Chapter 17

Blame and Punishment:
Two Distinct Mechanisms for Regulating Moral Behavior

Bertram F. Malle
Brown University

Morality and its Regulation ................................................................. 2
Cultural History of Blame and Punishment ........................................ 2
  Hunter-Gatherer Communities and Moral Criticism ......................... 3
  The Emergence of Punishment Post Settlement ............................... 3
  Escalation of Punishment ............................................................. 5
Psychology of Punishment and Blame .................................................. 6
  Distinguishing Punishment and Blame ............................................ 6
  Understanding Punishment ............................................................ 6
    Forces that Maintain Support for Punishment ............................... 7
    Does Punishment Foster Cooperation? ......................................... 8
    Punishment Judgments .............................................................. 9
Understanding Blame .......................................................................... 9
  Blame as Judgment ....................................................................... 10
  Blame as Moral Criticism .............................................................. 14
Conclusions ..................................................................................... 16
References ....................................................................................... 17
Morality and its Regulation

To regulate group living, humans have developed two related social-cultural tools: a system of norms and complex social practices of regulating those norms. Central to norm regulation are responses to violations of norms. They include conciliatory ones, such as tolerance or forgiveness, and corrective ones, such as blame and punishment. Blame is often treated as parallel to punishment, as two currencies of moral sanctions or norm enforcement (Ames & Fiske, 2013; Berger & Hevenstone, 2016; Cushman, 2008). Indeed, both blaming and punishing can be pedagogical (Cushman, 2015; Malle et al., 2014), and both impose costs on the offender. However, the two are also importantly distinct (Baumard, 2011; Buckholtz et al., 2015; Malle, 2021). Blaming a transgressor is both a particular type of moral judgment (Malle, 2021) and, when socially expressed, an overt act of moral criticism. Punishment is always an overt act—and not merely one of criticism but one of damage. Because it would, in other circumstances, violate a person’s rights, punishment must be warranted or legitimized, typically by roles or institutions, such as parent, teacher, or state. Blame, too, must be warranted (Coates & Tognazzini, 2012; Malle et al., 2022), but blame can be withdrawn when found unwarranted whereas punishment cannot. Once inflicted, the damage of punishment is rarely reversible.

This is just a rough comparison. In this chapter, I investigate blame and punishment from two perspectives: their cultural history and their psychology. I also hope to make plausible the hypothesis that the distinct cultural histories of blame and punishment have shaped their distinct psychology.

I should acknowledge at the outset that some literatures use the term punishment as another word for all forms of norm enforcement or sanctions and then subsume blame (as verbal criticism) under the label punishment. I agree that both blame and punishment are forms of norm enforcement, but I hope to convince the reader that distinguishing between the phenomena of blame and punishment as well as between the corresponding terms helps clarify the important differences among the various tools of human norm enforcement.

Cultural History of Blame and Punishment

In human cultural evolution, responses to an individual’s norm violations have undergone enormous changes, often characterized as overcoming a brutal past (Farrington, 1996) to more “humane” present times (Pinker, 2011). However, if we look closely, there were really two distinct phases of human social living (Dubreuil, 2010) that gave rise to two different responses to norm violations: blame and punishment.

The first phase was a biological evolution from often violent dominance hierarchies in our primate ancestors to more egalitarian forms of human social living (Boehm, 2000), beginning perhaps as early as homo erectus, 1 million years ago (Dubreuil, 2010). As humans evolved in small communities of hunter-gatherers, they enforced social-moral norms with a system of sanctions that were primarily informal, interpersonal, and relatively mild (Wiessner, 2005)—much like today’s informal acts of blame as moral criticism.

The second phase began with human settlements after 10,000 BCE, leading to the formation of chiefdoms and early states, rapid population growth, and eventually culminating in nations and empires with large populations. Crucially, these societies re-introduced dominance hierarchies but legitimized them, not through our primate ancestors’ raw strength and aggression, but
through wealth, inheritance, military, and religious structures. Many norms became codified into law, shaping a system of institutionalized punitive sanctions imposed on the rest of the community by those at the top of the dominance hierarchies (e.g., chief, king, state).

I now flesh out this sketch of how two phases of human evolution gave rise to two forms of sanctioning, which form the origins of today’s tools of blame and punishment.

**Hunter-Gatherer Communities and Moral Criticism**

Until about 10,000 years BCE, humans lived as hunter-gatherers in small bands of 25 to 50 (Boehm, 1999; Knauf, 1994). We know this from some archeological finds (Bandy, 2004; Enloe, 2003), population genetic analyses (Atkinson et al., 2008; Henn et al., 2011) but predominantly from ethnographic research of hunter-gatherer societies over the past 100 years (Lee & Daly, 1999; Wilson, 1988; Woodburn, 1982). According to this accumulated evidence, hunter-gatherer communities were nomadic and thus highly mobile, changing camp every few weeks to months (Lee, 1972), mostly lacking possessions, and having little sense of territorial boundaries (Wilson, 1988). They were highly egalitarian and, without one supreme ruler, lawmaker, or judge, leadership was provided by different members for different tasks (Service, 1966).

The hunter-gatherer norm system was grounded in reciprocity, which regulated both community living and vital activities such as food acquisition and consumption. When a hunter shared his yield with another hunter one day, he could expect to receive part of the other’s yield another day. Indeed, sharing large hunts counted as a “virtually universal rule” among hunter-gatherer societies (Wilson, 1988, p. 37), and violating this sharing norm was met with sanctions. In these highly interdependent bands, sanctioning was interpersonal. Most transgressions were easy to detect (Silberbauer, 1982) because life was public and transparent. But sanctions of norm violators rarely consisted of punishment (Baumard, 2010; Guala, 2012), since these egalitarian groups disdained assertions of power and coercion, and debilitating penalties would have hurt the community at least as much as the transgressor. Instead, communities favored communication, criticism, sometimes ridicule (Wiessner, 2005), and gossip (Dunbar, 1996). Repeat offenders may have been isolated or, as a last resort, expelled from the group (Woodburn, 1982). But the typical way of responding to norm violations was a form of public criticism, a threat to the person’s social standing and reputation—which, in small, interdependent groups largely ensured norm compliance.

These civil forms of sanctions, however, occurred within groups and did not prevent highly punitive actions between groups. How much warfare occurred in early human hunter-gatherer communities is debated. But few scholars doubt that intergroup violence gradually increased with population density and with the correlated expansion from extended families to clan-based networks (Flannery & Marcus, 2012). These growth patterns set the stage for a transformation in the social organization of human societies.

**The Emergence of Punishment Post Settlement**

The second phase in the cultural evolution of sanctioning methods occurred after humans settled down, beginning around 12,000 BCE. Norm enforcement turned from an interpersonal process of criticism and threat to reputation to an institutionalized process that confronted the perpetrator with a ruling entity and its henchmen. The transition occurred in different places at different times, but the endpoint was, with few exceptions, a hierarchical society that centralized and
legitimized harsh punishment. A few essential facts about sedentary life help explain this change (Lee, 1972; Peregrine et al., 2007; Redman, 1978).

**Property.** As hunter-gatherers, people possessed only tools, tents, and clothes; food was distributed and eaten right away; and land had no boundaries of ownership (Wilson, 1988). After settling down, more and more people had permanent housing, household effects, eventually land, livestock, and crops. Quickly, some gained more property than others. Such inequalities were enlarged by inheritance rights, intermarriage, and reciprocal trade and protection among property owners (Peregrine et al., 2007). The have-nots desired others’ property, and the crimes of robbery and burglary emerged. The have-nots feared losing their property, so the need for protection, law enforcement, and punishment increased.

**Population growth.** Between 12,000 and 8,000 BCE, population growth was spurred by environmental changes (the end of the Last Glacial) and geographic opportunities (e.g., the fertile Near East; Aurenche et al., 2013), turning camps into villages and towns (Atkinson et al., 2008; Gignoux et al., 2011). Settlements also increased the rates of offspring. Whereas nomad communities had to carry their newborns for thousands of kilometers a year, limiting child birth to once every 3-4 years (Lee, 1972), settled communities were able to dramatically increase the frequency of pregnancies (Buikstra et al., 1986). Furthermore, plant and animal breeding experiments as well as improved food production and storage technologies fed more people. In a positive feedback loop, the resulting population growth put pressure on food production and demand for labor. Consequently, a second population growth turned towns into cities, several of which grew to 30,000 inhabitants or more between 2000 BCE and 1200 BCE. Another feedback loop then emerged between rapidly increasing built structures (Wilson, 1988)—from private homes to royal monuments and sacral edifices—and the massive opportunities for work to sustain this building boom (Boserup, 1965). That work was likely done by those who didn’t own property—the poor, foreigners, and conquered populations; thus, dependent labor and class stratification accelerated.

**Intergroup conflict.** By and large, hunter-gatherer societies were peaceful (Kelly, 2000; Silberbauer, 1982). There was little reason to attack another band (what do they have that we don’t? There was also little reason to fight over territory that constantly shifted (Woodburn, 1982). But once there were settlements, some land was more fertile, some villages more productive, and fast-growing populations had a need to either intensify food production (Carneiro, 1970) or expand their territory (Dumond, 1972), especially because food resources may not have kept pace with population growth (Larsen, 1995). Thus emerged incentives for organized warfare (Milner et al., 1991; Redman, 1978) to gain land, food, raw materials, tools, and to acquire further labor by subjugating other communities. As potential targets for attack and raids, settled communities cared about their safety, developing weapons and defensive facilities, which further demanded an expanded labor force and eventually a professional army.

**Hierarchical social structure.** All these changes—growing and unequally distributed property, population growth, dependent labor, and intergroup hostilities—demanded organization and administration, taken on by a few that turned into elites and authorities (Redman, 1978). New crimes emerged when this authority was violated, such as treason, social revolt, or tax evasion. Leadership structures were also necessary to maintain within-group cohesion in light of new role differentiation and resulting conflict potential (White, 2007). Settled humans were arranged from slaves to freemen to land owners, eventually aristocrats and priests, and almost always a supreme ruler. In the hundreds of thousands of years prior to settlement, primate hierarchies were
transformed into egalitarian, cooperative human bands; within just a few thousand years, a new society was born, returning to hierarchical and often brutally competitive primate roots.

**Escalation of Punishment**

In small hunter-gatherer communities, a norm violator had equal-status relationships with other group members, and these relationships had to be maintained for continued group success. Sanction severity for norm violations was therefore restrained because harsh penalties might invite equally harsh responses in the future and harm community relationships. By contrast, in larger settlements, ties between norm violator and the ruling class became weak and asymmetric, promising scant reciprocal future interaction and setting few limits on sanction severity. Over thousands of years of growing social hierarchies, we therefore see both increasingly centralized and increasingly brutalized forms of punishment. Several causal processes contributed to such escalation of punishment.

Punishment became more severe when rulers defended themselves against threats to their position. For example, in 18th-century Tahiti, the chief left it to family heads to enforce rules about property or marriage but applied severe punishment when offenses were “sacred,” typically those against his interests (Claessens, 2003). In 19th-century despotic African Buganda, sacred crimes, especially those against the ruler, were met with punishments of maiming, cutting into pieces, and burning alive (Claessens, 2003). Political rulers in China’s hierarchical societies BCE (Bozan et al., 2001) also relied on stern rules and punishment, especially during times of civil unrest. The death penalty applied to many offenses against the emperor (e.g., treason, rebellion) and the state (e.g., malpractice, bribery, illicit sales). In the early Roman Republic, laws were in tight control of the ruling class and were even kept secret from the lower class. When the laws were finally publicized in the Twelve Tables (c. 450 BCE), they contained detailed civil statutes but also prescribed capital punishment for those who defamed others in song (VIII.1b)—to avert any chance of inciting the proletariat’s sentiments against the ruling elite. Over the centuries, wars and uprisings caused tyrannic rulers to take the reins of the Roman Republic, and punishments became excessively cruel, even public spectacles, to deter any threat to the elites (Moore, 2001).

War is another driver of severe punishment. War brutalizes relations between communities, and it brutalizes relations within. When exposed to war, violence is part of life, and aggressive behavior becomes disinhibited (MacManus et al., 2011). In war times, rulers and state governments also more easily justify violent punishment within the state, especially for crimes of treason or cowardice. For example, before and during World War II, Nazi Germany took capital punishment to incomparable extremes, with a huge increase in death warrants and executions (Messerschmidt, 2005). The death penalty was applied to a wide range of criminal acts, various forms of military disobedience, and many forms of political resistance.

Despite the brutality of punishment in many early societies, it did not succeed all that well in deterring crime, which then further incites harsher punishments. We see this, for example, in the “war on crime” of the 1980s (Pratt et al., 2005), reflecting the misguided hardening of punishment in response to (perceived) rising crime rates and the failure to loosen punishment severity when crime rates decline. To this day, increased punitiveness does not result in crime reduction (Cullen et al., 2011). The history of escalating punishment hardly illustrates a successful method of moral regulation but an expression of steep social hierarchies and their dominance-maintaining mechanisms (Sidanius & Pratto, 1999).
Psychology of Punishment and Blame

I now shift from a cultural history of blame and punishment to the contemporary psychology of blame and punishment. I begin with an outline of their broad differences and then devote more detailed attention to our current understanding of these two major forms of sanctioning.

Distinguishing Punishment and Blame

Everyday moral sanctions range from raised eyebrows to muttered disapproval, from emotional complaints to thoughtful criticism; but there are also acts of physical aggression, withholding of resources, public shaming, and social exclusion. The first set are more representative of moral criticism (blame, for short) whereas the second set is more representative of genuine punishment.

Even though blame and punishment are not categorically distinct, a bundle of features allows us to distinguish them. Blame has a dual nature as a particular kind of moral judgment and as communicated moral criticism; punishment is always an observable act. When expressed, blame often requests or demands that the violator change their (future) behavior; punishment coercively imposes costs on the transgressor, often relying on power or status to legitimize the imposition of costs. Blame is typically reversible—one can admit that one unfairly blamed the other and take back the criticism—whereas the tangible costs of punishment can rarely be taken back. Finally, blame can be contested, leaving room for justification and negotiation; punishment leaves little room for negotiation and is often the last resort when moral criticism has failed.

In light of these distinguishing properties, we would expect that in informal, everyday interactions, humans prefer to use moral criticism over punishment. Indeed, using daily surveys to study naturally occurring norm violations, Molho et al. (2020) found that people were generally unmotivated to engage in physically confronting punishment and somewhat more motivated to gossip or exclude the transgressor. Field studies show that people enforce norms with subtle verbal or nonverbal interventions but not active punishment (e.g., Przepiorka & Berger, 2016). In a cross-cultural study, rates of responding to violations with cost-imposing sanctions (e.g., yelling, insulting, hitting) were extremely low, between 1% and 6% (Pedersen et al., 2020). Even though a number of behavioral economics experiments have suggested that people are willing to punish norm violators, this form of punishment is mild and indirect (taking a little money from a stranger), and when given a choice, people prefer not to punish but to criticize the violators to their face or to others (Feinberg et al., 2012; Kriss et al., 2016; Xiao & Houser, 2005) or compensate the victim (Chavez & Bicchieri, 2013; Hershcovis & Bhatnagar, 2017), or even restore order by correcting the negative outcome (e.g., cleaning up other people’s litter; Berger & Hevenstone, 2016). As we move into more hierarchical relationships—parent and child, teacher and student, boss and employee—punishment becomes more likely, more consequential, and is often seen as more legitimate (Gershoff & Lee, 2020; Mooijman & Graham, 2018). Embedded in such hierarchies, “[p]unishment is not a behavior, but an institution” (p. 321, Binder, 2002). Indeed punishment is the predominant sanctioning instrument used by the state (Tonry, 2009).

Understanding Punishment

If interpersonal punishment is relatively infrequent in daily life, why is it so wide-spread in its institutionalized form? The brief cultural history of punishment has taught us that punishment has been used to maintain hierarchies, defend the powerful, and remove the unwanted. The severity of punishment has declined over time, but to this day, punishment is an instrument of
power, oppression, and the maintenance of dominance hierarchies (Redford & Ratliff, 2018; Sidanius et al., 2006). And not only is punishment discriminatory, it is also ineffective as an instrument of deterrence (Cullen et al., 2011). If punishment tends to fail in so many ways, what maintains the public support for institutional punishment (recent declines of such support notwithstanding; (Schwaeble & Sundt, 2020)?

**Forces that Maintain Support for Punishment**

**Retribution**

A first hypothesis is that people are fundamentally retributivists—desiring proportional, and painful punishment of norm violators, even when it has no deterrent function (Carlsmith et al., 2002; Goodwin & Gromet, 2014). But rather than explaining common support for institutional punishment, people’s retributive tendencies may be themselves explained by their exposure to institutions of punishment. Research on lay retributivism always probes whether people recommend or endorse retributive punishment for crimes, such as assault, rape, or murder. For those violations, people may endorse severe punishment because that is what they have learned to be the proper response. However, people are not merely reflecting back the severity of punishment in legal reality; they are often manipulated by the political rhetoric, more so from the right, that crime is rampant and society needs to be tougher on crime (Muhammad et al., 2015; Tonry, 2009).

However acquired, what is it that retributive tendencies really support? The current literature often defines retributive commitments as referring to “deserved punishment of a guilty offender” (Goodwin & Gromet, 2014, p. 562). People are retributivists, then, if they endorse a statement such as “by punishing offenders we give them what they deserve” (Nadelhoffer et al., 2013). In what sense is deserved punishment “retributive” but a deserved award is not? (Cottingham, 1979, p. 239). In both cases, the response is considered appropriate or justified, given what the person actually did. This notion of desert stands in contrast to the extreme consequentialist position that what the transgressor did (or thought while doing) is irrelevant; that the only thing that matters for punishment is its beneficial consequence (e.g., deterrence of crime, calming public outrage). Rejecting this position, people endorse retribution in the sense of paying back (Cottingham, 1979), which is what the original 12th-century English word referred to. Such retribution was a symmetric practice of moral accounting—what is paid back are both rewards and punishments.

If retribution is the mirror image of the human disposition toward reciprocity for good deeds (Gouldner, 1960), then being retributive need not imply cruel enjoyment of a wrongdoer’s suffering. Instead, a scale is being balanced between the transgressor’s behavior and the community’s norms. This is why goddess *Justitia* holds scales in her hand to symbolize justice. Whether the repayment is cruel or civil—mutilations and executions, monetary fines, or apology and atonement—depends on the historically and geographically specific currency in which social-moral debts are repaid. Thus, even if ordinary people’s support for retributive punishment is grounded in a general principle of reciprocity, it has undoubtedly been nourished by longstanding patterns of political and religious influence (Griffith, 2020).

**Delegation**

A second contributor to support for institutional punishment stems from the dangers of personal punishment. Punishment inflicts damage on the perpetrator and is typically an act of aggression; that person may therefore retaliate. Especially a perpetrator who has committed a serious
transgression has already shown little commitment to the community’s norms, so the person, when punished, will repay the inflicted damage. In hunter-gatherer societies, personal interventions were more pressing and more likely in response to extreme violations (Kelly, 2000, p. 27). However, they came with significant risk. Lee (1979) reported that almost half of the !Kung community’s fatalities over several decades were bystanders or “peacemakers,” who died in the attempt to stop a conflict. Retaliation against norm-enforcing punishment is even more likely to be committed by people outside one’s group because the perpetrators have no relationship with the norm enforcer or the community and thus no reason to show remorse, correct their behavior, and protect their reputation. Indeed, in human history, intergroup conflicts have often escalated from individual revenge to family- or group-based revenge, spiraling into blood feuds (Boehm, 2011).

In all these cases, the costs to directly punish a perpetrator are high, but the risks of retaliation can be reduced or even eliminated by handing the task of punishment to specific agents or institutions (Cushman, 2015). Such delegation of punishment protects ordinary people and allows them to safely condemn violators and support the punishment meted out by the authorities—even excessively harsh punishment. It is far easier to cheer on the torturer or vote for “tough on crime” policies than it is to break the prisoner’s arm oneself or stand guard at a solitary confinement cell. Delegated punishment, however, requires justification, which a ruler or state has in general and can award to particular officials, such as judges, prison guards, or executioners. Institutionalized punishment guarantees legitimacy without having to establish it anew each time (Binder, 2002; Garland, 1990). As long as the institution is seen as reasonably consistent and fair, the community will support it and grant it the permission to punish some of its members.

**Does Punishment Foster Cooperation?**

Leaving the institutionalized context, punishment in small groups can foster cooperation by decreasing the number of free riders and increasing the contributions individuals make to the group (Fehr & Gächter, 2000; Yamagishi, 1986). Punishment works best when it articulates community norms (Andrighetto et al., 2013; Xiao, 2018), teaches the violator (Cushman, 2015), brings about a response from the violator (Funk et al., 2014), and when it is gentle and deemphasizes power differences (Kochanska & Aksan, 2006).

But punishment can go awry. When cooperators are punished, costly punishment no longer promotes cooperation (Rand et al., 2010), and the option to punish can lead to retaliation (Hopfensitz & Reuben, 2009). When social communities view punishment as unwarranted, it is less effective at correcting social behavior (Herrmann et al., 2008). Furthermore, impressions of third-party punishers are less than favorable (Dhaliwal et al., 2021), and even victims who punish are not seen very positively; instead, victims who forego punishment are seen as moral, trustworthy, and altruistic (Heffner & FeldmanHall, 2019).

Punishment may not even be necessary for cooperation (Baumard, 2010). Punishment’s apparent power to increase cooperation in mixed-motive games is matched (Feinberg et al., 2014) or surpassed (Wu et al., 2016) by that of gossip, and it is also matched by communicated disapproval (Masclet et al., 2003). When compared in terms of overall payoffs in these game contexts, both gossip and disapproval lead to better collective outcomes than does punishment (Dugar, 2010; Wu et al., 2016).
Punishment Judgments

Research has examined punishment not only as a behavior of imposing costs on another person but as ordinary people’s recommendations for the legal punishment a transgressor should suffer. Researchers have studied the kinds of information people are sensitive to and the various judgment biases they may fall prey to. On the information side, there is consistent evidence that people assign different degrees of punishment (e.g., length of prison term or monetary fines) as a function of violation severity, intentionality, foreseeability, and justification (Kneer & Machery, 2019; Robinson & Darley, 1995), largely consistent with what the law prescribes. In addition, however, studies document that punishment recommendations are susceptible to factors that should arguably not influence legal judgments—for example, the defendant’s ideology (Sood & Darley, 2012), physical or social attractiveness (e.g., Nemeth & Sosis, 1973), stereotypically Black appearance (Eberhardt et al., 2006), or the juror’s prejudice (Gamblin et al., 2021).

But while several studies do confirm biased punishment recommendations, numerous other studies failed to find similar biases in people’s verdicts—that is, judgments of whether a person is guilty or not (Block, 1991; Stewart, 1985). Why might people be more calibrated in their judgments of guilt than in their punishment recommendations? First, punishment and sentencing recommendations are inherently more discretionary, comparative, and subjective than are guilty verdicts (Tata, 1997), creating more room for errors and biases. Second, at least in the United States, people don’t normally make punishment recommendations—as jurors, their role is to come to a verdict, whereas judges decide on punishment. Thus, research that measures legal punishment recommendations may ask people for unusual, unfamiliar judgments, which may be more susceptible to biasing information. Third, previous studies often treated past transgressions, character, and motive as “biasing” factors when in fact they often are legally admissible for sentencing, if not for guilty verdicts (Guglielmo, 2015); so by and large, people’s judgments are consistent with the law.

In no way does this imply that verdicts by research participants or real jurors are generally unbiased. One of the most enduring biases in the United States criminal system is racial discrimination, which exerts its impact through police work, prosecution decisions, pre-trial detention, guilty pleas, testimony, all the way to verdicts and sentencing (Kansal, 2005; Mitchell et al., 2005; Sutton, 2013). Research is also growing to document similar biases toward other marginalized groups (e.g., Mirabito & Lecci, 2020). Such discrimination replays the hierarchical dominance patterns that have tilted punishment systems since early human settlement.

Understanding Blame

We have seen that people rarely mete out punishment in everyday life, and punishment recommendations occur more often in research studies than in the courtroom. By contrast, blame judgments occur frequently and naturally in everyday life. Interestingly, however, blame does not have the best of reputations in psychological research. Acts of blaming are often portrayed as distorted and dysfunctional. Some philosophical work, by contrast, has characterized blaming as more constructive moral criticism that calls out unacceptable behavior, affirms norms and values, and demands justice (Bell, 2012; Ciurria, 2019). Research on “complaints” in sociology and pragmatics further shows the important function that criticism plays in regulating interpersonal behavior and the detailed work that goes into staging and warranting this kind of criticism (Morris, 1988). Finally, recent research highlights the sophisticated information
processing that seems to underlie blame judgments (Cushman, 2008; Monroe & Malle, 2019), at least when the community enforces standards of evidence (Malle et al., 2022).

I now take a closer look at these opposing facets of blame, dividing the review into cognitive blame (judgments in the head) and social blame (acts of moral criticism).

**Blame as Judgment**

*An information processing model of blame*

A substantial literature shows that blame assigned to a person varies as a function of a number of factors: the specific norm that was violated, the causal contributions the person makes, whether the person acted intentionally, what mental states or reasons the person had and how justified they were, and whether the person could have and should have prevented the norm violation (Cushman, 2008; Guglielmo & Malle, 2017; Monroe & Malle, 2019; Nadler & McDonnell, 2012).

My colleagues and I have aimed to integrate these findings in the Path Model of Blame (Malle et al., 2014), depicted in Figure 1. The model (a) specifies the various pieces of information that people normally process en route to blame; (b) identifies a systematic order of processing—such that, for example, a question of intentionality comes up only when a causal contribution has been perceived (Guglielmo & Malle, 2017); and (c) predicts patterns of blame updates when new information enters the processing stream (Monroe & Malle, 2019). A unique feature of the model is that intentionality does not merely influence blame judgments, like in other models (Alicke, 2000; Cushman, 2008) but bifurcates the processing into distinct search paths (Monroe & Malle, 2017): for the transgressor’s reasons that motivated and may justify an intentional violation; for evidence that the transgressor could have prevented an unintentional violation.

The Path Model provides an orienting framework that allows us to examine a number of important questions about blame: to what extent it is motivationally biased, whether it is “intuitive,” and how it relates to emotions.

*How much does blame suffer from motivated bias?*

All human information processing, moral judgment included, is imperfect: People have attentional lapses, forget information, or fail to integrate information under time pressure. Of greater concern in the moral psychology literature have been questions of *motivated* bias (Alicke, 2000; Ditto, 2009; Nadler & McDonnell, 2012).

Motivated moral judgment is usually described as a “desire to blame” (Alicke, 2000; Ames & Fiske, 2013), followed by information processing that strives to confirm the initial blame. However, when appropriate control conditions are implemented, people appear equally responsive to exacerbating information as to mitigating information (e.g., Monroe & Malle, 2019; Nadler & McDonnell, 2012, Experiment 2). But even without a general desire to blame, people may still show confirmation bias. Yet making an initial blame judgment and later revising it does not come with a confirmation bias (Monroe & Malle, 2019). Such a bias typically requires a more powerful source, such as the blamer’s relationship with the transgressor (Forbes & Stellar, 2021), the blamer’s ideology (Niemi & Young, 2016), or salient “extra-evidential” information (Alicke, 2000)—that is, information that one should not take into account when forming blame judgments.
Figure 1. The Path Model of Blame, adapted from Figure 2 of Malle et al. (2014). People form blame judgments by processing the depicted information components, typically in the order indicated. The fourth phase differs depending on whether the event is considered intentional (4_i) or unintentional (4_u).

The normative standards that determine what is extra-evidential, however, have been debated. Some scholars proposed that, even for everyday moral judgments, the standards should be set by “philosophers, legal theorists and psychologists” (Alicke, 2008, p. 179). Unfortunately, these experts do not necessarily agree on the pertinent normative standards, and such standards vary across cultures and historical times. So perhaps it is best to sidestep disputes over the normative standards themselves and ask, descriptively, which factors people take into account when forming blame judgments and how those factors influence blame. I will call these factors “exogenous” because they go beyond the canonical “endogenous” information components of blame, as assembled in the Path Model of Blame.

The current body of research appears consistent with the interpretation that, even when making blame judgments that some have called biased, people follow the information processing path depicted in Figure 1, though exogenous factors can alter steps in this path (Guglielmo, 2015). I briefly review research on two groups of exogenous factors.
As a first group, norms and values provide the standard against which people evaluate the initial violation (component 1 in Fig. 1), which shapes the degree of downstream blame. In Sood and Darley (2012, Study 3), participants disapproved of a man who went to the supermarket in the nude, and they did even more so when he handed out flyers promoting a position on abortion opposite to their own. Participants’ blame (and recommended punishment) was greater presumably because they saw more or stronger norms violated. Likewise, people seem to apply different norms to intimate others than to strangers, finding strangers’ transgressions less acceptable than intimates’ (Weiss & Burgmer, 2021).

Norms also influence what people accept as justifications for a given intentional norm violation (component 4 in Fig. 1). For example, the Southern and Northern U.S. tend to differ in how justified they find violent acts in service of protecting oneself or of defending one’s honor after an insult (Cohen & Nisbett, 1994); hence these groups will differ considerably in how much they blame such violent acts.

Finally, people’s norms guide the obligations people impose on a person to prevent certain unintentional violations. For example, people with hierarchy-legitimizing ideologies believe that someone with implicit racial bias is not obligated to prevent their own discriminatory acts that follow from this bias; as a result of this normative belief about prevention, they blame a person less who unintentionally committed such discrimination (Redford & Ratliff, 2016). Conversely, in liberal communities, people are obligated to avert microaggressions (Princing, 2019) whereas others criticize this obligation as “ridiculous” (Dickey, 2019). In all these cases, blame varies considerably across groups of people, not because of motivated bias but because they interpret the same events in light of different norms.

Beliefs and attitudes toward the transgressor constitute the second group of exogenous factors. Several studies have shown that judgments of causality or preventability can be affected by negative character information (Alicke & Zell, 2009) or the perceiver’s attitude (Niemi & Young, 2016). Moreover, character, past behavior, and stereotypes can affect people’s inferences of a transgressor’s intentionality or reasons for acting (Sood, 2019). So while there is good evidence that exogenous beliefs and attitudes can insert themselves into the blame processing path, some authors have proposed that these exogenous factors directly influence blame (Alicke & Zell, 2009; Ciurria, 2019, Chapter 4), effectively bypassing the endogenous blame processing. The current evidence, however, does not provide convincing support for such direct effects (Guglielmo, 2015). Exogenous and endogenous factors would have to be properly compared in one and the same study, but they are rarely jointly measured (Dukes & Gaither, 2017; Nadler & McDonnell, 2012); or, if they are measured, not all included in model tests (e.g., Alicke & Zell, 2009; Nadler, 2012); or, if included, do not yield the predicted direct effects (Mazzocco et al., 2004).

There is no doubt that people sometimes blame unfairly, to protect their self-esteem, enact their prejudice, or protect their group. Extant evidence, however, suggests that even in these cases people rely on the same information components—causality, intentionality, reasons, and so on—on which they rely when blaming fairly.

Is blame intuitive?

When scholars characterize moral judgments as intuitions, they rarely specify the type of moral judgment they are referring to. Some describe moral intuitions as feelings of “approval or disapproval” (Haidt, 2001, p. 818) or as some action being “bad” (Clark & Winegard, 2019, p.
14), which we may call evaluations (Malle, 2021). To be moral evaluations, these intuitions rely on the relevant moral norms being activated in the observer, who then judges the relevant behavior as violating the norm. Evaluations might code for simple causality (De Freitas & Alvarez, 2018) and potentially for the intentionality at least of visibly performed behaviors (Decety & Cacioppo, 2012). By contrast, blame judgments also incorporate information about the intentionality of unobserved actions, the agent’s specific, potentially justifying reasons, or the complex assessment of preventability obligations and capacities (Monroe & Malle, 2019)—the kind of processing that typically does not fall under “intuitions.”

Suggestions that moral judgments are intuitions are often tied to the claim that people are “dumbfounded” when trying to justify their judgments (Haidt & Bjorklund, 2008, p. 197). This claim faces a number of challenges. First, researchers routinely ask participants for simple wrongness judgments, which rely primarily on citing the relevant violated norm (Malle, 2021). But proponents of the dumbfounding hypothesis treat a norm statement (e.g., “because it’s incest”) as a dumbfounding response, whereas others insist that they constitute an actual justifying reason (Stanley et al., 2019). Second, proponents of the dumbfounding hypothesis instruct their experimenters to explicitly “undermine whatever reason the participant put forth in support of his/her judgment or action” (Haidt et al., 2000, p. 7), which arguably biases results in favor of the hypothesis (Gray et al., 2014; Royzman et al., 2015). Despite these favorable conditions, surprisingly small numbers of people give dumbfounding responses (e.g., 32% in McHugh et al., 2017; see Malle, 2021, p. 309); and in less coerced situations, dumbfounding drops below 20% (McHugh et al., 2020). These weak dumbfounding rates can hardly support the claim that moral judgments are by nature “intuitive.”

Is there any dumbfounding evidence for blame judgments? Quite the contrary. Bucciarelli et al. (2008, Study 3) showed that people had no trouble explicating their blame judgments in a think-aloud protocol, and Voiklis et al. (2016) documented that people provide rich and systematic explanations of their blame judgments. Importantly, these explanations referred to just the kinds of canonical information that are known to cause variations in blame judgments (e.g., the seriousness of the norm violation, intentionality, justified reasons).

**How does blame relate to emotion?**

Determining the role of emotion in blame requires heeding distinctions between affect, evaluations, and emotions. Affect is often understood as a nonrepresentational valenced feeling state (Neumann et al., 2001); evaluations are rapid valenced appraisals of some object (“this is bad”); and emotions are a large class of states (e.g., anger, resentment, or disgust) that are differentiated by appraisals—the cognitive processing of rich arrays of information (Scherer, 2013). Further, we must separate the different roles that these three phenomena can play in blame: they could fully constitute blame judgments, cause them, accompany (and possibly amplify) them, or be caused by them (Monin et al., 2007; Strohminger, 2017).

Affect cannot constitute or cause blame because it lacks any object; but it can accompany blame. Evaluations of norm violating events lie at the beginning of any processing stream toward blame; so they make a causal contribution; but they do not process much of the information that blame judgments are normally based on and are therefore insufficient to fully cause, let alone constitute blame. Emotions such as anger rely on a number of appraisals (Ellsworth & Scherer, 2003) of just the kind of information that blame judgments respond to (e.g., causality, intentionality), so the two may co-emerge from this information processing. However, one can arrive at blame
judgments by processing the relevant information without feeling angry; but it would be difficult to become angry before processing this kind of information (e.g., causality, intentionality). Emotions thus do not seem to constitute moral judgments but may accompany and amplify them.

Researchers have attempted to induce anger to test whether it indeed amplifies blame or related moral judgments. Although some studies showed such impact (Ask & Pina, 2011; Lerner et al., 1998; Seidel & Prinz, 2013), others did not (Gamez-Djokic & Molden, 2016; Gawronski et al., 2018). Conversely, judgments of blame and responsibility can alter emotions or mediate between norm violations and emotions (Quigley & Tedeschi, 1996; Zajenkowska et al., 2020). Studies also found expressions of anger and moral outrage to arise after judgments of moral violations (Sasse et al., 2020), and response times of expressing anger were no faster or even longer than response times to blame judgments (Cusimano et al., 2017).

We can conclude that when people encounter a norm violation, they are likely to morally evaluate it, may or may not experience feelings along with it, and routinely process a canonical set of information. This information processing guides blame judgments and may generate emotions, but which specific emotions arise (e.g., anger, resentment, indignation) will depend on the specific processed information. These emotion then help scale the intensity of a socially communicated blame judgment (Drew, 1998). And this brings us squarely to the phenomenon of blame as communicated moral criticism.

**Blame as Moral Criticism**

Social acts of blame, reproach, and rebuke serve several goals. They draw attention to a transgression (McGeer, 2012), signal the blamer’s commitment to a norm system (Shoemaker & Vargas, 2021) and try to enforce it (Bell, 2012); they try to change the transgressor’s ongoing or future behavior (Miller, 2003; Przepiorka & Berger, 2016), but they also simply express moral judgments and moral emotions (Sorial, 2016). Acts of blaming are costly, however, for all involved parties: the transgressor, the moral critic, and the community (Malle et al., 2022). The transgressor suffers a loss of public standing, damaged relationships, and just plain bad feelings. The moral critic faces potential transgressor retaliation (Balafoutas & Nikiforakis, 2012), damaged relationships, and possibly lack of support from the community. And the community itself carries the costs of escalating community strife (Allen, 2002).

To keep these costs in check, most social communities impose norms on moral criticism (Coates & Tognazzini, 2012; Eriksson et al., 2017; Voiklis & Malle, 2018), which regulate the standards of evidence and discourse of blame (Friedman, 2013; Malle et al., 2022). When moral critics comply with these norms, they are also more likely to achieve the goals of their criticism.

Complying with norms of blaming means that moral critics will process the available evidence and assign blame proportionally to the observed transgression. If they ignore evidence, they are bound to blame out of proportion—they will either overblame or underblame. Overblaming is itself a norm violation and will provoke rejection or retaliation from the transgressor and may cause long-term damage to a relationship (Fincham et al., 1987). Underblaming may downplay the violated norm and thus fail to change the offender’s behavior. However, mild, subtle forms of criticism for everyday transgressions (e.g., blatant littering; an able-bodied person using a disability parking spot) can be effective ways of changing behavior. In most social communities, even a cold stare or a snide remark may successfully communicate moral criticism (Miller, 2003;
Molho et al., 2020). The critic thus affirms the offender as a respectable member of the moral community who is nonetheless worth criticizing (Bennett, 2002).

The transgressor can of course influence the success of moral criticism. Blame demands a response (Drew, 1998; McGeer, 2012; Shoemaker, 2012), and transgressors have many options—on the one hand, they can deny, dismiss, or retaliate (Dersley & Wootton, 2000); on the other hand, they can explain, apologize, and offer repair (Walker, 2006; Watanabe & Laurent, 2020). Without reconciling responses, the regulation of social relationships is bound to fail (Laforest, 2002).

The community, finally, upholds the norms of blaming by combatting premature, biased, or inaccurate criticism, by demanding warrant from those who overblame (Voiklis & Malle, 2018). It thereby protects unfairly criticized community members but also maintains trust that one is treated fairly in one’s community. Conversely, it can ignore a moral critic’s warnings and allow transgressions to continue—for example, by suppressing moral criticism of those in power (Ciurria, 2019, Chapter 9).

**Forces that alter costs and success conditions**

The costs of moral criticism vary as a function of a number of factors, with implications for the norms and success conditions of the criticism (Malle et al., 2022). For example, 2nd-person blame (expressed directly to the transgressor) comes with the risk of retaliation and fails more easily when overblaming occurs; but when communicated appropriately it can be immediately effective in regulating behavior. 3rd-person blaming (expressed to other parties) is safer and effective in affirming the community’s norm system, but any change in the transgressor’s behavior can be achieved only indirectly and with likely delay.

Another influential variable that alters costs is the ingroup or outgroup status of the transgressor (Malle et al., 2022). When blaming ingroup members, moral critics face pressure to heed the community norms that stem against disproportionate blaming and strife. But when the transgressor is outside the community, these norms lose their force. Critics may get away with sloppy information processing and unfair accusations, because the costs have been shifted almost entirely from critic and community to the transgressor. In the punishment literature, when an outgroup member transgresses against an ingroup member, sanctions are greatest—from mild monetary penalties (Bernhard et al., 2006) to death by lynching (Equal Justice Initiative, 2017) or state execution (Eberhardt et al., 2006). And initial results suggest that blame, too, is less regulated when applied to outgroup members (Monroe & Malle, 2019).

A recently emerging and powerful factor that alters costs is online discourse, especially on social media, which can deliver public shaming or social exclusion in just a few words (Klonick, 2015). Online moral critics carry fewer costs (Crockett, 2017), most obviously when they are anonymous. Even when identifiable, they launch their blame from the safe distance of a keyboard, which restricts the other’s retaliation and poses fewer risks to valued relationship; after all, those who are blamed, reviled, and excluded are typically strangers and outgroup members. Online blaming also imposes fewer costs on the community because that community is indeterminate. Social bonds that normally tie community members are far weaker online, so people are less committed to norms (including norms of blaming) and put fewer constraints on derogations (Márquez-Reiter & Haugh, 2019).
The looser restrictions on online moral criticism, however, can empower critics who have previously been silenced. The MeToo movement, for example, enabled acts of calling out celebrities or companies who transgressed, or documenting microaggressions and outright discrimination. Ideological camps differ in their reception of such amplified criticism—which some call “cancel culture” (Republican National Committee, 2020). A warranted rebuke of hurtful insults (e.g., microaggressions) in one community is seen as a ridiculous demand of political correctness in another community (Dickey, 2019). Such polarized reception stems from community differences in the norms for the original actions (e.g., political correctness, microaggression) as well as in the norms that govern criticism itself—including who is granted credibility and standing to criticize and how. Ultimately, online communities will have to calibrate what is a proportionate response to violations. We see the difficulty of such calibration in the case of a critic who used Twitter to blame two offenders for using the word dongle in what was interpreted to be a sexually offensive way; which then caused one of the offenders to lose his job after the publicity; but the Twitter-posting critic was subsequently vilified and threatened by numerous people and fired from her own job for the act of public blaming (Brown, 2013).

Conclusions
I have tried to make plausible the hypothesis that the distinct cultural histories of blame and punishment have shaped their distinct psychological underpinnings and social expressions. Today’s blame arose long ago as the primary sanctioning behavior in small, tight-knit, egalitarian communities, where moral criticism is often mild but effective, built on the power of reputation and the need to belong. Because of the costs of damaged relationships and the potential for retaliation, such blame is regulated by a set of norms that demand evidence and fair treatment — at least for members of one’s own community. Moral critics are expected to communicate their disapproval and to invite a response from the transgressor, with the goal of repairing and continuing the relationship and averting a repeat transgression. Such blame can be contested and mitigated by the transgressor’s response, and the critic may even admit a mistake and take back the criticism. This ideal can be missed in numerous ways, social media being a salient contemporary case.

Punishment arose in larger, more distant, hierarchical communities. Acts of punishment would normally themselves be serious norm violations, were it not for the legitimacy granted to, or taken by, the punisher. Frequently from high up in the hierarchy, the punisher—a person or an institution—coercively imposes costs on the transgressor that cannot be taken back. Some forms of punishment step in where moral criticism has failed, but punishment still often stops communication rather than reinstating it, damages relationships rather than repairing them.

One major drawback of current institutional punishment is its ineffectiveness in deterring crime and reforming individuals who committed crimes (Yukhnenko et al., 2020). As one alternative to the standard punitive system, restorative justice is an attempt to reconcile perpetrator and crime victim through a mediated conversation in which all thoughts and feelings are expressed, often restitution is agreed on, and the matter laid to rest (Braithwaite, 1999). Evidence shows that recidivism decreases considerably when restorative justice replaces legal punishment (Kennedy et al., 2019; Kuo et al., 2010). Restorative justice procedures are noticeably close to ancient and modern processes of moral criticism and reconciliation. A hope for the future is therefore that institutions more widely adopt constructive moral criticism as the powerful form of norm enforcement that it once was.
References


Aurenche, O., Kozłowski, J. K., & Kozłowski, S. K. (2013). To be or not to be... Neolithic: “Failed attempts” at Neolithization in Central and Eastern Europe and in the Near East, and their final success (35,000–7000 Bp). *Paléorient, 39*(2), 5–45. JSTOR.


Brown, E. (2013, March 25). *Is Adria Richards a bully, or was she bullied by the internet?* ZDNet. https://www.zdnet.com/article/is-adria-richards-a-bully-or-was-she-bullied-by-the-internet/


Dickey, R. (2019, May 28). It’s not that people deny the existence (although some do) it’s just that it’s ridiculous to expect people to keep track of them [Comment]. Quora.Com. https://qr.ae/pGBil5


