Chapter 3

What are Norms and How is Norm Compliance Regulated?

Bertram F. Malle, Brown University

Most great apes lived in hierarchical groups, often engaged in fierce battles for position (Boehm, 1999). Even though a few exceptions, like the Bonobos, exist today, the main line of our ancestors lived by power and domination. Sometime in the past two million years, a remarkable change occurred: the *homo* evolution moved toward social group living that substantially increased cooperation, joint action, teaching, and learning (Boehm, 2000; Tomasello & Vaish, 2013). Perhaps precursors to empathy helped (de Waal & Preston, 2017), but empathy is reactive and local (Bloom, 2016) and can conflict with fairness and justice (Decety & Cowell, 2015). By itself, empathy does not provide a social organization that can sustain a group’s success in the most uncertain and harshest environments. Norms do.

Norms are guides to action. They constrain an individual’s options in a given context to act in ways that benefit the group, and groups that have a well-functioning norm system function well as groups (Bicchieri, 2006; Chudek & Henrich, 2011; Wilson, 2002). Norms provide organization and regulation of group members’ actions that can sustain a group’s success even in the face of serious threats—through ice ages, natural catastrophes, and wars.

But while norms enable group success, and thus ultimately individual success, norms impose costs on the individual, thus tempting them to avoid those costs. Two questions therefore arise: What motivates individuals to resist this temptation and comply with the group’s norms? And how does the group strengthen compliance? To answer these questions, I will first clarify what norms are. I will then examine how norms regulate behavior at the individual level, how they regulate behavior at the social level, and how these levels are intimately entwined.

What are Norms?

The importance of norms has long been recognized in sociology and economics (Elster, 1989; Hechter & Opp, 2001; Horne & Mollborn, 2020; Schelling, 1960). These literatures aim to explain how human cooperation is possible despite the individual’s self-interest. Norms are therefore treated as an external force that constrains human action. But the social sciences have rarely addressed how norms operate cognitively, in such a way that enable norms (as external forces) to succeed at guiding behavior. A person complying with norms must have something in their mind that allows their action to conform to the norm—they must have a *norm representation*. A few empirical studies have examined the automatic activation of such norm representations by situation cues—for example, garbage on the floor triggers the *Don’t litter* norm (Cialdini et al., 1990), or the sight of a library triggers the *Be quiet* norm (Aarts & Dijksterhuis, 2003). But exactly what is activated?

**Definition**

A cognitive and social model of norms must clarify what norms are. The following definition integrates several overlapping proposals (Bicchieri, 2006; Brennan et al., 2013; Cialdini et al., 1991; Malle et al., 2017, 2021):
A norm $N$ is a directive, in a given social community $S$, to (not) perform an action $A$ in a given context $C$, provided that (i) a sufficient number of individuals in the community demand of each other, with a certain deontic force $D_f$, to follow the directive and (ii) a sufficient number of individuals in the community do follow it with prevalence $P$.

Using this definition we can differentiate norms, norm representations, and norm-based motivation. First, for something to be a norm $N$ of $A$ in $C$, conditions (i) and (ii) must hold true. Second, for a community member to have a representation of $N$, the person must believe (consciously or not) that conditions (i) and (ii) hold. Third, for someone to have a norm-based motivation to $A$ in $C$, the person must both believe conditions (i) and (ii) and try to (not) $A$ because of conditions (i) and (ii).

**Elaboration**

A norm directs action, and this directive can be a prescription (directing the person to act a certain way) or a prohibition (directing the person to not act a certain way) (Folger & Whiting, 2020; Janoff-Bulman & Carnes, 2013; McNamara, 2006). In either case, people may not be aware that a norm influences their action, or the norm can be so internalized that it feels like (or cognitively becomes) a personal goal. In some cases, people are fully aware of the norm, such as when they struggle with a conflict between norms or between personal goals and norms.

Every norm has a deontic force $D_f$, indicating how strongly people demand of each other to (not) perform $A$ (Home & Mollborn, 2020; Malle, 2020). For example, a prescription may be weak (as in a suggestion) or strong (as in a requirement). The motivation to comply with norms will therefore vary roughly proportional to the norm’s deontic force. Compliance also varies with the prevalence of a norm—the (perceived) frequency with which community members actually follow the norm. High prevalence facilitates norm activation (Cialdini et al., 1990; Lindenberg, 2013) and signals the presence of a norm’s deontic force (Eriksson & Strimling, 2015). Seeing that everybody stands in line to order ice cream makes the strength of that norm clear; seeing that hardly any bicyclist stops at the stop sign highlights the weakness of that norm (Cullum et al., 2012). People can misperceive prevalence or deontic force (Prentice & Miller, 1996), which illustrates that behavior is caused by norm representations, not by some objective force of norms themselves.

**Community specificity.** Norms are rarely1 universal—most norms exist relative to a social community $S$, at a certain time and place. For example, the social norm of wearing a mask in public varies widely (as of this writing) in different regions of the U.S. Moreover, communities are nested, such that norms vary across nations, across states within nations, urban and rural regions within states, neighborhoods within regions, and so on. Even within any given community, individuals will vary to some degree—because of different learning histories, different competing personal interests, or different sensitivity to community demands. However, for a norm to exist, enough community members must reasonably agree on $D_f$ for $A$ in $C$.

**Context specificity.** Being on the phone is forbidden in a train’s quiet car but required on the trading floor; killing another person, prohibited in many contexts, is permitted in self-defense and prescribed in war and executions (for communities that allow them). What varies here, “context,” is a fascinating and complex construct. We can think of it as bundles of features that include time (e.g., evening, religious holiday), physical environments (e.g., bathroom, hallway),
role (e.g., friend, colleague), interaction type (e.g., first date, business negotiation), and more. Norm compliance therefore requires refined knowledge of contexts and the ability to recognize them quickly and accurately, and the norms that are governed by them (Malle et al., 2017). Thus, a norm violation can stem from misperceiving the context one is in (e.g., an informal dinner vs. an extended job interview), rather than from an intention to flout the norm.

**Related Concepts**

Norms differ from other action guides, such as preferences, goals, and collective habits. The key distinguishing feature is the deontic force condition (i), which ensures that the action in question reflects the community’s demands and interests, not merely the individual’s desires or benefits. That many people put milk in their coffee is a wide-spread preference, not a norm, because people do not demand it of each other, and they do it for their own pleasure, not others’. By contrast, getting in line to order coffee is a norm, because that is what people demand of each other, even though each individual would rather place their order without a wait.

Norms also differ from virtues and values, which have a deontic force but lack other properties of norms. Virtues (such as generosity or courage) are personality dispositions that a community appreciates in its members, but they are rather like ideals than required traits, and as dispositions they lack action- and context specificity. Values (such as fairness, freedom, respect) also lack action- and context specificity as they govern large classes of possible actions/outcomes across a wide range of contexts (Serramia et al., 2018), and it is difficult to teach them by instruction, as most norms can be. However, sets of norms can reflect or implement broader values, such as when sharing norms implement fairness, or when politeness norms implement respect (Stohr, 2012).

**Features not included in the norm definition**

*Moral vs. conventional.* The first feature not included in the above norm definition is the distinction between moral norms on one side and social or conventional norms on the other side (Aharoni et al., 2012; Turiel, 1983). The proposed definition applies to both classes of norms, and the bundle of features traditionally meant to distinguish norms from conventions (e.g., universality, authority independence) are loose and ill-defined (Shoemaker, 2014). Moreover, claims of norm universality seem untenable, as what is treated as moral rather than merely social-conventional varies significantly across time and communities (Kelly et al., 2007). Mild conventions in one group, such as regarding food or dress, are moral in other groups (e.g., orthodox religious communities). The primary distinguishing feature of moral and conventional norms is deontic force. “Moral” norms are consistently higher in deontic force, and because of that force, other features follow (e.g., that an authority cannot simply cancel the norm). We may want to treat moral norms and social conventions as two prototypes, but the definition of norms offered above would seem to hold for both; cognitively, the two prototypes still fall under a common construct (Roughley, 2019).

*The role of sanctions.* The second feature not included in the definition of norms is the presence of sanctions (such as a blaming, shaming, or punishment), which some have described as a defining feature of norms (Bendor & Swistak, 2001; Horne & Mollborn, 2020). Sanctions are undoubtedly important. They can be informative in revealing a norm’s deontic force (e.g., the most severe punishment is often reserved for violations of the strongest norms), and many communities use sanctions to respond to norm violations. However, sanctions are not the only
method to communicate the deontic force of a norm—for example, proactive exhortations and incentives or modeling costly compliance (e.g., in rituals), among others, help convey the strength of a norm. And even in response to violations, communities have other tools besides sanctions (see the later section on Alternative Regulation). While it would therefore seem misleading to make sanctions part of the definition of norms, sanctions are certainly one core element of the regulation and enforcement of norms, as I will discuss shortly.

**Injunctive and descriptive norms.** The third feature not included in the definition of norms is the distinction between injunctive and descriptive norms (Borsari & Carey, 2003; Cialdini et al., 1991; Lapinski & Rimal, 2005). The core of this distinction is between “people’s beliefs about what ought to be done” (injunctive) and “beliefs about what is actually done by most others in one’s social group” (descriptive; Lapinski & Rimal, 2005, p. 130). Reflecting this important distinction, the above norm definition contains the two conditions of deontic force (corresponding to the injunctive property) and prevalence (corresponding to the descriptive property). But these are two properties of norms, not two kinds of norms. If the prevalent pattern of behavior in a group (“descriptive norm”) is a norm at all, it will have to come with some deontic force, otherwise it is a mere statistical regularity, perhaps providing cues for personal benefits. Telling a homeowner that neighbors use 20% less energy (a classic “descriptive norm”) will not activate a genuine norm unless the homeowner infers that saving energy is the right thing to do (deontic force) and they should therefore try harder. Thus, what researchers studied as “descriptive norms” is prevalence information about a norm that allows people to infer its deontic force. By contrast, telling a homeowner that neighbors spend 20% less money by shopping at the nearby grocery store is merely descriptive information, and not a norm (and therefore not a “descriptive norm”) because it is beneficial to spend less money, not (typically) demanded by the community.

**Regulation of Norm Compliance**

With this conceptual grounding of what norms are, I turn to the question of what regulation mechanisms motivate individuals to comply with norms. A fundamental conflict necessitates this regulation: following norms may benefit the community but often imposes a cost on the individual (such as freedom, time, effort, money; Cooter, 2000), though typically the individual benefits, too, in the end (compared to a situation in which everybody is out for themselves). If community members already tend to do something, there is no reason for a norm to emerge—there are no norms that encourage eating or sleeping, going to restaurants or movie theaters. But if there is some cost on the individual, norms may have to come into play, such as for being quiet in a movie theater or (in the U.S.) to tip the wait staff in a restaurant. Because of these costs to the individual, norm compliance requires mechanisms of regulation.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proactive</strong></td>
<td><strong>Norm activation, inducing guilt, demanding justification</strong></td>
</tr>
<tr>
<td>Internalization, anticipatory guilt, justification</td>
<td></td>
</tr>
<tr>
<td><strong>Reactive</strong></td>
<td><strong>Sanctions, forgiveness, reintegration</strong></td>
</tr>
<tr>
<td>Guilt, justification</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1. Select mechanisms that regulate compliance with norms*
We might draw a distinction between regulation at the individual level (e.g., guilt) and regulation at the social level (e.g., ostracism). However, the subsequent discussion will show that individual mechanisms are very much the product of social processes and that social processes cannot succeed without the individual’s psychological readiness (e.g., a need to belong to groups, which makes reintegration valuable). Another distinction may be useful to bring some organization to norm regulation mechanisms, namely those that are proactive (encouraging compliance in advance) and those that are reactive (responding to violations after they occur). Figure 3.1 shows a selection of regulation mechanisms within this space of individual-social and proactive-reactive groupings that I will discuss in the following sections. Of course this is not an exhaustive list, and other authors in this volume discuss additional processes.

**Compliance Regulation at the Individual Level**

**Internalization**

Among proactive individual regulation mechanisms, one phenomenon has received a considerable amount of attention in the literature: the internalization of norms. With some variations, most scholars cast internalization as a process of adopting and embracing a social or moral norm so that it becomes an intrinsically motivating personal standard, to which people commit even if explicit demands or sanctioning possibilities are low (Bandura, 1991; Gintis, 2003; Horne, 2003; Schwartz, 1977). Communities have an interest in seeing norms get internalized (Gavrilets & Richerson, 2017), because individuals not only follow internalized norms but they are more likely to enforce them in others (Hannah et al., 2016; Peysakhovich & Rand, 2016). Socialization, the community’s teaching of norms, is therefore very much focused on internalization (Grusec & Kuczynski, 1997). If the community can get most individuals to be intrinsically motivated to act the “right” way, then the expected community benefits are likely to obtain.

The success of internalization attempts stems in part from the original norm learning situation. Kochanska and Aksan (2006) identify “committed compliance” as the willing, eager stance toward social (often parental) influence, such as when an adult teaches a new rule. This receptive stance predicts later internalization—that is, sustained compliance with a norm even when no audience enforces the norm (Kochanska et al., 1995). Internalization has its limits, as there is virtually always some level of personal benefit that can tempt the person into breaking a strongly endorsed norm (Baron & Leshner, 2000). Internalization can also go overboard, when it becomes an inflexible, intolerant moral conviction (Skitka et al., 2021)—of which we have seen ample evidence in recent decades of political polarization in the United States.

It is apparent that norm internalization, even though in one way a psychological mechanism at the individual level, is deeply entwined with the community’s socialization practices and facilitates the social process of communal norm enforcement, which the internalizers are more willing to take on.

**Guilt**

A second mechanism of individual-level compliance regulation is guilt, which can be reactive or proactive. Reactive guilt is the feeling of discomfort after a norm violation (Kochanska & Aksan, 2006; McGraw, 1987), and such feelings predict subsequent prosocial behavior—not only restitution for the one whom one wronged but even benevolence toward others (Freedman et al., 1967). Meta-analyses have shown that guilt induces norm compliance.
across a wide range of conditions and with a strong and robust effect size (Boster et al., 2016; O’Keefe, 2000). Guilt is proactive when a person considers a potential action that would violate a norm and imagines feeling guilty if taking the action. This form of anticipated guilt has been shown to be impactful in blocking various processes that lead to transgressive behavior, such as moral disengagement (Bandura et al., 1996) or the influence of deviant peers (Svensson et al., 2013). Anticipated guilt is parallel to the role of anticipated regret in nonmoral decision making (Zeelenberg, 2018), but while regret is largely a personal emotion, guilt has a strong social component. For one thing, anticipated guilt can activate internalized norms (De Groot & Steg, 2009; Schwartz, 1977), which reflect, as discussed, the impact of social forces. In addition, guilt often co-occurs with shame (Ferguson et al., 2007; Tangney & Dearing, 2002). As a result, anticipated guilt over one’s potential norm violation may also invoke the prospect of shame, which incorporates other people’s evaluations and one’s perceived social standing (Maibom, 2010).

**Justification**

All forms of justification—when reactive for a performed action, or proactive for a planned action—build on the fundamental human tendency to explain intentional actions with reasons: mainly beliefs and desires that provide grounds for acting a particular way (Davidson, 1963; Malle, 2004). In order for reasons to justify an action they have to meet a normative criterion—they have to show that the action is somehow in compliance with social-moral norms. Demands for justification arise only when an action in question violates a norm. The agent will try to demonstrate that, in the given context, the norm was suspended (e.g., injuring another person was justified in the context of self-defense) or that the action upheld more important norms than the one it violated. For example, yelling at a person in a group conversation may be justified if the person told a racist joke, and the norm to resist racism is more important than the norm to not yell at people.

Who decides whether a given explanation justifies the action in question? Often it is the relevant moral community, such as the audience to whom the agent offers the justification. If the audience denies that the context suspends the violated norm or that the action upholds more important norms, the justification will fail. An audience, or moral community, thus provides a check on the agent’s proffered justification.

At other times, justifications take place in the mind of the decision maker. They may still be guided by the social community, such as when the person imagines what reasons a relevant moral community would find justifiable (Scanlon, 1998). Such imagined responses, however, can be inaccurate, and the decision maker may rely solely on their own subjective judgment of whether a particular course of action is justifiable. In this case, self-serving and potentially community-damaging justifications may take hold (Bandura, 1999; Mullen & Monin, 2016). For example, decision makers might suppose that everybody else is acting the same way (misrepresenting norm prevalence), that it isn’t their duty to protect the other (misrepresenting deontic force), or that they have already done a lot of good things (“licensing” the action in question; Green, 1991; Polman et al., 2013; Shalvi et al., 2011). Justification becomes here the enemy of guilt, and only a challenge from the community may correct that.

Even when justifications pass the challenge by a relevant community, a moral decision may be considered justified by one community and wholly unjustified by another (Cohen & Nisbett, 1994). Human history is full of examples in which members of one group justify acts of oppression, discrimination, or terror that they inflict on another group (Fiske & Rai, 2015). Thus,
justifications represent a core tool of regulation, but their value and impact are a function of the particular community’s norm system.

**Compliance Regulation at the Social Level**

Just as we have seen that individual processes of regulating norm compliance rely on social processes, so do social processes rely on individual ones. Among these social processes I discuss measures of activating norms, instilling guilt, and the sanctions of blame and punishment, along with some alternatives to sanctions.

**Norm Activation**

For context-specific norms to operate efficiently, they need to be readily available when the person enters relevant contexts but not in other contexts (e.g., talk when meeting someone for coffee but not during a poetry reading). Somehow, the context must activate the relevant norms (Aarts & Dijksterhuis, 2003; Cialdini et al., 1991; Malle et al., 2017; Schwartz, 1977). Internalized norms might be reliably activated even when the person is merely thinking about the context. Other norms may need help from the environment, through physical or symbolic means, or from other people.

What does the activating? The most obvious triggers are signs like “Don’t lean out the window” or “The line begins here.” People also offer verbal reminders (“What’s the magic word?”, “This is a no-shoes household”), and they model norm compliance for each other; even just two people queuing will typically suffice to trigger compliance (at least in London). Less obvious triggers lie in the presence of certain objects (e.g., a basket of kippahs at the entrance of a Jewish temple) or the physical arrangements of objects in space (e.g., chairs that indicate where and how far apart to sit).

What is being activated? In most cases, context activates the prescribed or prohibited action representation, and if the norm representation as a whole is restored, then the norm’s deontic force will be activated as well. The individual may sense (not necessarily consciously) this deontic force (Folger & Whiting, 2020; Heider, 1958), and the motivational pull toward prescribed actions or away from prohibited actions can emerge fast and with a strong (dis)value association (Cushman, 2013; Folger & Whiting, 2020). In these cases, the person will comply with the norm, with little further reasoning, especially when the costs of compliance are low (Diekmann & Preisendörfer, 2003). Many everyday situations are of this kind. We enter a coffee shop and swiftly join the end of the order line; a colleague says “Good morning” and we immediately return the greeting.

**Inducing Guilt**

Earlier we saw that guilt as a within-person process relies heavily on social factors. Conversely, the social process of inducing guilt succeeds only if individuals are receptive to such intervention. Guilt induction can be used proactively to persuade others to do what the persuader desires (O’Keefe, 2000), including not to commit a planned transgression (“You will regret this your entire life!”). More common are reactive attempts to make the transgressor feel guilty about a past transgression, often with the aim of deterring the person from committing it again. Either way, inducing guilt is akin to a test—if the person responds as desired (with concern, remorse, perhaps apology), the deterrence attempt promises to be successful and blame is mitigated (Darby & Schlenker, 1989); if the person does not respond as desired, the planned or repeat transgressions are likely to occur, and more serious interventions may be needed. Importantly,
guilt inductions have teeth only if the moral decision maker in principle shares the relevant norm or cares about the inducer’s disapproval. As a result, guilt induction succeeds primarily in smaller communities and in closer relationships (Baumeister et al., 1994). Guilt induction also has its risks (Baumeister et al., 2001) and can be dysfunctional (Rakow et al., 2011), but modeling, evoking, and nurturing guilt in measured ways is part of ordinary socialization (Abell & Gecas, 1997).

Sanctions and Their Alternatives

When proactive deterrence fails, when no valid justifications are offered, and guilt and apology are not forthcoming, people turn to sanctions. I will focus on the primary moral sanctioning mechanisms featured in the literature: punishment and blame. Even though the two are often treated as highly similar, in reality they are quite distinct (Baumard, 2011; Buckholtz et al., 2015; Malle, 2021). I will first consider punishment, then blame, as potential forces that respond to norm violations and reinstate compliance.

The Prospects of Punishment for Regulating Norm Compliance

A widely touted proposal is that informal punishment, or even just the threat of punishment, fosters cooperation (Fehr & Gächter, 2000; Gintis, 2000; Yamagishi, 1986). Much of the evidence for this proposal comes from behavioral economics studies, in which people play monetary games that pit self-interest against prosociality. In the public goods game, for example, individuals in small groups have an endowment and can spend any part of it to invest in a group payoff. When most group members contribute, each contributor benefits because investments are multiplied, though free-riders benefit even more; but when too many members free-ride, investors lose money and soon nobody contributes or gains. Yamagishi (1986) showed that instituting punishment as a sanction for free-riding increased cooperation and pay-offs. The effect of punishment, however, was far greater in groups composed of low-trusting individuals (who were skeptical of others’ cooperation) than in high-trust groups. In another type of game, the Ultimatum game, a “proposer” receives a monetary endowment (e.g., $10) and offers a “responder” a cut of that endowment (say, $3). If the responder accepts the proposer’s offer, both receive the allotted money ($3 for the responder and $7 for the proposed); if the responder rejects the offer, neither of them receives any money. Such rejections are infrequent but increase with unfair offers (such as 8:2 or 9:1 splits). Rejections are taken to be expressions of costly punishment because people respond to the proposer’s norm violation (e.g., 9:1 split) by foregoing a small amount ($1) themselves and causing the proposer to lose a large amount ($9).

The evidence from behavioral economics games must be met with caution, however, both because of its monetary focus and because of the absence of social interaction or relationships (Pedersen et al., 2013). For punishment to be a major norm regulation mechanism outside of economic games, two things must be true: punishment must be frequent enough to do the hypothesized regulatory work, and it must be effective enough at fostering norm compliance. There are reasons to doubt both of these propositions. First consider frequency.

How frequent is punishment?

Before humans settled down about 12,000 years ago, they lived in small groups of nomadic hunter-gatherers (Boehm, 1999; Knauf, 1991). The few remaining hunter-gatherer bands studied in the 20th century use little to no punishment to regulate behavior (Baumard, 2010; Guala, 2012). They respond to norm violations with criticism, ridicule, or temporary
ostracism rather than with punishment (Wiessner, 2005). In everyday life of modern societies, punishment is infrequent as well. Using daily surveys to study naturally occurring norm violations, Molho et al. (2020) measured people’s motivation to engage in various kinds of punishment, including verbal or physical confrontation, gossip, and withdrawal. People were generally unmotivated to engage in confrontation and somewhat more motivated to gossip or exclude the transgressor. In a cross-cultural study of individuals recalling an instance in which they witnessed (as a third party) an attack, insult, or other mistreatment (Pedersen et al., 2020), rates of cost-imposing punishment (e.g., yell, insult, push, hit) were extremely low, between 1% and 6%, varying primarily due to caring about the victim of the violation. Hofmann et al. (2018) assessed people’s desire to punish a norm violator, a variable that should lead to higher rates than actual punishment behavior, because fear of retaliation and other obstacles do not depress the rate of punishment desire. However, whether as victim or third party, people indicated weak desires to punish, from a mean of just over 2 (on a 0-6 scale) when the violator was a romantic partner to just over 3 when the violator was a stranger.

Returning to behavioral economics experiments, which had provided the best evidence for people endorsing and sometimes executing punishment, research increasingly suggests that people have little natural inclination to punish if alternatives to punishment are available (which the original experiments did not offer). For example, when given a choice between punishing the perpetrator and compensating the victim, people prefer the latter (Chavez & Bicchieri, 2013; Ohtsubo et al., 2018). This finding also replicates in real life, where people who witness a restaurant customer mistreat wait staff will support the victim rather than directly confront the perpetrator (Hershcovis & Bhatnagar, 2017). Likewise, when given a choice not to punish but to warn others who will interact with the transgressor, people select the latter path (Feinberg, Willer, et al., 2012). People will also try to get out of punishing a perpetrator if they can do so without losing face (Kriss et al., 2016). And in the limited cases when people do punish others, they do so as the victim of the transgression or as a close ingroup member to the victim (Bernhard et al., 2006; Pedersen et al., 2018). Even victims of unfair offers in the Ultimatum game will often refrain from punishing the proposer when they are granted an opportunity to communicate their disapproval by sending a note (Xiao & Houser, 2005). Interestingly, people reject unfair offers even when the proposer does not learn they received a punishment or when the rejection has no monetary impact on the proposer. This pattern suggests that people’s rejection of unfair offers does not primarily function as inflicting punishment on the transgressor but as a signal to themselves and others that they are no suckers, that they have dignity. However, this signal can be sent just as well with a note of disapproval.

Punishment also does not elicit much social support. People who punish an unfair player are not trusted more by other participants in a subsequent game (Patil et al., 2018); victims who punish are not seen positively, whereas victims who forego punishing the perpetrator are seen as moral, trustworthy, and altruistic (Heffner & FeldmanHall, 2019). Kiyonari and Barclay (2008) showed that, in an economic game context, when given a choice between rewarding or punishing other players, punishers were not rewarded; instead, people overwhelmingly assigned rewards for both cooperators and for rewarders of cooperators; the only group that was punished were nonrewarders of cooperators. In a real-life setting of a staged act of littering (Balafoutas et al., 2014), only 17% of participants confronted a stranger who littered; they preferred indirect punishment by subsequently withholding help from the violator when he dropped some books. Importantly, experimenter-staged punishers of the original norm violator were not rewarded (received no more help than did control individuals). It should be noted that the observed rates of
“punishment,” low as they already were, relied on a very lenient punishment definition. Participants’ response counted as punishment if they “explicitly asked the violator to pick up the cup or expressed disapproval of the norm violation” (p. 15927), and the experimental manipulation of punishment behavior amounted to a speech act of “Would you please pick up your garbage? The platform is not a garbage bin.”

In sum, in pre-settlement human history and modern life, punishment is infrequent because people prefer other responses to norm violations: compensating victims, rewarding cooperators, warning community members. If they engage with the perpetrator at all, they prefer mild expressions of disapproval, such as verbal criticism, and often they simply ignore or stop interacting with the person (Baumard, 2010; Martin & Cushman, 2015). Whether the latter is self-protection or punishment depends on whether there are any costs to the perpetrator (e.g., true ostracism would carry such costs; Williams, 2009).

How effective is punishment?

The second assumption underlying the hypothesized significance of punishment for moral regulation is that it works—that it halts ongoing violations and deters future violations. We have seen earlier that weak, low-cost punishment via monetary fines can increase cooperation (Fehr & Gächter, 2000; Yamagishi, 1986), but often it does not (Xiao, 2018), and it can lead to retaliation (Bone et al., 2016). Punishment in behavioral economics games works best when the punishment expresses community norms (Balliet & Van Lange, 2013; Xiao, 2018)—a function, however, that can also be achieved by nonpunishing interventions (Xiao & Houser, 2005).

Research on whether everyday punishment deters repeat violations comes primarily from the educational and developmental literature. To be effective, punishment must have clear and reliable outcome contingencies (Cipani, 2004). Under these conditions, when the child chooses between the compliant and the violating action, the compliant one (resulting in desired benefits) will be dominant and punishment rarely has to be applied. When punishment does have to be applied, its form matters. Physical punishment is reliably ineffective (Gershoff & Lee, 2020), whereas fines, withheld benefits, and simultaneous teaching of norms can be effective (e.g., Azrin & Powers, 1975). For both child and adult transgressors, confidence in the prospect of punishment, should the norm violation occur, increases norm compliance (Cipani, 2004; Kura et al., 2015). But if norm internalization is the ultimate goal, disciplining children must be gentle, de-emphasize power assertions, and incorporate positive feelings (Kochanska & Aksan, 2006).

Further evidence on the lack of success of cost-imposing punishment comes from the justice system. Numerous studies on recidivism demonstrate clear failures of deterrence (Wright & Khade, 2017). Three to five years after being released from prison, 30-47% of past offenders are rearrested, and 25-43% are re-imprisoned (compared to the base rate of imprisonment in the USA as a whole of 0.4%; Wagner & Sawyer, 2018). Research on the death penalty, too, shows a lack of deterrent effects (Nagin & Pepper, 2012). In fact, the US South has the highest rate of executions and the highest murder rate; the US Northeast has the lowest rate of executions and the lowest murder rate (Death Penalty Information Center, 2019).

Thus, we face a paradox: If punishment is rare and ineffective, why does it have such a long and brutal human history (Lyons, 2017)? Part of the answer lies in the instrumental benefits of punishment. In a strongly hierarchical society, the powerful defend their power: rulers pass laws to their advantage, and the majority imposes selectively harsh punishments on the minority (as amply demonstrated in the USA; Tonry, 2009). In addition, over the course of history, the
agents of norm enforcement and conflict resolution have shifted from community members to sovereign institutions (Cushman, 2015). This shift may help prevent blood feuds among communities, but perhaps more importantly, it allows ordinary people to condemn violators and support their punishment without having to get their hands dirty and their conscience tainted. It is far easier to cheer on the torturer or vote for “tough on crime” policies than it is to break the offender’s arm oneself or stand guard at a solitary confinement cell. No doubt, there are vigilantes, militias, guards, and police officers who welcome opportunities for brutal submission, but most people are averse to harming others (Grossman, 1996; McDonald et al., 2017). They prefer to leave punishment to professionals and institutions; but violence abounds if the latter remain unchecked.

The systematic delegation of harsh punishment to institutions also explains why people are commonly portrayed as retributivists—desiring proportional, and often painful, punishment of norm violators, even when it has no deterrent function (Carlsmith et al., 2002; Keller et al., 2010). What the research actually shows is that people recommend or endorse proportional punishment for crimes. People mimic the familiar institutional punishment tendencies because that is what they have learned all their lives and presumably think is the proper response. People’s real attitudes toward punishment and their own punishment behavior may tell a different story. When there is less priming of crime and punishment institutions, people want a response from the perpetrator (Funk et al., 2014; Gollwitzer et al., 2011) and want the perpetrator to learn something (Cushman, 2015). People don’t necessarily want everyday norm violators to suffer deeply if a lesser intervention promises norm compliance. However, when people consider those who assault, rape and kill others, they see severe punishment as an acceptable response (O’Hear & Wheelock, 2020).

To summarize, evidence suggests that everyday punishment is infrequent, unpopular, and effective only under specific conditions. Thus, punishment is unlikely to be the primary social regulator of norm compliance in everyday life. Punishment, which arose over the past 12,000 years, may enjoy popular support, not because people are deep-seated retributivists but because they have adopted the logic and practice of institutional punishment. As a threat, punishment sometimes works (Gächter et al., 2008)—people do slow down (briefly) when they see a police car on the highway. But standard legal punishment is surprisingly ineffective (National Institute of Justice, 2016; Yukhnenko et al., 2020) and appears to serve more the continued affirmation of dominance hierarchies than the successful social regulation of norm compliance (Sidanius & Pratto, 1999).

The Prospects of Moral Criticism for Regulating Norm Compliance

The term blaming is often used to refer to unjustified, unfair attacks, such as in the “blame game” (Knobloch-Westerwick & Taylor, 2008), in which accusations of others are used to deflect fault from oneself. Dysfunctional forms of blame certainly exist (Douglas, 1995; Furlong & Young, 1996), but this is not the meaning of blaming I focus on here. The question I pursue is whether blaming as functional moral criticism, grounded in considerations of facts at hand (Malle et al., 2014), can and does regulate norm compliance.

Whereas punishment is a more recent phenomenon that emerged as part of the hierarchies and legal systems after human settlement, blame is an older form of regulation that predominated in small hunter-gatherer societies and continues to exist prominently today. We have seen that punishment is typically coercive regulation, often institutionally anchored, to correct or incapacitate transgressors. Blame is socially expressed moral criticism, often among equals,
typically with room for explanation, corrective action, and reconciliation (Voiklis & Malle, 2018). This gentler, more cooperative tool of regulation stems in part from the social demands in small groups. In a band of 30, each member is needed, and debilitating sanctions hurt the community. Further, an egalitarian social structure forbids assertions of power by coercively punishing a norm violator and instead favors communication, sometimes ridicule, and thus a threat to social reputation (Wiessner, 2005). Repeat offenders are ignored, isolated, or, as a last measure, expelled from the group.

This more measured approach to communicate moral criticism is grounded in blame judgments, which appear to be the most complex and well-calibrated moral judgments (Malle, 2021; Monroe & Malle, 2019). They are formed from a range of information: the norm that was violated; the agent’s causal contribution to the violation; whether the contribution was intentional; if it was perceived intentional, what the person’s reasons were and their potential justification; and if it was perceived unintentional, whether the agent should have and could have prevented the violation (Alicke, 2000; Laurent et al., 2016; Malle et al., 2014). Being sensitive to all this information enables blame to be a graded and evidence-based moral judgment.

My colleagues and I have proposed that part of the reason for the evidence-based information processing in blame judgments is the fact that people appreciate that even modest blame is costly (Malle et al., 2014; Malle et al., 2022). Costs for the perpetrator include loss of status, damage to a relationship, and bad feelings; costs for the blamer include risk of retaliation and embarrassment if blame is misplaced; and even for other community members, costs arise from the burden of damaged relationships within the group and the potential for unfair treatment (e.g., false accusations) that would be toxic to the community. In light of these costs, the community puts demands on expressed blame: to make likely that blame is backed by evidence and is proportional to the violation and its circumstances. When challenged, a blamer has to be able to present the relevant evidence and thus offer warrant for the act of blaming (Malle et al., 2014; Voiklis & Malle, 2018)—specify, through evidence, why the perpetrator deserves the judged amount of blame. Valid evidence lies in just the information that people process, such as causality, intentionality, reasons, etc. Such evidence can be contested by the blamer’s audience (including the perpetrator)—for example, that the violation wasn’t intentional, or that it was intentional but done for justified reasons. In the course of debating the evidence, the judgment can become refined, further attuned to the facts of the case. If demands for warrant hold, then blame judgments have what, in the law, has been called “procedural fairness” (Tyler & Jost, 2007) and hold promise for effective regulation of norm compliance.

How frequent is blame?

Researchers sometimes call “punishment” what seems, on closer inspection, to be moral criticism. In Molho et al. (2020), when participants recalled their responses to acts they deemed morally wrong, they reported “confronting” the perpetrator 35% of the time. However, about 78% of the confronting acts were communications, requests, or explanations, and the rest were shouting and yelling and one withdrawal of help (see Molho et al., 2020, supplementary materials, pp. 23-24). Contrast this to the recall study by Pedersen et al. (2020), where the rates of more explicit third-person punishment (e.g., hitting, insulting) were 1%-6%. In studies that observe naturally occurring moral criticism, we find rates that greatly vary by context. When alone in facing a stranger who littered, 4% of people in Greece (Balafoutas & Nikiforakis, 2012) and 17% in Germany (Balafoutas et al., 2014) asked the norm violator to pick up the garbage. In greater safety, around others in a train’s silent car, 50% of people told a stranger to turn off their
music (Przepiorka & Berger, 2016). And when in a crowd, 19% in Greece (Balafoutas & Nikiforakis, 2012) but up to 75% in Germany (Wolbring et al., 2013) verbally intervene on a stranger who blocks an escalator’s fast left lane. Once more, the most common interventions were requests or demands to move, verbally or with a tap on the shoulder.

Third-person blaming refers to telling other people about a norm violation when the norm violator is absent. Molho et al. (2020) found that people who witnessed a violation preferred to talk about (“gossip”) than to confront the violator, especially when they had less power and when the norm violations were severe. When people gossip to make norm followers known to others, it can guide community members’ own partner choice (Feinberg, Willer, et al., 2012; Smith, 2014). However, such warnings may veer into punishment when they actively damage reputation and impose opportunity costs (Feinberg, Cheng, et al., 2012).

**How effective is blame?**

The behavioral economics literature shows positive effects of blame in the form of expressed disapproval. In a voluntary contribution game, such expressions raised cooperation as much as monetary “punishment” (Masclet et al., 2003); in a coordination game, expressed disapproval led to near-optimal collective outcomes (Dugar, 2010); and in an Ultimatum Game, even just anticipating communication between proposer and responder increased proposers’ offer levels, presumably to avert potential moral criticism (Capizzani et al., 2017). Such anticipated blame is powerful in smaller communities, where a potential norm violator will interact with, and be criticized by, potential future witnesses or victims. In such communities, the power of gossip and reputation encourage people to selectively choose cooperators and ostracize defectors, which can bring those defectors back to normal levels of cooperation (Feinberg et al., 2014).

Dersley and Wootton (2000) reported that 95% of second-person complaints (many of which can be classified as blaming) are to some degree contested, and Alberts (1989) found that 65% of spousal responses to their partner’s complaints are denials and justifications. Such contested communication between victim and perpetrator can have benefits, however, because the purpose of blame is not only to change the other’s behavior but to mark a problem, reaffirm norms, and reconcile.

**Individual Requirements for the Success of Sanctions**

In the foregoing discussion it already became clear that social sanctions, in the form of punishment or blame, by no means automatically alter the perpetrator’s course of action. The individual must be sensitive to the sanction — fearing punishment, feeling hurt by criticism, worrying about one’s reputation. Indeed, except for psychopaths, people value being accepted in social communities, as they have a “need to belong” (Baumeister & Leary, 1995), strive to be “morally adequate” (Steele, 1988), and suffer pain from social rejection (Eisenberger et al., 2003; Williams, 2009). This, in short, is the contract: Behave according to our norms, and we will value you as one of us; violate our norms, and we will take from you what you value: us.

**Alternative Regulation Methods**

I extensively discussed punishment and moral criticism, in part because research has devoted so much attention to these two social regulation responses and in part because they do appear predominant in modern life. However, there are historic, indigenous, and novel social and legal practices that offer alternatives but are only slowly being recognized in the literature.

One of the more frequently discussed mechanisms is forgiveness. Rather than demanding
punishment, the victim or community express their grievance but forgive the perpetrator and allow victim-perpetrator relationships to be restored (Gobodo-Madikizela, 2008). There is evidence that forgiving a transgressor actually reduces the likelihood of repeat transgressions (Wallace et al., 2008), but acknowledgment of the wrongdoing appears to be necessary in many cases before forgiveness is granted (Hamber, 2007; Vollhardt et al., 2014). Also, for some atrocities, feelings of hurt and revenge simply do not allow forgiveness (Kurgan, 2001). When forgiveness is possible, it often has benefits for the victim: Those who forgive feel a sense of justice (Wenzel & Okimoto, 2010), are seen as moral and trustworthy (Heffner & Feldman-Hall, 2019), and receive more compensation from community members (Hershcovis & Bhatnagar, 2017).

A second mechanism is restorative justice (RJ), which has a good track record in lowering recidivism in the legal domain (Kennedy et al., 2019) and in improving psychological outcomes among crime victims (Goodmark, 2015; Lloyd & Borrill, 2019). Its success may in part stem from relying on the structure of moral criticism in its best form, where acknowledgement of the violation, norm affirmation, and reconciliation replace damaging penalties—practices that overlap with a number of indigenous justice systems (Chartrand & Horn, 2016; Wielenga et al., 2020). It should be noted that RJ still imposes costs on the perpetrator, such as facing the victim, addressing the violation and the damage it caused, experiencing guilt, and sometimes providing compensation. In this sense, RJ is not free of retribution (Goodwin & Gromet, 2014), and some “retributive” measures (but not classic punishment) may actually improve its outcomes (Gromet & Darley, 2006).

A final alternative, probably the least known and most ambitious, is transformative justice, TJ (Barrie, 2020; Generation FIVE, 2007). Like RJ, it rejects the destructive and ineffective carceral system and considers the perpetrator as part of a larger community. It goes further by aiming to not only restore but transform the relationships between perpetrator, victim, and community (Barrie, 2020). TJ does not force the victim to reconcile with or forgive the perpetrator. It demands insight, accountability, and change from the perpetrator but also from the community. It thereby tries to change the broader causal nexus that enabled violence in the first place (Howe, 2018), often with a long-term perspective, such as ending child sexual abuse within five generations (Generation FIVE, 2007). The transformative process includes reflecting on community norms, both identifying norms that are harmful (e.g., those enabling child sexual abuse) and those that create accountability, safety, and healing (Ansfield & Colman, 2012). TJ may not be a standard mechanism of regulating norm compliance, but it offers a community-level analysis that the common discourse of sanctions rarely considers.

Conclusion

No human community can exist without norms. As mental representations of directives that the community puts on individuals, norms steer their actions toward community benefits, with small sacrifices asked of the individual. The sacrifices may at times feel natural, even desirable in the case of internalized norms, but the tension between individual and community interests requires regulation to motivate individuals to comply with norms. I discussed guilt, felt and anticipated by the individual and encouraged by the community; action justifications, offered by the individual and challenged by the community; as well as sanctions and their alternatives, enacted by the community and enabled by the individual’s valuing of community reintegration. All of these mechanisms are imperfect, some are outright counterproductive; but from the early, perhaps simpler beginnings of hunter-gatherer groups, human norm regulation has evolved to
uphold social benefits even as our society of strangers has become larger and stranger.

Humans are deeply social learners; not merely in the sense that they excel in social forms of learning, but in the sense that they so easily learn what it means to be social, to respect and live within the norms of one’s community.

Endnotes

1 Or never. Some authors have proposed that a norm of reciprocity is universal and innate (e.g., Krebs & Janicki, 2004). But reciprocity is more akin to a principle or a social process than a norm. It would be meaningless as a concrete action direction: “Reciprocate!” How? To what? To make it meaningful would require context constraints, as it is not always prescribed to reciprocate (e.g., not when the other’s favor is strategic or when the other does something immoral); and many constraints will vary culturally (e.g., rules of gift giving and matching).

2 Though rewards are sometimes used to regulate behavior, mere norm compliance is not normally rewarded (unless it is the first compliance after one or more violations). We would have to reward people nearly all the time for complying with norms; in reality, reward applies to actions that exceed the normative expectations.

References


Ferguson, T. J., Brugman, D., White, J., & Eyre, H. L. (2007). Shame and guilt as morally


interaction. MIT Press.


https://doi.org/10.1177/1368430213517270


https://doi.org/10.1016/j.jesp.2007.02.012


https://doi.org/10.12688/wellcomeopenres.14970.3