Cognition 208 (2021) 104544

Contents lists available at ScienceDirect

Cognition

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

Punishment is Organized around Principles of Communicative Inference

Arunima Sarin [a,*], Mark K. Ho [b], Justin W. Martin [c], Fiery A. Cushman [a]

[a] Department of Psychology, Harvard University, Cambridge, MA 02138, United States of America
[b] Department of Psychology, Princeton University, Princeton, NJ 08540, United States of America
[c] Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA 02467, United States of America

ARTICLE INFO

Keywords:
Punishment
Communication
Social cognition
Mental state inference
Pragmatics
Learning

ABSTRACT

Humans use punishment to influence each other's behavior. Many current theories presume that this operates as a simple form of incentive. In contrast, we show that people infer the communicative intent behind punishment, which can sometimes diverge sharply from its immediate incentive value. In other words, people respond to punishment not as a reward to be maximized, but as a communicative signal to be interpreted. Specifically, we show that people expect harmless, yet communicative, punishments to be as effective as harmful punishments (Experiment 1). Under some situations, people display a systematic preference for harmless punishments over more canonical, harmful punishments (Experiment 2). People readily seek out and infer the communicative message inherent in a punishment (Experiment 3). And people expect that learning from punishment depends on the ease with which its communicative intent can be inferred (Experiment 4). Taken together, these findings demonstrate that people expect punishment to be constructed and interpreted as a communicative act.

1. Introduction

Unruly pets are scolded; disobedient children are put in time out; hardened criminals are imprisoned—when humans encounter bad behavior, they will often punish it. The prototypical punishment imposes a cost on the transgressor, and its primary function is to change the recipient's behavior (Boyd & Richerson, 1992; Clutton-Brock & Parker, 1995; Fehr & Gächter, 2002). But it is less clear exactly how punishment changes behavior. In principle, there are many different ways in which people might learn from punishment, and many corresponding ways in which punishment might be designed to generate this learning.

We focus on two broad possibilities. The first, obvious possibility is that the experience of punishment is designed to exploit simple reward learning. On this view, punishment is a "corrective whip that acts as a negative reinforcer of desired values" (Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018). Although humans have sophisticated and sometimes specialized learning mechanisms, punishment may not depend on these in order to work. Rather, it may require learning mechanisms no more complex than Thorndike's Law of Effect (1927): Repeat what has been rewarded; avoid what has been punished. This view of punishment is widely supported, and most psychological theories of punishment commit to it implicitly or explicitly. One of its principal merits is minimalism: by construing punishment as a direct incentive scheme constructed by the punisher, it seems to capture its essence. Parsimoniously, this view construes *learning from social punishment* as fundamentally similar to *learning from non-social experience*.

Yet, our everyday experiences belie this: when a person gets whipped by the wind, they avoid wind. But when a person feels the corrective whip of punishment, they do not simply avoid the punisher. Instead, they try to understand what the punisher meant. Thus, we suggest that at least between humans, punishment may be designed not only as an incentive, but also as a means for the punisher to convey their message of frustration and disapproval to the target. In keeping with this possibility—and building on prior philosophical (Bentham & Bowring, 1962; Ewing, 1943; Feinberg, 1965) and psychological research (Funk, McGeer, & Gollwitzer, 2014)—we suggest that punishment is designed to be interpreted as a form of communication (Cushman, Sarin, & Ho, 2019). On this view, transgressions are teachable moments that punishers exploit, using punishment to communicate their expected set of values and norms. Recipients, in turn, interpret the punishments as communicative acts and learn by both responding to the costs as well as by accurately inferring the punisher's intended message. This involves a sort of meeting-of-the-minds between the punisher and the recipient. It requires mental state inference on the part of both parties, as each of them attempts to guess how the other will interpret and anticipate the communicative act. This model therefore proposes that we learn from

* Corresponding author at: Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA, United States of America.
E-mail address: asarin@g.harvard.edu (A. Sarin).

https://doi.org/10.1016/j.cognition.2020.104544
Received 26 March 2020; Received in revised form 20 November 2020; Accepted 4 December 2020
Available online 28 December 2020
0010-0277/© 2020 Elsevier B.V. All rights reserved.

social punishment very differently than from non-social experiences (like a whipping wind), which are not typically interpreted as intentional communicative acts. Learning from social punishment therefore engages social cognition and inference, much like other forms of communication like language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Grice, 1957; Kao & Goodman, 2015; Kao, Wu, Bergen, & Goodman, 2014; Sperber & Wilson, 1986) or demonstration (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016; Shafto, Goodman, & Griffiths, 2014).

In summary, we contrast two models of how human punishment is designed. The first is as a "constructed incentive": The goal of the punisher is to set up a simple contingency between bad action and bad consequences, relying on the target's basic capacity for reward learning to adjust in response to this incentive. According to the second model, punishment is both a constructed incentive and a communicative act. The unique feature of this model is that punishment is designed to be interpreted by a learner who assumes communicative intent on the part of the punisher. Of course, this second model does not deny the incentive value of punishment, but instead augments it. Both models agree that punishers wish to modify the behavior of transgressors, and that transgressors wish to understand the behavioral modification intended by punishers. They disagree on a further question: whether behavioral modification is organized around simple principles of incentive (Sutton & Barto, 1998; Dayan & Niv, 2008), or around more complex principles of inferential communication akin to natural language pragmatics (Frank & Goodman, 2012; Grice, 1957; Sperber & Wilson, 1986). Our goal is to distinguish the key predictions of these models experimentally.

1.1. The psychology of punishment

At an ultimate level, one of punishment's functions is to modify the behavior of social partners (Boon, Deveau, & Alibhai, 2009; Cushman, 2015; Hampton, 1984; Heider, 1958; Martin & Cushman, 2016; Miller, 2001; Morris, 1981; Smith, 1869; but see Raihani & Bshary, 2019 for alternative functions). It typically occurs when somebody has caused harm or violated a norm, and it typically serves to both modify that person's future behavior and also indirectly influence others who observe it. The mechanisms responsible for accomplishing this function, however, are debated.

One body of evidence contends that punishment is motivated by a relatively simple taste for revenge (Carlsmith, 2006; Carlsmith, Darley, & Robinson, 2002). On this view, when people are harmed, they experience a direct motivation to cause harm in retaliation (McCullough, Kurzban, & Tabak, 2013). Thus, people often punish in situations where doing so does not serve them well rationally (Lerner & Clayton, 2011). For instance, people punish in one-shot interactions (Balafoutas, Grechenig, & Nikiforakis, 2014; Fehr & Gächter, 2002) and even when punishing makes them feel bad (Carlsmith, Wilson, & Gilbert, 2008). Such retributive instincts may serve the adaptive function of modifying others' behavior, but without much role for a proximate psychological motive of behavior modification (again, much like hunger motivates nourishment, but without much role for the proximate psychological motive of nourishment).

To the extent that human punishment amounts to simple and pure retribution, this aligns best with the "constructed incentive" model of punishment. The essential, "eye-for-an-eye" logic of retribution simply involves imposing a cost on others whenever a cost is experienced for oneself. This mechanism neither assumes nor requires that punishments be interpreted as communicative acts.

In contrast, a rich tradition in philosophy argues that a primary function of punishment is to express disapproval of the perpetrator (Durkheim, 2014; Feinberg, 1965; Stephen, 2014) and to educate both the offender and the public (Ewing, 1943). This view, collectively referred to as expressionism (Primoratz, 1989; Skillen, 1980), suggests that the cost imposed on a perpetrator is meaningful because it communicates society's condemnation of the crime. A growing body of experimental research examines this thesis by focusing on the thoughts and actions of punishers. This work suggests that human punishment is generated by more sophisticated motives than the simplest models of retribution suggest. For instance, punishers are satisfied with their punishments more when the target accurately understands their intent to punish and additionally, signals a change in his/her attitude towards his/her wrongdoing (Funk et al., 2014; Gollwitzer & Denzler, 2009; Gollwitzer, Meder, & Schmitt, 2011). In the absence of such an acknowledgment, punishers are dissatisfied with the punishment no matter how strong of a cost it levies on the target (Funk et al., 2014). In fact, even those punishers who typically punish harshly will willingly punish less severely if that would allow the target to better understand their intent (Molnar, Chaudhry, & Loewenstein, 2020). As these findings demonstrate, the desire to communicate one's beliefs, values, and disapproval plays a central role in the kind of punishment a punisher chooses to enact.

This accords with recent formal work showing that people seem to give our rewards and punishments to communicate and express approval rather than to simply incentivize an action. In one such set of experiments, participants played a computer game in which their job was to teach a child or puppy the right way to navigate to a goal (Ho, Cushman, Littman, & Austerweil, 2019a, 2019b). Participants could only teach by responding to the learner's behavior with rewards and punishments. Quantitative analysis showed that people had substantial difficulty teaching pure reward-maximizing agents (i.e., agents programmed to learn through the Law of Effect), but easily taught agents who interpreted feedback in terms of participants' communicative goals.

In sum, this body of work suggests that punishers want to be "understood", effectively modifying the behavior of those who they punish. Yet, it remains an open question exactly how they try to achieve this goal. Do they rely on a simple incentive scheme by imposing observable costs on bad behavior? Or do they rely on a more complex cognitive model of the inferences that an observer would draw in order to select specific punishments? Similarly, we know little about how people learn from punishment. Do they respond to the punishments merely as an incentive, or do they model the communicative intent of the punisher? The question at stake is not whether punishment, and learning from it, is designed to communicate *something* - it clearly is. It is also not whether people have the general capacity to express and infer communicative intent; we know this is essential to natural language, for instance (Frank & Goodman, 2012; Grice, 1957). The question is whether these aspects are connected. Is the "message" communicated by punishment limited to the immediate imposition of material costs? Or is punishment more like natural language: a form of communication organized around principles of communicative inference? On the latter view, punishers could express, and recipients could accurately infer, the "message" even in the absence of immediate, imposed incentives. This is the key idea motivating our experiments.

1.2. Experimental logic

A typical punishment is a mixture of two ingredients: the incentive it imposes and the message it communicates (see Fig. 1). Our goal is to examine which of these elements most accurately captures the necessary and sufficient conditions that make punishment work. To do so, we need to investigate the parts of this diagram that do not overlap. Do punishments that impose incentives but lack communicative elements work? Conversely, how effective are punishments that impose no cost but carry clear communicative elements?

Thus, we explore a setting in which punishment will only be successful if the recipient models the communicative intent of the punisher and cannot succeed based on simple incentive alone. We accomplish this by constructing cases where the logic of communication allows punishment to work not by imposing a cost on transgressions, but instead by providing an immediate benefit.

How can providing a benefit to somebody play the communicative

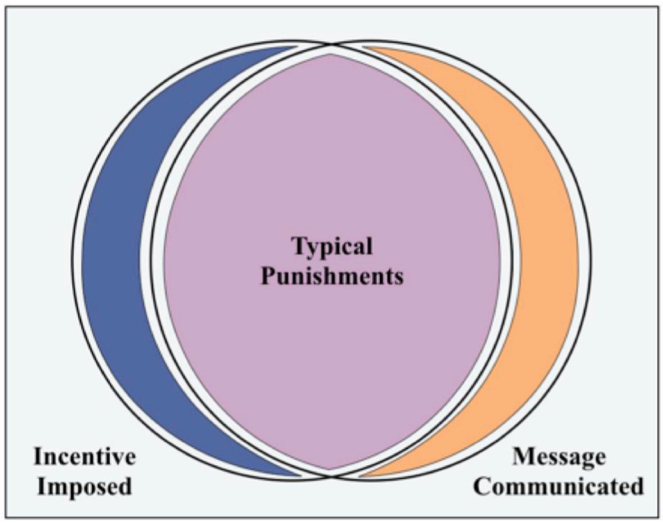


Fig. 1. Conceptual diagram of Experimental logic.

role of punishment? Our approach is inspired by models of figurative speech. Instances of figurative speech, such as verbal irony, hyperbole, etc. (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Kao et al., 2014; Kao & Goodman, 2015) entail statements whose literal meanings are different from the speaker's intended meaning (Colston & O'Brien, 2000; Roberts & Kreuz, 1994), something listeners seem to easily understand. Existing models of language suggest that listeners interpret statements not only by attending to their literal meaning, but also by inferring their potential figurative meaning from a model of the speaker's communicative intent. Therefore saying "What great weather we have today!" in response to a cold, stormy day literally conveys the speaker's enjoyment of the weather but figuratively and pragmatically communicates their displeasure.

Inspired by this work on figurative speech, we created instances of 'figurative punishment': punishments that could be interpreted appropriately as punishments only by considering the punisher's communicative intent. For example:

> *Alice lives in an apartment with three other roommates. One of them, Sandra, has a very frustrating habit of leaving her dirty dishes in the kitchen sink. The roommates have pointed out the dishes to Sandra, but they don't know exactly how to say the right thing, and she keeps doing it. Alice talks with her other roommates and together they decide to do something about it. The next day, before Sandra gets back home, Alice leaves a brand-new kitchen sponge and bottle of dish soap on her pillow with a ribbon around them and a tag that says: "Love, your roommates".*

Alice's gift to Sandra imposes no cost. On a model where recipients treat punishment as an action-incentive contingency to be maximized, Sandra should not change her behavior in light of the sponge and dish soap. If anything, she should double-down because the items impart a slight benefit. And yet, understood pragmatically, in light of her roommate's communicative intent, the sponge and dish soap feel like sanctions and lead Sandra to correct her behavior. This is because (as we show experimentally) the most plausible interpretation of Alice's behavior is that it communicates frustration with Sandra's failure to do the dishes, not approval of it. Of course, this presumably motivates Sandra because it implies that she will *eventually* pay costs for not doing dishes. Alice's action may therefore be effective because it communicates a threat. It may also engender feelings of guilt. However, the threat, the potential for guilt, and any associated social and emotional costs experienced stem from the inferred communicative intent of Alice's act and not from the act itself. In other words, the force motivating Sandra to change her behavior does not follow directly from the value of Alice's act, as the constructed incentive model demands. Instead, it follows from the inferred communicative intent of Alice's act, as our model allows.

We study these cases of figurative punishments not because they are normal, but because they provide us with a useful case study. Studying figurative punishments can shed light on the psychological underpinnings of typical "literal" punishment, much in the same way that studying figurative speech, such as irony and hyperbole, can inform our understanding of regular speech. According to our account, even ordinary punishments may be interpreted in light of their apparent communicative intent. But it is hard to know because a simple constructed incentive model also predicts that we can learn from these ordinary punishments. Studying figurative punishment is useful because, despite being uncommon, it is highly diagnostic of the underlying psychological mechanisms.

It is worth noting that our focus in this research is on acts and whether they are interpreted in light of the actor's communicative intent, not on speech. Even though we propose that interpreting social punishments is cognitively akin to understanding language, actions differ from speech in many ways (see General Discussion for a more in-depth discussion on this). One fundamental difference is that language relies on conventionalized semantics that relate meanings to words, whereas arbitrary actions do not have conventionalized meaning. Additionally, while much work investigates linguistic pragmatics, less work in social and cognitive science has examined the principles that ground inferences about communicative actions. Understanding this, within the context of punishments, is the focus of our research.

Our studies, all pre-registered[1,2] use several complementary methods to identify whether punishment is designed to impose immediate, direct costs on social partners who transgress, or instead to provide informative communicative signals. Experiment 1 asked whether people think that figurative punishments can be informative and effective. Experiments 2a and 2b asked whether people ever prefer figurative punishments to literal ones and how context influences their choice. Experiment 3 asked whether people infer the communicative intent of figurative punishments even in ambiguous cases. Finally, Experiment 4 asks whether "literal" punishments are less effective in the absence of clear communicative intent. We report on all the dependent variables that were collected for the experiments. Data and analysis scripts for all experiments can be found on OSF at https://osf.io/fxdwm/.

2. Experiment 1

We begin by testing whether people believe that a "figurative punishment" could ever possibly be an effective way of changing someone's behavior. In other words, do they believe that an immediate, direct "incentive" is necessary to make punishment work, or do they believe that a communicative act alone can be sufficient?

We described to participants several hypothetical cases of "literal" versus "figurative" punishments. We designed both types of punishment to have clear communicative intent to express displeasure with a transgression (e.g., leaving dirty dishes piled up in a communal sink). What differed was the actual, immediate consequences of the punishment: Literal punishments imposed a cost (e.g., dirty dishes left on a roommate's pillow), while figurative punishments imposed a slight benefit (e.g., a new sponge and bottle of dishwashing detergent left on the roommate's pillow).

In two separate conditions we also tested whether people believe that clear communicative intent is *necessary* for punishment to work, or

[1] Pre-registered documents can be found at the following links: Experiment 1 (https://aspredicted.org/9g5zj.pdf); Experiment 2a (https://aspredicted.org/8cn5k.pdf); Experiment 3 (https://aspredicted.org/vv5t3.pdf); Experiment 4 (https://aspredicted.org/7de6p.pdf)

[2] Sample size for each study was calculated based on pilot data and power analyses to obtain at least 80% power of detecting the reported effects.

whether they instead consider an imposed cost (i.e., incentive) to be necessary. To do this we constructed baseline conditions that were matched in the relative costs and benefits of the literal and figurative punishments, respectively, but were unclear in their communicative intent (that is, difficult to interpret given the context). For instance, in a case where the roommate has been leaving her dishes undone, the punisher either throws her mail into her room in a disorderly manner, or else stacks it very neatly in her room. These behaviors impose immediate costs and benefits, respectively. But, crucially, they have no transparent connection to dishes. Therefore, their communicative intent is obscure. We asked people whether behaviors like these would-be effective forms of punishment.

Thus, we have a 2 × 2 design: punishments either imposed a cost or were costless and they either clearly communicated the punisher's intent or not. This yielded the following four conditions: Literal (costly) and figurative (costless) punishments with transparent communicative intent, and matched non-communicative versions of each (costly and costless baselines). Participants read one of the four possible kinds of punishment and answered a series of questions meant to assess their expectations about the effectiveness of the punishment as third-party observers. If people believe that imposing a cost is sufficient to make punishment effective, then the two conditions imposing costs (literal punishments and baseline costly punishments) should be judged effective. If they believe that clear communicative intent is sufficient to make punishment effective, then the two conditions with clear communicative intent (literal and figurative punishments) should be judged effective. It is also possible that people will judge both of these factors sufficient—in this case, the only *ineffective* condition would be the baseline costless condition. Or it is possible that they will judge both factors necessary, in which only literal punishments will be judged effective. In addition to Experiment 1 that tested participants' expectations as uninvolved third-party observers, in two separate experiments we also manipulated perspective-Experiments S1 put the participants in the role of the punisher and Experiment S2 put them in the role of the recipient. Results were qualitatively similar across changes in perspective. As a result, we focus on (third-party) observer perspective results below and report results from Experiments S1 & S2 in the Supplementary Information.

2.1. Method

2.1.1. Participants

Participants were recruited using Amazon's Mechanical Turk. 2000 people completed the study and we retained 1953 after excluding those who failed our attention checks (2.4% excluded). 56% of the sample was female and the modal age range was 31–40 years. Participants were paid $0.33.

2.1.2. Design & materials

Eight everyday scenarios were created. In each scenario, one agent committed a norm violation (for instance, leaving her dirty dishes about) and another agent responded to the violation by punishing the transgressor (for instance, by stacking the dirty dishes on her pillow). The punishment always involved leaving some "item" for the transgressor. The item was either beneficial for the transgressor or costly (thus, "figurative" or "literal") and was either contextually related to their transgression or not (thus, "informative" or "uninformative"), leading to a total set of 8 × 2 × 2 = 32 vignettes (see Appendix A in the Supplementary Information for the list of scenarios used). Participants were randomly assigned to read one scenario.

In order to develop these stimuli, we created a set of candidate "items" (e.g., dirty dishes, or a brand-new sponge and dish soap) and pretested these on an independent sample of participants (N = 103). They rated how they would feel upon receiving each item (from terrible (0) to delighted (10)). We did not provide any context regarding a transgression—thus, these participants were unlikely to interpret any of the items as a form of punishment. Only items with mean negative (<5) or mean positive ratings (>5) were used in the final vignettes. To further control for variance caused by the items themselves, each pair of items was used twice: once in an "informative" condition (where they were semantically related to the transgression) and once again in an "uninformative" condition for another vignette context where they were semantically unrelated to the transgression (see Appendix A).

Following the presentation of a single vignette, all participants answered a series of questions in the order presented here. Participants first rated the interpretability of the punishment: "*Will Sandra get the message that she needs to start cleaning her dirty dishes?*" (0 = definitely not; 10 = definitely yes). Next, they judged the effectiveness of the punishment: "*How likely is Sandra to start cleaning her dirty dishes?*" (0 = very unlikely; 10 = very likely). Following this, they rated their perception of how the recipient of the punishment would feel: "*When Sandra sees what her roommates have left for her, how will she feel?*" (0 = terrible; 10 = delighted). Finally, they judged if the punishment would feel like a punishment to the recipient: "*Would this feel like a punishment to Sandra?*" The last rating was made on a binary scale with labels 'definitely yes' and 'definitely no'. Each scenario included one comprehension question and participants failing to answer this question correctly were excluded from all analyses.

2.1.3. Statistical methods

To analyze participant responses, we fit three linear mixed effects models and one generalized linear fixed effects model (Gelman & Hill, 2007; Jaeger, 2008) using the lme4 package in the R statistical computing environment (Bates, Mächler, Bolker, & Walker, 2014; R Core Team, 2019).

All the models were run with effect-codes for cost imposed (*costless* = −0.5, *costly* = 0.5), informativeness of the punishment (*uninformative* = −0.5, *informative* = 0.5) and their interactions. Following our pre-registered analysis, the analysis reported below has a random intercept only for the stimuli (DV ~ cost imposed * informativeness + (1| item), data = data). In the Supplementary Information, we present additional analysis with a full random effect structure (Barr et al., 2013) along with our rationale for choosing the analysis presented below (Bates et al., 2015) . The results from both sets of analysis (the one presented below and the maximal random effect model) converge on all qualitative findings.

2.2. Results

We first looked at participants' expectations about each punishment and how clearly it would allow the recipient to 'get the message' (Fig. 2 (a)). We found a main effect of cost imposed, $b = 0.50$, $t_{1942.03} = 3.90$, $p < .001$, a main effect of informativeness, $b = 3.77$, $t_{1942.17} = 29.53$, $p < .001$, and a significant cost × informativeness interaction, $b = -0.95$, $t_{1942.15} = -3.72$, $p < .001$. To interpret this predicted interaction, we carried out planned pairwise contrasts using the emmeans package (Lenth, 2020). When the punishment was informative and communicative, the cost it imposed did not matter, as participants rated the recipient to be equally likely to get the message from costly, literal punishment (e.g., dirty dishes; $M = 5.87$, $SE = 0.20$, CI: [5.44, 6.31]) as from a costless, figurative punishment (e.g., new sponge and dish-soap; $M = 5.85$, $SE = 0.20$, CI: [5.42, 6.28]), $t(1942) = -0.13$, $p = .90$. However, when the punishment was uninformative, participants judged the recipient to be more likely to get the message from a costly punishment (e.g., mail thrown around messily; $M = 2.58$, $SE = 0.20$, CI: [2.15, 3.01]), compared with a costless punishment (e.g., mail stacked away neatly; $M = 1.61$, $SE = 0.20$, CI: [1.17, 2.04]), $t(1942) = -5.38$, $p < .001$.

Next, we looked at the perceived effectiveness of each punishment in bringing about a change in the recipient's future behavior (Fig. 2 (b)). The results revealed a main effect of cost, $b = 0.55$, $t_{1942.11} = 4.71$, $p < .001$, of informativeness, $b = 2.53$, $t_{1942.51} = 21.49$, $p < .001$, and a significant interaction between the two, $b = -0.75$, $t_{1942.44} = -3.16$, $p =$

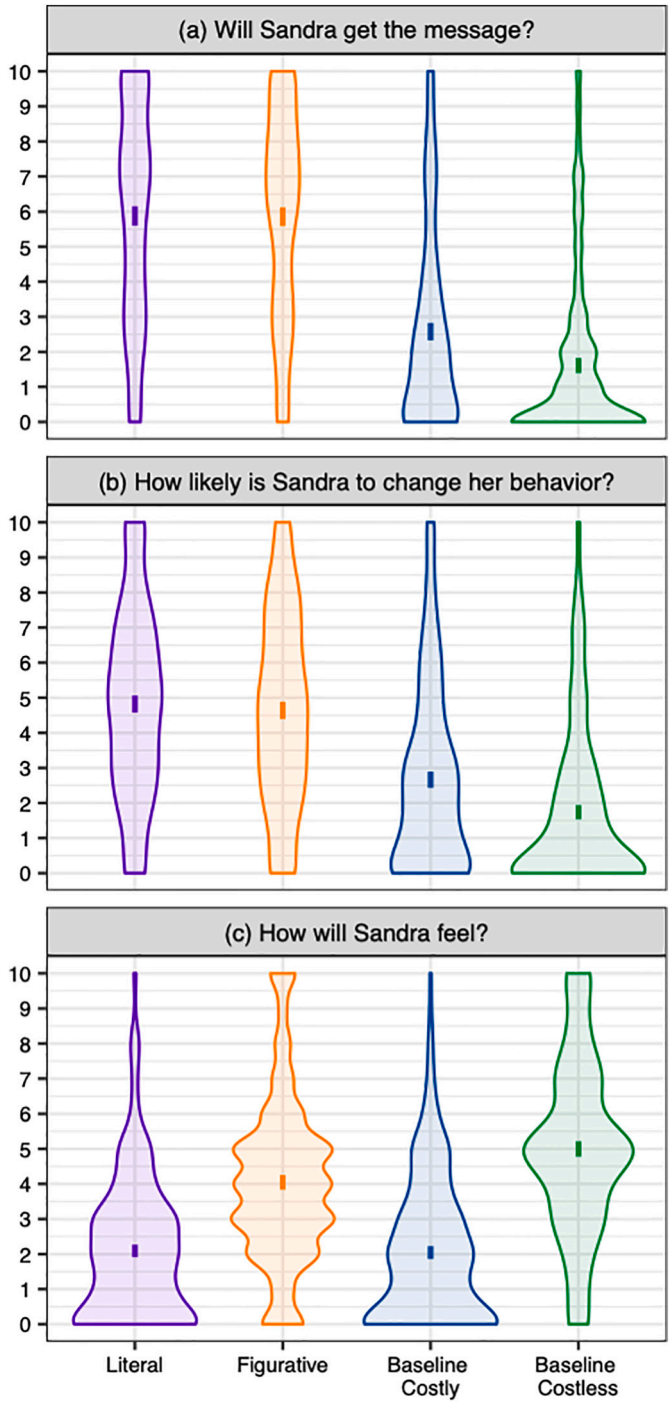


Fig. 2. Distribution of participant responses for three of the four dependent variables in Experiment 1. Lines inside the violins represent 95% CIs of means.

.002. Planned pairwise contrasts showed that when punishments were informative, costless punishments (M = 4.64, SE = 0.14, CI: [4.36, 4.92]) were perceived to be as effective in changing behavior as costly punishment (M = 4.82, SE = 0.14, CI: [4.54, 5.10]), $t(1942) = -1.09$, p = .27. However, when punishments were uninformative, the cost imposed made a difference as costly punishments (M = 2.66, SE = 0.14, CI: [2.39, 2.94]) were perceived to be more effective in changing behavior than costless punishments (M = 1.74, SE = 0.14, CI: [1.46, 2.02]), $t(1942) = -5.56$, $p < .001$.

Participants' judgments regarding the recipient's feelings upon receiving the punishments also showed a main effect of cost imposed, $b = -2.45$, $t_{1942.03} = -24.47$, $p < .001$, of informativeness, $b = -0.46$, $t_{1942.14} = -4.60$, $p < .001$, and a significant interaction, $b = 1.00$, $t_{1942.12} = 4.97$, $p < .001$ (Fig. 2 (c)). Planned pairwise contrasts revealed that, when the punishment imposed a cost, participants expected the recipient to feel bad, irrespective of whether the punishment was informative ($M = 2.09$, $SE = 0.17$, CI: [1.71, 2.46]) or uninformative ($M = 2.05$, $SE = 0.17$, CI: [1.68, 2.42]), $t(1942) = -0.26$, $p = .80$. However, when the punishment was costless, participants judged the recipient to feel significantly worse (as evident by the lower mean score) when the punishment was informative and directly related to the transgression ($M = 4.04$, $SE = 0.17$, CI: [3.67, 4.41]) than when it was uninformative and unrelated to the transgression (M = 5.00, SE = 0.17, CI: [4.63, 5.37]), $t(1942) = 6.75$, $p < .001$.

Finally, we examined ratings of whether the act in question would feel like a punishment (results are on the log odds ratio scale). We found a main effect of cost imposed, $b = 1.48$, $SE = 0.10$, $z = 14.70$, $p < .001$, informativeness, $b = 0.91$, $SE = 0.10$, $z = 9.05$, $p < .001$, and a significant interaction, $b = -0.50$, $SE = 0.20$, $z = -2.45$, $p = .01$ (see Fig. 3). Planned contrast revealed that, when the punishment was costly, informative punishments (M = 1.08, SE = 0.15, CI: [0.79, 1.36]) were judged to feel more like punishments than uninformative punishments ($M = 0.41$, $SE = 0.14$, CI: [0.14, 0.68]), $z = -4.76$, $p < .001$. A similar pattern of results was found for costless punishments, as informative punishments were judged to feel more like punishments ($M = -0.16$, $SE = 0.14$, CI: [−0.42, 0.11]) than uninformative punishments ($M = -1.32$, $SE = 0.15$, CI: [−1.61, −1.02]), $z = -8.00$, $p < .001$.

2.3. Discussion

What do people think is required in order for a punishment to be effective? We find that people think a punishment must be "informative"—i.e., sufficiently semantically related to the offense that its communicative intent is apparent. In contrast, we find that people do not think the punishment needs to impose a cost. Rather, participants in our experiment expected costless, benefit-imposing (figurative) punishment to bring about future behavioral change in a recipient, and in themselves when they are described as the recipients (Experiment S2). Although figurative punishments do not impose a literal cost, people think that the recipients of (informative) figurative punishments, including themselves, will *feel* bad, presumably because they will understand the underlying communicative intent. The same costless, beneficial punishments, however, were not judged to be effective

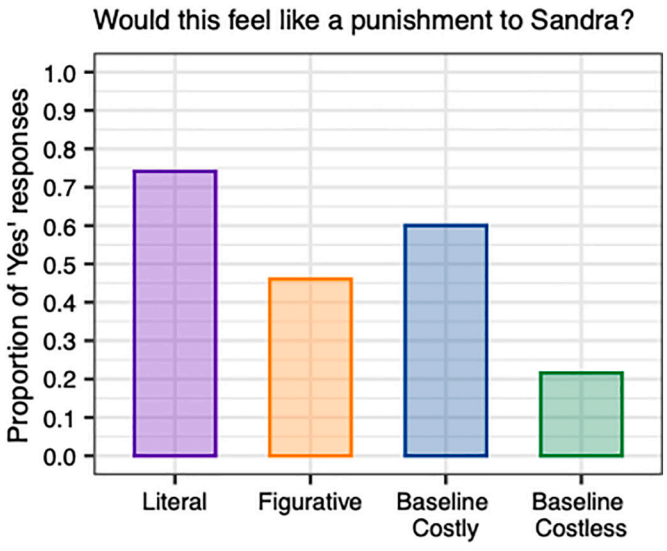


Fig. 3. Proportion of participants who think the action would feel like a punishment to the target.

sanctions if they were not semantically related to the transgression, and therefore uninformative. These inferences remain stable across changes in perspective – participants judge figurative punishments to be effective punishments as uninvolved third-party observers (Experiment 1), as punishers themselves (Experiment S1) and even as recipients of these punishments (Experiment S2). Taken together, these results suggest that the psychology underlying punishments is better understood as organized around principles of effective communication rather than principles of reinforcement-based learning.

While Experiment 1 investigates whether people can make accurate inferences from costless, figurative punishments, it does not tell us what kinds of punishment people would prefer to actually employ. This is our focus in Experiment 2.

3. Experiment 2

In Experiment 2a and Experiment 2b we investigated the kinds of punishments people voluntarily select and the role context plays in their selection.

3.1. Experiment 2a

We presented participants with the same set of vignettes and candidate punishments used in Experiment 1. We then asked them to indicate which punishment they would choose to perform, presenting them with four candidates in a 2 (costly vs. costless) × 2 (informative vs. uninformative) design. Our main goal was to see whether participants would select figurative punishments in the presence of literal alternatives. For instance, we wanted to see whether in Alice's position, they might prefer to leave a clean sponge and dish soap on Sandra's pillow rather than leave Sandra's dirty dishes on her pillow.

3.1.1. Method

3.1.1.1. Participants. Based on a power analysis reported in our preregistration, 88 participants were recruited using Amazon's Mechanical Turk and paid $0.25 for their participation. After excluding those who failed to pass the comprehension check, the final sample was made up of 81 participants (8.0% excluded; 56% female, modal age range: 31–40 years).

3.1.1.2. Design & materials. The eight base scenarios used in Experiment 1 were employed. As in Experiment 1, each scenario included a norm violating agent. Unlike Experiment 1, however, participants were asked to imagine that they were on the receiving end of the transgression and were asked how they would respond to the norm-violating agent. Participants were presented with four potential responses that corresponded to the four kinds of punishments used in Experiment 1: literal, figurative, baseline costly, and baseline costless. They were asked to select one of the four and then give reasons for their selection. Each participant was randomly assigned to read one vignette.

3.1.2. Results

Collapsing across different contexts, a chi-square test of goodness-of-fit was performed to determine whether the four kinds of punishments were preferred equally. We found evidence in favor of the alternative hypothesis X^2 (3) = 114.51, $p < .001$, $V_{Cramer} = 0.69$, *CI*: [0.58, 0.81], $n = 81$. Descriptive analysis revealed that overall, figurative punishments were selected most frequently (75%, *CI*: [67%, 84%]), followed by literal punishments (19%, *CI*: [10%, 27%]), baseline costly punishments (4%, *CI*: [0%, 13%]) and baseline costless punishments (2%, *CI*: [0%, 11%]) (see Fig. 4). This preference ordering held in every context (except for one, see Table 1). Chi-square tests conducted simultaneously within each scenario, using the ggstatsplot package (Patil, 2018), provided evidence for the alternative hypothesis and a preference for the figurative punishment for all but one context (see Table 1).

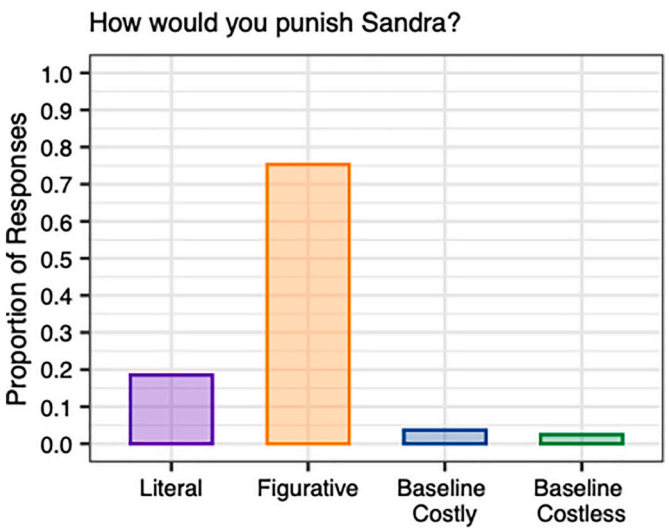


Fig. 4. Proportions reflecting which punishment participants would choose themselves as punishers.

3.2. Experiment 2b

Experiment 2a revealed that people prefer costless yet communicative punishments to costly, typical ones. This preference was consistent across the various scenarios we used. Yet, the only decision people could make in Experiment 2a was the kind of punishment they could impose. In contrast, in real life a frequent first step towards confronting norm-violating behavior is to talk directly with the transgressor. Moreover, though people displayed a consistent preference for figurative punishments in the cases we employ, we nevertheless suspected that some features of a context (punisher-transgressor relationship, frequency of the offense, etc.) would influence when figurative punishments are preferred over literal ones. In Experiment 2b we addressed both of these concerns. We manipulated the frequency of previously communicating the offense to see whether this impacted the response our participants chose. We also included the option to 'talk' to the perpetrator. In two additional experiments reported in the Supplementary Information, we tested whether the nature of relationship between the punisher and transgressor (friend vs strangers in Experiment S3a and landlord vs tenant in Experiment S3b) affected which punishment is selected.

3.2.1. Method

3.2.1.1. Participants. We recruited 161 participants on Amazon's Mechanical Turk in exchange for $0.25 for their participation. We excluded those who failed our attention check, leaving 158 participants in the final sample (1.9% excluded; 57% female, modal age range: 31–40 years).

3.2.1.2. Design & materials. We modified one of the vignettes used in Experiment 2a. Participants read about the norm-violating agent Sandra, who had not been doing her dishes. Then in a between-subjects design they read that Sandra's roommates had either asked her to clean up on three previous occasions or had never brought up the dirty dishes with her before. In other words, they had either responded to the norm-violating behavior many times or this was their first time addressing it. After reading the vignette we asked our participants how they would respond to Sandra. We gave them a choice between talking to her one-on-one, leaving a brand-new dish soap and kitchen sponge on her bed, or leaving a stack of her dirty dishes on her bed. Unlike Experiment 2a, we removed the two baseline conditions and focused

Table 1
Choice decision broken down by scenario.

Context	N	Percent selected				Chi-square statistic	p.value	df
		Literal	Figurative	Baseline costly	Baseline costless			
Sweaty clothes	8	12.50%	75%	12.50%	0	11	1.17e-2	3
Stationary	9	22.22%	66.67%	0	11.11%	9.22	2.65e-2	3
Loud music	12	8.33%	83.33%	0	8.33%	22	6.52e-5	3
Messy mail	9	0	100%	0	0	27	5.89e-6	3
Laundry	12	8.33%	83.33%	8.33%	0	22	6.52e-6	3
Hair	11	18.18%	72.73%	9.09%	0	14.1	2.78e-3	3
Fridge food	10	20.00%	80%	0	0	17.2	6.43e-4	3
Dirty dishes	10	60%	40%	0	0	10.8	1.29e-2	3

only on figurative and literal punishments. We additionally included the opportunity for the punisher to talk to the perpetrator.[3] Our motivation behind this was two-fold: (i) it better mimicked the kinds of options we have in our daily life, enhancing the ecological validity of our experiment, (ii) it allowed for a more stringent test of the usefulness of figurative punishment (will people ever select to punish figuratively when the alternatives are to talk or punish literally?).

3.2.2. Results

We ran a chi-square test of goodness-of-fit and found that the proportion of choice differed across the between-subjects condition X^2 (2) = 12.03, p = .002, V_{Cramer} = 0.25, *CI*: [0.08, 0.41], n = 158 (also see Fig. 5). When participants were in the condition where they had never before brought up the norm-violating behavior, 81% (*CI*: [73%, 89%]) wanted to talk to the perpetrator, 17% (*CI*: [9%, 25%]) wanted to punish her figuratively, and 3% (*CI*: [0%, 11%]) wanted to punish her literally. In contrast, when they were in the condition where they had previously spoken about the bad behavior before 60% (*CI*: [50%, 71%]) wanted to talk, 40% (*CI*: [30%, 51%]) chose to punish figuratively, and no one (0%, *CI*: [0%, 11%]) chose to punish literally.

3.3. Discussion

In Experiment 2a we presented our participants with four different kinds of punishments and asked them how they would punish the perpetrator if they were the punisher. People displayed a clear preference for costless yet communicative punishments (i.e., figurative punishments), even in the presence of costly alternatives (i.e., literal punishments). This result is difficult to reconcile with a model where the effectiveness of punishment is derived directly from its imposed incentive. However, the result is consistent with the hypothesis that human punishment is organized around communicative principles. Not only do people understand the logic of figurative punishments, they considered them to be viable and preferred alternatives to typical, cost-imposing punishments. This is most clearly evident in Experiment 2b. When participants are faced with a norm-violating agent they typically prefer to begin by talking to them about their behavior. But if this fails, their next choice is to punish figuratively, not literally. Furthermore, this pattern of choice appears consistent across two different relational contexts as evidenced from Experiments S3a and S3b.

An alternative interpretation for this preference, however, is that our "figurative" punishments often involve objects that are functionally appropriate for solving the underlying norm violation–for instance, a sponge can be used to do the dishes your roommate is neglecting. Possibly, then, people prefer figurative punishment not for its communicative value but for its practical value. To test for this possibility, we ran another experiment (Experiment S4 in the Supplementary Information) which was identical to Experiment 2a in every regard except that the items used for figurative punishments could not be used practically to solve the problem. For instance, instead of leaving a brand-new kitchen sponge and dish soap on her roommate's pillow, Alice leaves a toy, plastic kitchen sponge and dish soap. Changing the items to non-usable, toy items did not change our results: across all vignettes people preferred to punish figuratively, suggesting that the preference is indeed tied to the communicative inference of the punishment.

Why might this be? We asked participants to explain their judgments, and these free responses offer some insight. Participants who chose figurative punishments did so because they judged them to communicate the transgression clearly, but also in a non-aggressive manner. They often referred to communicative goals: to let the offender know that their action was unacceptable, but in a manner that was kind, minimized the risk of future retaliation, and did not jeopardize their ongoing relationship. As one participant put it, they allowed the punisher to "kill with kindness".

4. Experiment 3

Experiments 1 and 2 present situations in which a figurative interpretation of costless punishment is strongly contextually favored. Experiment 3 offers a robustness check. In a situation where the context provides a salient alternative interpretation of a "costless punishment", will people still infer that it may be designed as a communicative form of sanction?

To test this, we modified the figurative punishment vignettes used in Experiments 1 and 2 so that there was a plausible explanation for the focal behavior (e.g., leaving a new sponge and dish soap on a roommate's pillow) other than the figurative punishment of a semantically related transgression. For instance, in the "dishes" vignette we told participants that the person who left these items on her roommates' pillow worked at a company that manufactured cleaning products. Additionally, rather than specifying that her roommate had not been doing her dishes, we simply referred obliquely to the fact that each roommate had various tasks to do around the house. We implemented these changes to make the intention underlying the focal behavior relatively ambiguous, and to eliminate task demands. We then asked participants to explain why the protagonist might have performed the focal behavior in question.

4.1. Method

4.1.1. Participants

608 people, recruited using Amazon's Mechanical Turk, completed the experiment in exchange for $0.25. After excluding those who failed to pass either of the two comprehension checks, the final sample was made up of 425 participants (30% excluded; 57% female, modal age range: 31–40 years).

[3] Participants in Experiments S3a and S3b had the same three options - to talk, punish literally, or punish figuratively. The main difference was in the relationship they shared with the perpetrator. In Experiment S3a participants were described as being either friends with the perpetrator or strangers. In Experiment 3b participants were either the landlord (and the perpetrator was their tenant) or they were the tenant (and the perpetrator was their landlord).

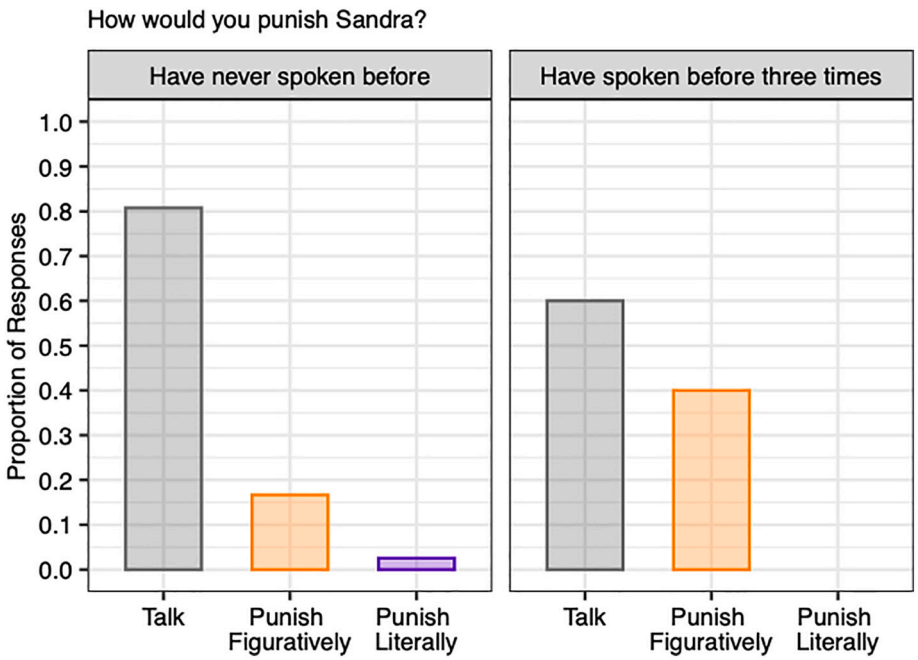


Fig. 5. Proportions reflecting which punishment participants would choose themselves as punishers, broken down by the two experimental conditions.

4.1.2. Design & materials

Our goal in this experiment was to determine how readily people will infer punitive communicative intent in ambiguous cases. That is, how strongly do we infer that punishment is communicative, such that we approach ambiguous acts with the possibility that they may be communicative sanctions? To this end, we presented participants with ambiguous acts that could be interpreted as figurative punishments but could also be interpreted as ordinary non-punitive acts. We asked them to explain the action in free response form in order to avoid experimenter demand.

These scenarios were modeled on those used in Experiments 1 and 2. Each scenario revolved around one agent leaving something positive (e.g., brand-new kitchen sponge) or negative (e.g., stack of dirty dishes) for another agent. Two critical changes rendered these acts ambiguous. First, we did not describe any norm violation (e.g., in the case of Sandra and Alice we mentioned that each roommate had some chores she was assigned to, but we mentioned nothing about Sandra having recently failed to do the dishes). Second, we introduced new information that provided an alternative explanation of the act (e.g., we told participants that Alice works at a company that makes sponges and dish soap). In each case, we asked our participants to explain why the item was left (see Appendix B for the list of scenarios used).

We developed a coding rubric to categorize participant responses. The rubric had 4 pre-defined categories. The first category was "Gift" explanations–i.e., those indicating a positive, gift-like inference about the item left (e.g., "*I think she did that to help Sandra out with her chores. She may have seen that the old sponge was dirty or needed to be replaced and wanted to do something nice to show appreciation for Sandra.*") The second category was "Disapproval" explanations—i.e., inferences that one agent was communicating disapproval of the second agent's norm-violation (e.g., "*I think that she left that as a subtle way of letting her know that the dishes need to be done.*"). The third category, "both", was used to classify those responses that explicitly mentioned both of the two previously mentioned interpretations (e.g., "*Sandra might have forgotten to do the dishes when she was supposed to, or Alice was showing her appreciation for Sandra's diligence in doing the dishes regularly by buying her some new cleaning products*"). Finally, the fourth category, "none", was used to classify responses that were either too vague to correctly classify (e.g., "*To make sure that Sandra found them, and that she noticed they were from Alice.*") or irrelevant (e.g., "*Perhaps she did it to go viral for her marketing campaign.*"). All coding was done by two coders (one of whom was the first author) independently. Our coders agreed on 399 of the 425 responses and obtained high inter-observer reliability (unweighted K = 0.90, 95% *CI*, 0.86 to 0.93). Disagreements regarding the remaining 26 responses were resolved by a third coder (the senior author), who independently reviewed the coding scheme and adjudicated between the divergent codes. The full list of scenarios can be found in Appendix B. The coding scheme can be found on OSF at https://osf.io/fxdwm/.

4.2. Results

For costly focal behaviors (e.g., leaving a stack of dirty dishes on roommate's pillow), 81% of responses were classified under "disapproval" (*CI*: [76%, 86%]), 15% were classified under "none" (*CI*: [10%, 20%]), 3% were classified under "gift" (*CI*: [0%, 9%]), and fewer than 1% were classified under "both" (*CI*: [0%, 6%]). Also see Fig. 6.

For costless focal behaviors (e.g., leaving a brand-new kitchen sponge and dish soap on roommates' pillow), 39% of the responses were classified under "disapproval" (*CI*: [33%, 47%]), 38% were classified under "gift" category (*CI*: [31%, 45%]), 18% were classified under "both" (*CI*: [12%, 26%]), and 5% were classified under "none" (*CI*: [0%, 12%]).

4.3. Discussion

In Experiment 3, we intentionally designed ambiguous cases in which a figurative punishment could instead be naturally interpreted instead as a genuine gift. In the presence of this plausible alternative would people even entertain the possible interpretation of it being a figurative punishment? In line with our model, we found that 39% of the participants solely inferred a punitive communicative intent while an additional 18% entertained the possibility that the ambiguous act was a sanction. Although none of our vignettes mentioned the violation of a norm (eliminating a possible task demand), and all of them explicitly provided participants with a credible basis to infer that "costless" items were intended as a gift, the most frequent inference (for participants in the costless condition) was that these items were in fact designed as a form of punishment, closely followed by the inference that the item was

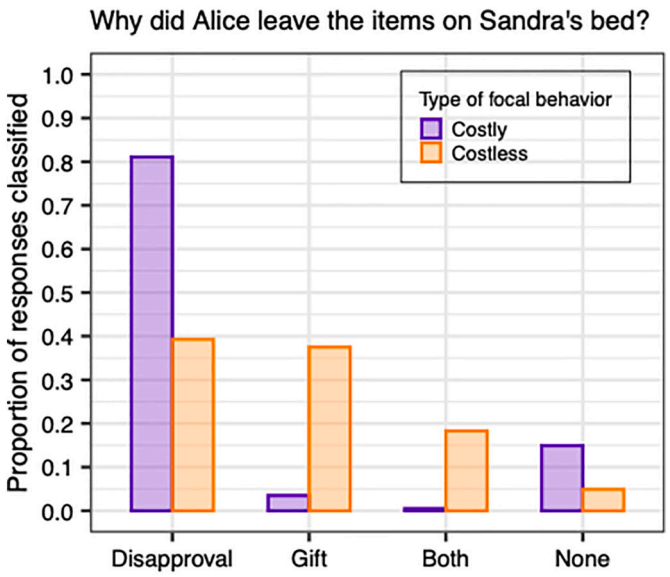


Fig. 6. Proportion of participants free-form responses classified into one of the four predetermined categories.

a gift. Unsurprisingly, this inference was even more common when items were costly to the recipient (e.g., dirty dishes on her pillow), foreclosing any plausible inference that they were intended as a gift. These results from the ambiguous, costless condition therefore indicate that people are highly attuned to seeking and inferring the communicative intent behind punishments.

5. Experiment 4

Our overarching goal is to understand whether punishment between humans is effective because it works as a constructed incentive or because it additionally acts as a signal of communication. Experiment 4 approaches this question from a different and complementary angle. Experiments 1, 2, and 3 ask whether benign acts can be useful punishments if they strongly convey communicative intent. In Experiment 4 we instead ask whether harmful acts can be ineffective punishments when the transparency of communicative intent is compromised by plausible alternative motives. We generate ambiguity by considering cases of "profitable punishment", in which the act of punishment imposes a cost on the person being punished, but also generates a benefit for the person doing the punishing. Such punishments impose a direct, apparent incentive, but any intention to influence behavior through incentives is "explained away" by the profits they generate for the punisher. Specifically, each scenario in Experiment 4 described three agents: a perpetrator, a victim, and an unaffected third-party. In each case, the perpetrator committed a norm-violation against the victim and the unaffected third-party punished the perpetrator. In half of the cases, the punishment benefitted the victim. In the other half of the cases, the punishment benefitted the punisher (i.e., the unaffected third-party). In both cases, the cost imposed on the perpetrator remained exactly the same. Here is an example:

> *Alice, Becky, and Cassandra work together at an office. They all work in the same division and have office desks close to one another. Alice and Becky both bring home cooked lunches to work. They both store their food in the common office fridge, along with all of the other employees.*
>
> *Cassandra does not bring any lunch to work. Initially, she used to buy her lunch from the office cafeteria. However, for some time now, she has been eating Becky's lunch. Everyday just before lunch time, Cassandra takes out Becky's lunch from the common fridge and eats it without asking Becky for permission. This has been going on for quite some time and Cassandra has eaten many of Becky's lunches, without buying her any lunch in return or paying her back in any way. At first Becky didn't know who was behind her missing lunch but recently, Becky found out that the culprit is Cassandra. She has tried to speak to Cassandra about it. But Cassandra continues to eat Becky's lunch without asking her and without paying her for the food. Alice is aware of the whole situation and decides to do something about it. Alice uses Cassandra's credit card to buy a month's worth of groceries for Becky (or herself). She gets all of the groceries delivered to Becky (or herself) at her home and leaves the receipt on Cassandra's desk.*

In both scenarios, Alice uses Cassandra's money to buy groceries; the key manipulation is whether these groceries are for Becky (Cassandra's victim) or Alice herself (a bystander). The two cases are therefore identical in the cost they impose on the perpetrator, Cassandra. They differ, however, in degree of ambiguity regarding Alice's motives. When the punishment benefits the victim (Becky), its motive is unambiguous: to dissuade Cassandra from future transgression. In contrast, when the punishment actually benefits the punisher herself, its motive is ambiguous: it may be to dissuade Cassandra from future transgression, but it may be simple self-interest.

On the simple constructed incentive model both of these punishments should be equally effective: each imposes an identical cost. In contrast, if punishment is structured to be interpreted in light of the punisher's apparent communicative intent then "profitable punishment" by third parties should be judged ineffective, at least relative to punishment that instead compensates the victim.

5.1. Method

5.1.1. Participants

437 people participated in the experiment. Of these, 418 completed the study and 334 (24% excluded; 55% women, modal age range was 31–40 years) passed all of the three comprehension checks. The participants were recruited through Amazon's Mechanical Turk and paid $0.35 for their participation.

5.1.2. Design & materials

We created four new sets of scenarios (see Appendix C for the full list). Each scenario described 3 agents: a perpetrator, a victim, and an unaffected third-party. In each case, the perpetrator transgresses against the victim and gets punished by the third-party. The punishment imposed is always costly for the perpetrator and it directly benefits either the second-party (victim) or the third-party (bystander), leading to a total of 8 vignettes (scenarios (4) x beneficiary (2)).

Participants were randomly assigned to read one of the vignettes, after which we asked them 2 questions, presented in the order mentioned here: *"Will Cassandra get the message that she needs to stop eating Becky's lunch?"* (0 = definitely not; 10 = definitely yes), and "*How likely is Cassandra to stop eating Becky's lunch?*" (0 = very unlikely; 10 = very likely). Each scenario included three comprehension questions and participants failing to answer any of the three correctly were excluded from all analyses.

5.2. Results

To analyze participant responses, we fit two linear mixed effects models (Gelman & Hill, 2007) using the lme4 package in the R statistical computing environment (R Core Team, 2019; Bates et al., 2014). For both models, there was a single dummy-coded fixed effect for beneficiary (*second party (victim)* = 0, *third party* (*bystander*) = 1), and random intercepts for the different scenarios.

We first analyzed participant responses to the clarity of each punishment. We found a main effect of beneficiary, $b = -1.12$, $t_{329.60} = -3.09$, $p = .002$. This means that when the punisher stood to profit personally from the punishment, participants judged the punishment to

be less clear and communicative to the recipient ($M = 4.96$, $SE = 0.26$) than when the exact punishment benefitted the victim of the transgression ($M = 6.15$, $SE = 0.26$). Also see Fig. 7(a).

Results for the perceived effectiveness of the punishment were similar (see Fig. 7(b)). There was a main effect of the beneficiary, $b = -0.76$, $t_{329.45} = -2.20$, $p = .03$, such that, punishments benefitting the unaffected third-party punisher were perceived to be significantly less effective ($M = 4.67$, $SE = 0.26$) in bringing about a change in future behavior than those benefiting the affected second-party victim ($M = 5.52$, $SE = 0.25$).

5.3. Discussion

Findings from Experiment 4 demonstrate that people expect profitable punishments to be relatively less effective at changing a perpetrator's behavior than traditional costly punishments. This is curious because the relevant "profit" and "costs" do not affect the perpetrator at all—rather, the question is whether the punishment generates profits or costs for the *punisher*. We can readily make sense of these data, however, by observing that the motives underlying profitable punishment are relatively less ambiguous than those underlying costly punishment. These results lend further support to the hypothesis that punishment is not best understood as a simple constructed incentive, but instead as a joint communicative and interpretive action.

6. General discussion

Humans are adept at using punishment to modify each other's behavior. We ask how this works. Many current approaches model punishment as a direct imposition of cost, and model learning from punishment as a simple association between the transgression and the imposition of cost. In contrast, our research suggests that people learn from punishment by inferring the punisher's communicative intent, and that punishers construct punishments in expectation of these inferences.

We show that people expect costless, yet communicative, punishments to be as effective as typical, punitive sanctions (Experiment 1). This expectation holds across changes in perspective: people endorse the effectiveness of communicative punishments as first-party punishers (Experiment S1), as second-party recipients of the punishments (Experiment S2), and as third-party observers (Experiment 1). We find that across a range of contexts and relationships (Experiments S3a and S3b) people prefer costless punishments over more canonical cost-imposing punishments, as long as the former clearly communicate the punisher's disapproval (Experiment 2a, Experiment 2b, and Experiment S4). These inferences about communicative intent occur spontaneously and even in the absence of apparent contextual cues (Experiment 3). Finally, even a punishment that imposes clear and direct costs is expected to be less effective if there is ambiguity regarding the punisher's intent (Experiment 4).

Our data thus favors a model of punishment according to which social punishment is structured not merely as an incentive, but also as a communicative act. These two dimensions of punishment are not, of course, mutually exclusive. Rather, we assume that (in addition to communicating the punisher's disapproval) an important part of what gets communicated is an implied incentive. The reason that a literal gift can act as a figurative punishment, and the reason that people expect that it will be effective, is because it is interpreted as a sanction and a threat.

Inferring and conveying communicative intent based on observable actions relies on our ability to model other people's minds (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Wellman, 1992; Young & Saxe, 2009). In this regard it is similar to inferring and conveying the communicative intent behind language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Grice, 1957; Kao et al., 2014; Kao & Goodman, 2015) or demonstrative action (Ho et al., 2016; Shafto et al., 2014). In order to understand instances of figurative acts as punishments, both recipients and observers must model the relationship between an actor's various possible intentions and the actions she performs, as well as the inferences that the recipient can draw based on this model.

This may involve several layers of embedded mental state inference: i.e., the punisher models the target modeling the punisher's mind. Take, for instance, the case of Alice leaving a new sponge for Sandra as "punishment". What is required for Alice to believe that Sandra will learn a lesson from this, and experience it as a punishment (as we show that participants do)? Alice must model Sandra performing an inference of communicative intent—i.e., extracting the punitive message from this act. And what is required for Sandra to perform this inference of communicative intent? Sandra must model the process by which Alice selects informative actions. Connecting these models, Alice is (choosing

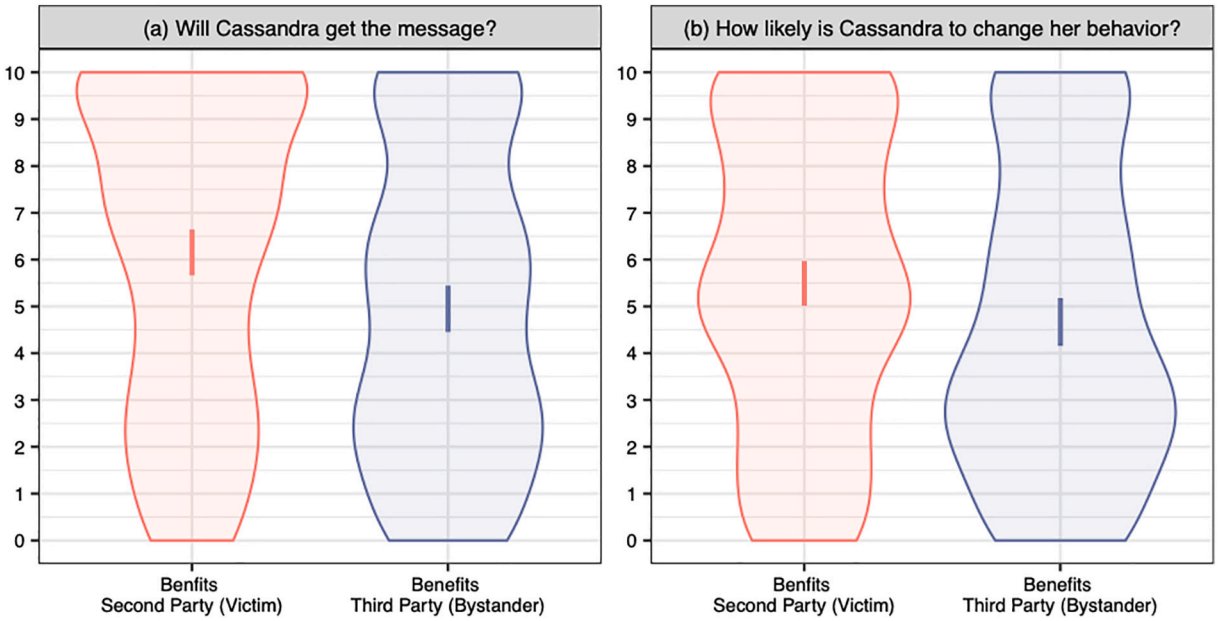


Fig. 7. Participant response densities and 95% CIs for (a) likelihood of 'getting the message' and (b) likelihood of changing future behavior.

an action by) modeling Sandra (inferring intent by) modeling Alice's choice of action. In this way, sending and receiving the kinds of communicative signals that we demonstrate here may involve higher-order theory of mind. This process is a form of pragmatic, inference-based communication (Grice, 1957; Scott-Phillips, 2015; Sperber & Wilson, 1986) that has been formalized for language (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Kao et al., 2014; Kao & Goodman, 2015) and demonstration (Ho et al., 2016; Shafto et al., 2014). Pragmatic inference grounded in natural language has been widely explored; our extension to a "pragmatics" grounded in reward and punishment indicates how the key ideas can be extended to many other non-linguistic human behaviors. An exciting direction for future research is to explore how people make rich social inferences from relatively sparse social observations of many different kinds (see, e.g., Lopez-Brau & Jara-Ettinger, 2020; Ho et al., 2019a, 2019b).

More broadly, by investigating the "pragmatics of reward and punishment", the findings presented here provide a valuable starting point for drawing out further connections between moral psychology and linguistic pragmatics. For example, linguists have extensively studied how polite speech plays a key role in regulating interpersonal relationships (Brown & Levinson, 1987) and how it requires computing and interpreting tradeoffs between communicative and social utilities (Yoon, Tessler, Goodman, & Frank, 2016). Like polite speech, punishment is a mechanism by which people regulate and influence others' behaviors, and, as we argue in this paper, it can exploit inferences about communicative intent to send a message through minimally socially offensive means. Relatedly, figurative punishment provides an interesting case of indirect, non-verbal communication with parallels to "off-the-record", indirect speech such as bribing a police officer by asking "Could we deal with this ticket without going through a lot of paperwork?" (Lee & Pinker, 2010). Both of these kinds of phenomena involve communicating while delicately navigating conflicting social and moral goals, and, as such, rely on context-sensitive inferences about intentions to be successful. Future work will need to explore these rich interactions between action, communication, and morality.

Of course, there may be many instances of "literal" punishment that do not require extensive mentalizing, just as the interpretation of much literal language does not. (If you reach to grab a French fry from your partner's plate and they slap your hand away, little mentalizing is required for effective learning.) In practice, we should often be able learn from punishment by merely tracking its "literal" incentives. Reliance on "literal" punishment may be especially common in situations where there are established norms of behavior and punishment, such as in institutional settings. Here, there are important reasons to adopt clear, consistent, and predictable sanctions for similar transgressions across individuals and across time. This may be best accomplished by direct and "proportional" sanctions for transgressions. On the other hand, mentalizing may be more essential to identifying and responding to social sanctions in less institutionalized settings—for instance, among friends and other peer groups. In these settings (as in Experiments 2a and 2b) we may often feel the imperative of sending a message without imposing an actual cost, perhaps even with plausible deniability (Lee & Pinker, 2010; Pinker, Nowak, & Lee, 2008). One reason for this could be the desire to preserve ongoing relationships. However, results from Experiments S3a and S3b fail to find evidence for increased use of figurative punishment in the context of especially close relationships. Instead, we find that people use figurative punishments with the same frequency irrespective of the relationship they share with the perpetrator (friend vs stranger; landlord vs tenant). Possibly this reflects a uniformly strong desire to preserve relationships regardless of their closeness, or possibly other motives explain the preference for figurative punishment. Resolving these possibilities is an important direction for future research. Lastly, insofar as the ability to employ and infer figurative communication is a specialized human ability, requiring higher order theory of mind, this may explain why sophisticated and subtle sanctioning systems of norms are more prevalent in humans than other species.

Why would people ever rely on actions like punishment to communicate when they could simply use language instead? In other words, why leave your roommate a sponge and dish-soap when we can much more easily just tell them, "Do your dishes!"? We consider several complementary explanations. First, action is generally more costly than speech, and thus can serve as a more 'honest' signal of an agent's intentions and feelings (Baumeister, Vohs, & Funder, 2007; Nock, 2008). Second, communicating through sanctions may also allow punishers to circumvent potentially awkward and uncomfortable spoken interaction. Third, communicative acts may simply reinforce the contents of speech (as we see in Experiment 2b), and vice versa—two channels of communication are sometimes better than one. Finally, our ability to communicate with punishment may be evolutionarily conserved from an ancestor that did not possess natural language.

We find that figurative punishments are anticipated to be effective, and sometimes preferred. They can avoid imposing direct costs, and we know from other recent research that people often seek to enhance the communicative value and reduce the costs imposed by their punitive acts (Molnar et al., 2020). Why, then, do people also sometimes make punishments very harmful to the target (Rai, Valdesolo, & Graham, 2017), preferring direct, immediate incentives over figurative acts? First, as Experiment 3 suggests, in ambiguous situations actual costs support more accurate inference. Second, actual costs do not merely imply, but rather demonstrate, the punisher's willingness to incentivize their desired pattern of behavior. Third, imposing certain kinds of costs—for instance, incarceration—may incapacitate the wrongdoer, or accomplish the ancillary goal of harming them for competitive reasons (Raihani & Bshary, 2019). Finally, as noted above, institutionalized punishment requires procedural justice and impartiality, treating violations consistently across people and across time. This may favor actual costs over carefully constructed veiled threats.

Finally, how can we reconcile the apparent role of communication in punishment with evidence that people punish in one-shot, anonymous settings (Balafoutas et al., 2014; Fehr & Gächter, 2002; Nadelhoffer, Heshmati, Kaplan, & Nichols, 2013)? We adopt two explanations that have been well explored in the literature. First, punishment may play a communicative role that does not benefit the punisher specifically, but rather serves the victim or the community at large (Fehr & Gächter, 2002; Balafoutas et al., 2014; Boyd, Gintis, Bowles, & Richerson, 2003; Boyd & Richerson, 1992; Gürerk, Irlenbusch, & Rockenbach, 2006; Henrich & Boyd, 2001)—i.e., as a form of general deterrence. Second, cognitive mechanisms supporting punishment may be adapted (by learning, or by cultural, or biological evolution) to contexts that are repeated and non-anonymous, and these mechanisms may over-extend to one-shot anonymous settings where they cannot serve those same adaptive ends (Delton, Krasnow, Cosmides, & Tooby, 2011; Jordan & Rand, 2020; Krasnow, Delton, Cosmides, & Tooby, 2016; Rand et al., 2014).

In sum, our work suggests re-examining a widely used, standard model of punishment. Together with other recent findings (Cushman et al., 2019; Funk et al., 2014; Ho et al., 2019a, 2019b; Ho, MacGlashan, Littman, & Cushman, 2017; Molnar et al., 2020) it suggests that punishment is structured not only to modify the behavior of social partners by its incentive value, but also to be understood and interpreted as a form of communication. Identifying the levers of what makes punishments effective will lead to better insight into the structure and function of punishment, as well as illuminate how punishment can be leveraged to facilitate the learning of social and moral norms in humans.

Author contributions

All authors contributed to the conceptions of experiments. AS designed, conducted, and analyzed the experiments. AS wrote the paper with assistance from the other authors.

Declaration of Competing Interest

The authors declare no competing financial interests.

Acknowledgments

We thank Tomer Ullman, Adam Bear, Indrajeet Patil, Regan Bernhard, and members of the Moral Psychology Research Laboratory for their advice and assistance. This research was supported by grant 61061 from the John Templeton Foundation to FAC.

Appendix. Supplementary data

Supplementary information (additional data and stimuli) to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104544.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 1–10.
Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.
Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters, 122*(2), 308–310.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.
Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint. arXiv:1506.04967.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint. arXiv:1406.5823.
Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403.
Bentham, J., & Bowring, J. (1962). *Principles of penal law, works. 1* p. 398). New York: Russell & Russell.
Boon, S. D., Deveau, V. L., & Alibhai, A. M. (2009). Payback: The parameters of revenge in romantic relationships. *Journal of Social and Personal Relationships, 26*, 747–768.
Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531–3535.
Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 195*, 171–195.
Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge Univ. Press.
Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437–451.
Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.
Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology, 95*(6), 1316–1324.
Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature, 373*, 209–216.
Colston, H. L., & O'Brien, J. (2000). Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse Processes, 30*(2), 179–199.
Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass, 10*(2), 117–133.
Cushman, F. A., Sarin, A., & Ho, M. K. (2019, December 11). *Punishment as communication*. https://doi.org/10.31234/osf.io/wf3tz.
Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*(2), 185–196.
Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108*(32), 13335–13340.
Durkheim, E. (2014). *The division of labor in society*. Simon and Schuster (Original work published in 1893).
Ewing, A. C. (1943). Punishment as viewed by the philosopher. *Can. B. Rev., 21*, 102.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.
Feinberg, J. (1965). The expressive function of punishment. *The Monist*, 397–423.
Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998.
Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40*(8), 986–997.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.
Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45*, 840–844.
Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41*, 364–374.
Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science, 5*(1), 173–184.
Grice, H. P. (1957). Meaning. *The Philosophical Review*, 377–388.
Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science, 312*(5770), 108–111.
Hampton, J. (1984). The moral education theory of punishment. *Philosophy & Public Affairs, 13*, 208–238.
Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79–89.
Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019a). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General, 148*(3), 520–549.
Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019, February 19). *Communication in action: Planning and interpreting communicative demonstrations*. https://doi.org/10.31234/osf.io/a8sxk.
Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems* (pp. 3027–3035).
Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition, 167*, 91–106.
Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44*(12), 1697–1711.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.
Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology., 118*(1), 57.
Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the thirty-seventh annual conference of the cognitive science society*.
Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences, 111*(33), 12002–12007.
Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science, 27*(3), 405–418.
Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review, 117*(3), 785.
Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means. R package version 1.5.1. Retrieved from https://CRAN.R-project.org/package=emmeans.
Lerner, M. J., & Clayton, S. (2011). *Justice and self-interest: Two fundamental motives*. Cambridge, UK: Cambridge University Press.
Lopez-Brau, M., & Jara-Ettinger, J. (2020, August 4). *Physical pragmatics: Inferring the social meaning of objects*. https://doi.org/10.31234/osf.io/mnf4y.
Martin, J. W., & Cushman, F. (2016). The adaptive logic of moral luck. In J. Sytsma, & W. Buckwalter (Eds.), *The blackwell companion to experimental philosophy* (pp. 190–202).
McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 36*(1), 1–15.
Miller, D. T. (2001). Disrespect and the experience of injustice. *Annual Review of Psychology, 52*, 527–553.
Molnar, A., Chaudhry, S., & Loewenstein, G. (2020). *"It's not about the money. It's about sending a message!": Unpacking the components of revenge*.
Morris, H. (1981). A paternalistic theory of punishment. *American Philosophical Quarterly, 18*, 263–271.
Nadelhoffer, T., Heshmati, S., Kaplan, D., & Nichols, S. (2013). Folk retributivism and the communication confound. *Economics and Philosophy, 29*(2), 235–261.
Nock, M. K. (2008). Actions speak louder than words: An elaborated theoretical model of the social functions of self-injury and other harmful behaviors. *Applied and Preventive Psychology, 12*(4), 159–168.
Patil, I. (2018). *ggstatsplot: "ggplot2" based plots with statistical details*. CRAN. Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html.
Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences, 105*(3), 833–838.
Primoratz, I. (1989). Punishment as language. *Philosophy, 64*(248), 187–205.
R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.
Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences, 114*(32), 8511–8516.
Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences, 1*.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*(1), 1–12.

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science, 5*(3), 159–163.

Scott-Phillips, T. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Palgrave Macmillan.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55–89.

Skillen, A. J. (1980). How to say things with walls. *Philosophy, 55*(214), 509–523.

Smith, A. (1869). *A theory of moral sentiments*. London, England: Ball & Daldy (Original work published 1759).

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.

Stephen, J. F. (2014). *A history of the criminal law of England* (Vol. 2). Cambridge University Press (Original work published in 1863).

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). Cambridge: MIT press.

Thorndike, E. (1927). The law of effect. *The American Journal of Psychology, 39*(1/4), 212–222. https://doi.org/10.2307/1415413.

Wellman, H. M. (1992). *The child's theory of mind*. The MIT Press.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2771–2776). Cognitive Science Society.

Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405.