Evolution of flexibility and rigidity in retaliatory punishment

Adam Morris[a,1], James MacGlashan[b], Michael L. Littman[c], and Fiery Cushman[a]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138; [b]Cogitai, Providence, RI 02912; and [c]Department of Computer Science, Brown University, Providence, RI 02912

Edited by Andrew W. Delton, Stony Brook University, Stony Brook, NY, and accepted by Editorial Board Member Michael S. Gazzaniga August 1, 2017 (received for review March 13, 2017)

Natural selection designs some social behaviors to depend on flexible learning processes, whereas others are relatively rigid or reflexive. What determines the balance between these two approaches? We offer a detailed case study in the context of a two-player game with antisocial behavior and retaliatory punishment. We show that each player in this game—a "thief" and a "victim"—must balance two competing strategic interests. Flexibility is valuable because it allows adaptive differentiation in the face of diverse opponents. However, it is also risky because, in competitive games, it can produce systematically suboptimal behaviors. Using a combination of evolutionary analysis, reinforcement learning simulations, and behavioral experimentation, we show that the resolution to this tension—and the adaptation of social behavior in this game—hinges on the game's learning dynamics. Our findings clarify punishment's adaptive basis, offer a case study of the evolution of social preferences, and highlight an important connection between natural selection and learning in the resolution of social conflicts.

punishment | evolution | reinforcement learning | game theory | commitment

Human social behavior is sometimes remarkably rigid, and other times remarkably flexible. A key challenge for evolutionary theory is to understand why. That is, when will natural selection favor "reflexive" social behaviors, and when will it instead favor more flexible processes that guide social decision-making by learning?

We investigate a case study of this problem that illuminates some general principles of the evolution of social cognition. Specifically, we model the dynamic between antisocial behavior and retaliatory punishment in repeated relationships. Our goal is to understand when natural selection will favor flexibility (e.g., "try stealing and see if you can get away with it") versus rigidity ("punish thieves no matter what"). We approach this question through both a game-theoretic model of punishment and agent-based simulations that allow for the evolution of the rewards that guide learning. We demonstrate that the evolution of punishment depends on the learning dynamics of competing flexible agents, and that this interaction between learning and evolution can produce individuals with innate "social preferences," such as a taste for revenge (1–4).

The Evolution of Retaliatory Punishment

Individuals often punish those who harm them, even at a cost to themselves (5, 6). The adaptive rationale of this behavior seems clear in repeated or reputational interactions: Punishment promises a long-run gain by deterring social partners from doing future harm. This logic was classically formalized with a simple two-party repeated game (5) (Fig. 1*A*). On each round, a thief has the option to either steal from a victim (earning s and inflicting a cost $-s$) or do nothing. In response, the victim may either punish (paying a cost $-c$ to inflict a cost $-p$) or do nothing. Formal analysis shows that "punish all theft/stop stealing from victims who punish" is evolutionarily stable. This model, and many others that followed, offered a straightforward explanation for retaliatory punishment in repeated/reputational interactions (7–10). As a consequence, much attention has shifted to the puzzle of "altruistic" punishment in one-shot, anonymous settings (11, 12).

Models of retaliatory punishment embody a peculiar assumption, however: Thieves can flexibly adjust to different victim types, but victims must commit to identical behaviors against all thief types (Fig. 1*B*). The first element is uncontroversial: A discerning thief will steal from victims who never punish (i.e., pushovers), while respecting the property of those who do. This form of flexibility is sometimes also called "facultative adjustment" (10) or "opportunism" (9). In fact, it is necessary for retaliatory punishment to be adaptive: Punishment only pays when people adjust their behavior in response (5, 7, 10).

Our focus is the second element of this assumption. In contrast to thieves, it is typically assumed that victims are relatively rigid—that is, unable to tailor punishment to different types of opponents. In the classic model (5), whereas thieves can play "steal from pushovers, don't steal from retaliators," victims can only pick between two relatively rigid strategies: "always punish theft" or "never punish theft." Similarly, while many models show that punishment can help maintain cooperation in public goods settings, they typically assume that free-riders can play flexible strategies (like "cooperate with retaliators, defect against pushovers"), while victims of free-riding can only pick between "always punish defection" or "never punish defection" (7–10).

What are the consequences of relaxing this assumption? In theory, victims could profit from playing a flexible strategy. When facing a thief who learns from punishment (i.e., a flexible opponent), the victim would punish theft. Facing a rigid thief who cannot learn, however, the victim would abandon this costly and ineffectual punishment. (The latter possibility is well understood by parents who capitulate to incorrigible toddlers or towns that accede to entrenched organized crime.)

Significance

Two aims of behavioral science, often pursued separately, are to model the evolutionary dynamics and cognitive mechanisms governing behavior in social conflicts. Combining these approaches, we show that the dynamics of proximate mechanisms such as reward learning can play a pivotal role in determining evolutionary outcomes. We focus on a widespread feature of human social life: People engage in retributive punishment with surprising disregard for its efficacy, yet they respond to punishment with behavioral flexibility finely tuned to costs and benefits. We explain this pattern and offer a general picture of when evolution favors rigid versus flexible social behaviors.

Author contributions: A.M., J.M., M.L.L., and F.C. designed research; A.M., J.M., M.L.L., and F.C. performed research; A.M. analyzed data; and A.M. and F.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.W.D. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: adam.mtc.morris@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1704032114/-/DCSupplemental.

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.

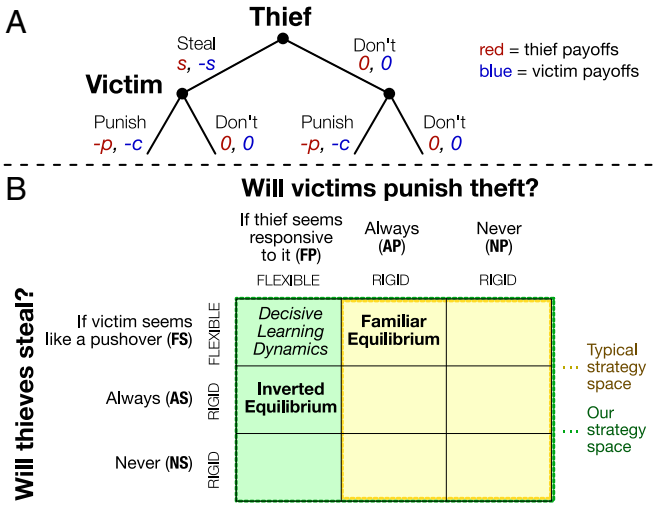

Fig. 1. (*A*) A round of the steal/punish game (5). s is the value of the stolen good, c the cost of punishing, and p the cost of being punished. $s, c > 0$ and $p > s$. The game is repeated for N rounds. (*B*) Pure strategies in the steal/punish game. Typical strategy spaces in models of retaliatory punishment, depicted by the yellow area, produce the familiar equilibrium of rigid punishment and flexible theft. Extending this strategy space to allow flexible victims reveals an inverted equilibrium: rigid theft and flexible punishment. The direction of selection hinges on the outcome of flexible thief against flexible victim: Whichever role tends to "back down" first evolves rigidity to compensate. (See Fig. S1 for payoffs.)

We introduce this possibility by incorporating a symmetric, role-neutral notion of flexibility into an evolutionary model of retaliatory punishment. Using this model, we demonstrate that there are two potential equilibria: the familiar equilibrium (in which victims rigidly punish all theft and thieves flexibly learn who to steal from) and an inverted equilibrium, in which thieves rigidly steal and victims flexibly learn who to punish. In the inverted equilibrium, thieves are incorrigible, and flexible victims cease punishing them when they learn that it is costly and useless.

These two rival equilibria establish a formal framework in which to investigate the adaptive underpinnings of flexible versus rigid social cognition. Prior work has identified numerous benefits of rigid "commitment" to behaviors like punishment (13–15). For instance, when agents can honestly signal a commitment to inflexibly punish in irrational settings (e.g., one-shot games), this motivates opponents to desist from theft (14, 16). Building on this tradition, we identify another strategic benefit of rigidity in competitive repeated games: It can compensate for systematic weaknesses in the learning mechanisms that enable flexibility. Some opponent strategies can bias flexible agents to represent a behavior as suboptimal, when, in fact, it would have been optimal in the long run. For an important class of learning algorithms, we show that flexible victims are more vulnerable to this weakness than flexible thieves. This asymmetry can drive victims to evolve rigid punishment—which, in the reward learning framework, corresponds with an innate "taste" for retaliation (1–3). We conclude by confirming a simple behavioral prediction of the model: People playing the steal/punish game rigidly persist in costly punishment, even when it appears ineffectual.

Model and Evolutionary Analysis

We begin by incorporating a role-neutral notion of flexibility into the steal/punish game and analyzing its evolutionary implications. We define a strategy space that encompasses the two rival equilibria described above (Fig. 1*B*), and determine the conditions under which each strategy is evolutionarily stable (17) and risk-dominant (18). We then support these static analyses by simulating the evolution of agents in a finite population under a wide range of parameter values. All results point to the same conclusion: Selection between the two equilibria hinges on the outcome of a multiagent learning scenario, flexible thief versus flexible victim. The role with the greater difficulty in learning an optimal strategy will evolve rigidity to compensate.

Strategy Space. In our model, random pairs of agents in a large, well-mixed population play the steal/punish game for N rounds (Fig. 1). Each agent inherits a strategy specifying its behavior in the game. We consider three pure strategies for each role: always, never, or flexibly steal (AS, NS, or FS, respectively) and always, never, or flexibly punish theft (AP, NP, or FP, respectively). The always/never strategies rigidly persist in their behavior, even when it is suboptimal against the current partner.

In contrast, flexible agents adopt the behavior of whichever inflexible strategy is optimal against the current partner. An agent playing FS never steals (i.e., adopts the inflexible strategy NS) when facing an opponent who punishes (i.e., AP), and always steals when facing an opponent who doesn't (i.e., NP). An agent playing FP never punishes when facing AS, and is indifferent when facing NS (in this case, both options—always or never punish theft—yield zero payoff).

The critical case is when two flexible agents meet (FS versus FP). Here, we assume there are two stable outcomes: Either the thief learns to steal and the victim gives up on punishing (i.e., thief adopts AS and victim NP), or the victim learns to punish and the thief gives up on stealing (thief adopts NS and victim AP). In other words, we assume that the flexible agents will fall into one of the two weak Nash equilibria among the inflexible strategies: AS/NP or NS/AP. We define θ as the probability of the former (the thief learns to steal), and $1-\theta$ as the probability of the latter (the victim learns to punish theft).

This case captures the vulnerability of flexible strategies. When flexible victims learn not to punish flexible thieves (high θ), they abandon a behavior that would be optimal in the long run and are disadvantaged. In contrast, if flexible thieves learn not to steal from flexible victims (low θ), then the thieves are disadvantaged. θ thus captures the relative vulnerability of flexible victims/thieves to learning suboptimal behavior. We find that θ is a critical determinant of the evolutionary outcome.

Population Diversity. Our aim is to identify the strategic benefits of rigid and flexible strategies. The benefit of flexibility is the ability to adjust to diverse opponents. In the static equilibrium analysis we use here, such diversity between individuals is represented by a single individual playing a mixed strategy [e.g., a population of 75% AS, 25% NS is represented by the strategy "play AS with probability $\frac{3}{4}$ and NS with $\frac{1}{4}$" (19)]. However, in games like ours with asymmetric roles, mixed strategies are always erased by selection (20). (See *SI Text* for proof.) Thus, the population has no permanent diversity, and there is no stable benefit to flexibility. If flexible strategies cannot be stable, neither can punishment: Punishing is only useful if opponents flexibly modify their behavior in response.

Consequently, past work implicitly assumed that ancestral populations were permanently diverse (5), or analyzed more complex evolutionary dynamics with mutation rates that ensured diversity (9). Our approach is to model diversity explicitly within a simpler, static analytic framework. We define an "epsilon game" (a second-order game parameterized by a constant ϵ), where each pure strategy is a probabilistic mixture of the pure strategies in the basic game. Specifically, if s is a pure strategy in the basic game (a "basic" strategy), then the corresponding pure strategy in the epsilon game (an "epsilon" strategy) is s_ϵ, which plays s with probability $1-\epsilon$ and the other two role-specific strategies each with probability $\frac{\epsilon}{2}$. For example, a thief playing FS_ϵ would play FS with probability $1-\epsilon$, and AS or NS each with probability $\frac{\epsilon}{2}$. By combining these epsilon strategies, agents can play any mixture of the basic strategies AS, NS, etc., but the probability of a basic strategy being played never falls below $\frac{\epsilon}{2}$, where $\epsilon > 0$. This approach allows us to model the effects of

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.

population diversity within a simple static framework, motivating the evolution of flexible strategies.

Evolutionarily Stable Strategy Analysis. Each agent inherits both a thief and victim epsilon strategy, plays both roles equally throughout its life, and reproduces in proportion to its accumulated payoffs. We write an agent's strategy as *(thief strategy, victim strategy)*. To explore the outcome of selection, we use the static solution concept of an evolutionarily stable strategy (ESS). Intuitively, a strategy is an ESS if, once it has taken over the population, it cannot be invaded by isolated appearances of other strategies ("mutations") (17). Using this rule, we derive the conditions of evolutionary stability for the two potential equilibria: rigid punishment/flexible theft and rigid theft/flexible punishment. In the epsilon game, these outcomes are represented by the strategies $(FS_\epsilon, AP_\epsilon)$ and $(AS_\epsilon, FP_\epsilon)$ (Fig. S2).

The conditions involve four parameters: θ (a flexible victim's relative vulnerability), ϵ (percent of the population that is permanently diverse), the ratio of c to s, and the ratio of p to s. We denote $r_{c:s} = \frac{c}{s}$ and $r_{p:s} = \frac{p}{s} - 1$ (the -1 is convenient for mathematical reasons). For exposition, assume that $\epsilon = 0.05$.

The resulting stability conditions are shown in Fig. 2. Roughly, when θ is high—i.e., flexible victims are relatively more vulnerable—only the familiar $(FS_\epsilon, AP_\epsilon)$ pair is stable. When θ is low—flexible thieves are more vulnerable—only the inverted $(AS_\epsilon, FP_\epsilon)$ pair is stable. [The conditions with arbitrary ϵ are derived in *SI Text* and conform to the same pattern (Fig. S3).]

In other words, rigid punishment/flexible theft is the equilibrium strategy specifically when flexible victims find it difficult to learn to punish flexible thieves. This result holds as long as punishment is not exceedingly strong (with our default parameters, no more than 38 times the value of the stolen good; see *SI Text* for robustness across parameter settings).

Risk-Dominance Analysis. When θ is moderate, both equilibria are stable (middle region in Fig. 2). To analyze this case, we use a standard criterion for selecting between two equilibria: "risk-dominance" (18, 21). Roughly, one equilibrium risk-dominates the other if it has a larger basin of attraction. The condition for risk-dominance here is simple: When both outcomes are stable, rigid punishment is risk-dominant if and only if $\theta > \frac{r_{c:s}}{r_{c:s}+r_{p:s}}$. Otherwise, rigid theft is risk-dominant.

Moran Process Simulations. Next, we simulate the system's evolution with a Moran process (22). This tests the robustness of our result to a relaxation of assumptions. The simulated agents live in a small population and use strategies from an extended space that includes all four "reactive" strategies [where a player's move can be conditioned on her opponent's last move (23)] and a flexible strategy (Fig. S4). Each generation, all agents play each other, one agent "dies" at random, and its replacement inherits a new strategy with probability proportional to the strategies' accumulated payoff. The process repeats many times.

We vary θ from 0 to 1, sampling the other parameters from a large space (Table S1). As an alternate method of incorporating population diversity, we include stochastic mutation. The result confirms our formal analysis (Fig. 3): As flexible victims become more vulnerable, agents become more likely to evolve to rigidly punish. [This finding holds for a broad range of mutation rates and selection intensities (Fig. S5).]

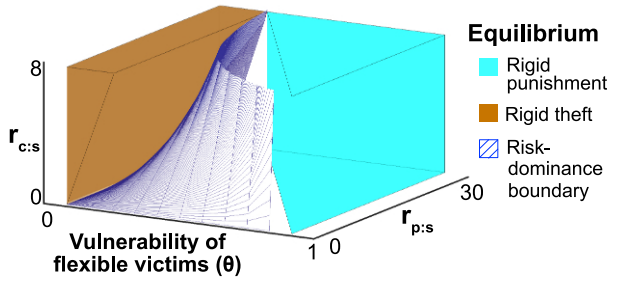

Fig. 2. Stability conditions in the steal/punish game (assuming $\epsilon = 0.05$, $r_{p:s}, r_{c:s} > \frac{1}{38}$, and $r_{p:s} + r_{c:s} < 38$). When θ is high, only the familiar $(FS_\epsilon, AP_\epsilon)$ is stable (blue-green region). When θ is low, only the inverted $(AS_\epsilon, FP_\epsilon)$ is stable (orange region). In the middle region, both are stable, and risk-dominance is determined by the boundary [$(FS_\epsilon, AP_\epsilon)$ to the right and $(AS_\epsilon, FP_\epsilon)$ to the left]. The evolution of rigid punishment depends on the relative vulnerability of flexible strategies in each role (see *SI Text*).

How Learning Dynamics Determine Relative Vulnerability

Our model suggests that, for a broad parameter space, the evolution of punishment in repeated relationships hinges on the outcome of a multiagent learning scenario: a flexible thief (FS) playing against a flexible victim (FP). In this competitive setting, flexible strategies have a vulnerability. They may learn to give up on a behavior (i.e., punishment or theft) that would have been optimal in the long run. Whichever role is more vulnerable in this scenario will evolve rigidity. We captured the outcome of this multiagent learning process with a parameter θ. Next, we seek a more principled way to predict the outcome of this process by considering a plausible cognitive model of behavioral flexibility.

Flexible behavioral control in humans is often accomplished via reward learning, including in social contexts (3, 24). To model this, we adopt the reinforcement-learning (RL) framework (25), a formalization of reward learning widely used in computer science, neuroscience, and psychology. In brief, RL agents choose actions by estimating their value: the sum of their expected future rewards. If agents experience rewards equal to the payoffs of the game, then they learn to choose payoff-maximizing actions against diverse opponents (26–28). We focus on a popular class of RL algorithms ("model-free") that directly estimate the value of actions based on historical returns (*Materials and Methods*). Using these algorithms, we aim to identify factors influencing the relative vulnerability of flexible strategies in each role. In other words, we ask: What happens when a reward-learning thief meets a reward-learning victim?

Learning Dynamics for RL Agents. We perform Monte Carlo simulations of two RL agents playing each other in the steal/punish game, with payoffs as the reward function and randomly sampled s, c, and p (Table S1). [We use the popular algorithm Q-learning (29) with eight internal states. See *Materials and Methods* for details.] In 72% of games, the thief learns to steal and the victim to not punish; in the remaining 28% of games, the victim learns to punish theft and the thief to stop stealing. In other words, under these conditions, $\theta = 0.72$.

Why are victims more vulnerable? Across games, variability in three parameters all strongly predict the outcome: As s and c increase, the victim becomes less likely to learn to punish (i.e., becomes more vulnerable), and as p increases, the victim becomes more likely to learn to punish (i.e., the thief becomes more vulnerable; logistic regression, $ps < 0.001$).

To illustrate how these parameters influence the learning dynamics, we highlight the effect of just one: the cost of punishing c. Intuitively, it is hard to learn that punishing has long-term value because it carries short-term costs. There is not a comparable obstacle to learning the long-term value of theft: Indeed, the most immediate consequences of theft are positive (you obtain the stolen good). Thus, when a reward-learning victim meets a reward-learning thief, the victim is biased against learning to punish, while the thief is biased toward learning to steal. All else being equal, the most probable outcome of these learning dynamics is persistent theft and no punishment (i.e., high θ).

To demonstrate this effect, we rerun the RL simulations above while varying c, and holding the other parameters constant. RL victims are more vulnerable when punishment costs are high, but not when those costs are negligible (Fig. 4*A*). [This result holds as long as s is low; otherwise, flexible thieves have too strong a learning advantage, and θ is always high (*SI Text*).]

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.

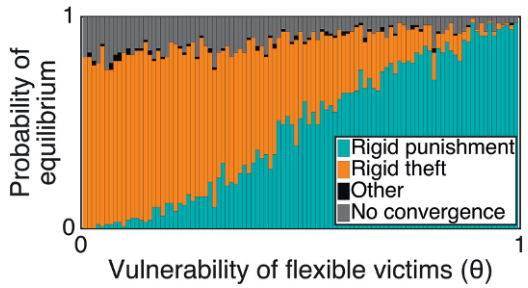

Fig. 3. The simulated evolution of strategies in the steal/punish game for different values of θ (see *Materials and Methods*). As flexible victims become more vulnerable, the population is more likely to converge to the familiar equilibrium with rigid punishment and flexible theft.

This clarifies why flexibility bears a strategic risk, beyond its previously identified relation to signaling (14). Although RL algorithms are guaranteed to converge to an optimal policy in stationary environments (25), there is no such blanket guarantee for competitive multiagent settings. In the steal/punish game, each learning agents is in a "race" to discover the policy that establishes their preferred behavioral equilibrium before their opponent does. Punishment's initial costs can bias learning away from its preferred equilibrium, exposing a systematic strategic risk of flexibility for this role. (See *SI Text* for details.)

The Evolution of Social Rewards. In our evolutionary analysis, we show that when flexible victims fail to learn to punish against flexible thieves (i.e., θ is high), victims evolve rigid punishment to compensate. We also show that when flexibility is accomplished via reward learning, punishment's short-term cost causes flexible victims to fail to learn to punish against flexible thieves. Combining these results, punishment's cost should bias selection toward the equilibrium with rigid punishment, which will evolve to compensate for the disadvantage of flexible punishment.

We test this prediction by embedding the RL agents described above in an evolutionary agent-based simulation. To model the flexible strategies (FS and FP), we again use Q-learning and set rewards equal to the game's payoffs.

To model the rigid strategies (AS, AP, NS, and NP), we use Q-learning agents with additional rewards conditioned on performing certain actions (30). These agents find stealing or punishing theft intrinsically rewarding or aversive, over and above the objective fitness payoffs. A thief who finds stealing sufficiently intrinsically aversive would never steal; a victim who finds punishing sufficiently intrinsically rewarding would rigidly punish theft; etc. In other words, agents can evolve social rewards—strong intrinsic rewards for social behaviors like retribution that can dominate their other costs. This technique incorporates a notion of rigidity into the RL framework, and affords a natural computational interpretation of economists' notion of social preferences as tastes for retribution, fairness, and so forth (3).

An agent's genotype, then, comprised two values subject to mutation: an intrinsic bias for or against theft and an intrinsic bias for or against punishing theft [denoted "(theft bias, punish-theft bias)"]. For example, an agent with genotype $(0, +)$ would play like (FS, AP). We simulate the evolution of these agents with the same Moran process as before, with reproductive success proportional to accumulated payoff and stochastic mutation ensuring permanent population diversity (*Materials and Methods*). To elucidate the effect of punishment's cost on selection, we vary c while holding the other payoffs constant.

The results support our analysis (Fig. 4*B*). When punishment is costly (and thus flexible victims are more vulnerable), agents evolve an intrinsic hedonic bias for punishing theft. However, when the cost is negligible (and thus flexible thieves are more vulnerable), agents evolve an intrinsic bias for stealing, not punishing. Ironically, then, for reward learning agents, selection can favor rigid punishment precisely when it is costly.

Of course, other learning algorithms exist that might perform differently in this game. Because reward learning is a key basis for much human social behavior (3, 24), it is a particularly interesting and important case study.

Rigid Punishment in Humans

Our analyses suggest that, when punishment carries a short-term cost, evolution will favor the familiar equilibrium of rigid punishment and flexible theft. In the context of RL, people will find punishing, but not stealing, intrinsically rewarding. Consistent with these results, people seem to find punishment rewarding (1, 2, 31) and rigidly punish in irrational settings (32, 33). Past work has not, however, explicitly tested for a behavioral asymmetry in the flexibility of punishment and theft. We conclude by demonstrating this asymmetry.

Human participants from Mechanical Turk are endowed with money and play repeated steal/punish games against real opponents (Fig. 5*A*). Participants are randomly assigned to the thief or victim role. To facilitate the procedure, one player in each game secretly precommits to one of several strategy choices, including two crucial options: rigidly steal and rigidly punish theft. These participants are matched with freely acting participants, who choose their actions sequentially with full knowledge of the opponent's prior actions. We focus on the behavior of the freely acting participant when facing rigid opponents.

We designed the payoffs such that, absent an asymmetric bias, participants in both roles would show identical learning curves: The thief would learn to stop stealing, and the victim to stop punishing, at similar rates. (See *Materials and Methods* for details.) Yet people persist much longer in punishing than stealing (Fig. 5*B*). We fit a mixed-effects model, regressing participant choices on their role, the round number, and the interaction between role and round. The interaction between role and round is significant (likelihood ratio test, $\chi^2(1) = 24.3$, $P < 0.001$; $\eta^2_G = 0.03$, 95% CI = [0.024, 0.24]). Consistent with our predictions, in repeated games against inflexible opponents, thieves act relatively flexibly and victims act relatively rigidly. [This result holds for a variety of payoff settings (*SI Text*).]

Although consistent with an asymmetry in the intrinsic reward of punishment versus theft, this result has several alternative

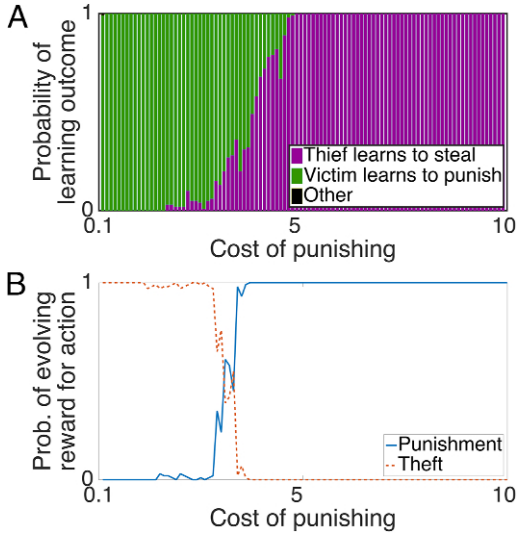

Fig. 4. (*A*) Outcome of two RL agents in the steal/punish game. As the cost of punishing increases, victims become less likely to learn to punish (i.e., become relatively more vulnerable). (*B*) We embedded the RL agents in an evolutionary simulation, allowing selection for hedonic biases for or against stealing/punishing. As the cost of punishing increases, selection increasingly favors an intrinsic hedonic bias for punishing (rather than stealing) (see *Materials and Methods*). Prob., probability.

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.

explanations. People may have learned different instrumental values for theft/punishment from experience, or different expectations about the pliability of thieves versus victims. To adjudicate between these explanations, future work should investigate the neural and psychological mechanisms that underlie this asymmetry, and test for the asymmetry in children and across cultures.

Discussion

Prior models of punishment assume that only "thieves" (harmdoers or free-riders), and not victims, can flexibly adapt their behavior to different opponent types (5, 7, 10). By relaxing this assumption, we show that there are two potential equilibria: one with rigid punishment/flexible theft and another with rigid theft/flexible punishment. Evolutionary outcomes hinge on the multiagent learning dynamics when two flexible players interact. Modeling this in the RL framework (25), we find that punishment's initial cost can make flexible victims more vulnerable to suboptimal learning than flexible thieves. This asymmetry favors the evolution of inflexible punishment—in the RL framework, an innate taste for retribution.

Our model makes three significant contributions. First, it helps explain the relatively inflexible nature of human punishment. People clearly exhibit some flexibility in their punishment choices, often tailoring punishment to minimize the risk of retaliation (e.g., they punish a coworker, but not a mob boss). However, people are remarkably insensitive to the contextual potential for effective deterrence (6, 32)—punishing, for instance, in one-shot anonymous settings (33). We find that people also persist in punishment, but not theft, for several rounds of a repeated game against opponents who never learn.

Why do people possess a deterrence mechanism (punishment) that persists in contexts where it is ineffectual? Some prior analyses have posited that inflexibility is strategically beneficial because it signals your commitment to punish even in "irrational" contexts (e.g., one-shot interactions) (13–15).

We identify an additional strategic benefit of inflexibility: By committing to punish in "rational" contexts, it prevents proximate learning mechanisms from converging on suboptimal behavior in repeated games. Moreover, our model explains why evolution would commit people to punishment, but not theft: Punishment's immediate cost makes flexible victims asymmetrically vulnerable to this weakness. This result extends our understanding of the adaptive underpinnings of social inflexibility. It also complements the prior suggestion that inflexibility can compensate for weakness in planning due to temporal discounting [as opposed to learning, as we investigate here (14)].

Second, our model highlights an ironic (but perhaps common) interaction between learning and evolution: A social role that tends to "lose" in the learning dynamic may consequently "win" in the evolutionary dynamic. For instance, when victims are relatively handicapped in learning the benefits of punishment, they evolve a rigid punishment strategy that ultimately achieves their preferred in-game pattern: no theft. This parallels prior research demonstrating that, in mutualisms between two species, the species which is slower to adapt ends up receiving more benefits in the long run, because it is more committed to its preferred outcome (34). In both cases, relative weakness at a shorter timescale fosters relative strength at a longer timescale (35).

Third, our work is a step toward reconciling proximate and ultimate models of social behaviors like punishment. A wealth of evidence suggests that punishment decisions are guided by systems of value and reward. People often report hedonic satisfaction from punishing (31) (i.e., "revenge is sweet"), and punishing wrongdoers is associated with activation in the striatum (2) and orbitofrontal cortex (1, 36), brain regions central to reward-based decision-making. The same is true for fairness, cooperation, and other social behaviors, suggesting that people have a suite of evolved tastes that guide their social decisions (3, 24).

Evolutionary models, however, typically rely on abstracted versions of these behaviors, and rarely incorporate details about proximate mechanisms (37); they model "eats apples," not "loves fructose." By embedding RL agents in an evolutionary model, we attempt to bridge this gap. Our model offers a precise account of how and why evolution would make punishment intrinsically rewarding. This basic approach can be used to predict other social behaviors that people will find rewarding and to explain why.

Our analysis has several limitations. People are relatively rigid when punishing, but they still exhibit some flexibility; our model does not explain when or how this occurs. Our model does not allow agents to abstain from future interactions (38) or to counterpunish (4). Our formal evolutionary analysis uses a static framework, which is only an approximation for more accurate dynamic models. Finally, we interrogate one plausible reward-learning algorithm, but learning comes in many varieties. Future work must fill these gaps.

Despite these limitations, our case study of punishment highlights the utility of evolutionary models defined over plausible psychological mechanisms (37, 39), a research path promising fruitful insight into the origins of social behavior.

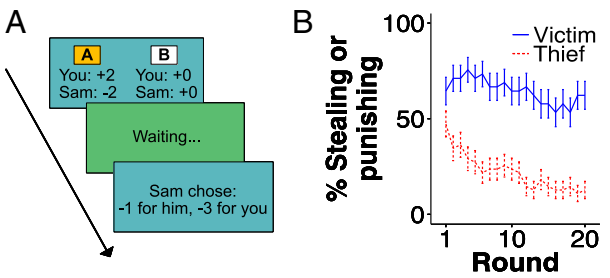

Fig. 5. (*A*) Example round of the behavioral experiment. (*B*) Rate of theft/punishment against rigid opponents. Thieves learn to stop stealing, but victims do not learn to stop punishing theft. (Bars are ± SEM.)

Materials and Methods

All code and data can be found at https://github.com/adammmorris/rigid-punishment. See *SI Text* for ESS and risk-dominance calculations.

Moran Simulations. To simulate the evolution of this system, we use a Moran process with selection and mutation. A population of K agents evolves over 10^3 generations. Each agent A_i inherits both a thief and victim strategy (see *SI Text* for the expanded strategy space), plays each other agent as both thief and victim, and receives a fitness score f_i equal to its average payoff. Then, one agent dies at random, and another agent is chosen to reproduce according to a softmax function: $P(A_i) = e^{wf_i} \Big/ \sum_{j=1}^{K} e^{wf_j}$, where w controls selection intensity. We also include a mutation rate μ. When an agent reproduces with probability $1 - \mu$, it passes on its strategy pair, and with probability μ, it passes on a random different strategy pair. (This models a death–birth process with exponential payoff to fitness mapping.)

We vary θ from 0 to 1 and run 100 simulations for each value. We classify each simulation as "converging" to a strategy when, averaged across all generations past 1,000, $>1-\mu-0.1$ of the population inherit that strategy. (If no strategy meets that criterion, we classify the simulation as having no equilibrium.) See *SI Text* for the parameters we use, along with a detailed analysis of how the results vary with different parameter values. Notably, the results are robust to variation in w and μ (Fig. S5).

RL Simulations. We adopt the RL framework to model the multiagent learning scenario of flexible thief versus flexible victim. The steal/punish game can be conceptualized as a Markov decision process (MDP), with a set of states S that agents can be in (e.g., "I was punished last turn"), a set of actions A_s each agent can take in each state (e.g., steal or do nothing), a reward function that maps state–action pairs onto payoffs, and a transition function between states. RL algorithms learn to choose the behaviors in a MDP that maximize their total reward and can be used as a model of learning in games like steal/punish (28).

Specifically, we take Q-learning as our model of learning (29). In Q-learning, the agent estimates the long-term expected value of each action a in each state s, $Q(s, a)$. When a game starts, these Q-values are initialized

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.

to zero. At each time step, the agent selects an action a from state s based on its current Q-value estimates (using a softmax function with parameter β) and then updates those estimates according to:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a' \in A_{s'}} Q(s',a') - Q(s,a))$$

where α is a learning rate, γ is a discount rate, r is the observed reward, and s' is the subsequent state.

The learning agent must represent what state of the game it is in. Following prior work (28), our agents keep track of the last two moves. For instance, one state might be, "On the last two moves, I stole and was punished." A Q-learner reacts to its partner's type by estimating the value of being in the states following its actions. For example, if a thief is facing a victim who always punishes, then it will learn a negative value for the state following stealing. This allows a Q-learner to mold its behavior to different opponents without building an explicit model of the opponent's behavior (hence its designation as model-free RL).

First, we simulate 10,000 Q-learning agents playing each other in the steal/punish game, randomly sampling s, c, and p from the distributions given in Tables S1 and S2. For each simulation, we analyze all turns past 1,000 and categorize it as "punisher exploited" if the thief stole and the victim refrained from punishing on >95% of turns, and as "thief exploited" if the thief refrained from stealing on >95% of turns. Then, we rerun the simulations, varying c from 0.1 to 10 while fixing $s = 5$, $p = 15$ (Fig. 4*A*). [We choose these values for s and p because they highlight the illustrative effect that c can have on punishment's learning dynamics. However, in other ranges, s and p can drown out the effect of c (*SI Text*).]

Finally, we embed the RL agents in an evolutionary simulation where agents have a heritable genotype (theft bias, punish-theft bias). As a baseline, the reward an agent receives for a behavior is equal to the behavior's payoff, equivalent to its fitness consequences. A nonzero bias alters the agent's reward function: An agent with a steal bias of +2, e.g., would experience an additional 2 units of reward upon stealing. This extra reward only affects the agent's experience during learning, not its fitness.

There are three possible steal biases and punishing biases (corresponding to the three thief and victim strategies in the evolutionary analysis). The numerical magnitude of each bias is chosen as the smallest integer necessary to guarantee the appropriate behavior (*SI Text*). We simulate the evolution of these agents with the same Moran process as above. To make the simulations tractable, we first cache the results of each genotype playing each other genotype in 100 full games. Then, in each generation, we sample one of those results for each genotype–genotype match.

We again vary c from 0.1 to 10 (with $s = 5$ and $p = 15$), run 100 Moran simulations for each value, and record the percentage in which agents evolve an intrinsic reward for stealing or punishing theft (using the convergence criterion described above; Fig. 4*B*). In *SI Text*, we describe the assumptions, design, and results of these simulations in more detail. All results are robust to variation in parameters (Fig. S6). We also describe an analysis which suggests that the results are due to the effect of c on the learning dynamics, not the change in the payoffs themselves.

Behavioral Experiment. One hundred participants were recruited on Amazon Mechanical Turk. All gave informed consent, and the study was approved by Harvard's Committee on the Use of Human Subjects. Each participant plays one focal game against an opponent who always steals/punishes (as either thief or victim) and two background games against the other opponent types (one as thief, one as victim). Game order is varied between participants and controlled for in the analysis. The focal game lasts 20 rounds; the background games last a random number of rounds, chosen uniformly between 10 and 20. Participants do not know the game lengths. In each round, participants are presented with a choice of two (neutrally labeled) actions: steal (+2 cents to you and −2 to partner) or do nothing (0 to both) when thief, and punish (−1 cent to you and −3 to partner) or do nothing when victim. They are then informed of their partner's decision. (When playing as victim, participants are told whether their partner stole before deciding whether to punish.)

The payoffs of $s = 2$, $c = 1$, $p = 3$ present identical pecuniary incentives to thieves and victims: Compared with the alternative of doing nothing, thieves receive a net −1 cent for stealing, and victims −1 cent for punishing (when facing opponents who rigidly punish theft or steal). Thus, without some asymmetric bias, people should show identical learning curves in the two roles. [The result is robust to variation in payoffs (Fig. S7).]

ACKNOWLEDGMENTS. We thank Adam Bear, Moshe Hoffman, and Drew Fudenberg for feedback on the manuscript. This research was supported by Office of Naval Research Grant N00014-14-1-0800.

1. Seymour B, Singer T, Dolan R (2007) The neurobiology of punishment. *Nat Rev Neurosci* 8:300–311.
2. de Quervain DJ-F, et al. (2004) The neural basis of altruistic punishment. *Science* 305:1254–1258.
3. Fehr E, Camerer CF (2007) Social neuroeconomics: The neural circuitry of social preferences. *Trends Cogn Sci* 11:419–427.
4. Fehl K, Sommerfeld RD, Semmann D, Krambeck HJ, Milinski M (2012) I dare you to punish me—vendettas in games of cooperation. *PLoS One* 7:e45093.
5. Clutton-Brock TH, Parker GA (1995) Punishment in animal societies. *Nature* 373:209–216.
6. McCullough ME, Kurzban R, Tabak BA (2013) Cognitive systems for revenge and forgiveness. *Behav Brain Sci* 36:1–15.
7. Sigmund K, Hauert C, Nowak MA (2001) Reward and punishment. *Proc Natl Acad Sci USA* 98:10757–10762.
8. Hauert C, Haiden N, Sigmund K (2004) The dynamics of public goods. *Discrete Contin Dyn Syst B* 4:575–587.
9. Hilbe C, Sigmund K (2010) Incentives and opportunism: From the carrot to the stick. *Proc R Soc Lond B Biol Sci* 277:2427–2433.
10. Gardner A, West S (2004) Cooperation and punishment, especially in humans. *Am Nat* 164:753–764.
11. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137–140.
12. Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. *Proc Natl Acad Sci USA* 100:3531–3535.
13. Schelling TC (1980) *The Strategy of Conflict* (Harvard Univ Press, Cambridge, MA).
14. Frank RH (1988) *Passions Within Reason: The Strategic Role of the Emotions* (W W Norton, New York), Vol xiii.
15. Nesse R (2001) *Evolution and the Capacity for Commitment* (Russell Sage Foundation, New York).
16. Barclay P (2017) Bidding to commit: An experimental test of the benefits of commitment under moderate degrees of conflict. *Evol Psychol* 15:1474704917690740.
17. Smith JM, Price GR (1973) The logic of animal conflict. *Nature* 246:15–18.
18. Harsanyi JC, Selten R (1988) *A General Theory of Equilibrium Selection in Games* (MIT Press, Cambridge, MA).
19. Weibull JW (1997) *Evolutionary Game Theory* (MIT Press, Cambridge, MA).
20. Selten R (1980) A note on evolutionarily stable strategies in asymmetric animal conflicts. *J Theor Biol* 84:93–101.
21. Kandori M, Mailath GJ, Rob R (1993) Learning, mutation, and long run equilibria in games. *Econometrica* 61:29–56.
22. Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* 37:17–23.
23. Rand DG, Ohtsuki H, Nowak MA (2009) Direct reciprocity with costly punishment: Generous tit-for-tat prevails. *J Theor Biol* 256:45–57.
24. Lee D (2008) Game theory and neural basis of social decision making. *Nat Neurosci* 11:404–409.
25. Sutton RS, Barto AG (1998) *Introduction to Reinforcement Learning* (MIT Press, Cambridge, MA), 1st Ed.
26. Erev I, Roth AE (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am Econ Rev* 88:848–881.
27. Roth AE, Erev I (1995) Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ Behav* 8:164–212.
28. Masuda N, Ohtsuki H (2009) A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bull Math Biol* 71:1818–1850.
29. Watkins CJCH, Dayan P (1992) Q-learning. *Mach Learn* 8:279–292.
30. Ackley D, Littman M (1992) Interactions between learning and evolution. *Artificial Life II, Santa Fe Studies in the Science of Complexity Proceedings*, eds Taatgen N, van Rijn H (Westview, Boulder, CO), Vol X, pp 487–510.
31. Gollwitzer M, Meder M, Schmitt M (2011) What gives victims satisfaction when they seek revenge? *Eur J Soc Psychol* 4:364–374.
32. Carlsmith KM, Darley JM (2008) Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, ed Zanna MP (Academic, New York), Vol 40, pp 193–236.
33. Camerer C (2003) *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ Press, Princeton).
34. Bergstrom CT, Lachmann M (2003) The Red King effect: When the slowest runner wins the coevolutionary race. *Proc Natl Acad Sci USA* 100:593–598.
35. Haken H (1987) *Synergetics in Self-Organizing Systems, Life Science Monographs*, eds Yates FE, Garfinkel A, Walter DO, Yates GB (Springer, New York), pp 417–434.
36. Singer T, et al. (2006) Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439:466–469.
37. Rand DG, Nowak MA (2013) Human cooperation. *Trends Cogn Sci* 17:413–425.
38. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K (2007) Via freedom to coercion: The emergence of costly punishment. *Science* 316:1905–1907.
39. Bear A, Rand DG (2016) Intuition, deliberation, and the evolution of cooperation. *Proc Natl Acad Sci USA* 113:936–941.
40. Ellison G (2000) Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Rev Econ Stud* 67:17–45.
41. Maruta T (1997) On the relationship between risk-dominance and stochastic stability. *Games Econ Behav* 19:221–234.
42. Kelley K (2007) Methods for the behavioral, educational, and social sciences: An R package. *Behav Res Methods* 39:979–984.

Downloaded from https://www.pnas.org by 146.115.149.241 on March 28, 2022 from IP address 146.115.149.241.