Neuroscience and Biobehavioral Reviews 157 (2024) 105525

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

The science of justice: The neuropsychology of social punishment

Qun Yang[a,*], Morris Hoffman[b,*], Frank Krueger[c,*]

[a] Department of Psychology, Jing Hengyi School of Education, Hangzhou Normal University, Hangzhou, China
[b] Second Judicial District (ret.), State of Colorado, Denver, CO, USA
[c] School of Systems Biology, George Mason University, Fairfax, VA, USA

ARTICLE INFO

Keywords:
Social punishment
Second-party punishment
Third-party punishment
Motivation
Neuropsychology
Justice

ABSTRACT

The social punishment (SP) of norm violations has received much attention across multiple disciplines. However, current models of SP fail to consider the role of motivational processes, and none can explain the observed behavioral and neuropsychological differences between the two recognized forms of SP: second-party punishment (2PP) and third-party punishment (3PP). After reviewing the literature giving rise to the current models of SP, we propose a unified model of SP which integrates general psychological descriptions of decision-making as a confluence of affect, cognition, and motivation, with evidence that SP is driven by two main factors: the amount of harm (assessed primarily in the salience network) and the norm violator's intention (assessed primarily in the default-mode and central-executive networks). We posit that motivational differences between 2PP and 3PP, articulated in mesocorticolimbic pathways, impact final SP by differentially impacting the assessments of harm and intention done in these domain-general large-scale networks. This new model will lead to a better understanding of SP, which might even improve forensic, procedural, and substantive legal practices.

1. Introduction to social punishment

Social justice is an ideal shared across many human societies, and punishing norm violators (what we will call "social punishment" or SP) is a key component of it (Gintis et al., 2008). Evolutionary biologists have posited that SP was critical to the development of the social norms that arguably allowed our genetically heterogeneous emergent ancestors to survive in small intensely social groups (Boyd et al., 2003; Fehr and Fischbacher, 2004a; Fehr and Fischbacher, 2004b; Fehr and Gächter, 2002). Experimenters and theorists distinguish between two types of SP: second-party punishment (2PP) and third-party punishment (3PP), which differ depending on whether the punisher is the victim of the norm violation (2PP) or merely an observer (3PP) (Fehr and Fischbacher, 2004a; Fehr and Fischbacher, 2004b; Henrich et al., 2006) (Fig. 1).

Both types of SP arguably conveyed fitness advantages by reducing social norm violations and stabilizing our ancestral groups, by increasing the expected costs of violating social norms (Boyd et al., 2003). SP increased those costs both for norm violators considering future violations (what criminologists and other legal scholars call "special deterrence") and for everyone else in the group considering future violations ("general deterrence") (Bellucci et al., 2020; Fehr and Gächter, 2002; Henrich et al., 2006).

Second-party punishment is no evolutionary mystery; it is a form of self-defense that gave its wielders overt survival advantages in addition to deterring future violations (Cushman, 2014). However, there has been controversy about whether costly 3PP could have evolved, given that third-party punishers incur direct present costs but only indirect and remote future benefits in the form of group stability (Cushman, 2014; Fowler, 2005). This is a special case of the more general controversy over the evolution of altruism, which has generated several competing accounts, including group selection (Haidt, 2007), kin selection (Hamilton, 1964a; b), and reciprocal altruism (Trivers, 1971). Third-party punishment may well have evolved out of 2PP (Bellucci et al., 2020; Buckholtz and Marois, 2012; Marlowe, 2009; Marlowe et al., 2011), and some phylogenetic, ontogenetic, and ethnologic evidence supports this argument.

Phylogenetically, 2PP in its preventive form of self-defense is ubiquitous across the animal kingdom, exhibited in species as varied as insects (Wenseleers and Ratnieks, 2006), fish (Bshary and Grutter, 2005), and non-human primates (Hauser, 1992). Indeed, self-defense may have cellular roots in the form of the immunological response (Krueger and Hoffman, 2016). By contrast, 3PP seems unique to humans (Riedl et al., 2012), though there are some tantalizing hints of precursor behaviors in

[*] Corresponding authors.
E-mail addresses: qunyang@hznu.edu.cn (Q. Yang), morris.hoffman27@gmail.com (M. Hoffman), FKrueger@gmu.edu (F. Krueger).

https://doi.org/10.1016/j.neubiorev.2023.105525
Received 18 September 2023; Received in revised form 21 December 2023; Accepted 23 December 2023
Available online 27 December 2023

some non-human primates (Flack et al., 2006) and even dogs (Anderson et al., 2017).

Developmentally, self-defense behaviors emerge in human infants, as in all animals, virtually from birth (Wiedenmayer, 2010). Social evaluations of good and bad, upon which 3PP is based, emerge just a few months later. Infants as young as 3-month-olds start to show an aversion to antisociality in others as a third-party observer (Hamlin et al., 2010). Around 5 to 8 months of age, the capacity to distinguish between prosocial and antisocial signals, and preferences towards prosocial others against antisocial others, seem to become stable and prevalent (Hamlin and Wynn, 2011; Hamlin et al., 2007; Tan and Hamlin, 2022; Van de Vondervoort and Hamlin, 2018). Preverbal infants around 4–9 months of age already expect and prefer equal distribution (Buyukozer Dawkins et al., 2019; Geraci et al., 2022; Geraci and Surian, 2023a). By the end of the first year and the start of the second, infants develop the concept of fairness beyond egalitarian considerations (DesChamps et al., 2015; Rabinowitz et al., 2018; Ziv and Sommerville, 2016). Nine-month-olds expect unfair distribution to be punished in third-party contexts (Geraci and Surian, 2023b). Toddlers not only expect corporal punishment of an indifferent bystander who does not defend a victim hit by an aggressor (Geraci, 2021; Geraci and Surian, 2021), but also are willing to punish antisocial individuals by themselves as third-parties (Hamlin et al., 2011). By using a gaze-contingency technique, eight-month-old infants have been shown to engage in third-party punishment towards antisocial agents who inflict harms on others (Kanakogi et al., 2022). One-year-old infants show a willingness to pay a cost to avoid interacting with a wrongdoer (Tasimi and Wynn, 2016) while three-year-old children can overcome self-interests to punish unfair proposers in resource distribution and reject unfair allocations in the ultimatum game (Wu and Gao, 2018). Costly 2PP is commonly seen among children around 6–8 years of age (Gummerum and Chu, 2014; Jaroslawska et al., 2020). However, children do not begin to engage reliably in costly 3PP until one or two years after they start engaging in costly 2PP (Bernhard et al., 2020; McAuliffe et al., 2015). The neurodevelopmental evidence, though sparse, generally conforms with this behavioral trajectory (Grayson and Fair, 2017; Guroglu et al., 2011; McAuliffe et al., 2017; Supekar et al., 2009; van den Bos et al., 2014). All these findings related to the early emergence of social norm enforcement provide ontogenetic evidence for the evolution of the two types of punishment.

Ethnologists have found some form of institutionalized 3PP across virtually all human societies (Cushman, 2014; Henrich et al., 2006; Singh and Garfield, 2022). One key ethnological piece of evidence connecting 2PP with 3PP is the fact that universal legal notions of self-defense have almost always included not just the right to defend one's self but also the right to defend others (Hoffman, 2014).

The significance of SP for evolutionary theory and its potential policy implications for contemporary issues such as justice reform have made it a topic of great interest across many disciplines. Recent neuroscientific approaches have been particularly informative about the underlying neuropsychological mechanisms of SP. A number of meta-analyses (Bellucci et al., 2020; Feng et al., 2021; Feng et al., 2014; Gabay et al., 2014; Zinchenko and Arsalidou, 2017) or theoretical papers (Buckholtz and Marois, 2012; Decety and Yoder, 2017; Krueger and Hoffman, 2016; Seymour et al., 2007) have deepened our understanding of the affective, cognitive, and motivational processes of SP. However, none of these studies, or the models proposed in some of them, has offered a systematic review of SP, or made a theoretical effort to integrate the affective, cognitive and motivational processes of SP by connecting the second-party and third-party perspectives.

Here, we review the neuropsychological literature on SP, beginning with the overarching psychological model postulating that all human decisions are a confluence of affect, cognition, and motivation. We then apply that model to SP and to its two main drivers: the amount of harm caused by the norm violation and the intention of the norm violator. Next, we review the literature on the neural substrates of SP, organizing those substrates into affective, cognitive, and motivational circuits. We then present three recent neuroscience models of SP, discuss the limitations of those models, and present our proposed model. We finish with

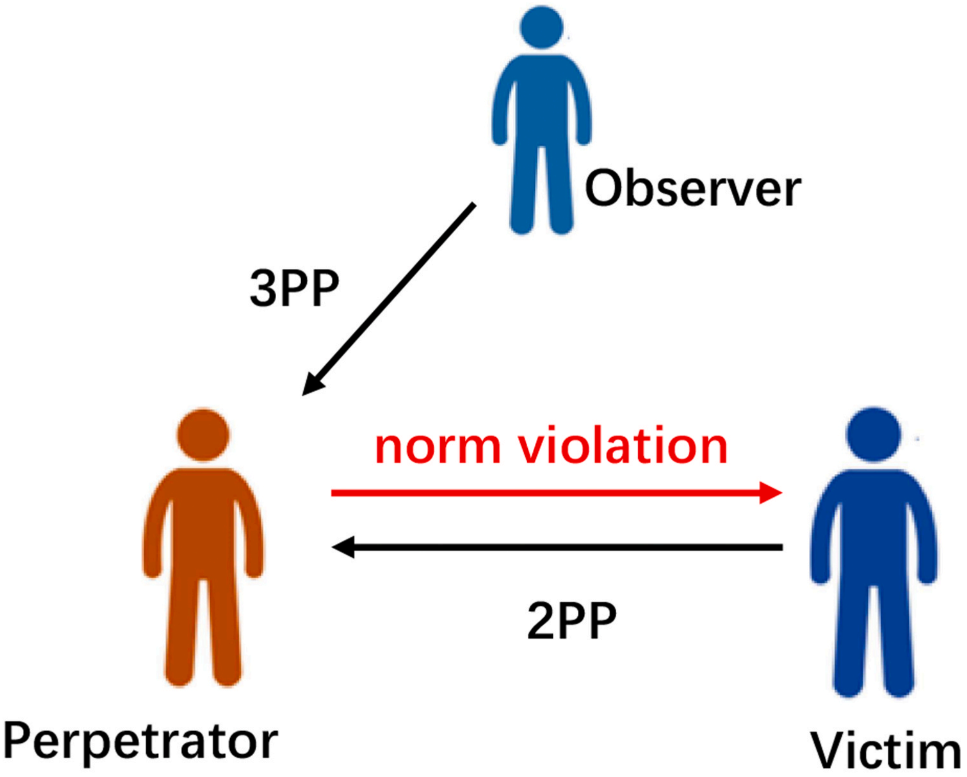

Fig. 1. The two types of punishment. Second-party punishment (2PP), where a victim punishes a perpetrator for violating a social norm and harming the victim (lower); and third-party punishment (3PP), where an observer punishes a perpetrator for violating a social norm and harming a victim (upper).

a discussion of future research directions and how the results of that future research might impact legal practice and policy.

2. The psychology of SP

Psychologists have long posited that human decision-making is driven by three core psychological processes: motivation, affect, and cognition (Alves et al., 2017; Dai and Sternberg, 2004; Hoffman, 1986; Kim et al., 2016). These processes are richly interdependent; our emotions and motivations can impact how we reason, and they can also impact each other (Dai and Sternberg, 2004; Hoffman, 1986). Likewise, we can sometimes reason our way to mitigate some of the behavioral impacts of emotions and motivations (Philippot and Feldman, 2004). SP is of course just one kind of decision-making, and it has also been analyzed as involving these three interacting processes. As we will see in the balance of this section: evaluating the harm that norm violations cause is primarily affective; evaluating the norm violator's intention and integrating intention and harm into blame are primarily cognitive; perceived harm and blame—integrated into what researchers call "motivational readiness"—are the primary motivators of SP, but whether motivational readiness to punish turns into a decision to punish is a second, primarily cognitive, step at which the costs and benefits of SP are weighed by executive and control processes.

Experimental data on SP is gathered using two different paradigms: economic games and hypothetical criminal scenarios. The economic game modality uses versions of classic exchange games like the ultimatum, dictator, and trust games sometimes modified to allow the other player (2PP) or a third-party observer (3PP) to punish unfair offers. For example, in an unmodified ultimatum game (Güth et al., 1982), Player 1 (the potential norm violator) is endowed by the experimenter with a sum of money and directed to split the money with Player 2. Player 2 then is given the option to accept or reject the split; if Player 2 rejects, then neither player receives anything, Player 2 thus punishing Player 1 at the cost of giving up his/her own share of the proposed split. In this version of the game, an unfair offer is considered to be the norm violation, the amount of the unfairness is considered to be the measure of the harm, and the amount Player 2 gives up is considered the punishment cost. In a modified version of the game (Leibbrandt and López-Pérez, 2012), Player 2 is not required to accept or reject the offer in its entirety, but can instead spend personal resources to reduce the proposer's payoff, operationalizing 2PP as the amount Player 2 is willing to spend to reduce Player 1's payoff. In the 3PP version of the game (Leibbrandt and López-Pérez, 2012), a third party, Player 3, observes the unmodified version of the game between Player 1 and 2 and may spend resources to punish Player 1 for unfair offers. In the hypothetical crime scenario methodology, subjects are presented with written scenarios depicting hypothetical perpetrators committing various crimes. The crimes, the amount of harm, and the perpetrator's intention (purposeful, accidental, or something in between), can all be varied. After reading the scenarios, subjects are asked to make hypothetical sentencing decisions or impose scaled punishment (Alter et al., 2007; Buckholtz et al., 2008; Treadway et al., 2014) or to rank order the scenarios according to the magnitudes of punishment they would impose (Robinson and Kurtzban, 2007).

These two methodologies have reciprocal advantages and disadvantages. Economic games can investigate both 2PP and 3PP, but they generally do so only over the single norm of monetary fairness. Scenario-based experiments use criminal narratives that can explore a broader range of social norm violations, but they are used almost exclusively in 3PP. It is significantly more difficult to model costly punishment using scenario-based experiments and significantly more difficult to vary the norm violator's intention in economic games. Most economic games, when they vary intention at all, do so only in a binary manner—intended or unintended (Feng et al., 2022; Nelissen and Zeelenberg, 2009; Yu et al., 2015). The richer states of mind commonly recognized by the law (purposeful, knowing, reckless, and negligent, discussed in the next section) are typically examined only in scenario-based experiments (Ginther et al., 2016; Ginther et al., 2014; Shen et al., 2011).

SP starts with the detection of norm violations, which is characterized by identifying deviations from expected social values. These expectations are informed by sets of prescribed and proscribed rules widely acknowledged as the standards of acceptable behaviors during social interaction (Buckholtz and Marois, 2012). The scenario-based studies of SP usually include social norms that are universally endorsed across cultures with a strong moral valence, such as prohibitions against intentionally harming others either physically, emotionally or economically (Buckholtz et al., 2008; Buckholtz and Marois, 2012) while the economic-games-based studies normally include social norms related to trust, fairness, and reciprocity of resource distribution (Bicchieri and Chavez, 2010; Crockett et al., 2014; Fehr and Fischbacher, 2004a; Fehr and Fischbacher, 2004b). Equity norms in these economic games serve as a cognitive heuristic for SP. People naturally expect the distribution of resources to be equal when there are no reasons to be unequal. Detecting violations of expectations may trigger and motivate punitive responses; the more a behavior deviates from equity, the more punishment is implemented (Civai et al., 2013). SP decreases when the expectations of participants about the equity rules change, for example, after learning that the majority would behave unfairly (Sanfey, 2009). Unlike daily economic social decision-making, people in some situations may need to apply explicit rules to detect norm violations. For instance, professional judges are trained to decide whether crimes have been proved to have been committed by using explicit rules of evidence and procedure, though even judges may be relying on heuristic strategies in the evaluation of legal cases (Li, 2013).

After detecting norm violations, the amount of harm and the norm-violator's intention become the primary drivers of both kinds of SP. Holding harm constant, experimental subjects punish purposeful norm violations more than accidental ones (Buckholtz et al., 2008; Ginther et al., 2016; Yu et al., 2015); holding intention constant, their punishments increase as harm increases (Crockett et al., 2014; Ginther et al., 2016; Yu et al., 2015). These two drivers are robustly reflected in how most legal systems and moral and legal philosophers conceptualize criminal responsibility and punishment (Krueger and Hoffman, 2016; Young et al., 2007). On the harm side, murder is of course typically punished more seriously than assault and assault more seriously than trespassing. On the intention side, the idea that criminal liability usually requires some level of culpable intention is well-rooted in English law, embodied by the Latin phrase "mens rea," which means "guilty mind." But in fact, it is much older, dating back to Augustine, and is now widespread across the world in common law and civil law systems alike. To give just one oft-cited example of the taxonomies of mental states that can arise from this principle, most American jurisdictions follow the Model Penal Code ("MPC"), which recognizes the following four mental states, in decreasing order of culpability: purposeful (specifically desiring to cause the harm); knowing (realizing the harm is practically certain to result, but willing to risk causing it for some other purposes); reckless (taking a substantial and unjustified risk of harm); and negligent (taking an unreasonable risk of harm) (American Law Institute, 1962).

There is a significant interaction effect between harm and intention, in both 2PP and 3PP. The greater the harm the more likely second and third parties will conclude the norm violator's intention was purposeful (Knobe, 2003). Conversely, subjects perceive otherwise identical harm as being more severe when they believe its infliction has been purposeful compared to when they believe it was accidental (Ames and Fiske, 2013; Ames and Fiske, 2015). These interaction effects appear to be super-additive, meaning the two factors weigh more heavily together (high harm, high intent) than either of them contributes separately (Ginther et al., 2016).

There are differences between 2PP and 3PP. Some experiments have found that at equivalent levels of harm and intention, second parties punish more often and more harshly than third parties (Civai et al., 2019b; Fehr and Fischbacher, 2004b; Stallen et al., 2018), although other experiments have not detected significant differences (Leibbrandt and López-Pérez, 2012). Still others have found that second parties punish attempted norm violations more frequently than third parties do, but both punishers punish completed norm violations with equal frequency (Feng et al., 2022). Second parties seem to value severe punishment more than third parties, although they show no differences in their willingness to impose it (Stallen et al., 2018).

Different individual patterns also emerge when differences have been detected between 2PP and 3PP. In one economic study using the dictator game, 26% of second-party punishers never punished and 39% punished only if the offer was below 50 points (of a 100-point endowment). But these numbers reversed for 3PP: 39% never punished and 26% punished only unfair offers (Fehr and Fischbacher, 2004b). The interaction between harm and intention appears greater in 3PP than in 2PP (Ginther et al., 2022). Many of these differences between 2PP and 3PP could be partly explained by the fact that those who suffer the harm (second parties) presumably experience a greater level of pain than those (third parties) who merely observe the same amount of pain being inflicted on others. But as we will see below, this is not a complete explanation.

2.1. *Affective and cognitive processes of SP*

The detection of norm violations, like the detection of any unexpected behaviorally salient event, also seems to be an affective process (Uddin, 2015) and may trigger an evaluation of the harm caused by norm violations. Exposure to harm from a norm violation is associated with several different aversive emotional states. In the ultimatum game, research consistently demonstrates that responders who receive unfair offers experience unpleasant feelings (Civai et al., 2010; van 't Wout et al., 2006), and as the severity of the unfairness increases, recipients reject the offers more quickly (Ma et al., 2012). Even when responders know their rejection of unfair offers will not be communicated to the proposers, and that the proposers will be able to keep their share of the split, a substantial level of rejection of unfair offers persists (Yamagishi et al., 2009). These findings suggest that emotion plays a role in driving 2PP. However, research into whether negative emotional experiences similarly fuel 3PP remains inconsistent. Some studies observe similar rejection of unfair offers in both 2PP and 3PP but increased negative emotional arousal only in 2PP (Civai et al., 2010). Others show a stronger correlation between the level of aversion and the amount of punishment in 2PP than in 3PP (Gummerum et al., 2022). Still others have found that equally strong negative emotions emerge from norm violations in both 2PP and 3PP (Hartsough et al., 2020; Nelissen and Zeelenberg, 2009), that affective experiences not only act as the necessary antecedents of 3PP but in fact predict the amount of 3PP in economic games (Lotz et al., 2011) and in criminal scenarios (Yang et al., 2019) and that magnifying or inhibiting emotional responses to the wrongdoings increases or decreases the amount of 3PP punishment in economic games (Nelissen and Zeelenberg, 2009).

Anger is the aversive emotion most frequently linked to SP (Fehr and Fischbacher, 2004b; Gummerum and Chu, 2014; Gummerum et al., 2016; Nelissen and Zeelenberg, 2009). Researchers surmise that people feel personal anger towards others who harm them directly and "empathic anger" towards those who inflict harm on others (Heckler and Kessler, 2018). Some researchers contend that "empathetic anger" is better described as "moral outrage" (Batson et al., 2007), and in fact there is some evidence to support this distinction (Hartsough et al., 2020; Pedersen et al., 2018).

In addition to anger or moral outrage, other emotions, such as fear (Taylor and Uchida, 2022), disgust and envy (Pedersen et al., 2013; Pedersen et al., 2018; Yang et al., 2019), guilt and shame (Nelissen and Zeelenberg, 2009), have been found to be associated with SP. Fear arguably reflects the very fact that these norms are ones most of us follow, so their violation raises concerns about whether we will be the target of further violations. Disgust and contempt seem to be similar forms of revulsion at the norm violator's willingness to violate norms the rest of us feel bound to follow; envy is a kind of unfairness version of disgust and contempt. Guilt and shame are what we feel when we have done nothing in response to a norm violation (Fig. 2). Just like the SP decisions they help drive, these emotions are driven both by the amount of harm and by the level of the norm violator's intention (Nelissen and Zeelenberg, 2009; van 't Wout et al., 2006; Yang et al., 2019).

While the harm caused by norm violations is usually concrete and emotionally accessible, the norm violator's mental states must almost always be inferred, and inference generally requires cognition. Psychologists call this ability to assess what another person is thinking "theory of mind" (ToM) or "mentalizing." Like the SP it enables, ToM is believed to have been a significant development in human evolution (Brüne and Brüne-Cohrs, 2006; Frith and Frith, 2005; Sodian and Kristen, 2010). Evolutionary theorists believe that being able to read others' intentions significantly improved our ancestors' fitness across several social domains, from intragroup and intergroup conflicts to sex, and enabled the kind of social cognition that resulted in several other core human advantages, including the ability to learn from one another (Brüne and Brüne-Cohrs, 2006; Frith and Frith, 2005; Sodian and Kristen, 2010).

As introduced above, the law has refined the intention/accident dichotomy into four varieties of intention—purposeful, knowing, reckless, and negligent—with purposeful harms punished the most and negligent ones punished the least, if at all (Fig. 3). A purposeful norm violator's specific intention is to cause the harm he or she causes. At the other extreme, a negligent actor has no intention of causing any harm, but has taken an objectively unreasonable risk of harm. The two intermediary states of mind are also risk-based. A knowing norm violation is committed as a side-effect to another purpose, but in circumstances where the actor knows the harm is "practically certain" to occur. A reckless actor takes a substantial and unjustified risk that harm will occur (American Law Institute, 1962).

In scenario-based experiments, subjects are reliably able to distinguish between most of these legally-relevant mental states, and reliably punish them just as the law predicts: purposeful harm most and negligent harm least (Ginther et al., 2014; Shen et al., 2011). Even in economic games, individuals can distinguish purposeful unfairness from accidental unfairness, punishing the former more than the latter (Falk et al., 2008). The sole exception is at the boundary between knowing and reckless states of mind—subjects can distinguish knowing from reckless actors but do not punish them any differently (Ginther et al., 2014; Ginther et al., 2018; Shen et al., 2011).

Psychologists have identified three aspects of the norm violator's

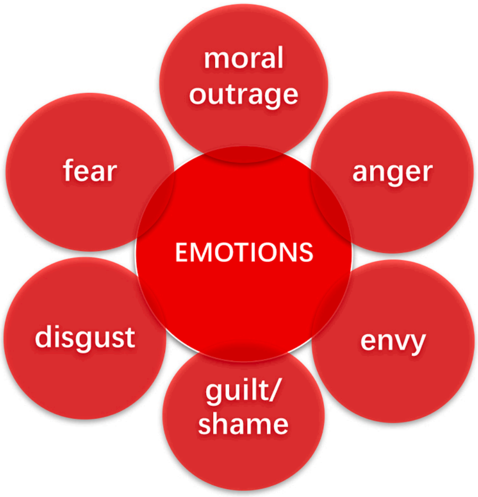

Fig. 2. The affective components of SP. Major emotions elicited by exposure to social norm violations.

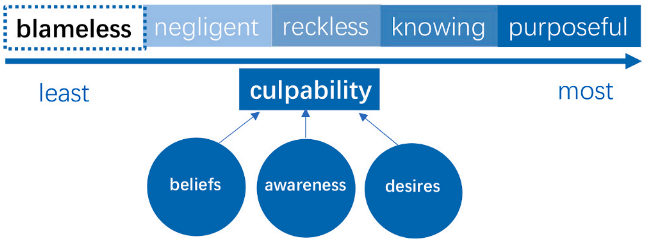

Fig. 3. Different types of culpability under Model Penal Code. States of mind hierarchy from least culpable on left (blameless) to most culpable on right (purposeful). Bottom: Factors contributing to punisher's assessment of a norm violator's mental state: what does punisher think a norm violator believes, knows, and wants?

state of mind that social punishers (and others to whom the norm violator's state of mind might be pertinent) must be able to understand before being able to reason about the norm violator's mental state: 1) the norm violator's beliefs (about how likely it is that an act will lead to harm); 2) the norm violator's awareness (of performing an action and the presence of the victim and his/her vulnerability to harm); and 3) the norm violator's desires (for the harmful result to occur to the victim, or for some other desired result to which harm is a side effect) (Laurent et al., 2016) (Fig. 3). Note that the two most confounding mental states—knowingly and reckless—are the two that rely on the norm violator's beliefs and awareness about the risk of harm. It appears people are generally quite sensitive to the risks and rewards attending their own decisions under uncertainty, even across the knowing/reckless boundary (Vilares et al., 2017), but are very likely to exaggerate risk assessments they attribute to others (Mueller et al., 2012). When assessing risks of harm post-hoc, as the legal system does, there is a ToM version of the hindsight effect: subjects tend not only to exaggerate the risks of harm the norm violator should have perceived (treating negligent acts as reckless, and reckless acts as knowing) but they even sometimes jump across the risk-based states of mind into purposefulness, concluding that hypothetical norm violators purposefully desired harm even though the risk of such harm was low (Mueller et al., 2012).

Except for the knowing and reckless states of mind, subjects across many different studies reliably and predictably punish depending on the amount of harm and the norm violator's state of mind (Cushman, 2008; Ginther et al., 2014; Gummerum and Chu, 2014; Shen et al., 2011). But the amount of this punishment, in both 2PP and especially 3PP contexts, can vary substantially between individuals (Jordan et al., 2016; Li et al., 2022a) and between societies (Balliet and Van Lange, 2013). Part of the individual variation may be due to differences in the relative weights that punishing experimental subjects give to harm and to intention, and those individual differences seem to correlate with differences in reasoning style: more deliberative reasoners tend to give more weight to intention than to harm; more intuitional reasoners tend to give more weight to harm than intention (Schwartz et al., 2022).

In addition to understanding the mind of others, people need to cognitively integrate harm and intention into a pre-punishment cognitive state, which some researchers have labeled "blame" (Krueger and Hoffman, 2016) and implement the punishment decision later in time, which also depends on the effective functioning of cognitive processes. Blame assignment and punishment decisions can be affected by working memory capacity (dos Santos et al., 2014; Goldinger et al., 2003). Poorer executive functioning task performance is associated with less 3PP in patients with Huntington's disease (Brüne et al., 2021), and punishment frequencies significantly change in both second-party and third-party contexts under cognitive resources depletion in normal populations (Achtziger et al., 2016; Liu et al., 2015).

Despite involving similar cognitive process, subjects can and do distinguish between the terms "blame" and "punishment" (Cushman, 2008). When subjects are descriptively presented with a number of purposeful norm violations across a wide variation in harm, and asked only to order them by the amount they would punish (arguably an act of blaming), but not asked for specific amounts of punishment, they order them with an impressive degree of concordance, substantially more concordant than if they are asked to impose specific punishment amounts (Robinson and Kurtzban, 2007).

This conceptualization of blame comports with long-standing principles of criminal responsibility in virtually all legal systems: defendants are criminally responsible when they commit a prohibited act that causes harm to the victim, coupled with some level of intention (Shen et al., 2011). Punishment is not only conceptually separate from blame in most legals systems; it is procedurally separate as well. If a factfinder (usually a jury in common law jurisdictions) convicts a defendant of a crime, then and only then is a separate sentencing proceeding held (in most American states before the judge, except in death penalty cases), after which punishment is imposed (Hoffman, 2014).

In both 2PP and 3PP contexts, harm and intent determine only blame; whether blame turns into punishment depends in the first instance on whether the punisher is sufficiently motivated to punish, which in turn depends not only on the amount of blame but also on the salience of the punisher's goals, the costs of punishment, and other contextual facts.

2.2. Motivational processes of SP

Psychologists have conceptualized motivation as a kind of energizer that drives our behaviors in pursuit of cognitive representations of a desired state, a "goal" (Fishbach and Ferguson, 2007; Kruglanski et al., 1996). They call the state of desiring a goal "wanting" (Berridge and Robinson, 2003; Kruglanski et al., 2014a). Wanting can be affected by "incentive salience," learned experiences about reaching a desired goal in the past, often triggered by cues that result in arousal (Kruglanski, 2017). When goals compete, some creating a wanting signal with respect to contemplated action (appetitive wanting) but others a not wanting signal because that action would interfere with other goals (aversive wanting), these conflicting goals and wanting signals must be accommodated into a net wanting signal with respect to the contemplated action (Elliott and Niesta, 2009). The strength of the net wanting signal, coupled with assessing the probability the desired goal can be achieved, is called the pre-action state of "motivational readiness" (Kruglanski et al., 2014a). Whether motivational readiness turns into action depends on comparing the forces driving and restraining action. The driving force is determined primarily by the strength of the motivational readiness to achieve the desired goals, while the restraining force is determined primarily by the size of the impediments to those goals, such as the cost and effort needed to achieve them or other aversive consequences of acting (Kruglanski et al., 2014b). All of the contextual circumstances surrounding a contemplated action can drive or restrain action (Kruglanski et al., 2012). When the driving force exceeds the restraining force, people act on their motivations (Kruglanski et al., 2014b).

Humans can have a rich set of goals beyond mere economic gain, and in fact the pursuit of normative values themselves can have its own intrinsic reward (Gabay et al., 2014; Sanfey, 2007; Tabibnia et al., 2008). It is this intrinsic reward that triggers the motivational aspects of SP. Social punishers start with the aversive feelings they have after they detect a violation of social expectations and then experience a harmful norm violation (either directly as the victim or remotely as a third party), which, depending on its seriousness, may trigger an initial state of wanting to punish, which we will call "retributive urge." The term "retribution," however, is more than a description of the retributive urge; it is its own theory of punishment. Just as norm compliance can be its intrinsic reward, norm violation can deserve its intrinsic punishment and this is what legal philosophers call retribution. According to the retributivist perspective, norm violators should receive their "just deserts" and suffer reciprocal punishment for the harm or loss they caused to be restored to the social fold (Kant, 1797; Keller et al., 2010). Retributivist SP should therefore fit the seriousness of the crime (Carlsmith et al., 2002).

In addition to retribution, there are extrinsic goal-oriented theories of punishment, which scholars label "utilitarian." The utilitarian goals of punishment include special deterrence (deterring the norm violator from future violations), general deterrence (deterring other group members from future violations), incapacitation (temporarily preventing the norm violator from committing future violations by incarcerating him), and rehabilitation (treating the norm violator so he will not desire to commit future crimes) (Darley, 2009; Vidmar and Miller, 1980).

Despite explicitly claiming to support the utilitarian perspective in justifying SP, experimental subjects punish norm violators in a manner primarily consistent with the retributivist perspective (Carlsmith, 2006; Carlsmith et al., 2002; Crockett et al., 2014; Keller et al., 2010; Ouss and Peysakhovich, 2015), although when experimenters make utilitarian information about the costs and effectiveness of SP more salient, subjects are more likely to endorse a future-oriented utilitarian SP perspective on SP (Aharoni et al., 2019; Twardawski et al., 2020).

There are three additional SP goals that are largely orthogonal to the retributive and utilitarian theories: self-defense, retaliation, and reputation. Nothing is more tightly tied to fitness than survival, and nothing more tightly tied to survival than defending oneself against physical injury. Retaliation is the non-self-defense form of 2PP: a kind of time-delayed self-defense too late to prevent harm to oneself, but not too late to balance the interpersonal books. Maintaining one's reputation is a form of social self-defense. Reputation is an important evolutionary metric in all primate interactions (Manrique et al., 2021), and defending it against injury is almost as important as defending one's body against injury (Jordan et al., 2016).

Motivational readiness in the SP context is determined by integrating two streams of motivationally-informed blame: 1) a purely and primarily affective retributive stream with 2) a goal-directed and primarily cognitive stream representing a mix of retributive, utilitarian, and orthogonal punishment goals. (Fig. 4**.**) These goals can reinforce or conflict with one another. For example, consider a norm violator whose violation we are confident is a one-off; we are reasonably sure he will never violate any important norm again. The goals of special deterrence and incapacitation would argue against punishment, but the goals of retribution and general deterrence would argue in favor of punishment. These appetitive (punish) and aversive (don't punish) goal-directed motivations are integrated with one another and with the retributive urge to produce a single level of motivational readiness to punish.

Whether motivational readiness to punish turns into punishment depends on the forces driving and restraining SP. The primary driving force is the magnitude of the motivational readiness to punish; the primary restraining force is punishment cost. Cost seems to be a critical determinant of whether prospective punishers turn motivational readiness to punish into actual punishment. The need to invest personal resources to sanction norm violators significantly reduces how individuals punish in both 2PP and 3PP (Anderson and Putterman, 2006; Cheng et al., 2022). When transgressions are sufficiently minor, these costs drive some experimental subjects to select less costly non-punitive options, such as compensating the victim with public funds (Arini et al., 2023; Heffner and FeldmanHall, 2019).

There can be other costs to SP, including retaliation by the norm violators or their families and friends, and lost opportunity costs when one's attention is devoted to punishment. There is a subtler but just as important moral cost: punishment, by definition, inflicts harm on the norm violator, which itself violates universal norms against harming others. That is, our SP motivations, driven by norms against harmful conduct, are themselves restrained by those very norms (Li et al., 2022b).

There are a host of additional contextual factors that can impact the SP decision both at the blame/motivational readiness stage by directly influencing the assessments of harm and intent, and at the punishment action stage as driving or restraining forces, including the circumstances surrounding the victim (age, vulnerability, gender, race, relationship to violator, whether victim forgave violator), the norm violator (age,

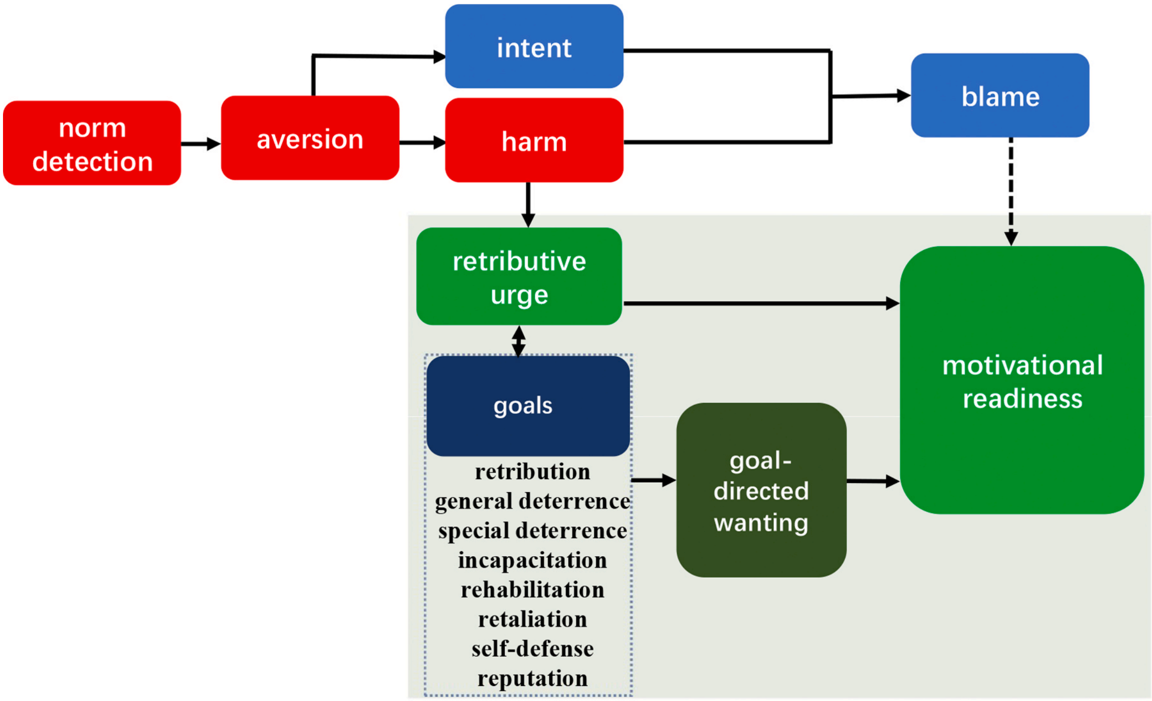

Fig. 4. The motivational readiness process of SP. Aversion to social norm violation triggers the affective stream involved in the evaluation of harm and the cognitive streams involved in the evaluation of intent. The harm evaluations trigger an initial retributive urge which then activates utilitarian punishment goals, producing goal-directed motivations. The retributive and goal-directed motivations are integrated into a net motivational readiness to punish, which could be affected by blame.

gender, race, prior norm violations, physical appearance, whether violator apologized) (Bushway and Piehl, 2007; Eriksson et al., 2017; Heffner and FeldmanHall, 2019; Hoffman et al., 2020; Johnson and King, 2017; Peay and Player, 2018; Robinson and Kurtzban, 2007; Tang et al., 2023; Yu et al., 2015; Schwartz et al., 2022), the time delay, if any, between a transgression and the implement of punishment (Kundro et al., 2023), and any ambiguity of norm violation (Toribio-Flórez et al., 2023).

Integrating affect, cognition and motivation into a SP decision takes time, and there is some evidence that that decision is itself comprised of two decisions: should I punish?; and if so, how much? (Civai et al., 2019b; Stallen et al., 2018). The willingness to punish and the severity of punishment are negatively correlated in both 2PP and 3PP—subjects are less likely to impose harsh punishment when they punish frequently (Stallen et al., 2018).

There is a separate though related motivational process devoted to measuring whether our goals, once reached, have met our expectations, a state psychologists call “liking” (what some behavioral economists would call “experienced utility”) (Tibboel et al., 2015). The relationship between wanting and liking has become an important clue to understanding the brain’s reward system and its implications for behavior, decision-making, incentive salience, and addiction (Berridge and Robinson, 2016). Although this relationship appears to be less important when it comes to non-repeat SP, institutional SP is of course repeated SP, and the interplay between liking and wanting could thus have some significance in institutional SP.

3. The neural mechanisms of SP

SP’s affective and cognitive processes appear to engage three domain-general large-scale networks: the salience network (SAN), the default-mode network (DMN), and the central-executive network (CEN) (Bellucci et al., 2020). With key nodes including the anterior insula (AI) and dorsal anterior cingulate cortex (dACC), and with extensive connections to subcortical structures, SAN is generally associated with the detection of behaviorally relevant events and the processing of affective experiences (Menon, 2015; Uddin, 2015), and in the case of SP with signaling a norm violation and any resulting harm, and generating emotional responses to the harm caused by the violation (Bellucci et al., 2020; Feng et al., 2018; Krueger and Hoffman, 2016). DMN, with key nodes including the medial prefrontal cortex (mPFC), posterior cingulate cortex (PCC), and temporoparietal junction (TPJ), is generally associated with cognitive tasks, including ToM evaluations (Buckner et al., 2008; Hyatt et al., 2015). In SP, it is associated with assessing the norm violator’s intention and integrating intention and harm into blame (Bellucci et al., 2020; Krueger and Hoffman, 2016). CEN, with key nodes including the dorsolateral prefrontal cortex (dlPFC) and posterior parietal cortex (PPC), is generally associated with high-order cognition including planning, rule-based problem solving, and goal-directed decision-making (Menon, 2011). In SP, it is associated with integrating blame, motivational readiness, costs, and context into a punishment decision (Bellucci et al., 2020; Buckholtz and Marois, 2012; Krueger and Hoffman, 2016). Indeed, the topological properties of these large-scale brain networks predict the heterogeneity of costly 2PP (Feng et al., 2018), and the functional connectivity of CEN is especially related to individual differences in the propensity for 3PP in economic games (Yang et al., 2021) and in criminal scenarios (Bellucci et al., 2017).

SP’s motivational processes seem to involve mesocorticolimbic pathways comprised of separate mesocortical and mesolimbic pathways originating from the midbrain’s VTA (Yetnikoff et al., 2014). The mesocortical pathway projects to prefrontal areas, including the vmPFC, OFC, and dlPFC (Arias-Carrión et al., 2010; Bittar and Labonte, 2021; Kim et al., 2016) and the mesolimbic pathway connects to core structures of the striatum (mainly nucleus accumbens (NAcc), caudate, and putamen) and to amygdala (Arias-Carrión et al., 2010; Kim et al., 2016). The mesolimbic pathway is associated with generating reward whereas the mesocortical pathway is involved in valuation, goal-directed control and transforming motivational readiness into action in general (Kim et al., 2016; O’Doherty, 2016; Yetnikoff et al., 2014). Although these motivational pathways have generally garnered less attention in the specific domain of SP, in the next sections evidence is provided that they are involved in representing, evaluating, and accommodating SP goals, evaluating the risks and benefits of SP in light of those goals, and instantiating motivations to optimize those goals. These motivations appear to impact SP decisions by interacting with the affective and cognitive pathways involved in the assessment of harm and intent, and the integration of different streams of information.

3.1. Neural correlates of affective and cognitive processing during SP

Early fMRI studies have consistently shown activation of AI during detection of deviations from equity or fairness norms in economic games (Hsu et al., 2008; Sanfey et al., 2003). Rejection of an unfair offer as a second-party receiver is associated with increased AI activation (Sanfey et al., 2003), while acceptance of an unfair offer is associated with decreased AI activation (Tabibnia et al., 2008). Activation of AI not only predicts the decision of whether to reject an unfair offer but also mediates the relationship between participants’ emotional states and their acceptance rates of unfair offers (Harle et al., 2012). The involvement of the AI in response to economic unfairness is seen in both 2PP and 3PP (Civai et al., 2012). AI has been implicated in integrating visceral experiences received from posterior insula and in particular mediating the translation of subjective feeling states into cognitive and motivational processes (Namkung et al., 2017). Dorsal AI (dAI) and ventral AI (vAI) are consistently found to be recruited in social exchange games, with dAI functionally associated with cognitive networks (DMN, CEN) and vAI functionally associated with the affective network (SAN) (Bellucci et al., 2018). In social norm enforcement, vAI is believed to signal violations of expected outcomes by involving affective and motivational processes through the co-activation of limbic regions, including the amygdala; dAI is believed to signal violations of expected norms by involving social-cognitive processes through the co-activation of regions in DMN and CEN (Krueger et al., 2020).

The insula has anatomical and functional connection with the amygdala, which has been widely recognized to encode the harmful consequences inflicted by norm violations and to generate aversive experiences which may be used later as a source of information to guide intuitions about punishment severity (Buckholtz and Marois, 2012; Krueger and Hoffman, 2016). The amygdala and the insula appear to be associated with separate SP functions (Gospic et al., 2011), with AI correlated with the willingness to punish and the amygdala linked to the severity of punishment (Civai et al., 2019; Stallen et al., 2018).

Another key SAN region co-activated with the insula during SP is the ACC, which is sensitive to the level of unfairness (Sanfey et al., 2003). Some studies suggest that ACC detects violations of expected social norm compliance and emotional appraisal (Chang and Sanfey, 2013; Etkin et al., 2011; Harle et al., 2012). Others indicate that ACC is primarily engaged in higher-level cognitive processing, such as tracking violations of social expectations (Chang and Sanfey, 2013; Guroglu et al., 2014) and monitoring motivational conflicts (Feng et al., 2014; Sanfey et al., 2003). Similar to the pattern observed in the insula, distinct subdivisions of ACC are associated with separate affective and cognitive processes: the rostral ACC (rACC) appears more specialized for affective processing whereas the dorsal ACC (dACC, also known as midACC or MCC) appears more specialized for cognitive processing (Shackman et al., 2011). The dACC interacts more frequently than rACC with regions involved in motivational and cognitive processes, such as the ventromedial prefrontal cortex (vmPFC), TPJ, and dlPFC (Baumgartner et al., 2011; Feng et al., 2014; Treadway et al., 2014). Additionally, studies using electro-encephalography (EEG) or event-related potential (ERP) technique to decode the time course of SP consistently reveal larger amplitudes of early negative-going potentials—medial frontal negativity

(MFN) or feedback-related negativity (FRN) — in response to the evaluation of negative outcomes and violations of the fairness norm (e.g., unfair offer) than compliance with the norm (e.g., fair offer) (Civai et al., 2020; Wang et al., 2016; Yoder and Decety, 2020). The MFN is distributed in the medial frontal scalp, particularly above the ACC area, and has been frequently associated with the detection of social expectancy violation (Boksem and De Cremer, 2010; Boksem et al., 2011; Gehring and Willoughby, 2002). The fMRI results combined with the ERP findings suggest a critical role of the salience network in detecting deviations from expected social norms and in identifying the behaviorally and motivationally relevant signals in the early stage of social punishment processes.

DMN and CEN are the two large-scale domain-general networks associated with the cognitive processing of SP. DMN's initial role seems to be the evaluation of the norm violator's intent through its mentalizing network, including primarily TPJ (Bellucci et al., 2020; Buckholtz and Marois, 2012; Krueger and Hoffman, 2016). As the difficulty of processing mentalizing information increases, TPJ activation increases (Feng et al., 2022; Ginther et al., 2016). Whether sympathizing with norm violators who have mitigating circumstances underlying their intentions (Buckholtz et al., 2008; Yamada et al., 2012; Yang et al., 2019) or antisocially punishing people who behave prosocially (Lo Gerfo et al., 2019), experimental subjects show increased differential activation in TPJ, as participants seemingly need to spend more mentalizing resources to interpret the norm violator's mind in these ambiguous contexts. TPJ activation levels have also been used to help predict whether experimental subjects are in knowing or reckless mental states (Vilares et al., 2017). Observations from 2PP in economic games and 3PP in scenario experiments show that TPJ sends regulatory signals to the amygdala suppressing its activation when the harm is unintended (Treadway et al., 2014; Yu et al., 2015). In the converse case of attempt (where harm is intended but does not result), TPJ shows greater differential activation during 3PP than 2PP (Feng et al., 2022). Disabling TPJ function by transcranial magnetic stimulation (TMS) causes third-party decision-makers to punish attempts less harshly than accidental harms, presumably because the norm violator's intention cannot be assessed or is assessed less reliably (Young et al., 2010). TPJ seems to communicate directly with dlPFC to adjust punishment in circumstances where the information about the intention behind the norm violation and the harm is incongruent, for example, attempt and harms inflicted with good intentions (Feng et al., 2022; Yang et al., 2019).

TPJ works in concert with other DMN regions during the mentalizing task, including vmPFC, dmPFC, PCC, and superior temporal sulcus (STS), and 3PP engages more of these mentalizing regions than 2PP does (Bellucci et al., 2020; Ginther et al., 2016). Effective connectivity experiments show that during 3PP, dmPFC receives inputs from other mentalizing regions, including the temporal pole, and that the degree of this directional connectivity is positively correlated with the amount punishment (Bellucci et al., 2017).

Before intention is integrated with harm into blame, these two streams interact when they are in conflict—that is when harm is high but intention low (accidents) or the reverse (attempts). The neural circuitry involved in mental states seems to gate the neural responses to affective experiences. In the case of high harm accidents, when the harm signal would ordinarily be high and the intention signal low, second parties in economic games and third parties in criminal scenarios exhibit enhanced directional connectivity from TPJ to amygdala, suggesting that the intention-sensitive regions play the regulatory role of suppressing the harm signal in these circumstances (Treadway et al., 2014; Yu et al., 2015).

The higher-order cognitive processes of CEN are involved in integrating blame, motivational readiness, costs, and context into a SP decision, although the exact role of its central region, dlPFC, has been disputed, giving rise to two different theories. One theory is an integration-and-selection hypothesis, which contends that dlPFC accomplishes this integration of affective, cognitive, and motivational signals, and then determines a punishment within a punishment scale set by other CEN regions (Buckholtz and Marois, 2012; Krueger and Hoffman, 2016). This integration-and-selection hypothesis is supported by data from scenario-based 3PP experiments showing dlPFC is more activated in scenarios where the offender is fully responsible than those where the offender has diminished responsibility, and that participants who choose to punish in the diminished-responsibility condition show greater activation in the dlPFC than those who decide not to punish (Buckholtz et al., 2008). Additionally, inhibitory TMS on the dlPFC reduces punishment in the full-responsibility condition without changing ratings of blameworthiness, providing neurobiological support both for the two-step punishment model (blame then punishment) and for the specific role of dlPFC in determining punishment (Buckholtz et al., 2015).

However, proponents of the competing theory—called the cognitive control theory—point out that the experiments on which the integration-and-selection hypothesis relies all used 3PP scenarios where few motivational and cognitive conflicts are involved. Some argue that SP is largely an impulsive reaction to norm violations primarily driven by emotions, subject to top-down cognitive control by dlPFC. They point to the fact that costly 2PP increases following the artificial depletion of serotonin, which is known to be important in maintaining self-control (Carver et al., 2008), and decreases once serotonin levels are restored (Crockett et al., 2010a; Crockett et al., 2010b; Crockett et al., 2008). They also note that dlPFC is recruited even more when offenders purposefully inflict harm out of helping motivations, suggesting that dlPFC is actually being recruited to resolve motivational conflicts between moral and legal values (Yang et al., 2019). Consistent with this, dlPFC is relatively more responsive to accepting an unfair offer than other punishment-related areas such as the insula (Sanfey et al., 2003). Moreover, applying TMS to temporarily disrupt the dlPFC in subjects playing a 3PP economic game increased the average punishment magnitudes, suggesting that the SP role of dlPFC is primarily to regulate the impulsive choice to punish (Brune et al., 2012; Muller-Leinss et al., 2018). Others argue that economic games often require participants to incur personal costs to punish norm violations and therefore dlPFC is needed to resolve the central conflicts between self-interest and norm enforcement. They also point to strong evidence for the role of dlPFC in controlling over selfish instinct to maximize economic personal interests in such economic games. When subjects play 2PP games where fairness and economic self-interests are in conflict, temporary disruption of dlPFC using TMS or transcranial direct current stimulation (tDCS) significantly reduce participants' propensity to punish unfair offers, indicating the role of this region in overcoming selfish preferences for material gains in order to make more normatively sensitive choices (Baumgartner et al., 2011; Knoch et al., 2008; Knoch et al., 2006).

3.2. Neural correlates of motivational processing during SP

Mesolimbic pathways are well-known to be associated with reward, and in the SP context the reward is the satisfaction we feel when we punish social norm violations (Sanfey, 2007; Tabibnia et al., 2008). For most people, a retributive urge may arise when a social norm has been violated and harm inflicted on a person. Punishing a norm violator can be socially rewarding (Gabay et al., 2014) by satisfying our retributive urge, and therefore could activate the reward processing in the mesolimbic pathway. VTA is activated in the ultimatum game when subjects are prompted to respond to offers deviating from their expected fairness norm, suggesting a role for VTA in incentive salience (Hetu et al., 2017). Striatal structures are involved in encoding social norm values and in the motivational reactions to violations of those norms. The striatum is activated more when SP reduces the norm violators' economic payoff than when SP is purely symbolic; and the degree of this differential activation predicts the cost second-party punishers are willing to pay in order to punish (de Quervain et al., 2004). Simply watching norm violators who receive high intensity shock as SP for defecting in a prisoner's dilemma game differentially activates NAcc (Seymour et al., 2007). These striatal involvements have been found both in 2PP and 3PP, using both task-based and task-free studies (Bellucci et al., 2020). Moreover, structural MRI evidence shows that third parties who are willing to initiate costly SP without peer encouragement have larger gray matter volume in the bilateral caudate—another core nuclei within the striatum—than those who punish only if peers do (Baumgartner et al., 2021). Interestingly, the electrophysiological evidence reveals that FRN time-locked to the presentation of an offer in UG not only acts as an index of violations of expectancy toward social norms (Wu et al., 2011) but also predicts the decision to reject — larger positive amplitudes of FRN to fair offers predict higher rejection rates, reflecting its potential involvement in encoding reward-related feedback (Hewig et al., 2011). In line with this, a combined EEG-fMRI study showed direct trial-by-trial coupling between the relative positive feedback signals associated with FRN and activations of reward circuits including the ventral striatum (Becker et al., 2014). It is possible that detecting deviance from social norms might be captured by the negative feedback signal of FRN generated from ACC, while processing the salience of norm violations may further engage the reward processing in the mesolimbic pathway (motivating individuals to act towards a desired outcome) coupled to the relative positive feedback signal of FRN.

Mesocortical pathways appear to be involved in the representation of appetitive and aversive attributes of punishment-related goals, accommodation of conflicting goals, and the integration of appetitive and aversive wanting signals into motivational readiness, which is eventually translated into an actual punishment decision. Valuation and integration seem to occur primarily in ventral and dorsal parts of the prefrontal cortex. The experienced value of each contemplated punishment decision seems to be calculated and constantly updated in vmPFC, suggesting its central role in the representation of motivational goals in a choice space, and the production of an overall motivational readiness value (O'Doherty, 2016). The detection of any motivational conflicts among the competing goals occurs mainly in the dACC (Feng et al., 2014) and the resolution of these conflicts and eventually translating the motivational readiness into a punishment decision occurs in the dlPFC (O'Doherty, 2016). Inhibitory stimulation of dlPFC by TMS decreased the connectivity between the dlPFC and vmPFC, and lowered rejection rates of unfair economic game offers (Baumgartner et al., 2011), demonstrating the significance of the communication between the valuation circuits at the motivational readiness stage and the control circuits at the punishment action stage.

In addition to interacting with the cognitive system, motivational processing also works in concert with the affective system. There is evidence that the willingness to punish is associated with AI activation (Civai et al., 2019a; Stallen et al., 2018), presumably because AI detects norm violations and generates aversive experiences, which are essential for triggering motivational processing. Moreover, the co-activation of affective circuits during motivational processing may provide some neural-level confirmation of the psychological model that some behavioral differences between 2PP and 3PP are bottomed, partly, on different levels of retaliative motives. Pain accidentally delivered by an anonymous person in an interpersonal game activates AI in the hurt person, and the activation levels are positively correlated with the hurt person's measured personality trait of revenge (Yu et al., 2015). Meta-analytic evidence also demonstrates more activation in bilateral AI during 2PP than 3PP (Bellucci et al., 2020). The posterior part of the midcingulate cortex (pMCC)—implicated in pain processing—is preferentially activated in retaliatory 2PP (Boccadoro et al., 2021), indicating that people may need to process the painful provocation before initiating reactive punishment responses.

3.3. *Current neuropsychological models of SP*

Although there were several earlier descriptive efforts (e.g., de Quervain et al., 2004; Seymour, 2007), the first fully neuropsychological model of SP, published by Buckholtz and Marois (2015), links the psychological component of 3PP to the core structures implicated in the emotional processing, mentalizing and executive functions. Harm is encoded primarily in the amygdala, intention primarily in TPJ, and these two signals integrated primarily in mPFC (into what others would later call blame but which these authors left unnamed); a context-dependent response space is created in the intraparietal sulcus (IPS), and a punishment decision made in dlPFC. The Buckholtz-Marois model was significant in that it made the first attempt to associate an affective structure with the assessment of harm (the amygdala) and cognitive areas with the assessment of intent (TPJ), to posit an integration of intent and harm (mPFC), and to identify a separate punishment action (IPS, dlPFC). It was also significant in that it adopted the integration-and-selection model of the role of dlPFC, as opposed to the cognitive control model.

A somewhat different model was published by Krueger and Hoffman (2016), which retained the prior model's essential hypothesis that harm and intent are assessed in affective and cognitive regions, respectively, and its integration-and-selection approach, but which was the first to model the neuropsychological mechanism of 3PP in terms of the large-scale brain networks: the SAN is responsible for signaling a norm violation and generating an emotional signal to the harm severity of the violation; the DMN encodes the information about the intention of the norm violators and then integrates affective signals related to the harm from the SAN into an evaluation of blame; eventually, the CEN converts the blame signal into a punishment response. Moreover, their model expanded the affective regions to include AI and ACC, and functionally differentiated between the roles of vmPFC and dmPFC, positing that vmPFC is the conduit of the harm signal, dmPFC the conduit of the intention signal, and the two signals are integrated into blame in mPFC.

Although they both mention 2PP, the Buckholtz-Marois and Krueger-Hoffman papers model 3PP, not 2PP, and they do not attempt to account in any comprehensive way for the observed differences between 2PP and 3PP. This gap was addressed by Bellucci et al. (2020), who built what they called a "hierarchical punishment model" aimed specifically at explaining the differences between 2PP and 3PP. Bellucci et al. posit that 3PP is a functional extension of 2PP, and specifically that, while both forms of SP recruit many common regions across the three large-scale networks noted in the prior models, 3PP preferentially engages TPJ and 2PP preferentially engages AI.

4. A new model of SP

None of the prior models accounts in any systematic way for the role of motivation. Although Bellucci et al. (2020) observe affective and cognitive differences between 2PP and 3PP, and posit that those affective and cognitive differences drive the behavioral differences, they do not provide any underlying explanation of *why* 2PP seems more sensitive to the harm caused by the norm violation and 3PP more sensitive to the intention of the norm violator. We suggest that the answer lies in the fact that second parties and third parties, because of their profoundly different relationships to the norm violator, have profoundly different punishment motivations. We therefore develop a neuropsychological model of SP that takes motivational processes into account, integrating them with affective and cognitive streams, and dub it the motivation-affect-cognition nexus (MACN) model of SP (Fig. 5).

We have made several assumptions in constructing the MACN model, all of which we believe are established by or are reasonable extensions of the psychological and neuroscience literature reviewed above. We have assumed, as have prior models, that the harm and intention components of SP are addressed separately in large-scale domain-general networks, with harm assessed primarily in SAN and intent primarily in DMN, and that there are significant interaction effects between harm and intention, but we have added the complication that motivation can also interact with the assessment of harm and intention.

We have followed prior models in assuming these two streams are

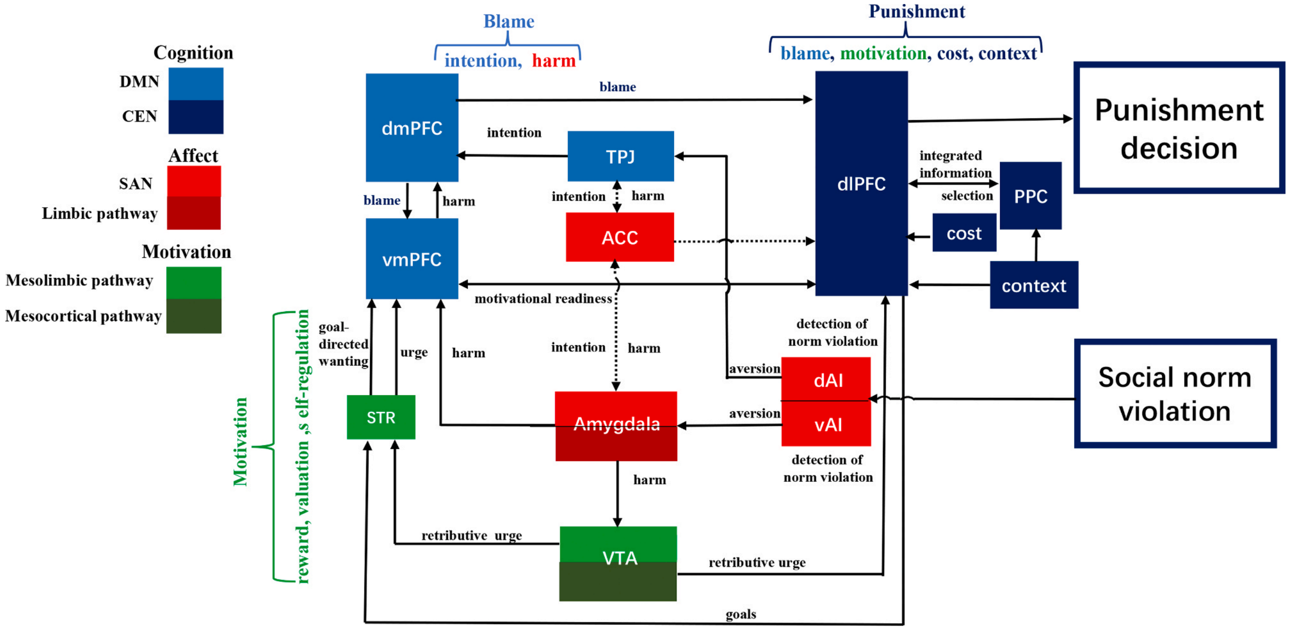

Fig. 5. The MACN model of SP. Social punishment (SP) is driven by three domain-general large-scale networks—the salience network (SAN) (red), default mode network (DMN) (light blue), and central executive network (CEN) (dark blue)—working with mesolimbic (light green) and mesocortical (dark green) motivational pathways. Social norm violations are detected in anterior insula (AI), which sends simultaneous aversion signals dorsally to temporoparietal junction (TPJ), which begins the assessment of intention along with other theory of mind regions, and ventrally to amygdala, which assesses harm then sends harm signal to ventral tegmental area (VTA) to begin motivational assessment. Harm signal is motivationally adjusted in ventromedial prefrontal cortex (vmPFC)—second-party punishers experiencing more harm than third-party punishers. Incongruent harm and intention signals are adjusted through interaction mediated by anterior cingulate cortex (ACC) (dotted lines), high intent signal up-regulating low harm signal (attempt), and low intent signal down-regulating high harm signal (accident). Harm and intent are then integrated into blame in dorsomedial prefrontal cortex (dmPFC) and sent to vmPFC for integration with motivational signals into a motivational readiness signal. Mesolimbically, VTA sends retributive urge signal to striatum (STR) and striatum sends it to vmPFC. Mesocortically, VTA first projects to dorsolateral prefrontal cortex (dlPFC) to activate punishment goals, which are then sent to STR, later converted into goal-directed motivations, and sent to vmPFC. Conflicts between goals detected in ACC are resolved by dlPFC using executive functions. vmPFC then applies a blame signal to the matrix of remaining retributive and goal-directed motivations to create a single motivational readiness signal, which it sends to dlPFC. dlPFC integrates motivational readiness, blame, with costs and other contextual facts into a selected punishment decision in a decision space created by posterior parietal cortex (PPC).

integrated into blame in DMN, in our model primarily in dmPFC, and that the decisional role of CEN is primarily through dlPFC in an integration-and-selection process. The MACN model departs from prior models by including motivational pathways and by positing that AI performs a critical role in triggering cognitive, affective and motivational circuits, which later interact significantly with each other to produce a punishment decision.

We begin the description of our MACN model at AI, which is an anchor in SAN, and where we believe violations of our expectations that others will comply with social norms are detected, and which subsequently acts as a critical affective switch between motivational and cognitive circuits. We posit that vAI sends an aversive signal to amygdala, triggering its representation of harm which then triggers the motivations to punish, and that dAI simultaneously engages TPJ to evaluate the norm violator's intention.

The amygdala sends the harm signal to VTA, which triggers motivational circuits that begin the assessment of goals with respect to the harmful norm violation. The mesolimbic pathway (VTA-striatum) converts the harm signal into a retributive urge to punish, sending that retributive motivation to vmPFC. The mesocortical pathway (VTA-prefrontal cortex) converts punishment goals stored in dlPFC into a set of goal-directed motivations, sending those to vmPFC. vmPFC accumulates the retributive and goal-directed motivations into a motivational readiness signal which is the main driving force of punishment.

While vAI is sending the aversive signal to amygdala to trigger the assessment of the harm, dAI sends the aversive signal to TPJ, triggering the assessment of the norm violator's intention in ToM regions. Incongruent harm and intention signals (accidental or attempted harm) are adjusted through interaction mediated by ACC. These two signals are integrated into a blame signal in dmPFC, which then communicates with vmPFC to adjust motivational readiness, and to dlPFC to integrate the contextualized driving forces, restraining forces, and blame into a punishment decision framed by a response space constructed in PPC.

Our model is unified both in the sense that it accounts for motivational circuits in mesocorticolimbic pathways, and in the sense that by accounting for motivation it helps explain some of the observed differences between 2PP and 3PP. It is also unified in another sense: we attempt to accommodate the conflict between the integration-and-select theory of decision-making and the cognitive control theory, discussed in part 3.1 above, by hypothesizing that cognitive control processes are used to produce motivational readiness, and that integrate-and-select processes are used to convert motivational readiness into a punishment decision. In particular, we hypothesize that cognitive control suppresses some goals and wanting signals to allow others to emerge into a motivational readiness signal. Integration-and-select processes integrate motivational readiness with affective, cognitive, and contextual streams into a punishment decision. Indeed, gray matter structures in anterior dlPFC have been associated with regulation (Schmidt et al., 2018), while more posterior and dorsal parts are associated with cognitive processing during value-based evidence accumulation and selection (Hutcherson and Tusche, 2022). ACC may play a role in detecting and monitoring any conflicts before different streams of information are integrated and a punishment decision is eventually selected in dlPFC.

Our motivationally-informed MACN model, unlike prior models, accounts for many of the behavioral differences between 2PP and 3PP. The crucial difference between second- and third-party punishers is motivational, driven by their very different relationships with the norm violator. These different relationships present different punishment risks and benefits and therefore animate different punishment goals. In self-defense 2PP, victims are just trying to survive, so punishment is the affective default to achieve this unambiguous and evolutionarily powerful goal. Even retaliatory 2PP, which we have described earlier as a desire to "even the interpersonal books," is substantially more self-centered than the subtler and more deeply conflicting cognitively-driven goals of the third-party punisher, who in ancestral times had to ask whether risking immediate injury was worth remote pro-social benefits. Institutional punishers, like judges, are even further removed from the clear choices presented in 2PP.

These different sets of goals drive different motivational sensitivities to harm and intention. In 2PP, harm is substantially more important than intention; it won't matter much to a victim facing life-threatening harm whether his would-be killer is acting purposefully, knowingly, recklessly, or negligently. In 3PP, the most salient harm is the harm punishers risk by inserting themselves into a situation that does not directly or immediately involve them. For them, non-punishment is the default, overcome only if weak retributive urges and remote utilitarian motivations are powerful enough. Those motivations depend very much on the norm violator's intention and much less on the amount of harm. Accidental norm violators need not be incapacitated and need only be deterred in the sense that we may want to impose some sanction (damages, in civil cases) to make them and others more careful in the future. Purposeful norm violators, by contrast, are acting to satisfy their desires. They will have desires in the future and, therefore, are more likely to act on those future desires just as they have acted on the present ones—by violating a norm. Now, utilitarian goals like incapacitation and deterrence are substantially more relevant. As we march down the levels of culpability from purposeful to knowing to reckless to negligent, an argument can be made that what these different states of mind really convey is the likelihood that the norm violator will harm others in the future. In this account, third-party sensitivity to the norm violator's intention is a naturally-selected proxy for recidivism.

The motivational aspects of our MACN model also account for the wide variance in punishment behaviors across individuals. While there is tight concordance in the assessment of blame, there is wide variance in punishment (Robinson and Kurtzban, 2007), and we would argue that that is because there is wide variation in individual goal-directed motivations, driven by a wide variation in goals and the way we order goals and resolve conflicts between them.

One word about retribution, which we have included as one of the goals of SP, although it may seem less a goal than a stand-alone appetite; indeed, evolutionary theorists have argued that the retributive urge that drives SP, especially 3PP, was natural selection's ultimate solution to the proximate problem of calculating special and general deterrence (Carlsmith et al., 2002; Gintis et al., 2008; Hoffman and Goldsmith, 2004). As we mentioned in the context of conflicting punishment goals, if we could somehow be sure that a group member's norm violation was a one-off and that other group members will not be tempted to copycat if we don't punish, then 3PP carries all risk and no benefits. But because we cannot be sure of that, evolution favored a retributive urge to get us over the seemingly impossible hump of immediately risking our own lives in the hope that third-party punishment might one day have some salutary group benefits (Haselton and Buss, 2009). And, as we have seen, studies consistently show that most of us have a retributive core to which we return in hard cases, even when we profess to be utilitarians (Carlsmith et al., 2002).

5. Future research and implications

Our MACN model, taking into account motivational as well as affective and cognitive processing, suggests many lines of future research, and could also drive some clinical applications in forensic psychiatry, some applications to legal process, and even some substantive legal reforms.

More behavioral and neuroscience work needs to be done to verify many of the assumptions contained in our model. We know very little at the neural level about how different sets of goals are represented, activated, and matched with different motivations and how those motivations are represented and accommodated into motivational readiness. Much remains to be done to confirm our suspicions that core differences between 2PP and 3PP are driven by differences in motivation, operating primarily by differentially impacting assessments of harm and intention. Few details are known about the neural transition from motivational readiness to an actual punishment decision or the driving and restraining forces that have been theorized to animate that decision. Additional research will also be needed to confirm our contention that the two models of dlPFC function—integrate-and-select versus cognitive control—can be harmonized in the way we propose, with cognitive controls operating motivationally and integrate-and-select processes operating at the decision stage, in different parts of dlPFC.

Very little is known, psychologically or neurologically, about what drives individual differences in SP. One of the most widely-known psychological measurements consists of four subscales to measure what researchers call "justice sensitivity" from the perspective of a victim, observer, perpetrator, and third-party beneficiary (Schmitt et al., 2010), and it appears all subscales but the one from the victim's perspective predict the amounts of SP in economic games (Baumert et al., 2014; Fetchenhauer and Huang, 2004; Zhen and Yu, 2016). These results would be consistent with all SP models to the extent we can assume wider individual variance in cognitive streams than in affective ones and greater sensitivity to cognitive streams in 3PP than 2PP, but we do not know whether these assumptions are true. Limited work has been done comparing subjects' scores on an instrument designed to measure beliefs in free will to their amounts of 3PP in criminal scenario experiments but with mixed results (Krueger and Hoffman, 2016; Nettler, 1959; Viney et al., 1988). Individual differences in the propensity of both 2PP and 3PP were found to be predicted by differences in resting-state functional connectivity in certain regions (Feng et al., 2018; Yang et al., 2021), though this experiment was limited to economic games and therefore to the norm of fairness. Other studies show moderate levels of genetic correlation to individual differences in SP and to differential activations in AI and NAcc, but again these used economic games and therefore tested only fairness (Strobel et al., 2011; Wang et al., 2019).

Although much work has been done on the impact that various contextual factors have on 3PP, very little has systematically tested these impacts across levels of harm and intention, and virtually nothing is known about whether any of these contextual factors have differential impacts as between 3PP and 2PP. Likewise, it is unclear whether these contextual factors are impacting SP at the stage of harm and intention assessment, during motivational processing, or as part of the final cognitive weighing of driving and restraining forces, or at each of these points (as we surmise). One contextual area that has received some attention is the impact apology and forgiveness have on SP (Martinez-Vaquero et al., 2015) (e.g., Martinez-Vaquero et al., 2015), but very little has been done systematically across the two forms of SP or between different levels of harm and intention. There have been a few attempts to examine the neural correlates of apology and forgiveness, but they have been concentrated on economic games and thus on the single norm of monetary fairness (Fourie et al., 2020; Strang et al., 2014).

Another potentially fruitful and critical area of inquiry has to do with potential differences between the two main methodologies of investigating punishment behavior: economic games and criminal scenarios. Both are used because both have complementary advantages and disadvantages. Economic games allow researchers to look at both 2PP and 3PP in tightly controlled experimental conditions, while scenarios allow researchers to test norms beyond mere fairness but typically only in 3PP contexts. The external validity of economic games is known to be poor in some domains: individuals' decisions to accept unfair offers in the ultimatum game are not related to real-effort helping or donating behaviors in naturalistic field situations (Galizzi and Navarro-Martinez, 2019). Other work shows little correlation between 2PP and 3PP on the one hand and cooperation games on the other (e.g., dictator, public goods, and trust games) (Peysakhovich et al., 2014). All of this raises a significant question about the translatability of the SP results across these two modalities, and there is virtually no literature on this question. The conflicting results on the behavioral differences between 2PP and 3PP, namely, whether second parties do in fact punish equal levels of blame more frequently and more harshly than third parties, may well be an artefact of these methodological differences, as may the apparent conflict between integrate-and-select theories of dlPFC function and cognitive control theories.

An additional unexplored line of research suggested by our model involves the relationship between norm compliance and norm enforcement. Some scholars have posed the question indirectly by describing the forces that drive norm compliance—conscience and guilt—as "first-party punishment" (Hoffman, 2014), but very little is known about the relationship between norm compliance and SP. We do not know whether those who punish more severely as second or third parties are more likely to follow norms as first parties. There are some confounding results. One fMRI study showed activations in areas associated with empathetic pain were greater than areas associated with personal pain when subjects were engaged in tasks accidentally hurting themselves and others, suggesting that at least in the domain of accidental norm violations we are harder on ourselves than on others (Hirschfeld-Kroen et al., 2021). But it remains unknown whether these supposed pain differentials translate into actual self-punishment at a real cost, and whether they persist in non-accidental contexts.

The developmental aspects of the psychology of SP have been well investigated, as we summarized in the Introduction, but much less is known about SP's neurodevelopmental trajectory, and much less still, in both psychology and neuroscience, about the end stages of that development, namely, how and whether SP changes as we become elderly. Given that the average age of federal judges in the U.S. is 69 (Shen, 2020), this is an area that could have important practical applications, as we touch on below.

Finally, and perhaps most importantly, enormous amounts of research have been done looking at out-group punishment biases in general, and race, gender, and ethnic biases in particular, and although similarly enormous amounts of empirical work have been done examining biases in our legal system, very little has been done to put these two research streams together. *Why* do we impose harsher SP on out-group members—is the effect driven by different harm, intention, or motivational assessments, or some combination of all three?

Although models of SP focus on punishers and not on the punished, a better understanding of how we react to norm violators may indirectly lead to a better clinical understanding of them. For example, punishers must first recognize that a norm has been violated, and the diminished ability to do so arguably contributes to the criminal behaviors of a significant segment of our incarcerated population who are psychopaths (Kiehl and Hoffman, 2011; Koenigs et al., 2010). Understanding how all of us internalize norms and recognize their violation might help us treat psychopaths, who have been famously immune to treatment (Caldwell and Van Rybroek, 2001; Kiehl and Hoffman, 2011), and others with criminogenically relevant mental conditions. Insights into how we process norm violations may also have clinical impacts in our understanding of legal issues that remain troublingly straddled across psychiatry and law, such as competency to stand trial and insanity (Diamantis, 2021; Morse and Hoffman, 2007). Learning more about the four legally-recognized mental states could even have important policy impacts. Currently, intoxication is a defense in most common law jurisdictions only to purposeful crimes, but it is conceivable the law has this exactly wrong: perhaps some psychotropic drugs impair our risk-assessment machinery but not our purpose-forming machinery.

Of course, neuropsychological insights into SP are insights about how *individuals* behave, while legal systems are the products of social consensus. Disciplines like public choice theory are attempting to connect the individual insights of behavioral economics with public policy choices made institutionally (Buchanan & Tollison 1984), and similarly thoughtful and careful efforts are beginning in the law and evolutionary psychology/neuroscience realm (Ben-Nur & Putterman 2000). Still, the law is not entirely or even mostly about public policy; much of it involves individual legal actors making individual decisions within the confines of settled public policy. Indeed, SP is a perfect example: individual judges (and occasionally juries) impose criminal sentences within the ranges set by legislatures. Knowing more about how we punish wrongdoers could have institution-shaking impacts on the legal process. Imagine how jury and judicial selection systems might handle the prospect of being able to separate harsh punishers from lenient ones, or deal with knowledge about what happens to our SP faculties as we age. Insights into how we treat out-group wrongdoers could lead to strategies to detect, and perhaps even inoculate against, police, judge, and jury biases, including racial, ethnic, and gender biases.

More broadly, future research into the mechanisms of SP may help legal systems return to more controlled forms of retribution, recognizing it as a limiting principle with respect to other utilitarian policies. Doing so may even help us reduce exploding prison populations while remaining true to our evolved natures as retributive social punishers.

Acknowledgements

The authors want to thank Abby Perkins and Zidi Li for their invaluable research help.

This work was supported by the Humanity and Social Science Foundation of the Ministry of Education of the People's Republic of China (22YJAZH130).

References

Achtziger, A., Alós-Ferrer, C., Wagner, A.K., 2016. The impact of self-control depletion on social preferences in the ultimatum game. J. Econ. Psychol. 53, 1–16.

Aharoni, E., Kleider-Offutt, H., Brosnan, S., Watzek, J., 2019. Justice at any cost? The impact of cost-benefit salience on criminal punishment judgments. Behav. Sci. Law 37, 38–60.

Alter, A.L., Kernochan, J., Darley, J.M., 2007. Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. Law Hum. Behav. 31, 319–335.

Alves, H., Koch, A., Unkelbach, C., 2017. Why good is more alike than bad: processing implications. Trends Cogn. Sci. 21, 69–79.

Ames, D.L., Fiske, S.T., 2013. Intentional harms are worse, even when they're not. Psychol. Sci. 24, 1755–1762.

Ames, D.L., Fiske, S.T., 2015. Perceived intent motivates people to magnify observed harms. Proc. Natl. Acad. Sci. USA 112, 3599–3605.

Anderson, C.M., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games Econ. Behav. 54, 1–24.

Anderson, J.R., Bucher, B., Chijiiwa, H., Kuroshima, H., Takimoto, A., Fujita, K., 2017. Third-party social evaluations of humans by monkeys and dogs. Neurosci. Biobehav. Rev. 82, 95–109.

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-González, M., Pöppel, E., 2010. Dopaminergic reward system: a short integrative review. Int. Arch. Med. 3, 24.

Arini, R.L., Mahmood, M., Bocarejo Aljure, J., Ingram, G.P.D., Wiggs, L., Kenward, B., 2023. Children endorse deterrence motivations for third-party punishment but derive higher enjoyment from compensating victims. J. Exp. Child Psychol. 230, 105630.

Balliet, D., Van Lange, P.A., 2013. Trust, punishment, and cooperation across 18 societies: a meta-analysis. Perspect. Psychol. Sci. 8, 363–379.

Batson, C.D., Kennedy, C.L., Nord, L.-A., Stocks, E.L., Fleming, D.Y.A., Marzette, C.M., Lishner, D.A., Hayes, R.E., Kolchinsky, L.M., Zerger, T., 2007. Anger at unfairness: is it moral outrage? Eur. J. Soc. Psychol. 37, 1272–1285.

Baumert, A., Schlösser, T., Schmitt, M., 2014. Economic games: a performance-based assessment of fairness and altruism. Eur. J. Psychol. Assess. 30, 178–192.

Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., Fehr, E., 2011. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat. Neurosci. 14, 1468–1474.

Baumgartner, T., Hausfeld, J., Dos Santos, M., Knoch, D., 2021. Who initiates punishment, who joins punishment? Disentangling types of third-party punishers by neural traits. Hum. Brain Mapp. 42, 5703–5717.

Becker, M.P.I., Nitsch, A.M., Miltner, W.H.R., Straube, T., 2014. A single-trial estimation of the feedback-related negativity and its relation to bold responses in a time-estimation task. J. Neurosci. 34, 3005–3012.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K.M., Grafman, J., Krueger, F., 2017. Effective connectivity of brain regions underlying third-party punishment: functional MRI and granger causality evidence. Soc. Neurosci. 12, 124–134.

Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S.B., Krueger, F., 2018. The role of the anterior insula in social norm compliance and enforcement: evidence from coordinate-based and functional connectivity meta-analyses. Neurosci. Biobehav. Rev. 92, 378–389.

Bellucci, G., Camilleri, J.A., Iyengar, V., Eickhoff, S.B., Krueger, F., 2020. The emerging neuroscience of social punishment: Meta-analytic evidence. Neurosci. Biobehav. Rev. 113, 426–439.

Bernhard, R.M., Martin, J.W., Warneken, F., 2020. Why do children punish? Fair outcomes matter more than intent in children's second- and third-party punishment. J. Exp. Child Psychol. 200, 104909.

Berridge, K.C., Robinson, T.E., 2003. Parsing reward. Trends Neurosci. 26, 507–513.

Berridge, K.C., Robinson, T.E., 2016. Liking, wanting, and the incentive-sensitization theory of addiction. Am. Psychol. 71, 670–679.

Bicchieri, C., Chavez, A., 2010. Behaving as expected: public information and fairness norms. J. Behav. Dec. Mak. 23, 161–178.

Bittar, T.P., Labonte, B., 2021. Functional contribution of the medial prefrontal circuitry in major depressive disorder and stress-induced depressive-like behaviors. Front. Behav. Neurosci. 15, 699592.

Boccadoro, S., Wagels, L., Puiu, A.A., Votinov, M., Weidler, C., Veselinovic, T., Demko, Z., Raine, A., Neuner, I., 2021. A meta-analysis on shared and distinct neural correlates of the decision-making underlying altruistic and retaliatory punishment. Hum. Brain Mapp. 42, 5547–5562.

Boksem, M.A., De Cremer, D., 2010. Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. Soc. Neurosci. 5, 118–128.

Boksem, M.A., Kostermans, E., De Cremer, D., 2011. Failing where others have succeeded: medial frontal negativity tracks failure in a social context. Psychophysiology 48, 973–979.

van den Bos, W., Vahl, P., Guroglu, B., van Nunspeet, F., Colins, O., Markus, M., Rombouts, S.A., van der Wee, N., Vermeiren, R., Crone, E.A., 2014. Neural correlates of social decision-making in severely antisocial adolescents. Soc. Cogn. Affect. Neurosci. 9, 2059–2066.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Natl. Acad. Sci. USA 100, 3531–3535.

Brune, M., Scheele, D., Heinisch, C., Tas, C., Wischniewski, J., Gunturkun, O., 2012. Empathy moderates the effect of repetitive transcranial magnetic stimulation of the right dorsolateral prefrontal cortex on costly punishment. PLoS One 7, e44747.

Brüne, M., Brüne-Cohrs, U., 2006. Theory of mind–evolution, ontogeny, brain mechanisms and psychopathology. Neurosci. Biobehav. Rev. 30, 437–455.

Brüne, M., von Hein, S.M., Claassen, C., Hoffmann, R., Saft, C., 2021. Altered third-party punishment in Huntington's disease: a study using neuroeconomic games. Brain Behav. 11, e01908.

Bshary, R., Grutter, A.S., 2005. Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. Biol. Lett. 1, 396–399.

Buckholtz, J.W., Marois, R., 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat. Neurosci. 15, 655–661.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60, 930–940.

Buckholtz, J.W., Martin, J.W., Treadway, M.T., Jan, K., Zald, D.H., Jones, O., Marois, R., 2015. From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron 87, 1369–1380.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann. N. Y. Acad. Sci. 1124, 1–38.

Bushway, S.D., Piehl, A.M., 2007. The inextricable link between age and criminal history in sentencing. Crime. Delinq. 53, 156–183.

Buyukozer Dawkins, M., Sloane, S., Baillargeon, R., 2019. Do infants in the first year of life expect equal resource allocations? Front. Psychol. 10.

Caldwell, M., Van Rybroek, G., 2001. Efficacy of a decompression treatment model in the clinical management of violent juvenile offenders. Int. J. Offender Ther. Comp. Criminol. 45, 469–477.

Carlsmith, K.M., 2006. The roles of retribution and utility in determining punishment. J. Exp. Soc. Psychol. 42, 437–451.

Carlsmith, K.M., Darley, J.M., Robinson, P.H., 2002. Why do we punish? Deterrence and just deserts as motives for punishment. J. Exp. Soc. Psychol. 83, 284–299.

Carver, C.S., Johnson, S.L., Joormann, J., 2008. Serotonergic function, two-mode models of self-regulation, and vulnerability to depression: what depression has in common with impulsive aggression. Psychol. Bull. 134, 912–943.

Chang, L.J., Sanfey, A.G., 2013. Great expectations: neural computations underlying the use of social norms in decision-making. Soc. Cogn. Affect. Neurosci. 8, 277–284.

Cheng, X., Zheng, L., Liu, Z., Ling, X., Wang, X., Ouyang, H., Chen, X., Huang, D., Guo, X., 2022. Punishment cost affects third-parties' behavioral and neural responses to unfairness. Int. J. Psychophysiol. 177, 27–33.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., Rumiati, R.I., 2010. Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. Cognition 114, 89–95.

Civai, C., Crescentini, C., Rustichini, A., Rumiati, R.I., 2012. Equality versus self-interest in the brain: differential roles of anterior insula and medial prefrontal cortex. Neuroimage 62, 102–112.

Civai, C., Rumiati, R.I., Rustichini, A., 2013. More equal than others: equity norms as an integration of cognitive heuristics and contextual cues in bargaining games. Acta Psychol. 144, 12–18.

Civai, C., Huijsmans, I., Sanfey, A.G., 2019a. Neurocognitive mechanisms of reactions to second- and third-party justice violations. Sci. Rep. 9, 9271.

Civai, C., Huijsmans, I., Sanfey, A.G., 2019b. Neurocognitive mechanisms of reactions to second- and third-party justice violations. Sci. Rep. 9, 9271.

Civai, C., Teodorini, R., Carrus, E., 2020. Does unfairness sound wrong? A cross-domain investigation of expectations in music and social decision-making. R. Soc. Open Sci. 7, 190048.

Crockett, M.J., Clark, L., Tabibnia, G., Lieberman, M.D., Robbins, T.W., 2008. Serotonin modulates behavioral reactions to unfairness. Science 320, 1739.

Crockett, M.J., Clark, L., Hauser, M.D., Robbins, T.W., 2010a. Serotonin selectively influences moral judgment and behavior through effects on harm aversion. Proc. Natl. Acad. Sci. USA 107, 17433–17438.

Crockett, M.J., Clark, L., Lieberman, M.D., Tabibnia, G., Robbins, T.W., 2010b. Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. Emotion 10, 855–862.

Crockett, M.J., Özdemir, Y., Fehr, E., 2014. The value of vengeance and the demand for deterrence. J. Exp. Psychol. Gen. 143, 2279–2286.

Cushman, F., 2008. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. Cognition 108, 353–380.

Cushman, F., 2014. Punishment in humans: from intuitions to institutions. Philos. Compass 10, 117–133.

Dai, D.Y., Sternberg, R.J.E., 2004. Motivation, Emotion and Cognition: Integrative Perspectives on Intellectual Functioning and Development. Lawrence Erlbaum Associates.

Darley, J.M., 2009. Morality in the law: the psychological foundations of citizens' desires to punish transgressions. Annu. Rev. Law Soc. 5, 1–23.

Decety, J., Yoder, K.J., 2017. The emerging social neuroscience of justice motivation. Trends Cogn. Sci. 21, 6–14.

DesChamps, T.D., Eason, A.E., Sommerville, J.A., 2015. Infants associate praise and admonishment with fair and unfair individuals. Infancy 21, 478–504.

Diamantis, M.E., 2021. The corporate insanity defense. J. Crim. Law Criminol. 111, 1–92.

dos Santos, M., Braithwaite, V., Wedekind, C., 2014. Exposure to superfluous information reduces cooperation and increases antisocial punishment in reputation-based interactions. Front. Ecol. Evol. 2.

Elliott, A.J., Niesta, D., 2009. Goals in the context of the hierarchical model of approach-avoidance motivation, in: Moskowitz, G., Grant, H. (Eds.), The psychology of goals Guilford, pp. 56–76.

Eriksson, K., Andersson, P.A., Strimling, P., 2017. When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance. Judgm. Decis. Mak. 396–407.

Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn. Sci. 15, 85–93.

Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness-Intentions matter. Games Econ. Behav. 62, 287–303.

Fehr, E., Fischbacher, U., 2004a. Social norms and human cooperation. Trends Cogn. Sci. 8, 185–190.

Fehr, E., Fischbacher, U., 2004b. Third-party punishment and social norms. Evol. Hum. Behav. 25, 63–87.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Feng, C., Luo, Y.J., Krueger, F., 2014. Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. Hum. Brain Mapp. 36, 591–602.

Feng, C., Zhu, Z., Gu, R., Wu, X., Luo, Y.J., Krueger, F., 2018. Resting-state functional connectivity underlying costly punishment: a machine-learning approach. Neuroscience 385, 25–37.

Feng, C., Eickhoff, S.B., Li, T., Wang, L., Becker, B., Camilleri, J.A., Hetu, S., Luo, Y., 2021. Common brain networks underlying human social interactions: Evidence from large-scale neuroimaging meta-analysis. Neurosci. Biobehav Rev. 126, 289–303.

Feng, C., Yang, Q., Azem, L., Atanasova, K.M., Gu, R., Luo, W., Hoffman, M., Lis, S., Krueger, F., 2022. An fMRI investigation of the intention-outcome interactions in second- and third-party punishment. Brain Imaging Behav. 16, 715–727.

Fetchenhauer, D., Huang, X., 2004. Justice sensitivity and distributive decisions in experimental games. Pers. Individ. Differ. 36, 1015–1029.

Fishbach, A., Ferguson, M.F., 2007. The goal construct in social psychology. In: Kruglanski, A.W., Higgins, E.T. (Eds.), Social psychology. Handbook of Basic Principles Guilford Press, New York, pp. 490–515.

Flack, J.C., Girvan, M., de Waal, F.B.M., Krakauer, D.C., 2006. Policing stabilizes construction of social niches in primates. Nature 493 426–429.

Fourie, M.M., Hortensius, R., Decety, J., 2020. Parsing the components of forgiveness: Psychological and neural mechanisms. Neurosci. Biobehav. Rev. 112, 437–451.

Fowler, J.H., 2005. Altruistic punishment and the origin of cooperation. Proc. Natl. Acad. Sci. USA 102, 7047–7049.

Frith, C., Frith, U., 2005. Theory of mind. Curr. Biol. 15, 644–645.

Gabay, A.S., Radua, J., Kempton, M.J., Mehta, M.A., 2014. The ultimatum game and the brain: a meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. 47, 549–558.

Galizzi, M.M., Navarro-Martinez, D., 2019. On the external validity of social preference games: a systematic lab-field study. Manag. Sci. 65, 976–1002.

Gehring, W.J., Willoughby, A.R., 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. Science 295, 2279–2282.

Geraci, A., 2021. Toddlers' expectations of corporal third-party punishments against the non-defender puppet. J. Exp. Child Psychol. 210, 105199.

Geraci, A., Surian, L., 2021. Toddlers' expectations of third-party punishments and rewards following an act of aggression. Aggress. Behav. 47, 521–529.

Geraci, A., Surian, L., 2023a. Intention-based evaluations of distributive actions by 4-month-olds. Infant Behav. Dev. 70, 101797.

Geraci, A., Surian, L., 2023b. Preverbal infants' reactions to third-party punishments and rewards delivered toward fair and unfair agents. J. Exp. Child Psychol. 226, 105574.

Geraci, A., Simion, F., Surian, L., 2022. Infants' intention-based evaluations of distributive actions. J. Exp. Child Psychol. 220.

Ginther, M.R., Shen, F.X., Bonnie, R.J., Hoffman, M.B., Jones, O.D., Marois, R., Simons, K.W., 2014. The language of mens rea. Vanderbilt Law Rev. 67, 1327–1372.

Ginther, M.R., Bonnie, R.J., Hoffman, M.B., Shen, F.X., Simons, K.W., Jones, O.D., Marois, R., 2016. Parsing the behavioral and brain mechanisms of third-party punishment. J. Neurosci. 36, 9420–9434.

Ginther, M.R., Shen, F.X., Bonnie, R.J., Hoffman, M.B., Jones, O.D., Marois, R., Simons, K.W., 2018. Decoding guilty minds: How jurors attribute knowledge and guilt. Vanderbilt Law Rev. 71, 241–283.

Ginther, M.R., Hartsough, L.E.S., Marois, R., 2022. Moral outrage drives the interaction of harm and culpable intent in third-party punishment decisions. Emotion 22, 795–804.

Gintis, H., Henrich, J., Bowles, S., R., B., Fehr, E., 2008. Strong reciprocity and the roots of human morality. Soc. Justice Res. 21, 241–253.

Goldinger, S.D., Kleider, H.M., Azuma, T., Beike, D.R., 2003. Blaming the victim" under memory load. Psychol. Sci. 14, 81–85.

Gospic, K., Mohlin, E., Fransson, P., Petrovic, P., Johannesson, M., Ingvar, M., 2011. Limbic justice–amygdala involvement in immediate rejection in the Ultimatum Game. PLoS Biol. 9, e1001054.

Grayson, D.S., Fair, D.A., 2017. Development of large-scale functional networks from birth to adulthood: a guide to the neuroimaging literature. Neuroimage 160, 15–31.

Gummerum, M., Chu, M.T., 2014. Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. Cognition 133, 97–103.

Gummerum, M., Van Dillen, L.F., Van Dijk, E., López-Pérez, B., 2016. Costly third-party interventions: the role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. J. Exp. Soc. Psychol. 65, 94–104.

Gummerum, M., López-Pérez, B., Van Dijk, E., Van Dillen, L.F., 2022. Ire and punishment: Incidental anger and costly punishment in children, adolescents, and adults. J. Exp. Child Psychol. 218, 105376.

Guroglu, B., van den Bos, W., van Dijk, E., Rombouts, S.A., Crone, E.A., 2011. Dissociable brain networks involved in development of fairness considerations: understanding intentionality behind unfairness. Neuroimage 57, 634–641.

Guroglu, B., Will, G.J., Crone, E.A., 2014. Neural correlates of advantageous and disadvantageous inequity in sharing decisions. PLoS One 9, e107996.

Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388.

Haidt, J., 2007. The new synthesis in moral psychology. Science 316, 998–1002.

Hamilton, W.D., 1964a. The genetical evolution of social behaviour. I. J. Theor. Biol. 7, 1–16.

Hamilton, W.D., 1964b. The genetical evolution of social behaviour. II. J. Theor. Biol. 7, 17–52.

Hamlin, J.K., Wynn, K., 2011. Young infants prefer prosocial to antisocial others. Cogn. Dev. 26, 30–39.

Hamlin, J.K., Wynn, K., Bloom, P., 2007. Social evaluation by preverbal infants. Nature 450, 557–559.

Hamlin, J.K., Wynn, K., Bloom, P., 2010. Three-month-olds show a negativity bias in their social evaluations. Dev. Sci. 13, 923–929.

Hamlin, J.K., Wynn, K., Bloom, P., Mahajan, N., 2011. How infants and toddlers react to antisocial others. Proc. Natl. Acad. Sci. USA 108, 19931–19936.

Harle, K.M., Chang, L.J., van 't Wout, M., Sanfey, A.G., 2012. The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. Neuroimage 61, 32–40.

Hartsough, L.E.S., Ginther, M.R., Marois, R., 2020. Distinct affective responses to second- and third-party norm violations. Acta Psychol. 205, 103060.

Haselton, M., Buss, D., 2009. Error management theory and the evolution of misbeliefs. Behav. Brain Sci. 32, 522–523.

Hauser, M.D., 1992. Costs of deception: cheaters are punished in rhesus monkeys (Macaca mulatta). Proc. Natl. Acad. Sci. USA 89, 12137–12139.

Heckler, S., Kessler, T., 2018. On the difference between moral outrage and empathic anger: anger about wrongful deeds or harmful consequences. J. Exp. Soc. Psychol. 76, 270–282.

Heffner, J., FeldmanHall, O., 2019. Why we don't always punish: preferences for non-punitive responses to moral violations. Sci. Rep. 9, 13219.

Henrich, J., McElreath, R., Abigail Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312.

Hetu, S., Luo, Y., D'Ardenne, K., Lohrenz, T., Montague, P.R., 2017. Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. Soc. Cogn. Affect. Neurosci. 12, 1972–1982.

Hewig, J., Kretschmer, N., Trippe, R.H., Hecht, H., Coles, M.G., Holroyd, C.B., Miltner, W.H., 2011. Why humans deviate from rational choice. Psychophysiology 48, 507–514.

Hirschfeld-Kroen, J., Jiang, K., Wasserman, E., Anzellotti, S., Young, L., 2021. When my wrongs are worse than yours: Behavioral and neural asymmetries in first-person and third-person perspectives of accidental harms. J. Exp. Soc. Psychol. 94, 104102.

Hoffman, M., 2014. The Punisher's Brain: The Evolution of Judge and Jury, 1 ed. Cambridge University Press New York.

Hoffman, M.B., Goldsmith, T.J., 2004. The biological roots of punishment. Ohio State J. Crim. L. 1, 627–641.

Hoffman, M.B., Shen, F.X., Iyengar, V., Krueger, F., 2020. The intersectionality of age and gender on the bench: are younger female judges harsher with serious crimes? Columbia J. Gend. L 40 (1), 128–165, 40.

Hoffman, M.L., 1986. Affect, cognition, and motivation. In: Sorrentino, R.M., Higgins, E. T. (Eds.), Handbook of Motivation and Cognition: Foundations of Social Behavior. Guilford Press, pp. 244–280.

Hsu, M., Anen, C., Quartz, S.R., 2008. The right and the good: distributive justice and neural encoding of equity and efficiency. Science 320, 1092–1095.

Hutcherson, C.A., Tusche, A., 2022. Evidence accumulation, not 'self-control', explains dorsolateral prefrontal activation during normative choice. Elife 11, e65661.

Hyatt, C.J., Calhoun, V.D., Pearlson, G.D., Assaf, M., 2015. Specific default mode subnetworks support mentalizing as revealed through opposing network recruitment by social and semantic FMRI tasks. Hum. Brain Mapp. 36, 3047–3063.

Jaroslawska, A.J., McCormack, T., Burns, P., Caruso, E.M., 2020. Outcomes versus intentions in fairness-related decision making: School-aged children's decisions are just like those of adults. J. Exp. Child Psychol. 189, 104704.

Johnson, B.D., King, R.D., 2017. Facial profiling: race, physical appearance, and punishment. Criminol.: Interdiscip. J. 55, 520–547.

Jordan, J.J., Hoffman, M., Bloom, P., Rand, D.G., 2016. Third-party punishment as a costly signal of trustworthiness. Nature 530, 473–476.

Kanakogi, Y., Miyazaki, M., Takahashi, H., Yamamoto, H., Kobayashi, T., Hiraki, K., 2022. Third-party punishment by preverbal infants. Nat. Hum. Behav. 6, 1234–1242.

Kant, I., 1797. The Philosophy of Law: An Exposition of the Fundamental Principles of Jurisprudence as the Science of Right. T&T Clark 1887, Edinburgh.

Keller, L.B., Oswald, M.E., Stucki, I., Gollwitzer, M., 2010. A closer look at an eye for an eye: Laypersons' punishment decisions are primarily driven by retributive motives. Soc. Justice Res. 23, 99–116.

Kiehl, K.A., Hoffman, M.B., 2011. The criminal psychopath: history, neuroscience, treatment, and economics. Jurimetrics 51, 355–397.

Kim, S., Reeve, J., Bong, M., 2016. Introduction to motivational neuroscience. Recent Developments in Neuroscience Research on Human Motivation. Emerald Publishing Limited, Bingley, pp. 1–19.

Knobe, J., 2003. Intentional action and side-effects in ordinary language. Analysis 63, 190–193.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E., 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314, 829–832.

Knoch, D., Nitsche, M.A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., Fehr, E., 2008. Studying the neurobiology of social interaction with transcranial direct current stimulation–the example of punishing unfairness. Cereb. Cortex 18, 1987–1990.

Koenigs, M., Kruepke, M., Newman, J.P., 2010. Economic decision-making in psychopathy: a comparison with ventromedial prefrontal lesion patients. Neuropsychologia 48, 2198–2204.

Krueger, F., Hoffman, M., 2016. The emerging neuroscience of third-party punishment. Trends Neurosci. 39, 499–501.

Krueger, F., Bellucci, G., Xu, P., Feng, C., 2020. The critical role of the right dorsal and ventral anterior insula in reciprocity: Evidence from the trust and Ultimatum Games. Front. Hum. Neurosci. 14, 176.

Kruglanski, A.W., 1996. Goals as knowledge structures. In: Gollwitzer, P.M., Bargh, J.A. (Eds.), The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford Press, New York, pp. 599–618.

Kruglanski, A.W., 2017. Motivational phases on the road to action. Motiv. Sci. 3, 196–207.

Kruglanski, A.W., Belanger, J.J., Chen, X., Kopetz, C., Pierro, A., Mannetti, L., 2012. The energetics of motivated cognition: a force-field analysis. Psychol. Rev. 119, 1–20.

Kruglanski, A.W., Chernikova, M., Rosenzweig, E., Kopetz, C., 2014a. On motivational readiness. Psychol. Rev. 121, 367–388.

Kruglanski, A.W., Chernikova, M., Schori-Eyal, N., 2014b. From readiness to action: how motivation works. Pol. Psychol. Bull. 45, 259–267.

Kundro, T.G., Nurmohamed, S., Kakkar, H., Affinito, S.J., 2023. Time and punishment: time delays exacerbate the severity of third-party punishment. Psychol. Sci. 9567976231173900.

Leibbrandt, A., López-Pérez, R., 2012. An exploration of third and second party punishment in ten simple games. J. Econ. Behav. Organ. 84, 753–766.

Li, A., 2013. Intuition and its bias control in the judicial process. Soc. Sci. China 5, 207–208, 142-161+.

Li, T., Yang, Y., Krueger, F., Feng, C., Wang, J., 2022a. Static and dynamic topological organizations of the costly punishment network predict individual differences in punishment propensity. Cereb. Cortex 32, 4012–4024.

Li, Y., Hu, J., Ruff, C.C., Zhou, X., 2022b. Neurocomputational evidence that conflicting prosocial motives guide distributive justice. Proc. Natl. Acad. Sci. USA 119, e2209078119.

Liu, Y., He, N., Dou, K., 2015. Ego-depletion promotes altruistic punishment. Open J. Soc. Sci. 03, 62–69.

Lo Gerfo, E., Gallucci, A., Morese, R., Vergallito, A., Ottone, S., Ponzano, F., Locatelli, G., Bosco, F., Romero Lauro, L.J., 2019. The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. Neuroimage 501–510.

Lotz, S., Okimoto, T.G., Schl?Sser, T., Fetchenhauer, D., 2011. Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. J. Exp. Soc. Psychol. 47, 477–480.

Ma, N., Li, N., He, X.S., Sun, D.L., Zhang, X., Zhang, D.R., 2012. Rejection of unfair offers can be driven by negative emotions, evidence from modified ultimatum games with anonymity. PLoS One 7, e39619.

Manrique, H., Zeidler, H., Roberts, G., Barclay, P., Walker, M., Samu, F., Fariña, A., Bshary, R., Raihani, N., 2021. The psychological foundations of reputation-based cooperatio. Philos. Trans. R. Soc. B 376, 20200287.

Marlowe, F.W., 2009. Hadza cooperation: second-party punishment, yes; third-party punishment, no. Hum. Nat. 20, 417–430.

Marlowe, F.W., Berbesque, J.C., Barrett, C., Bolyanatz, A., Gurven, M., Tracer, D., 2011. The 'spiteful' origins of human cooperation. Proc. R. Soc. B 278, 2159–2164.

Martinez-Vaquero, L.A., Han, T.A., Pereira, L.M., Lenaerts, T., 2015. Apology and forgiveness evolve to resolve failures in cooperative agreements. Sci. Rep. 5, 10639.

McAuliffe, K., Jordan, J.J., Warneken, F., 2015. Costly third-party punishment in young children. Cognition 134, 1–10.

McAuliffe, K., Blake, P.R., Steinbeis, N., Warneken, F., 2017. The developmental foundations of human fairness. Nat. Hum. Behav. 1, 0042.

Menon, V., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn. Sci. 15, 483–506.

Menon, V., 2015. Salience Network. In: Toga, A.W. (Ed.), Brain Mapping. Academic Press, Waltham, pp. 597–611.

Morse, S., Hoffman, M., 2007. The uneasy entente between insanity and mens rea: beyond Clark v. Arizona. J. Crim. Law Criminol. 97, 1071–1149.

Mueller, P., Solan, L., Darley, J.M., 2012. When does knowledge become intent? Perceiving the minds of wrongdoers. J. Empir. Leg. Stud. 9, 859–892.

Muller-Leinss, J.M., Enzi, B., Flasbeck, V., Brune, M., 2018. Retaliation or selfishness? An rTMS investigation of the role of the dorsolateral prefrontal cortex in prosocial motives. Soc. Neurosci. 13, 701–709.

Namkung, H., Kim, S.H., Sawa, A., 2017. The insula: An underestimated brain area in clinical neuroscience, psychiatry, and neurology. Trends Neurosci. 40, 200–207.

Nelissen, R.M.A., Zeelenberg, M., 2009. Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanction. Judgm. Decis. Mak. 4, 543–553.

Nettler, G., 1959. Cruelty, dignity, and determinism. Am. Soc. Rev. 24, 375–384.

O'Doherty, J.P., 2016. Multiple systems for the motivational control of behavior and associated neural substrates in humans. Curr. Top. Behav. Neurosci. 27, 291–312.

Ouss, A., Peysakhovich, A., 2015. When punishment doesn't pay: Cold glow and decisions to punish. J. L. Econ. Org. 58, 625–655.

Peay, J., Player, E., 2018. Pleading guilty: Why vulnerability matters. Mod. Law Rev. 81, 929–957.

Pedersen, E.J., Kurzban, R., McCullough, M.E., 2013. Do humans really punish altruistically? A closer look. Proc. R. Soc. B: Biol. 280, 20122723-20122723.

Pedersen, E.J., McAuliffe, W.H.B., McCullough, M.E., 2018. The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. J. Exp. Psychol. Gen. 147, 514–544.

Peysakhovich, A., Nowak, M.A., Rand, D.G., 2014. Humans display a 'cooperative phenotype' that is domain general and temporally stable. Nat. Commun. 5, 4939.

Philippot, P., Feldman, R.S.E., 2004. The Regulation of Emotion. Psychology Press,.

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305, 1254–1258.

Rabinowitz, M., Lucca, K., Pospisil, J., Sommerville, J.A., 2018. Fairness informs social decision making in infancy. In: Plos One, 13.

Riedl, K., Jensen, K., Call, J., Tomasello, M., 2012. No third-party punishment in chimpanzees. Proc. Natl. Acad. Sci. U. S. A. 109, 14824–14829.

Robinson, P.H., Kurtzban, R.O., 2007. Concordance and conflict in institutions of justice. Minn. Law Rev. 91, 1829–1907.

Sanfey, A.G., 2007. Social decision-making: insights from game theory and neuroscience. Science 318, 598–602.

Sanfey, A.G., 2009. Expectations and social decision-making: biasing effects of prior knowledge on Ultimatum responses. Mind Soc. 8, 93–107.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 300, 1755–1758.

Schmidt, L., Tusche, A., Manoharan, N., Hutcherson, C., Hare, T., Plassmann, H., 2018. Neuroanatomy of the vmPFC and dlPFC predicts individual differences in cognitive regulation during dietary self-control across regulation strategies. J. Neurosci. 38, 5799–5806.

Schmitt, M., Baumert, A., Gollwitzer, M., Maes, J., 2010. The justice sensitivity inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. Soc. Justice Res. 23, 211–238.

Schwartz, F., Djeriouat, H., Trémolière, B., 2022. Judging accidental harm: Reasoning style modulates the weight of intention and harm severity. Q. J. Exp. Psychol. (Hove), 17470218221089964.

Seymour, B., Singer, T., Dolan, R., 2007. The neurobiology of punishment. Nat. Rev. Neurosci. 8, 300–311.

Shackman, A.J., Salomons, T.V., Slagter, H.A., Fox, A.S., Winter, J.J., Davidson, R.J., 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nat. Rev. Neurosci. 12, 154–167.

Shen, F., 2020. Aging judges. Ohio St. L. J. 81, 235–314.

Shen, F.X., Hoffman, M.B., Jones, O., Greene, D., Marois, R, J.D., 2011. Sorting guilty minds. N. Y. Univ. Law Rev. 86, 1306–1360.

Singh, M., Garfield, Z.H., 2022. Evidence for third-party mediation but not punishment in Mentawai justice. Nat. Hum. Behav. 930–940.

Sodian, B., Kristen, S., 2010. Theory of mind. In: Glatzeder, B.M., V., G., von Muller, A. (Eds.), Towards a Theory of Thinking. Springer, Berlin, Germany, pp. 189–201.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C.K.W., Sanfey, A.G., 2018. Neurobiological mechanisms of responding to injustice. J. Neurosci. 38, 2944–2954.

Strang, S., Utikal, V., Fischbacher, U., Weber, B., Falk, A., 2014. Neural correlates of receiving an apology and active forgiveness: an FMRI study. PLoS One 9, e87654.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., Kirsch, P., 2011. Beyond revenge: neural and genetic bases of altruistic punishment. Neuroimage 54, 671–680.

Supekar, K., Musen, M., Menon, V., 2009. Development of large-scale functional brain networks in children. PLoS Biol. 7, e1000157.

Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol. Sci. 19, 339–347.

Tan, E., Hamlin, J.K., 2022. Mechanisms of social evaluation in infancy: A preregistered exploration of infants' eye-movement and pupillary responses to prosocial and antisocial events. Infancy 27, 255–276.

Tang, Z., Qu, C., Hu, Y., Benistant, J., Moisan, F., Derrington, E., Dreher, J.C., 2023. Strengths of social ties modulate brain computations for third-party punishment. Sci. Rep. 13, 10510.

Tasimi, A., Wynn, K., 2016. Costly rejection of wrongdoers by infants and children. Cognition 151, 76–79.

Taylor, P., Uchida, Y., 2022. Horror, fear, and moral disgust are differentially elicited by different types of harm. Emotion 22, 346–361.

Tibboel, H., De Houwer, J., Van Bockstaele, B., 2015. Implicit measures of "wanting" and "liking" in humans. Neurosci. Biobehav. Rev. 57, 350–364.

Toribio-Flórez, D., Saße, J., Baumert, A., 2023. Proof under reasonable doubt": Ambiguity of the norm violation as boundary condition of third-party punishment. Pers. Soc. Psychol. Bull. 49, 429–446.

Treadway, M., T.J., Buckholtz, J.W., Martin, J.W., Jan, K., Asplund, C.L., Ginther, M.R., Jones, O., D., Marois, R., 2014. Corticolimbic gating of emotion-driven punishment. Nat. Neurosci. 17, 1270–1277.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46, 35–57.

Twardawski, M., Tang, K.T.Y., Hilbig, B.E., 2020. Is it all about retribution? The flexibility of punishment goals. Soc. Justice Res. 33, 195–218.

Uddin, L.Q., 2015. Salience processing and insular cortical function and dysfunction. Nat. Rev. Neurosci. 16, 55–61.

Van de Vondervoort, J.W., Hamlin, J.K., 2018. The early emergence of sociomoral evaluation: infants prefer prosocial others. Curr. Opin. Psychol. 20, 77–81.

van 't Wout, M., Kahn, R.S., Sanfey, A.G., Aleman, A., 2006. Affective state and decision-making in the Ultimatum Game. Exp. Brain Res. 169, 564–568.

Vidmar, N., Miller, D.T., 1980. Social psychological processes underlying attitudes toward legal punishment. Law Soc. Rev. 565–602.

Vilares, I., Wesley, M.J., Ahn, W.Y., Bonnie, R.J., Hoffman, M., Jones, O.D., Morse, S.J., Yaffe, G., Lohrenz, T., Montague, P.R., 2017. Predicting the knowledge-recklessness distinction in the human brain. Proc. Natl. Acad. Sci. U. S. A. 114, 3222–3227.

Viney, W., Parker-Martin, P., Dotten, S., 1988. Belief in free will and determinism and lack of relation to punishment rationale and magnitude. J. Gen. Psychol. 115, 15–23.

Wang, G., Li, J., Li, Z., Wei, M., Li, S., 2016. Medial frontal negativity reflects advantageous inequality aversion of proposers in the ultimatum game: An ERP study. Brain Res. 1639, 38–46.

Wang, Y., Zheng, D., Chen, J., Rao, L.L., Li, S., Zhou, Y., 2019. Born for fairness: evidence of genetic contribution to a neural basis of fairness intuition. Soc. Cogn. Affect. Neurosci. 14, 539–548.

Wenseleers, T., Ratnieks, F.L., 2006. Enforced altruism in insect societies. Nature 444, 50.

Wiedenmayer, C., 2010. Plasticity of defensive behavior and fear in early development. Neurosci. Biobehav. Rev. 33, 432–441.

Wu, Y., Zhou, Y., van Dijk, E., Leliveld, M.C., Zhou, X., 2011. Social comparison affects brain responses to fairness in asset division: An ERP study with the Ultimatum Game. Front. Hum. Neurosci. 5.

Wu, Z., Gao, X., 2018. Preschoolers' group bias in punishing selfishness in the Ultimatum Game. J. Exp. Child Psychol. 166, 280–292.

Yamada, M., Camerer, C.F., Fujie, S., Kato, M., Matsuda, T., Takano, H., Ito, H., Suhara, T., Takahashi, H., 2012. Neural circuits in the brain that are activated when mitigating criminal sentences. Nat. Commun. 3, 759.

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., Cook, K.S., 2009. The private rejection of unfair offers and emotional commitment. Proc. Natl. Acad. Sci. U. S. A. 106, 11520–11523.

Yang, Q., Shao, R., Zhang, Q., Li, C., Li, Y., Li, H., Lee, T., 2019. When morality opposes the law: An fMRI investigation into punishment judgments for crimes with good intentions. Neuropsychologia 127, 195–203.

Yang, Q., Bellucci, G., Hoffman, M., Hsu, K.T., Lu, B., Deshpande, G., Krueger, F., 2021. Intrinsic functional connectivity of the frontoparietal network predicts inter-individual differences in the propensity for costly third-party punishment. Cogn. Affect. Behav. Neurosci. 21, 1222–1232.

Yetnikoff, L., Lavezzi, H.N., Reichard, R.A., Zahm, D.S., 2014. An update on the connections of the ventral mesencephalic dopaminergic complex. Neuroscience 282, 23–48.

Yoder, K.J., Decety, J., 2020. Me first: Neural representations of fairness during three-party interactions. Neuropsychologia 147.

Young, L., Cushman, F., Hauser, M., Saxe, R., 2007. The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. U. S. A. 104, 8235–8240.

Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R., 2010. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proc. Natl. Acad. Sci. U. S. A. 107, 6753–6758.

Yu, H., Li, J., Zhou, X., 2015. Neural substrates of intention-consequence integration and its impact on reactive punishment in interpersonal transgression. J. Neurosci. 35, 4917–4925.

Zhen, S., Yu, R., 2016. Tend to compare and tend to be fair: the relationship between social comparison sensitivity and justice sensitivity. PLoS One 11, e0155414.

Zinchenko, O., Arsalidou, M., 2017. Brain responses to social norms: meta-analyses of fMRI studies. Hum. Brain Mapp. 39, 955–970.

Ziv, T., Sommerville, J.A., 2016. Developmental differences in infants' fairness expectations from 6 to 15 months of age. Child Dev. 88, 1930–1951.