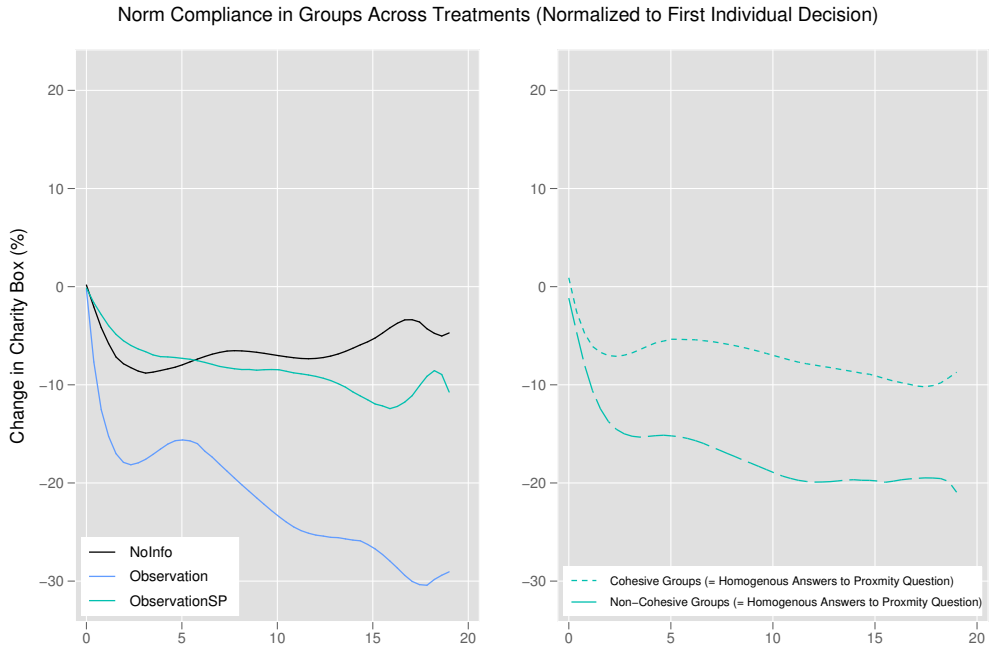

Figure 5: Left panel: Erosion in norm compliance across treatments and periods. Normalized to first individual decision (period 0 in the graph). Right panel: Erosion of norm compliance for ObservationSP treatment based on cohesive and non-cohesive groups. Figure originally appeared in Bicchieri et al. (2019a).

These insights are important from a policy perspective because they can improve the effectiveness of norm-based interventions and help to advance our understanding regarding the role of social proximity (identity) in the dynamics of norms and behavioral change. Group identity can be a powerful tool to induce pro-social behavior, but our results suggest that – to have an effect on behavior – the group should be perceived as cohesive and individuals should be able to observe *also* examples of positive behavior. If simply observing (or be informed about) negative behavior drives norm erosion by helping the spread of norm violation, introducing group identity (social proximity) mutes this effect. Norm-nudging should thus not only specify the appropriate reference network for a particular target population, but also, if possible, point to (or even create) some form of group identity.

4.4. Some Benefits of Social Norms Priming

Up to now, we have highlighted the potential pitfalls of social information. Here we present two experiments in which norm-nudging has been successful. In both cases, there has been norm elicitation, though the nature of the norm involved and the elicitation methodologies are very different. In the experiments discussed above, we have seen that participants either directly observed peer behavior, or were provided with information about what other participants had done or believed to be appropriate in similar circumstances. Even when there was uncertainty as to what others had truly done or believed, the experimenters were directly supplying empirical and/or normative information. An alternative would be to let participants themselves come to think about what their peers have done in the same situation, without any information provided by the experimenter.

A recent experiment by Bolton, Dimant and Schmidt (2019) adopts this methodology, priming participants to think about what others have done in the same circumstances. The context of the experiment is one in which there is social observability. Social observability is seen as a nudge intended to make people aware of the social consequences of their behavior. This has become a particularly popular approach due to its fairly frugal implementation, since it is often assumed that observability of one's actions by third parties promotes prosocial behavior (Ernest-Jones et al., 2011; Ekström, 2012; Rogers et al., 2018). Yet the existing literature shows mixed results and a confusing picture of when and why nudges that rely on observation do or do not work (e.g., Damgaard and Gravert, 2018). Bolton et al. (2019) show that simple observability can indeed have a negative effect, but this effect is reversed when observability is combined with being induced to think about what others have done in the same situation.

The experimental setup of Bolton et al. (2019) is a one-shot game in which decision-makers can give money to or take money from a charity (List, 2007; Dimant, 2019) The experiment varies the observability of one's actions by others, as well as the (non)monetary relationship between observer and observee. The key result is displayed in Figure 6. We can see that focusing on what is thought as common behavior increases donations to the charity box, both in situations with or without observation.[9] Here, focusing participants on a descriptive norm simply means asking them to think about what others have done in

[9]Note that compared to a setting in which behavior cannot be observed by a third party, anonymous observation without economic consequences leads to an inferior aggregate outcome and a significant decrease in the charity account. These results are disussed in more detail in the original Bolton et al. (2019) paper.

the same situation right before they make their choice to give/take from the charity.

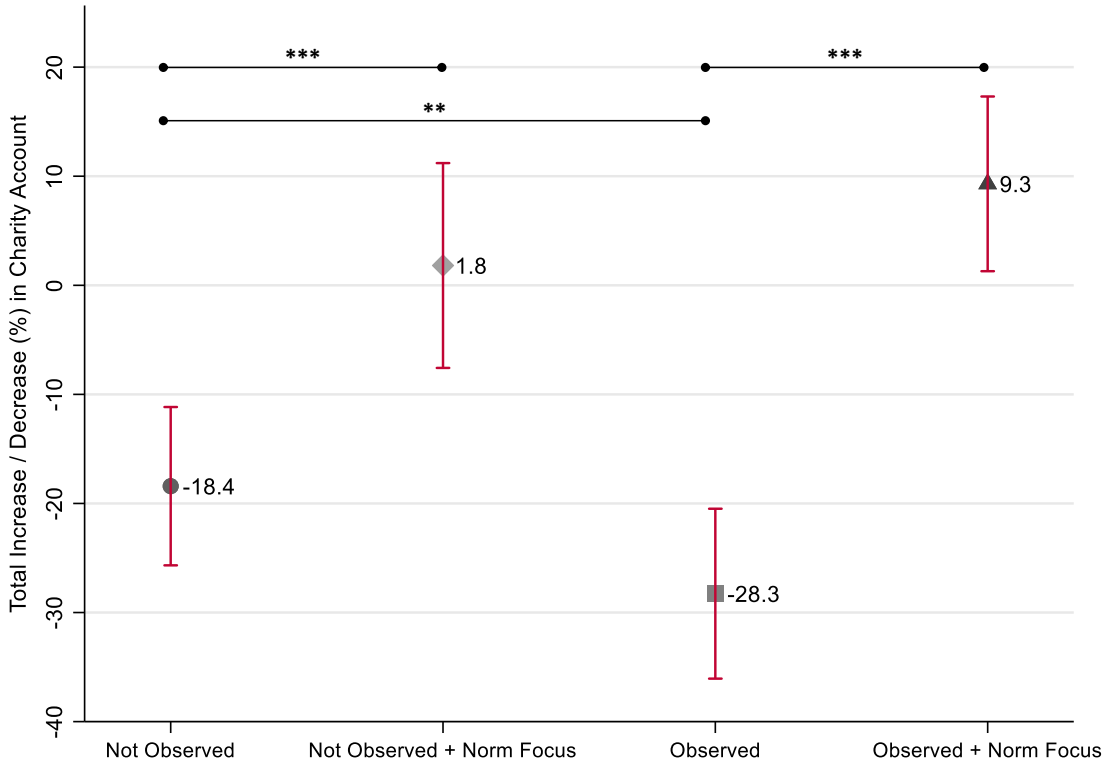

Figure 6: Change in charity account (compared to initial endowment) across treatments. Stars indicate significant differences at the conventional levels of *p<0.1, **p<0.05, and ***p<0.01. Figure originally appeared in Bolton et al. (2019).

The result shows that inducing participants to think about what most peers have done leads those who would otherwise take from the charity to take less.[10] In this case, the negative effect of being observed is reversed. Since taking from a charity is particularly egregious, empirical expectations will focus individuals on positive behavior (giving or at least not taking). As we already discussed, in a context where actions can be pro- or anti-social, empirical expectations can lead people to the normative conclusion that common behavior is also approved behavior. In this case, the intervention focuses participants on a (positive) social norm. We may conclude that if making actions observable by a third party

[10]This finding is in line with Allcott (2011) who finds that norm-messaging primarily affects the most deviant individuals with the highest energy use (also see Ferraro and Price, 2013)

could have a detrimental effect, merely inducing individuals to think about what others in the same position have done may reverse it, probably because the pro-social action (donate to charity, saving the environment) exerts a powerful pull, especially when one is not 'fed' information that one may not trust, or feel pressured to respond to..

We have discussed how providing single normative messages may not work, often because being told about what others think is appropriate behavior does not lead us to infer that the behavior is in fact common. However, there is some indirect and suggestive evidence that eliciting normative expectations can have a similarly powerful effect as *eliciting* empirical expectations (as shown in Bolton et al., 2019). For example, Jachimowicz et al. (2018) tested whether stronger second-order normative beliefs made the Opower energy conservation treatment more effective. In addition, they have causal evidence that the combination of providing descriptive social info and normative expectation is associated with higher energy conservation. While this study does not explicitly test the idea that eliciting normative beliefs alone would create a behavioral pull, it is a prudent assumption in light of the Bolton et al. (2019) findings and remains an open empirical question.

5. Conclusion

Interdependent behaviors are very common and come in different types. Behavior may be just conditional on empirical expectations, as is the case with descriptive norms. In this case, norm-nudging may simply induce different empirical expectations so as to steer individuals towards more beneficial behavior. For example, relying on imitation may induce individuals to behave like highly regarded others. In many other cases, behavior depends on both empirical and normative expectations. Important examples include social dilemmas and tragedies of the commons. Everybody benefits from the provision of public goods, such as flood control systems and street lighting, but everyone is better off free riding on others' contributions. Public water, the earth's atmosphere, and fisheries are natural resources shared by many individuals. Everyone has an incentive to maximize their use of the shared resource, eventually depleting it. In all these cases, the myopic rational choice is to act in one's self-interest, leading to suboptimal collective outcomes. Social norms exist precisely to solve such collective action problems.

Social norm-nudging to induce pro-social actions is particularly important when behavior is conditional on both empirical *and* normative expectations. In this paper, we highlight both the importance of distinguishing between different types of interdependent behaviors,

as well as some common pitfalls of norm-nudging. Designing norm interventions always includes providing social information in order to elicit social expectations. The effectiveness of these interventions will depend, among other things, on avoiding uncertainty about reference networks, relying on credible, trusted sources of information, and pointing to examples of positive behavior. A future venue of research may bring behavioral economics and the norm-nudge research closer together by studying the exact mechanisms underlying the processing. Exemplary, recent evidence suggests that the effectiveness of default rules, a popular nudge intervention, depends on the individual's pre-existing preference and that a stark contrast between the two can render the intervention ineffective (Dinner et al., 2011; Jachimowicz et al., 2019). Understanding such reference-dependence in the context of norm-nudging has the potential to improve the effectiveness of interventions that rely on social information.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Allcott, H. (2011). Social Norms and Energy Conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allen, V. L. (1965). Conformity and the Role of Deviant. *Journal of Personality*, 33(4):584–597.

Bardsley, N. (2008). Dictator Game Giving: Altruism or Artefact? *Experimental Economics*, 11(2):122–133.

Berkowitz, A. D. and Perkins, H. W. (1987). Recent Research on Gender Differences in Collegiate Alcohol Use. *Journal of American College Health*, 36(2):123–129.

Bhanot, S. P. (2018). Isolating the effect of injunctive norms on conservation behavior: New evidence from a field experiment in california. *Organizational Behavior and Human Decision Processes*.

Bicchieri, C. (2006). *The Grammar of Society*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Bicchieri, C., Ashraf, S., Das, U., Kohler, H.-P., Kuang, J., McNally, P., Shpenev, A., and Thulin, E. (2018a). Phase 2 Gates Project Report. Social Networks and Norms: Sanitation in Bihar and Tamil Nadu, India.

Bicchieri, C. and Chavez, A. (2010). Behaving as Expected: Public Information and Fairness Norms. *Journal of Behavioral Decision Making*, 23(2):161–178.

Bicchieri, C. and Chavez, A. (2013). Norm Manipulation, Norm Evasion: Experimental Evidence. *Economics & Philosophy*, 29(2):175–198.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2019a). Social proximity and the evolution of norm compliance. Working Paper Available at SSRN: https://ssrn.com/abstract=3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2019b). It's not a lie if you believe it: Lying and belief distortion under norm-uncertainty. Working Paper Available at SSRN: https://ssrn.com/abstract=3326146.

Bicchieri, C., Dimant, E., and Xiao, E. (2018b). Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment. PPE Working Paper 0016, University of Pennsylvania.

Bicchieri, C. and Ganegonda, D. (2016). Determinants of Corruption: A Socio-Psychological Analysis. *Thinking About Bribery, Neuroscience, Moral Cognition and the Psychology of Bribery*, pages 179–205.

Bicchieri, C., Jiang, T., and Lindemans, J. W. (2014). A social norms perspective on child marriage: The general framework.

Bicchieri, C. and Mercier, H. (2014). Norms and beliefs: How change occurs. In *The Complexity of Social Norms*, pages 37–54. Springer International Publishing.

Bicchieri, C. and Xiao, E. (2009). Do the Right Thing : But Only if Others Do So. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bolton, G., Dimant, E., and Schmidt, U. (2019). When a Nudge Backfires: Using Observation with Social and Economic Incentives to Promote Pro-Social Behavior. PPE Working Paper 0017, University of Pennsylvania.

Brent, D. A., Lott, C., Taylor, M., Cook, J., Rollins, K., Stoddard, S., et al. (2017). Are Normative Appeals Moral Taxes? Evidence from a Field Experiment on Water Conservation. Technical report.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., and Rogers, F. H. (2006). Missing in Action: Teacher and Health Worker Absence in Developing Countries. *Journal of Economic Perspectives*, 20(1):91–116.

Cialdini, R. B., Kallgren, C. A., and Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology*, volume 24, pages 201–234. Elsevier.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and social psychology*, 58(6):1015.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Damgaard, M. T. and Gravert, C. (2018). The Hidden Costs of Nudging: Experimental Evidence from Reminders in Fundraising. *Journal of Public Economics*, 157:15–26.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness. *Economic Theory*, 33(1):67–80.

Deutsch, M. and Gerard, H. B. (1955). A Study of Normative and Informational Social Influences Upon Individual Judgment. *The Journal of Abnormal and Social Psychology*, 51(3):629.

Dimant, E. (2019). Contagion of Pro-and Anti-Social Behavior Among Peers and the Role of Social Proximity. Working paper, The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham.

Dimant, E. and Schulte, T. (2016). The Nature of Corruption: An Interdisciplinary Perspective. *German LJ*, 17:53.

Dinner, I., Johnson, E. J., Goldstein, D. G., and Liu, K. (2011). Partitioning default effects: why people choose not to choose. *Journal of Experimental Psychology: Applied*, 17(4):332.

Ekström, M. (2012). Do Watching Eyes Affect Charitable Giving? Evidence From a Field Experiment. *Experimental Economics*, 15(3):530–546.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional Associations Between Descriptive and Injunctive Norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Ernest-Jones, M., Nettle, D., and Bateson, M. (2011). Effects of Eye Images on Everyday Cooperative Behavior: A Field Experiment. *Evolution and Human Behavior*, 32(3):172–178.

Fehr, E. and Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14(3):159–181.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The Persistence of Treatment Effects with Norm-Based Policy Instruments: Evidence From a Randomized Environmental Policy Experiment. *American Economic Review*, 101(3):318–22.

Ferraro, P. J. and Price, M. K. (2013). Using Nonpecuniary Strategies to Influence Behavior: Evidence From a Large-Scale Field Experiment. *Review of Economics and Statistics*, 95(1):64–73.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association*, 11(3):525–547.

Frey, B. S. and Meier, S. (2004). Social Comparisons and Pro-Social Behavior: Testing" Conditional Cooperation" in a field experiment. *American Economic Review*, 94(5):1717–1722.

Gächter, S., Molleman, L., and Nosenzo, D. (2018). The Behavioral Logic of Rule Following and Social Norm Compliance. Unpublished Manuscript.

Gino, F., Hauser, O. P., and Norton, M. I. (2019). Budging beliefs, nudging behaviour. *Mind & Society*, pages 1–12.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A Room With a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research*, 35(3):472–482.

Hallsworth, M., Chadborn, T., Sallis, A., Sanders, M., Berry, D., Greaves, F., Clements, L., and Davies, S. C. (2016). Provision of Social Norm Feedback to High Prescribers of Antibiotics in General Practice: A Pragmatic National Randomised Controlled Trial. *The Lancet*, 387(10029):1743–1752.

Hogg, M. and Turner, J. (1987). Social Indentity and Conformity. In *Current Issues in European Social Psychology*, volume 2. Cambridge University Press.

Jachimowicz, J. M., Duncan, S., Weber, E. U., and Johnson, E. J. (2019). When and why defaults influence decisions: a meta-analysis of default effects. *Behavioural Public Policy*, pages 1–28.

Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E., and Galinsky, A. D. (2018). The critical role of second-order normative beliefs in predicting energy conservation. *Nature Human Behaviour*, 2(10):757.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American economic review*, 90(4):1072–1091.

Krupka, E. L. and Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

Lindström, B., Jangard, S., Selbing, I., and Olsson, A. (2018). The Role of a "Common is Moral" Heuristic in the Stability and Change of Moral Norms. *Journal of Experimental Psychology: General*, 147(2):228.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, 115(3):482–493.

List, J. A., Berrens, R. P., Bohara, A. K., and Kerkvliet, J. (2004). Examining the Role of Social Isolation on Stated Preferences. *American Economic Review*, 94(3):741–752.

Löfgren, Å. and Nordblom, K. (2019). A theoretical framework explaining the mechanisms of nudging.

Mas, A. and Moretti, E. (2009). Peers at Work. *American Economic Review*, 99(1):112–45.

Miller, D. T. and McFarland, C. (1987). Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity. *Journal of Personality and Social Psychology*, 53(2):298.

Mols, F., Haslam, S. A., Jetten, J., and Steffens, N. K. (2015). Why a Nudge is not Enough: A Social Identity Critique of Governance by Stealth. *European Journal of Political Research*, 54(1):81–98.

Reijula, S., Kuorikoski, J., Ehrig, T., Katsikopoulos, K., Sunder, S., et al. (2018). Nudge, Boost, or Design? Limitations of Behaviorally Informed Policy Under Social Interaction. *Journal of Behavioral Economics for Policy*, 2(1):99–105.

Rivis, A. and Sheeran, P. (2003). Descriptive Norms as an Additional Predictor in the Theory of Planned Behaviour: A Meta-Analysis. *Current Psychology*, 22(3):218–233.

Rogers, T., Goldstein, N. J., and Fox, C. R. (2018). Social mobilization. *Annual Review of Psychology*, 69:357–381.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science*, 18(5):429–434.

Shpenev, A., Kohler, H.-P., and Bicchieri, C. (2019). Sanitation behavior in urban and rural India: a networks and norms approach.

Spiekermann, K. and Weiss, A. (2016). Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room'. *Games and Economic Behavior*, 96:170–183.

Stibe, A. and Cugelman, B. (2016). Persuasive Backfiring: When Behavior Change Interventions Trigger Unintended Negative Outcomes. In *International Conference on Persuasive Technology*, pages 65–77. Springer.

Stuntz, W. J. (2000). Self-Defeating Crimes. *Va. L. Rev.*, 86:1871.

Tajfel, H. (1982). Social Psychology of Intergroup Relations. *Annual Review of Psychology*, 33(1):1–39.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

The Guardian (2013). Antanas mockus: Colombians fear ridicule more than being fined.

Turner, J. C. (1985). Social Categorization and the Self Concept: A Social Cognitive Theory of Group Behavior. In Lawler, E. J., editor, *Advances in Group Process*, volume 2, pages 77–122. JAI, Greenwich, CT.

Yoeli, E., Hoffman, M., Rand, D. G., and Nowak, M. A. (2013). Powering Up With Indirect Reciprocity in a Large-Scale Field Experiment. *Proceedings of the National Academy of Sciences*, 110(Supplement 2):10424–10429.