From Human Morality & Sociality: Evolutionary & Comparative Perspectives, edited by Henrik Høgh-Oleson, Palgrave MacMillan, New York, 2010.

5

Evolutionary Psychology and Criminal Justice: A Recalibrational Theory of Punishment and Reconciliation

Michael Bang Petersen, Aaron Sell, John Tooby, and Leda Cosmides

Introduction

Exploitation – the imposition of costs on another for one's own benefit – was a major and ongoing adaptive problem during human evolution. What elements of the human mind evolved in response to the problems posed by exploitation? In this chapter, we analyze these problems, and derive predictions about the evolved design of psychological adaptations that allow us to detect, conceptualize, and respond adaptively to exploitation. We argue that our species-typical psychological architecture includes evolved programs whose function is to (1) evaluate the prospective return of continued association with the perpetrator of exploitive acts and, if the value is positive, (2) motivate actions that reduce the problems posed by the prospect of future exploitation. More specifically, the function of these actions is to recalibrate certain behavior-regulating variables in the minds of the perpetrator and other potential exploiters. Based on this argument, we analyze punitive and conciliatory strategies,

outlining how each strategy targets different regulatory variables in the motivational systems of perpetrators. Evolutionary analysis and empirical evidence both suggest that the strategy deployed against an exploiter is governed by cues that were correlated during our evolutionary past with the future net value of the exploiter. As we will show, this framework has implications for understanding the political attitudes and moral intuitions that humans have toward issues and events involving criminal justice.

Evolution, Exploitation, and Crime

In a highly social species like ours, there are reasons to expect that exploitation will not take the form of a Hobbesian war of all against all: Biologists have identified a number of selection pressures that favor the evolution of mechanisms designed to restrain an organism (under some conditions) from imposing costs on conspecifics for its own benefit. These include kin selection (Hamilton 1964; Williams and Williams 1957), reciprocation and exchange (Trivers 1971; Cosmides and Tooby 1992), the existence of positive externalities (Tooby and Cosmides 1996) and the avoidance of aggressive countermeasures from the exploited organism or its allies (Maynard Smith and Parker 1976; Sell, Tooby, and Cosmides 2009, forthcoming). Nevertheless, these selection pressures only carve out exceptions to the general selective gradient. Outside of the scope of such exceptions, organisms are selected to benefit themselves, regardless of the consequences of such acts on others. Accordingly, throughout our evolution, the average person was situated in a world full of individuals poised to impose costs on him or her if such acts were self-beneficial (Cosmides and Tooby 1992; Duntley 2005; Trivers 1971).

As many researchers have recognized, the risk of being exploited poses a major, chronic family of adaptive problems for humans (Daly and Wilson 1988; Duntley and Buss 2004). Organisms equipped with adaptations to prevent, deter, or productively respond to the threat of exploitation by others would be favored by selection. As expected, humans appear to have evolved adaptations to mitigate at least some varieties of exploitation. For example, humans appear to have evolved reasoning specializations to detect cheaters in contexts of social exchange (Cosmides and Tooby 1992; 2005). Similarly, the existence of patterned responses to exploitation in collective action supports the view that a motivational specialization deploying

punitive sentiment evolved as a defense against free riders (Price, Cosmides, and Tooby 2002).

In this chapter, we argue that intuitions, attitudes, sentiments, and moral discourse that spontaneously emerge when people confront crimes (whether directly or in the context of political questions about criminal justice) are, in part, the expressions of mechanisms that evolved to defend humans ancestrally against exploitative behavior. Here, we focus on the distinctive psychology that evolved to respond to the problems posed by acts of exploitation perpetrated by co-members of the small residential groups in which ancestral humans lived. Exploitation by members of outgroups typically activates a different suite of evolved defenses – coalitional psychology – that lies largely beyond the scope of this chapter. In short, as far as our evolved psychology goes, "crimes" are (certain) acts of exploitation by ingroup members, while the same acts by outgroup members (especially sets of outgroup members) will often be categorized as "attacks" – that is, represented and responded to using a different evolved psychology.[1]

In what follows, it is important to keep in mind that these mechanisms evolved to assume the causal and statistical structure of the ancestral world, and make functional sense in that context. Although our species-typical set of complex neurocomputational adaptations changes only very slowly (because species-wide gene substitution takes a great deal of time), our social environment has been changing rapidly since the rise of agriculture. Today, in an environment with nation-states comprising millions of people and sophisticated technology, we cannot assume that our evolved computational equipment is operating functionally. What we can assume, however, is that despite possible dysfunctionalities, our motivations, categories, intuitions, and moral concepts are still guided in part by evolved adaptations whose outputs would have been adaptive under ancestral conditions. In other words, we argue that modern crimes exhibit cues that satisfy the input conditions of the mechanisms in our evolved counter-exploitive psychology. We expect, therefore, that our opinions about crime should be guided by the evolved neural programs whose design we outline below. This allows us to test some of our hypotheses about the nature of these mechanisms by drawing upon the criminological literature involving attitudes toward crime.

One central theme in our argument is that the mind evolved to produce not only punitive responses to crime but also reconciliatory responses as well – a topic that has been overshadowed by treatments

that focus on punishment. For example, the literature on cheating has mainly focused on cost-infliction, punishment, and exclusion (Price, Cosmides, and Tooby 2002; Frank 1988). But given the costs of inflicting punishment, practical limitations on its use in many contexts, and the availability of other avenues of social influence, we expect a variety of other counter-exploitive strategies to evolve in addition to punishment. Evidence suggesting that natural selection may have designed nonpunitive means of conflict resolution comes from primatologists who explore the role of reconciliatory behaviors in managing conflicts and aggressive encounters (Aureli and de Waal 2000; de Waal 1996; see also Trivers 1971 on reparative strategies in reciprocation). Similarly, anthropologists have documented restorative sanctions across diverse agricultural and small-scale societies (Braithwaite 2002; Fry 2000). Nevertheless, a position, which was most famously formulated by Freud (1961) but runs through large parts of twentieth century social science, depicts humankind as by nature punitive – only resorting to restorative reactions because of countervailing socialization processes that have emerged for culturally contingent reasons in "advanced" civilizations (Durkheim 1998; Elias 1994; Spierenburg 1984; Garland 1990). The renewed emphasis on punishment in recent evolutionary approaches leaves Freud's portrait of the inherent punitiveness of human nature relatively uncontested. Below, however, we will develop an alternative to the Freudian vision, in which we argue that the mind contains evolved programs that deploy both punitive and reparative strategies to deal with transgressors.

The dissection of exploitation and counter-exploitive strategies derives from new proposals about the computational architecture that evolved among humans to regulate social behavior (e.g., Tooby, Cosmides, Sell, Lieberman and Sznycer 2008). We will outline relevant features of the proposed architecture, and will then develop concepts of exploitation, punishment, and reconciliation anchored in this analysis. These proposals delimit the conditions under which punishment and reconciliation, respectively, should have been the best available responses to exploitive acts.

Internal Regulatory Variables

Given an evolutionary and computational approach, it is useful first to consider how certain evolved mental programs dissect human sociality, and then to locate exploitation, punishment, and reconciliation

within this framework. One divergence from traditional approaches to psychology that we will review and then apply is the idea of *internal regulatory variables* and their relationship to emotion programs (Tooby and Cosmides 2005, 2008; Tooby, Cosmides, Sell et al. 2008). The claim is that the architecture of the human mind, by design, contains registers for evolved computational variables whose function is to store summary magnitudes that are necessary for regulating behavior and making inferences involving valuation. These are not traditional and familiar psychological constructs, such as concepts, representations, goal states, beliefs, or desires (although they may contribute to the emergence of any of these). Instead, they are underlying indices that acquire their meaning from the evolved behavior-controlling and information-processing procedures that access them. That is, each has a location embedded in the input-output relations of other evolved programs, and the function of the internal regulatory variables is to regulate the decision flow of those programs.

Easy examples include a variable tracking the intensity of hunger, and a different one tracking the intensity of fatigue. These are increased or decreased by various input systems, and, when integrated with other processes, regulate choice behavior. Regulatory variables that have more interesting properties seem required when one attempts to model in detail how the human psychological architecture must be designed to respond successfully to recurrent problems posed by the social world.

For example, in the recent mapping of the architecture of the human kin detection system, research identified a series of regulatory variables needed to make the system work functionally and to explain the data (Lieberman, Tooby, and Cosmides 2007). In this case, it appears that for each familiar individual *i*, the system computes and updates a continuous variable, the *kinship index* K_i, that corresponds to the system's pairwise estimate of genetic relatedness between self and *i*. When the kinship index is computed or updated for a given individual, the magnitude is taken as input to procedures that are designed to regulate kin-relevant behaviors in a fitness-promoting way. In the case of altruism, the kinship index is fed as one of many inputs to an estimator, whose function is to compute a magnitude that regulates the weight placed on the welfare of *i* (see next section). A high kinship index upregulates the weight put on *i*'s welfare, while a low kinship index has little effect on the disposition to treat *i* altruistically. This is one element that upregulates the emotion of love, attachment, or caring.

In parallel, the kinship index is also fed into the sexual value estimator as one of many inputs. The function of the sexual value circuitry is to compute a magnitude, sexual value (SV_i), that regulates the extent to which the actor is motivated to value or disvalue sexual contact with individual *i*. As in the case of altruism, many factors (e.g., health, age, symmetry) affect sexual value. But a high kinship index renders sexual valuation strongly negative, making the idea of sex with individual *i* disgusting and aversive. In contrast, a low kinship index is expected to have no impact on the other factors leading to sexual valuation.

Empirical work (Lieberman, Tooby, and Cosmides 2007) has demonstrated that the kinship index is set by two cues: (1) whether an older sibling observes the mother caring for his or her younger sibling as an infant, and (2) the duration of coresidence between birth and the end of the period of parental investment. This system was designed by natural selection to detect which familiar others were close genetic relatives, create a magnitude corresponding to the degree of genetic relatedness, and then to deploy this information to motivate both a sexual aversion between brothers and sisters, and to motivate a disposition to behave altruistically toward siblings.

An internal regulatory variable – like the kinship index, the sexual value index, or the welfare trade-off ratio (see below) – acquires its meaning and functional properties by its relationship to the programs that compute it, and by the downstream decisions or processes that it regulates. It is clear from both research and introspection that such computations and their embedded variables are usually nonconscious or implicit, and express themselves as feelings, intuitions, and inclinations. Outputs of the nonconscious processes that access these variables may be consciously experienced, as (for example) disgust at the prospect of sex with a sibling, affection for, or indifference to, a sibling, fear on their behalf, grief at their loss, and so on.

Welfare Trade-Off Ratios in the Mind and in Behavior

Decisions involving welfare trade-offs are ubiquitous. Life in modern industrial societies involves innumerable daily dilemmas, such as, should I pick up trash that fell to the street? Should I let another car ahead of me? Should I accept the cost of babysitting my neighbors' children so that they can go out? Should I forego the benefit of playing loud music to avoid imposing the cost of it on my neighbor? Should I do the dishes, or let my spouse do them? In all such contexts

trade-offs between one's own welfare and the welfare of others have to be made. The ancestral world of hunter-gatherers would also have been permeated with contexts in which our ancestors were forced to make trade-offs between the welfare of self and other. As an illustration drawn from modern foragers, Hill (2002) lists a range of contexts for cooperative behavior in the Ache of Paraguay, including everything from clearing camp spots for others, to feeding another's offspring, to entertaining others by singing. Equally, deciding when to be selfish or aggressive also requires making trade-offs. In short, making welfare trade-off decisions adaptively constituted an important and pervasive adaptive problem for our ancestors.

In making a choice that impacts another, in principle an individual could weight a specific other's welfare not at all, moderately, strongly, or could self-sacrificially place weight only on the other's welfare. It would be an odd human being who was completely indifferent to her own welfare, however. Obviously, natural selection favored the evolution of motivational systems that favored acting in one's own self-interest (i.e., to proximate cues that would have predicted fitness enhancement ancestrally). Selection also favored the evolution of motivational systems designed to modify the individual's behavior based on its effects on others (e.g., kin selection, reciprocation, fear of retaliation, fitness interdependence, changes in externalities, acquiescence to extortion). This is why it is also rare to find humans who uniformly act with total disregard for the impacts of their acts on the welfare of others. In short, our evolved decision-making architecture must have components designed to weight the welfare of self versus the welfare of others, and to balance them in ways that would have promoted fitness ancestrally.

How is this accomplished? It is our belief that a substantial proportion of human social interactions are regulated, in part, by evolved circuitry that includes a particularly important family of variables: welfare trade-off ratios, or WTRs (Tooby, Cosmides, Sell, et al. 2008). A WTR indexes the degree to which one's valuation of another's welfare is expressed in choices and behavior – i.e., the extent to which you are disposed to trade-off your own welfare against another person's welfare when you take action. According to recent research (Sell 2006b; Sell et al. 2009, forthcoming; Tooby, Cosmides, and Price 2006a; Tooby, Cosmides, Sell, et al. 2008), decisions that impact the welfare of others appear to reflect the operation of evolved circuitry that embeds an internal

threshold magnitude, the WTR, and its associated welfare trade-off functions. The proposal is that these variables are magnitudes, instantiated in neural tissue, that function as control elements accessed during decision-making processes that impact welfare. Independent circuits – like the human kin detection system or the social exchange system – take in information about a person (e.g., cues of relatedness; or, did they reciprocate recently?) and use it over the life course to upregulate or downregulate the magnitude of the person-specific WTR – increasing or decreasing the disposition to help, for example.

During the decision-making process, the WTR variable between the self and the person impacted by the potential act is accessed to see whether the course of action being considered should be carried out, or whether it should not because it places too little value either on oneself or on the other, given the magnitudes of the costs and benefits to all affected parties (see Delton, Sznycer, Robertson, Lim, Cosmides, and Tooby, forthcoming). That is, emotions and decisions involving trade-offs exhibit a series of evolutionarily predicted, lawful patterns that parsimoniously implicate the existence of this family of variables. The WTR is a person-specific variable, which sets the threshold for acceptable cost-benefit transactions between the relevant person and the self – i.e., the threshold at which willingness becomes unwillingness with respect to the particular person. The level of my WTR toward another person thus guides how large a cost I will voluntarily incur in order to secure a benefit for the specific person; and, how large a cost I am willing to impose on that other to secure a benefit for myself. All else equal, the larger the WTR between myself and the specific other, the larger costs I am willing to incur and the smaller the costs I am willing to impose, respectively. Thus, if individual X has a welfare trade-off ratio of 1:1 toward individual J, that means that X values J's welfare equal to her own, while a WTR of 5:1 means that X would be willing to impose a cost of 4 on J to gain a benefit of 1, but not a cost of 6 to gain a benefit of 1. The welfare trade-off ratio constitutes the computational basis for the intuitive concept of the value of a particular person to oneself.

The circuitry within which welfare trade-off ratios are embedded is hypothesized to have several design features geared toward the problem of regulating cost-benefit transactions among individuals of differential value to the decision maker (see Sell, Tooby and Cosmides, forthcoming; Sell 2006b; Tooby, Cosmides, Sell, et al.

2008). First, our minds should intuitively and automatically assign a cost-benefit interpretation (including valuations from the perspective of self and other) to events and potential choices. For example, a formerly unacceptable cost might become acceptable to you if you receive new information about the increased benefit the other person might derive from an act or resource (i.e., their need has become greater). While we might not offer a person our newly bought sweater to clean his dirty hands, we might be willing to accept just the same cost (a ruined sweater) if the same person's child suffers a severe wound and the sweater could be used to stop the bleeding. The latter entails a far greater benefit to the recipient than the former.

Second, the agent parameter of the WTR system should be flexible enough to allow WTRs to be computed for a range of different types of agents, not just individual humans (Tooby, Cosmides, and Price 2006). There is no reason to expect that our ancestral social world was solely comprised by dyadic interactions. Triadic social exchange, as well as exchanges involving even larger numbers of individuals, would have been frequent (Tooby, Cosmides, and Price 2006). In order to engage in such n-person exchange and make decisions about allocating time and resources to collective action, we need to be able to, first, represent our own welfare and the welfare of the collective, and second, trade-off our own welfare relative to the welfare of the group. To the extent that our minds contain systems for representing multiple persons as a single entity (a group), the necessary computations can be performed by feeding this entity into the agent parameter of the regular WTR system – albeit with some specialized machinery regulating group trade-offs. Thus, we propose that the computational system that governs individual-level welfare trade-offs also enables us to form welfare trade-off ratios involving groups.

As is the case with other internal regulatory variables, the psychological architecture in which the WTR is embedded will contain information-processing mechanisms that continuously scan situations for the existence of relevant ancestral cues, which in the environment of evolutionary adaptedness were reliably correlated with either increases or decreases in welfare. Upon encountering such information, these mechanisms recalibrate the WTR. But to understand what precise kinds of information are responsible for setting the level of a person's WTR toward another, we need to dissect the concept of WTRs further.

Monitored and Intrinsic WTRs

There are two adaptively distinct contexts in which we trade-off our welfare relative to the welfare of others. First, there are situations in which our behavior is either directly monitored by the person whose welfare is affected, or where it is highly probable that the person (or their allies) will become aware of our agency. Second, there are situations in which the affected individuals are not present or capable of defending their interests, as when the choices we make are private. In the former context it is necessary to assess the potential responses from other persons and to factor in these responses in making our decisions. When decisions are not likely to be public, then only intrinsic reasons for weighting the other person's welfare need to be integrated into the decision. Hence, there are at least two parallel, independent WTRs: (1) the intrinsic WTR ($_{intrinsic}$WTR), which sets an altruistic floor in the weighting of the other party's welfare, even when the actor's choices are not being observed; and (2) the public or monitored WTR ($_{monitored}$WTR) that guides an individual's behavior when the recipient (or others) can observe the behavior (Sell et al. 2009; Sell, Cosmides, and Tooby, forthcoming, a, b; Tooby and Cosmides 2008). Some altruism is motivated through love (involving a high intrinsic WTR), and some through fear, shame, or hope of reward – and the mechanisms involved are different. Kin-selected mechanisms produce intrinsic WTRs – these make you want to help your brother (at least in part) for his own sake. In contrast, threats from powerful others select for low intrinsic WTRs toward them (better for you if they did not exist), but significant monitored WTRs toward them (as when societies are organized around catering toward powerful and oppressive elites). Of course, monitored WTRs occur not just in such dramatic conditions, but ubiquitously through life (as when a person does something to win approval from a friend, spouse, or coworker).

Accordingly, for each individual J with whom individual X interacts, we should expect the mind to compute both an intrinsic $WTR_{x,j}$ and a monitored $WTR_{x,j}$. The level of the intrinsic $WTR_{x,j}$ governs how X trade-offs his or her welfare relative to the welfare of J when J's responses to the act do not need to be considered (when they do not know about the act; when they have no power to respond, etc.). The level of the intrinsic $WTR_{x,j}$ should be set by computations of the basic interdependence of X's own welfare and the welfare of J. That is, it can be reproductively advantageous for X to benefit J regardless

of whether or not she finds out, if she is his sister, best friend, or someone else whose existence, health, and capacity to act are valuable to him (Tooby and Cosmides 1996).

In contrast, the level of the monitored $WTR_{x,j}$ should be influenced not only by all of the same parameters as the intrinsic $WTR_{x,j}$, but also by a range of other cues which would have correlated with J's ability to inflict costs on or generate benefits for X upon detecting his or her actions. In general, examples of evolutionarily recurrent cues to J's ability to affect our welfare are (cf. Sell 2006b; Sell, Tooby, and Cosmides 2009):

- J's physical strength
- the size and degree of coordination of J's coalition
- J's social status, J's skills and competences
- J's access to resources.

This theoretical analysis predicts that, other things being equal, the human mind should be designed to place less weight on another person's welfare and more on one's own to the extent the other party is physically weaker. Sell and colleagues demonstrated that this relationship not only exists but is robust (Sell et al. 2009; Sell, Tooby, and Cosmides 2009).

Anger and WTRs

The usefulness of these tools in the analysis of human sociality can be seen, for example, in the analysis of anger (Sell 2006b; Sell, Tooby, and Cosmides 2009, forthcoming). In this view, anger (in addition to being an experienced psychological state) is the expression of an evolutionarily organized neurocomputational system whose design features evolved to regulate thinking, motivation, and behavior adaptively in the context of resolving conflicts of interest in favor of the angry individual. Two negotiating tools regulated by this system are the threat of inflicting costs (e.g., aggression, punishment) and the threat of withdrawing benefits (e.g., the downregulation of cooperation, relationship termination, ostracism). Humans differ from nearly all other species in the number, intensity, and duration of close cooperative relationships, so traditional models of animal conflict must be modified to more fully integrate this cooperative dimension.

If the human mind really contains welfare trade-off ratios as regulatory variables that control how well one individual treats another,

then evolution can build emotions whose function is to alter welfare trade-off ratios in others toward oneself. Anger is conceptualized as a mechanism whose functional product is the recalibration in the mind of another of this other person's welfare trade-off ratio with respect to oneself. That is, the goal of the system (its evolutionarily designed product, not its conscious intention) is to change the targeted persons' disposition to make welfare trade-offs so that they more strongly favor the angered individual in the present and the future. As in animal contests, the target of anger may relinquish a contested resource, or may simply be more careful to help or to avoid harming the angered individual in the future.

In human cooperative relationships, there is the expectation that the cooperative partner will spontaneously take the welfare of the individual into account. Hence, in cooperative relationships, the primary threat from the angered person is the signaled possibility that the angry individual will withdraw future benefits if the unsatisfactory welfare trade-off ratio is not remedied. If the withdrawal of this cooperation would be more costly to the target of the anger than the burden of placing greater weight on the welfare of the angry individual, then the motivational system in the target should be induced to recalibrate – that is, to increase her welfare trade-off ratio toward the angry individual, and so treat her better in the future.

The anger program is designed to recalibrate the angry individual's own WTR toward the target of the anger for two functional reasons. The first is that it curtails the wasteful investment of cooperative effort in individuals who do not respond with a sufficient level of cooperation in return. The second is that the potential for this downward recalibration functions as leverage to increase the WTR of the target toward the angry individual. In the absence of a cooperative relationship, the primary threat is the infliction of damage. In the presence of cooperation, the primary threat is the withdrawal of cooperation. Concepts that are anchored in the internal regulatory variable $_{\text{monitored}}$WTR include respect, consideration, deference, status, rank, and so on.

We call the ability to inflict costs to enforce welfare trade-off ratios in one's favor *formidability*, and brains should have evolved a set of programs that (1) evaluates one's own and others' formidabilities, (2) transforms each of these evaluations into a magnitude (a *formidability index*) associated with the person, and, in situations where cooperation is not presumed, (3) implicitly expect or accord some level of deference based on relative formidability. Among our ancestors

(see above), one major cue that an individual would have been able to inflict costs was his/her physical strength. Furthermore, in the human social world, the ability to inflict costs should correlate with social status, position in a hierarchy, economic resources, social allies. The programs estimating formidability should assess all such cues.

The approach briefly sketched above can be unpacked into a large number of empirical predictions about anger. For example, it was predicted that physical strength in men would be a partial cause of individual differences in the likelihood of experiencing and expressing anger. Other things being equal, stronger men are predicted to be more likely to experience anger and express anger; they should feel more entitled; they should expect others to give greater weight to their welfare, and become angrier when they do not. Furthermore, arguments precipitated by anger should reflect the underlying logic of the welfare trade-off ratio: The complainant will emphasize the cost of the other's transgression to her/him and the value of one's cooperation to the transgressor, and will feel more aggrieved if the benefit the transgressor received (the justification) is small compared to the cost inflicted. A series of empirical studies support both sets of predictions of this theory about the design of anger (Sell 2006b; Sell, Tooby and Cosmides 2009, forthcoming).

In the next section, we will describe how the more general notion of a WTR allows us to outline an evolutionary and computational concept of exploitation. After this examination, we use the distinction between monitored and intrinsic WTRs to analyze punishment and reconciliation as different evolved strategies to deal with exploitive persons and crimes.

Exploitation as a Display of a Low WTR

With this approach in mind, it is possible to examine exploitation more generally, as well as some possible adaptations to it. To begin with, fitness benefits flow from being with people who care about your welfare and attend to the welfare consequences of their actions. For this reason, selection should favor motivations to increase the degree to which you surround yourself with people who express a high WTR toward you in their actions. Conversely, there are potentially large fitness costs related to being around people who do not attend to or care about your welfare. An individual J with a sufficiently low WTR toward another individual X will not hesitate to use X as nothing more than a means to realize J's own ends

(as evidenced by rape, murder, slavery, the subjugation of women, economic exploitation, etc.).

Ancestrally, therefore, there was selection for avoiding engaging in social interactions with people with a low WTR toward oneself, and/or for limiting the power of such individuals to make trade-offs unfavorable to you (and your family and friends). To achieve this, one must be able to gauge the WTR that others express in their actions, especially toward you and those you most value. Relevant information will in part come from behaviour directed against third parties. Hence, to the extent that their WTRs toward third parties predict their dispositions to act toward those you value (including yourself), then indeed the mind should in general track the WTR in the interactions it is exposed to. In general, we should be interested in behavior that reveals WTRs. That is, in addition to whatever else we note and remember, behavior should be interpreted and remembered in terms of the WTR relationships it reveals. Whenever an action affects the welfare of both the actor and someone else, the act expresses (some) information about how much the actor values herself versus the impacted person – how much they are willing to trade off their own welfare to increase the welfare of someone else.

Accordingly, across cultures, acts that indicate a low welfare trade-off ratio toward oneself or a group to which one belongs (family, coalition, band, ethnicity) ought to be distinguished from other kinds of actions by our psychology, and should be viewed as problematic or *wrong* (an evolved conceptual primitive). Exploitative acts – those in which one party imposes large costs on another to gain a much smaller benefit – fulfill these criteria. Indeed, empirical studies provide evidence that an intuitive concept of exploitation exists as a cross-culturally universal feature of human social life. That is, low WTR acts by others toward ourselves or toward others we are positively involved with are viewed as morally wrong. Stylianou (2003) thus summarizes the current criminological data by arguing that there is consensus both within and between cultures about how to rank different harmful crimes. Of course, crimes do not exhaust acts that indicate a low WTR. Cues to low WTRs include inattention, failure to be aware of another's interests, lack of empathy when another experiences a gain or loss, a failure to remember the person, an unwillingness to listen to their statements about their welfare, ridicule, insults, emotional expressions of hostility, and so forth (for discussion, see Sell 2006b).

The Computation of Baselines and a Definition of Exploitation

Before a definition of exploitation can be reached, it is necessary to reflect on a final complexity with regards to the computational processes that tag acts as 'exploitive'. It is important to recognize that it is cognitively impossible for an individual to anticipate the impact of every act on the welfare of every local individual. The computational scope of such an evaluation is unbounded. For this reason, we did not evolve to respond to behavioral choices as if they were the product of unbounded computation. Instead, humans implicitly accept scope limitations on their definitions of acceptable versus wrong conduct, and do so because we evolved mutually consistent cognitive procedures for framing situations.

For one thing, this means that an act of exploitation cannot simply be defined as acting with a low WTR toward another. This is because most of our acts do not and could not take into account the cacophony of mutually irreconcilable needs of everyone in our social universe. Instead, our minds evolved the automatic practice of setting the starting point (baseline) from which welfare is viewed as having been increased or decreased at the level that would have existed prior to or in the absence of the act. For instance, if I have gathered food, I might feed myself instead of feeding a random stranger without this being regarded as an instance in which I exploited the stranger. This is because, given this cognitive system for implicitly generating baselines, by eating food I have gathered I have not lowered the stranger's pre-existing welfare. Rather, I have chosen to not increase it. It is important to recognize that this is not a fact about the world, but the one framing out of many that evolution has caused our minds to adopt. From a physicist's perspective, this case is similar to a case in which I eat food that someone gave to the stranger: in both cases, my eating the food reduces the stranger's welfare by the same amount when compared to other acts I might have committed. However, the mind evolved to compute baselines in a way that leads these two cases to be treated differently.[2] Consequently, in one case the welfare of the other is seen as unchanged (an absence of exploitation – and an absence of charity), and in the other case the mind computes that the welfare of the other person is lowered (an act of exploitation). Additional rules for setting baselines exist as well, as in social exchange and collective actions (in which a failure to act can be reframed as imposing a cost below a baseline that is computed by considering the state of affairs if expected or required cooperative

acts had been mutually carried out). The key point is that what counts as an exploitive transgression is defined with respect to these computed welfare baselines.

Hence, to a first approximation, exploitation can be defined as an act that (1) expresses a welfare trade-off ratio that is either too low or negative toward an individual or group, and (2) imposes a cost that reduces the welfare of the impacted individual or group below the baseline they were entitled to. Infliction of minor costs to gain major benefits (for someone within one's cooperative sphere) is less often viewed as exploitation because it clashes with the input conditions for recognizing low WTRs. A situation that would render such behavior more problematic is if one attempts to conceal it – thereby covertly removing it from the reciprocity system.

Crime as Exploitation

There is a substantial body of evidence indicating that natural selection has fashioned anti-exploitive psychological mechanisms (on cheater detection, Cosmides and Tooby 1992 and 2005: on punitive sentiment as an anti-free-riding device, Price et al 2002 and Tooby et al. 2006a). More relevant to criminal justice is evidence about possible evolved psychological responses and defenses to the infliction of costs outside of the context of defection in cooperative endeavors (for an overview, see Duntley 2005; for empirical studies on rape avoidance mechanisms, see Chavanne and Gallup 1998; Petralia and Gallup 2002; Thornhill and Palmer 2000). The adoption of a WTR-oriented approach enables us to specify adaptive problems that cut across such types of cost-infliction. Despite their differences, theft, rape, hit and run driving, assault, and synagogue, church, or mosque desecration are all reliable signals of low WTRs, and therefore probabilistic signals that their perpetrators are more likely to inflict future damage than others who have not committed these acts. In this section, we will review a number of ways in which different crimes can be viewed as acts of exploitation in this sense.

Essential to exploitation is the imposition of high costs on a person or group, or the imposition of costs outside of a cooperative relationship (Duntley and Buss 2004; Sell 2006b; Tooby, Thrall, and Cosmides 2006). Harmful acts cross-culturally perceived as antisocial or crimes (see Ellis and Hoffman 1990), such as rape, theft, assault, and murder, all satisfy this requirement. Modern criminology has repeatedly documented that the perceived seriousness of

such acts are highly influenced by the degree of damage done to the victim's welfare (Stylianou 2003; Warr 1989). This is in line with the evolutionary concept of exploitation as expressing a low WTR by means of imposing a cost.

However, from an evolutionary perspective, the perception of an act as exploitive is not just governed by the magnitude of the cost, but also by the relationship between the imposed cost and the benefit gained by the perpetrator. Thus, keeping the costs constant, we should also expect the level of benefits generated for the perpetrator to negatively influence the degree to which acts are perceived as exploitive. Sell et al. predicted and found this relationship in the degree to which anger is provoked by a situation (Sell, Tooby and Cosmides, forthcoming). Rossi, Simpson, and Miller (1985) found that people rate the crime of theft more mildly when it is motivated by an attempt to provide food for the perpetrator's family in need. Such intuitions are ubiquitous, as when, for example, they are scathingly invoked by Anatole France when he remarked that "The law, in its majestic equality, forbids the rich as well as the poor to sleep under bridges, to beg in the streets, and to steal bread." (France 1894).

The majority of crimes we are confronted with (e.g., in the mass media) are directed against others than ourselves. Importantly, acts expressing low welfare trade-off ratios among third parties are also adaptively significant, and so it is expected that our evolved machinery respond to (some of) them as well. Harmful acts committed by a perpetrator P and directed against a victim V can be consequential for a third individual T in at least two ways. First, to the extent that the victim is a genetic relative or valuable social partner, T suffers directly by the lowering of the fitness of the relative or the capacity to act of the social partner. Second, the exploitive acts might raise the threat that T will be a future direct victim of the perpetrator. The adaptive problems associated with each scenario contain specific computational problems, which a third party witnessing some kind of cost-infliction needs to solve.

The estimation of the direct costs requires an assessment of the value of the victim V to the observer T. In line with this, Hembroff (1987) finds that people react more strongly to crimes directed against well-integrated members of the community. (Such acts might also provoke a stronger response because the collective reputation of the community for deterrence is more strongly threatened when central figures are attacked.) Landsheer and Hart (2000) found similar

effects among adolescents, who react more strongly to an offense when they know the victim than when they do not.

Having observed an act of exploitation, estimating whether the same perpetrator is likely to victimize oneself is a more cognitively complex task. As Panchanathan and Boyd (2003) have pointed out, it requires the observer to assess whether the perpetrator's infliction of costs on the victim was motivated by a general exploitive disposition or whether the specific act was part of a more local conflict between the perpetrator and the victim. In terms of WTRs, it requires the observer to assess whether the act expresses a low WTR solely toward the victim, or a low group-WTR toward a set that includes both the observer and the victim as members. If it expresses the latter, the observer is confronted with a potentially severe adaptive problem even if she is not intrinsically invested in or otherwise allied with the victim.

Moreover, such an act of exploitation can have another important negative consequence: Other parties are observing the act. Some of these parties express acceptable WTRs purely for prudential reasons, restraining themselves because of the potential consequences to them if they attempted exploitation. If acts toward members of a set or coalition are not responded to, this invites these others to exploit members of the set (Tooby, Cosmides, and Price 2006). Deterrence is an evolved function of revenge, and operates in a parallel fashion for individuals and groups. It is important to note that coalitional identities are automatically categorized by the mind (Kurzban et al. 2001). It is this automatic categorization and generalization that makes victimization of one member of a coalition an advertisement of the vulnerability to exploitation of other members of the coalition. Therefore, other members of a coalition (or their social allies) have an interest in advertising that victimizations of its members will not go unpunished (Tooby, Cosmides, and Price 2006). It strikes the mind as intuitively reasonable to recognize so-called hate crimes as a category, and to treat them more punitively; this reaction makes sense given that observers can use acts committed against a single group member to infer that other members of the group are now vulnerable to exploitation.

Due to these selection pressures, we should expect our minds to (1) be able to navigate adaptively between individual- and group-WTRs, (2) assess whether specific behaviors are guided by one or the other, and (3) be highly sensitive to cues that, under ancestral conditions, reliably predicted a low WTR toward groups, coalitions,

or social categories that we either belong to or intrinsically value. For example: If a harmful act occurs randomly without prior provocation, it could constitute one important cue that the threat the perpetrator poses generalizes to others beyond the victim. In line with this, Rossi, Waite, Bose, and Berk (1974) show that violence between people who have not interacted prior to the incident is seen as more serious than violence between individuals who knew each other beforehand. Similarly, classical studies in the anthropology of law indicate that, in some societies, collective reactions against violent persons are carried out only when these are seen as a potential danger to all community members (Hoebel 1964).

Criminological research strongly suggests that acts do not need to be physically harmful to be perceived as seriously offensive (Warr 1989). To the extent that an otherwise harmless act violates some locally accepted moral norm, it can generate strong indignation and punitive sentiment (Tooby, Thrall, and Cosmides 2006; Tooby, Cosmides, and Price 2006). Flag burning or naming a stuffed animal "Mohammed" are modern examples, but the anthropological literature contains numerous other accounts of strong disapproval or punitive responses to norm violations (Hoebel 1964). In general, showing intense disregard for those symbolic markers or emblems that function as indices of the status of the group is cross-culturally viewed as an "outrage" (Tooby, Thrall, and Cosmides 2006). Mistreatment of a group status index powerfully communicates that the perpetrator does not value the group and its members, and does not believe that members of that group have sufficient collective support or formidability to defend their interests. It presages future exploitation. This can mobilize a collective response to advertise the group's strength and vigilance in defending their status and their members. In a similar vein, symbolic transgression directed against a specific individual should be viewed as exploitive by that individual and others valuing her, because it involves the public devaluation of this person's status (and an advertisement of her low formidability).

Summing up, the theory that the human motivational system contains welfare trade-off ratios, which are embedded as control elements in decision-making systems, enables us to understand crime as exploitation and, hence, as acts expressing a low WTR by means of the imposition of a cost. From the victim's perspective, the harmful act's degree of seriousness or perceived wrongness should, to a large extent, be determined by the ratio between the harm caused and the benefits generated. From a third-party perspective, the seriousness

should be also influenced by the value of the victim to the observer, the extent to which the exploitive act was motivated by exploitive tendencies that can be generalized to the observer or the observer's group, and the precedent the act sets for the future if it is not responded to. If the tendency to act exploitatively is clearly limited to a specific perpetrator-victim pair (e.g., a husband and wife), then responses of others are expected to be more limited. In contrast, a general disregard for the welfare of the members of a group can be displayed by publicly exploiting an individual bearing a coalitional marker or having a widely-known coalitional identity. Table 5.1 displays an overview of these varieties of exploitation, which intuitively are perceived as "wrong."

Based on these and other cues, we expect the mind to automatically compute an index of the seriousness of exploitive acts; i.e., to calculate the degree to which the expressed WTR of an individual is discrepant from an acceptable WTR, given the parameters outlined above. This index is one element that underlies the intuitive reaction that a particular action is wrong. The function of the feeling of wrongness is to motivate individual or collective action to redress the fitness threat posed by those who commit certain acts. For this reason, we are designed to feel that something is more intensely wrong if it happens to us, our family, our friends, or our group; that it is more wrong to the extent that it predicts or invites bad future outcomes (even if the actual damage, as with status outrages, is minor); that it is more wrong if the benefit gained was small in proportion to the cost inflicted; and so on.

Table 5.1 A partial list of factors setting WTRs and indicating low WTRs.

Factors inducing X to "value" Y (Factors that set WTRs)	Acts intuitively perceived as "wrong" by person X (Indicators of a low WTR relevant to person X)
X's and Y's relative fighting ability	Intentional imposition of a large cost for small benefit on X
Size and "strength" of X's and Y's coalitions	Direct challenge to X's status or authority
Y's social skills and competences	Lack of empathy for X
Y's access to resources	Symbolic transgressions against X or X's group
Degree of relatedness between X and Y	Acts imposing large costs on someone valuable to X
Y's mate value	Costly acts motivated by low-group WTRs toward a group in which X is member
...	...

Evolved Strategies for Responding to Exploitation: An Overview

Because our ancestors were continuously subject to threats of exploitation during their evolutionary history, we expect that selection favored the evolution of neurocognitive programs to respond adaptively to exploitation. In other words, the relevant adaptive problem is the existence in the local social world of individuals who hold low WTRs – a state that reliably predicts future cost-infliction. How should humans be designed to respond to this problem? Responses (each of which is reflected in the formal or informal criminal justice systems of various cultures) include:

1. killing the perpetrator, which permanently removes the threat
2. expelling the perpetrator from the social world through ostracism or confinement; or, on an individual basis, not engaging in future cooperative endeavors with those who have cheated
3. punishing the perpetrator through infliction of costs or withdrawal of benefits
4. reconciling with the perpetrator.

There are two major kinds of benefits that result from active responses targeting the perpetrator: (1) the direct benefit that arises from a response's impact on the perpetrator (e.g., the perpetrator is deterred from misbehaving in the future), and (2) the indirect benefit that arises from the impact of the response on third parties (i.e., others are deterred from behavior that is exploitive). Importantly, these responses largely presume a sufficient nucleus of like-minded individuals to enforce them, compared to a sufficiently low number of individuals who would oppose them. We will only touch briefly on the effects of such populational characteristics. If, however, there is not the collective strength to actively operate on the perpetrator, remaining responses are: (1) do nothing and put up with it; (2) avoid the perpetrator, to the extent possible; or (3) leave the social group (hunter-gatherers frequently settle long-simmering conflicts by group fission; Lee and DeVore 1968). The rest of the chapter is oriented toward discussing each of the four categories of responses listed above. While we begin by discussing killing and ostracism, our main focus is on punishment and reconciliation. Killing and ostracism as evolved responses to exploitation will only be dealt with on a more cursory level.

If, over evolutionary time, individuals frequently encountered others whose prospective existence was a net fitness cost (as exploiters, competitors, or impediments to realizing fitness gains), then evolution should plausibly have favored circuits that motivate killing, when the costs and risks are not too great (Buss, and Duntley 2003; Daly and Wilson 1988). Collective action reduces the per individual costs of killing as a tool to remove social threats, and so sentiments favoring the social deployment of killing (execution) are expected to be a widespread feature of the ethnography of the collective treatment of exploiters (i.e., criminal justice). One important element that moderates such motivational outputs is the connection that some community members (such as family members, friends, or sympathizers) may have with the perpetrators. To the extent that perpetrators are connected, social conflict is engendered over inflicting such serious and irremediable harm on individuals who are valued by at least some others in the community.

Ostracism (as with group expulsion or confinement) is another way of limiting the costs inflicted by exploitive individuals. One benefit is that the potential malefactor is physically prevented from reaching potential victims during the enforced absence – like execution, it incapacitates them, but only for the duration of the confinement. It also has significance as an act of punishment, which brings us to the recalibrational theory of punishment and reconciliation as evolved responses to social exploitation, which continue to shape attitudes toward criminal justice. Our next objective is to outline this theory.

A Recalibrational Theory of Punishment and Reconciliation

If the problem is exploitation, and the threat of future exploitation is exhibited in acts that express low welfare trade-off ratios, then what does this imply for how our minds ought to be structured to respond to this threat? One strategic component acts by solving the problem physically or spatially. This creates the basis for the above-discussed responses such as capital punishment, the physical restraint of potential malefactors (e.g., confinement), and the ejection of the malefactor from the social group. Another strategic component acts by solving (or mitigating) the problem motivationally. In this case, our evolved responses should constitute strategies that are organized to act through the evolved computational architecture of the malefactor's motivational system. If the problem is that the WTRs in the malefactor's minds are too low, then the solution should be to

take actions that recalibrate the WTRs in the malefactor upward. Hence, we expect reactions to displays of cost-imposition guided by low WTRs, i.e., exploitation, to be strategies with the goal of upregulating the level of the exploitive person's WTR toward relevant potential victims (Sell 2006b; Sell et al. 2009, forthcoming; Tooby et al 2006a). Below, we will outline punishment and reconciliation from this recalibrational perspective. Although both strategies seek to upregulate the exploitive person's WTRs, punishment primarily targets the monitored WTRs, while reconciliation targets intrinsic WTRs (sometimes along with the recalibration of monitored WTRs by positive incentives).

Punishment and the Recalibration of Monitored WTRs

In general, punishment can be defined as the conditional imposition of costs on an agent who has committed an act because the agent has committed that act. Punitive strategies emerge from the logic of conflict, and so are not recent cultural inventions. A range of nonhuman animals (including higher primates) display punitive tendencies against exploiters or antagonists (Clutton-Brock and Parker 1995; de Waal 1992), suggesting that this form of interaction has been with our evolutionary lineage for tens of millions of years. Similarly, accounts of revenge as a motivation are cross-culturally and historically ubiquitous (Daly and Wilson 1988; Jacoby 1983).

More specifically, circuits that motivate revenge evolved as part of a system for defending the organism's interests against acts of exploitation that would occur in the absence of revenge circuitry. There are a number of elements to the revenge system:

- an element that computes baselines
- an element that detects cost-imposition with respect to those baselines and implicated WTRs
- an element that generates a calibrated intensity of punitive sentiment – specifically, with goals (1) to cause the experience of suffering in the malefactor and (2) to pair the inflicted suffering with the communication to the malefactor of what acts the delivered suffering is repayment for
- an element that modulates the intensity and expression of the infliction of suffering in light of the relative formidabilities of the potential punishers and the potential targets of revenge.

The latter elements relate to the fact that, on the one hand, the system must provide its own strong and distinct motivational imperative,

because carrying out punishment is often costly and painful, lowering the achievement of other competing goals and motivations. Hence, the core of this system is the generation of punitive sentiment (i.e., the motivation to inflict suffering as a response to a prior bad action). This core, we argue, evolved specifically to recalibrate the WTR in the psychological architecture of the malefactor (although, when the net value of the perpetrator is negative, then the revenge system may generate the motivation to kill rather than to recalibrate). On the other hand, the link between felt punitive sentiment and actual punitive behavior is affected by several social contingencies (cf. the final element in the above list). As the social scale gets larger, for example, one complication is the negotiation of the transfer of the agency of revenge from harmed families to the community – many of whom may be allies and supporters of the perpetrator. The fluid and contentious concept of "fairness" emerges from the dynamics of integrating the voices of other community members, in proportion to their power and influence in the community.

There is now a sizeable experimental literature documenting that individuals are indeed sometimes punitively motivated and, hence, are willing to incur costs to punish others, including (at least in experimental games) the impulse to punish those who treat third parties unfairly (Cameron 1999; Fehr and Fischbacher 2004; Fehr and Gächter 2002). This also holds true to a greater or lesser extent for games conducted in a range of nonindustrial small-scale societies (Henrich et al. 2004). Neuroscience results suggest that the motivation for punishing cheaters is in part created by a heightened activation in the brain's reward centers – rewards that (under certain conditions) can outweigh financial disincentives to punish (de Quervain et al. 2004). Furthermore, experimental studies indicate that punitive sentiments appear to be specifically designed to be elicited by exploitive acts (Price et al. 2002; Tooby, Cosmides, and Price 2006).

This punitive sentiment has been explored from a number of theoretical evolutionary standpoints (Boyd and Richerson 1992; Boyd, Gintis, Bowles, and Richerson 2003; Frank 1988; Trivers 1971; Price et al. 2002; Tooby, Cosmides, and Price 2006). However, it seems worthwhile to explore whether analyses applying the WTR framework might offer additional clarity. How can we make sense of punishment in the light of the theory of welfare trade-off ratios? Given that exploitive acts are the result of low WTRs, punitive counterstrategies can be interpreted as efforts designed to recalibrate the

decision variables in the minds of the targets of punishment – that is, the goal of punishment is the upregulation of the exploitive person's WTR toward prospective victims.

In this regard, we can conceptualize actual punitive cost-infliction or credible threats about it as (1) signals about a socially created contingency ("engage in prohibited behavior, and your welfare will be lowered"), and (2) signals about formidability, which makes this contingency real ("we have the ability to inflict costs, if you don't defer") (Sell 2006a; Sell, Tooby, and Cosmides 2009). Therefore, punishment taps directly into the input conditions of the systems that function to calibrate the target's monitored WTR toward the punisher(s).

Notice the consequentialist nature of the recalibration, if indeed successful: The adverse consequences of not caring about the welfare of the punisher (or those the punisher is acting to protect) are fully contingent on the punisher's ability to monitor the exploiter's behavior. Accordingly, we should expect the effects of cost-infliction to be confined to recalibrating the monitored WTR, to the extent that targets can confidently distinguish when they are being monitored or not. At the level of conscious experience, the successful punitive strategy should instill fear of the consequences of future exploitation in the perpetrator.

Experimental economic games do show that punishment powerfully reduces free-riding in a way consistent with the recalibrational perspective (Fehr and Gächter 2000; Ostrom, Gardner, and Walker 1994). The set-up of these repeated games lets players have full information about the behavior of other players, whereby any punishment-induced recalibrations of monitored WTRs can take full effect on the punished player's subsequent behavior (as he or she will in fact be monitored). Importantly, if the possibility of punishment is later removed in these games, we see rapid increases in free-riding (Fehr and Gächter 2000; 2002). Consistent with the claim that punishment recalibrates monitored WTRs, the effects of past punishment on social behavior are, thus, contingent on the possibility of future punishment. Similarly, both experimental studies and criminological macro-studies of crime trends seem to indicate that the effectiveness of deterrence is primarily conditioned by the certainty of detection and punishment (Klepper and Nagin 1989; Hirsch et al. 1999). This seems to indicate that punishment-induced recalibrations only guide behavior to the extent that this behavior is experienced as being monitored by punitive agencies.

The socially replicated elements (rituals) of punishment in formal and informal criminal justice contexts should in important ways be deducible from the input conditions of the evolved motivational and cognitive architecture responsible for resetting monitored WTRs (especially the revenge subsystem). The structure of the architecture suggests that punishment ought to have the following characteristics:

- The act that is being punished needs to be specified to the perpetrator – that is, it is not sufficient to harm the individual. The individual must know why he is being made to suffer.
- The harm inflicted must be communicated as having been caused by the perpetrator's prior exploitive act – that is what defines it as fitting the evolved mental category *punishment*. You are punished for what you did. (In modern societies, there is a judicial indictment that specifies the act. The absence in some contexts of this communicative step can make justice appear Kafkaesque.)
- The perpetrator is confronted with why the act was harmful to other persons or to some other entity (the group, deities, etc.), why the perpetrator's justification (if any) is insufficient to excuse it, and the guilty finding is often accompanied by a description about what was "outrageously" exploitive about the crime.
- The process of punishment is communicated as being "right" or legitimate. That is, that there is an agreement among enough social actors to define this social reality as the proper outcome. This agreement is recognized to reset baselines about what is legitimate, so that the punishment itself is not something that properly could provoke a retaliatory response from the punished individual. (Ordinarily, a person might reasonably attempt to nullify the confiscation of his property, or to retaliate against someone who has inflicted pain. But punishment resets the baselines after the punishing act, so that everything is then viewed as being at a new equilibrium that replaces previous arrangements).
- There is an attempt to communicate that the agencies of punishment are formidable – powerful and to be feared. Thus, another important element of punishment is that it requires a power differential that allows the punishers to inflict greater costs than they incur in the effort. This is why collective punishment looms larger than individual revenge. Hence, the infliction of punishment signals a dominance-subordination relationship, and the acceptance of punishment by the target signals the target's ratification that the subordination is legitimate, along with the punishment's new baselines.
- The punishers expect the punished individual to signal his or her social subordination to (deference to, respect for) the punishing entity.
- There is the attempt to communicate that the agencies of punishment will remain vigilant about future behavior (i.e., monitoring will continue).

- Finally, the magnitude of the inflicted cost used as punishment is modulated by the magnitude of the exploitation. Intuitions should set the magnitude of the punishment so that it is sufficient to reset the perpetrator's WTR system, so that future exploitation would no longer be considered as worthwhile by the perpetrator. If previous punishment was discovered to be insufficient to reset a repeat offender's WTR system, the collective or individual punisher feels the impulse to increase the intensity of punishment until it does work. If this process is a persistent failure, then this may activate sentiments favoring return to the physical strategies of incapacitation, such as execution or permanent confinement.

In short, there is an organized communicative component to the process of punishment derived from the structure of the punisher's motivational system. This communication goes from the parties enacting the punishment toward both detected perpetrators of exploitation, and toward undetected potential perpetrators. Both common experience and the philosophy of punishment support the view that punishment has an expressive or communicative aspect, which takes it beyond simple cost-infliction (see, for example, Duff 1996; Nozick 1981).

It is important to recognize that this system works if the target of punishment can correctly interpret the full range of future behavior that the punishers intend to proscribe. Computationally, this is an important issue, because possible behaviors are infinite. Moreover, the contents of lengthy legal codes are largely unknown to citizens in modern societies, and in small-scale societies legal systems are generally not codified. We suggest that what allows this system to work is that the minds of perpetrators and punishers alike intuitively grasp that what is most commonly at issue are welfare trade-off ratios expressed in actions outside of converged upon baselines. This is experienced as implicit and self-evident, and encapsulates (but does not exhaust) many widely shared intuitions about what is right and wrong.

Reconciliation and the Recalibration of Intrinsic WTRs

Punitive sentiments evolved because punishment works – but only within certain limits. If punishment is effective because it upregulates monitored WTRs, then punishment should leave unaffected behavior that potential exploiters confidently believe is unmonitored by punitive agencies. Because a great deal of behavior is potentially unobserved, pure reliance on punitive strategies leaves exploitation

that takes place in a broad range of conditions uncountered. That is, resetting only monitored WTRs is the grave defect of punitive strategies, because it only works for acts likely to be monitored. (Indeed, the actual decision function in potential exploiters ought to approximate the probability of being monitored, scaled by the magnitude of the costs inflicted if detected, compared to the benefit to the perpetrator of exploitation.) Viewed this way, the core of the problem of exploitation is that potential exploiters have too low an intrinsic welfare trade-off ratio toward potential victims. If they had high intrinsic WTRs toward others, they would not be motivated to injure them in pursuing their own interests, and there would be no problem.

This raises the possibility that another family of evolved strategies, complementary to punishment, could be designed to manipulate and upregulate the intrinsic WTRs of potential exploiters toward their victims (or groups and communities). If intrinsic WTRs could be generally increased, this would prevent exploitation, even during situations that punitive agencies could not monitor at sufficient frequencies. Here we will outline how nonpunitive strategies against exploitation might operate, and how their properties derive from the computational architecture of the intrinsic WTR motivational subsystem. This set of strategies can be subsumed under the heading "reparative strategies". A virtue of reparative strategies is they maintain and uphold deep social relationships, while punitive strategies rupture them. The limitation on this family of strategies is, however, that the long-term nature of payoffs over evolutionary time should make the upregulation of intrinsic WTRs difficult to achieve. The difficulty is especially aggravated in mass societies where people routinely interact with large numbers of socially distant individuals who they have little reason to have an intrinsic interest in.

Theoretically, an intrinsic WTR from X to J should be set to reflect how much changes in the welfare of J affect the fitness of X. Biologists recognize that genetic relatedness selects for an evolved system of family sentiments that makes humans have high intrinsic WTRs toward their children, and other close relatives (Hamilton 1964; Lieberman et al. 2007). Kin selection places some limits on exploitation toward kin, but not on the far larger category of non-kin. However, individuals who share interests or who broadcast positive externalities to others are commonly also bound together in relationships of fitness interdependence – indeed, often more strongly than kin are (Tooby and Cosmides 1996). These relationships – termed

deep engagement relationships – are an important feature of human sociality, and are believed to underlie friendship, romantic love, several added components of family sentiment (over what kin selection explains), and a general appetite to cultivate relationships in which one is valued in a way that makes the self difficult to replace to engaged others. Ideally, the well-situated ancestral human was best off when enmeshed in a network of relationships in which he was valued as irreplaceable by others (i.e., had low substitutability). Such relationships often provoke mirroring: that is, one of the things that make individuals value others is the fact that these others value them in the first place. Information signaling this kind of friend-, mate-, family-, coalition-, or community enmeshment ought to set intrinsic WTRs upwards toward other individuals in those relationships.

Humans are intensely social and cooperative, and throughout our evolution it was critically important to have reliable associates who valued you strongly enough to assist you when you were in dire need. This appetite for others who can be relied on is powerful. Exploitative behavior reduces the degree to which others value the exploiter, making highly exploitive individuals particularly vulnerable to risks raised by a lack of sufficient intrinsic social support. When individuals who rely on exploitive formidability in their dealings get sick, or find themselves outnumbered or ambushed, their fortunes can reverse rapidly. This leads to the expectation that such individuals will be (1) more paranoid or suspicious of the motives and support of others; and (2) hungry for evidence that they are valued. (Individuals who are unusually antisocial may have become that way because they (1) received signals throughout their life history that no one intrinsically valued them, (2) chose to pursue exploitive benefits that accrue to differential formidability over the gains of cooperation, or (3) have an impaired ability to detect or respond to signals designed to regulate the intrinsic WTR system – perhaps from developmental anomalies or genetic noise.)

If the emotion program of anger is triggered by the recognition that another has engaged in an act that expresses too low a WTR toward you, there is a reciprocal emotion program that is triggered by the converse: guilt. Guilt is an emotion program that is triggered when you receive information that you have engaged in an act that expresses too low a WTR toward another, given their value to you In this view, guilt is a recalibrational emotion program whose function is to upregulate your intrinsic WTR toward an individual (and/or your monitored WTR toward an individual, if the action is

detected). This recalibrational process is triggered when the interpretive system in your brain detects that your prior WTR (or its expression in action-decisions) has been too low, given the value of the victim to you. If you carelessly back your car into your mother, breaking her legs and rendering her amnesic, you feel guilt, even if no one but you knows what happened and, therefore, there is no punishment to fear.

Hence, we should expect that reparative strategies should involve conveying information to the exploitive person that he or she has underestimated the true magnitude of the harm inflicted; or underestimated the true value of the relationships jeopardized; or overestimated the gain to the exploiter of acting selfishly, when compared to the magnitude of the loss inflicted on the other party. If the intervention is successful, the target should realize how his or her own welfare is causally connected to the welfare of the person the target has been damaging (Tooby and Cosmides 1996). Guilt provides data formats in which recalibrational upregulations of intrinsic WTRs are made accessible to other behavior-regulating algorithms (Tooby, Cosmides, Sell, et al. 2008; Sznycer, Price, Tooby, and Cosmides 2007). In other words, strategies to upregulate intrinsic WTRs should be guilt-inducing strategies. (Of course, if the "debt" or guilt is too great ever to be discharged, and the net future payoff of the relationship will be negative, then guilt-inducement may sever rather than repair social ties.)

Because the level of X's intrinsic WTR toward J is set in part by the degree to which J values X in return (Tooby and Cosmides 1996), we should expect guilt-inducing strategies to make the perpetrator's stake in the exploited person more salient. Provoking guilt in the face of a transgression through confronting or reminding is a common reparative strategy. Participants should emphasize the high costs inflicted by the transgressor and the paltry or transient benefits reaped. Most important of all, the value of the victim to the transgressor can be stressed. This might be done by dwelling on the history of benefits the transgressor has received from the victim (she is your mother, after all), or by stressing the previous commitment of the victim to the relationship (see, Sell et al., forthcoming, a, for discussion of WTRs in arguments). In line with this, empirical analyses of verbal strategies to elicit guilt show that the predominant strategy is to remind the other how his or her behavior violates obligations central to a relationship (Vanglesti, Daly, and Rudnick 1991). Another possibility is to signal that the exploitive path threatens to terminate the relationship (in a small social group, this can mean ostracism). These signals should

motivate the exploitive person to reconsider the benefits associated with the relationship, and to act on that appreciation.

Reciprocally, the emotional display accompanying the emotion of guilt – the expression of suffering by the perpetrator at the contemplation of the harm he has inflicted – is a form of evidence that is relevant to computing how much recalibration is required by the transgression. Indeed, if the remorse is both genuine and great enough, this may indicate that the intrinsic WTR may be high enough after all. That is, it may not need recalibration to prevent the perpetrator from transgressing again. This would occur, for example, if the transgression reflected a lack of understanding or forethought, and was not a reflection of permanent comfort with imposing costs on others. In contrast, punitiveness by judges and juries is intensified if the perpetrator expresses no remorse – a situation indicating that values of regulatory variables in the perpetrator's cognitive architecture are set to commit the same act again, should the opportunity arise.

One component of the reparative approach focuses on facilitating recalibration in the perpetrator by emphasizing neglected value information about the victims, while another component focuses on generating new information by changing the external situation the malefactor is embedded in. For example, a direct provision of added benefits from victims to the malefactor advertises a change in the value of the victim-malefactor relationship to the malefactor. Under the right circumstances, this can trigger reparative guilt, revising the malefactor's intrinsic WTR upwards. This may be the moral intuition behind the Biblical Sermon on the Plain, where the following strategy against exploitation is suggested: "Love your enemies, do good to them that hate you. Bless them that curse you, and pray for them that despitefully use you." (Luke 6:27–36; see also the parallel Sermon on the Mount, Matthew 5–7). The benefits could take a wide variety of forms from actual material benefits to immaterial ones such as providing comfort, help or support. For example gift giving, the sharing of resources and physical contact are cross-culturally recurrent elements of reconciliation rituals (Fry 2000). Marriages create a confluence of fitness interests, and are commonly used to end feuds; similarly, if more rarely, there are a variety of ethnographic cases of reparative cross-adoptions. Observations of children similarly indicate that transferring benefits indeed facilitates reconciliation (see Fujisawa, Kutsukake, and Hasegawa 2005; Ljungberg, Horowitz, Jansson, Westlund, and Clarke 2005). Finally, there is evidence suggesting that this kind of strategy may be used among nonhuman primates. Thus, grooming – a benefit normally exchanged among social partners – plays an important role in conciliatory practices in certain species (de Waal 1996). It is important to recognize, however, that the actual function of primate conciliatory practices may be quite limited (Silk 2000; 2002); that is, the signal may be restricted to: "The fight is over for now", rather than reflecting the start of a more expansive social repair process.

This latter component underscores some of the inherent evolutionary problems of reparative strategies. Thus, if increasing the delivery of benefits to exploiters were an unconstrained general strategy, it would be a failing strategy, since it would simply provide another incentive for exploiters to exploit. Consequently, this strategy is deployed only under narrow circumstances, when the goal is to re-enmesh the target into a more pro-social relationship or relationships. If, for example, the victim is too weak to be punitive, attempts to increase how much the exploiter values the victim may be the best of a poor set of options. Such a response is on a continuum with appeasement, the psychological phenomenon of "identification with the aggressor," and Stockholm Syndrome – responses elicited by the relative weakness of the victim compared to the perpetrator. It is a high-risk strategy, prone to failure. Accordingly, we predict that when individuals favor reparative strategies they should do so less confidently than when individuals favor punishment.

That conciliatory strategies are risky is revealed (for example) in feuding, which is ethnographically ubiquitous. Thus, feuding and many exploitive interactions are more symmetric than asymmetric and involve patterns of adversarial interaction in which both parties have been negatively impacting the other, leading to chronic losses for both sides. Attempts to change such a dynamic are often hampered, because pro-social acts from one side can be mistaken for weakness, and hence invite heightened extortive efforts. Indeed, the enactment of punitive strategies signals a dominance relationship, and so the acceptance of the punishment may be resisted because it signals acquiescence to social subordination (the underlying theme acted out when defendants refuse to recognize the authority of a court to try them). To overcome such deadlocks, third parties may attempt to organize reconciliatory events in which the acts are designed to heighten *mutual* valuation of the parties to each other, leading to mutual changes in WTRs.

Reconciliation and Recalibration of Monitored WTRs

We have now discussed how reparative strategies might recalibrate intrinsic WTRs. However, it is important to recognize that the term "reconciliation" may refer to two distinct, if related, phenomena, which the recalibrational framework may help to clarify, and which reveal how reconciliation also might affect monitored WTRs. In general, the evolved organization of recalibration involves acts or signals that are designed to initiate a process of recalibration in a target. When the function of recalibration is finished, the process of recalibration should terminate. Thus, the social process triggered by transgression moves through sequential phases: (1) recalibrational efforts (in the senders) linked to processes of WTR recalibration in the target; (2) termination of recalibrational efforts and return to normal relations once recalibration has occurred (or is presumed to have occurred). One temporal boundary that might get referred to as "reconciliation" is the signal that the recalibration effort is finished – a termination point that is more clearly recognized if the target gives signals of successful recalibration, such as remorse, subordination, or an intention to act with higher WTRs in the future. This kind of reconciliation forms the endpoint of a reparative process that, as discussed, attempts to recalibrate someone's intrinsic WTRs.

A second, related meaning of reconciliation is the attempt to lower someone's disposition to inflict costs by embedding them in conditional cooperative relationships from which the target has the potential to either derive large benefits or to lose them. Being subject to new conditional cooperative relationships revises monitored WTRs, not intrinsic WTRs, because subsequent exploitation will lead to the withdrawal of conditionally delivered benefits if it is detected. It is important to recognize that efforts to domesticate a transgressor by embedding him in conditional cooperative relationships are not themselves punishment – the transgressor is better off. Both reparative strategies are closely related to each other, and are often found together because they mutually reinforce each other, and often depend on bringing about the same types of social arrangements.

To Punish or to Reconcile?

It is important to recognize that punitive and reparative strategies exist in tension with each other. The infliction of costs – even when those costs are inflicted as retaliation for prior exploitive acts – sends

a signal that the punishers do not feel inhibited in inflicting costs on the target of punishment. In other words, severe punishment is itself a signal of a low intrinsic WTR. This information should be picked up by the representational systems responsible for calibrating reciprocal intrinsic WTRs in the punished individual. Thus, while punishment might productively upregulate monitored WTRs, and hold the threat of future inflictions, they might at the same time lower intrinsic WTRs in the target of punishment – an outcome that at least partly offsets its advantages. While experimental economic games show that punishment indeed increase cooperation in environments with perfect opportunities for monitoring behavior (cf. above), other studies indicate that fear of punishment might in fact lower more general pro-social tendencies (Caprara, Barbaranelli, Pastorelli, Cermak, and Rosza 2001). The distinction between different kinds of WTRs explains such otherwise conflicting observations.

Based on the preceding discussion, it is clear that certain factors increase the probability a punitive strategy will be used, and others increase the probability that a reparative strategy will be used. Punitive strategies are more favored when (1) monitoring is possible and not too costly; (2) the actors are formidable enough with respect to the exploiter (or can withhold valuable enough future cooperation) that they can punish successfully and without too much cost; and/or (3) the exploiter has little potential for intrinsically valuing the set of potential victims (because of personality, or a lack of connections). Reparative strategies are more favored when (1) monitoring is impossible, or too costly and unreliable; (2) the actors are not formidable enough with respect to the exploiter and/or his allies (or have no bargaining power deriving from benefits that might be withheld), so that attempted punishment is too costly and/or insufficiently injurious; (3) the exploiter has substantial potential to intrinsically value the set of potential victims; and (4) the exploiter shows evidence of remorse – that is, shows some intrinsic valuation of the victims. If the malefactor's remorse is great enough, or opposition to punishment by his allies is too strong, then malefactors may be subject only to reparative acts. In sum, these factors revolve around the value of maintaining interactions with or a relationship with the malefactor, and we expect this to be the critical factor when a decision maker is to assess whether punishment or reparation are to be deployed. Table 5.2 provides an overview of some of the factors being processed in such assessments.

Table 5.2 A partial list of hypothesized (interrelated) factors used by the punishment and reconciliation systems.

Factors favoring reconciliation between X and Y	Factors favoring punishment of Y by X	Factors favoring execution/ expulsion of Y by X
Y has a high Association Value to X	Y has a low Association Value to X	Y has a negative Association Value toward X
Monitoring Y is difficult or costly	Monitoring Y is possible and low cost	X's previous recalibrational strategies against Y have failed
Y is relatively high in formidability	Y is relatively low in formidability	Y has no or relatively weak coalitional allies
X and Y are related	X's previous conciliatory strategies against Y have failed	...
Y shows signs of remorse	...	...
...	...	...

The Evolutionary Benefits of Being Punitive Versus being Reparative

Upon detection of an exploitive individual whose acts express low WTRs toward others, our minds need to solve the problem of choosing the best of the two types of response. Of course, all else equal, it is more adaptive to be valued intrinsically rather than extrinsically, because intrinsic valuation will result in benefits even in one's absence or temporary states of ineffectiveness. Also, reparative strategies leave the productive relationships of a selective exploiter intact. While reparative strategies potentially yield more benefits, such strategies may often be less effective by themselves, because it is hard to upregulate intrinsic WTRs broadly to all potential victims. Such strategies may also involve subjecting people to more potential exploitation, because attempted repair involves continued social contact with, and exposure to, an individual who knowingly has caused harm in the past. In contrast, punitive strategies can revise monitored WTRs easily, protecting broad classes of potential victims – but only so long as monitoring is effective.

To solve the problem of choosing between reconciliation, punishment, execution, or ostracism/confinement, our minds evolved to weigh a number of factors against each other. These factors include the relative formidability of the punishers compared to the punished, the likelihood of recidivism and future harm, and the future benefits

of continued interactions with the malefactor – including the malefactor's enmeshment with others in deep and productive social relationships. The key decision element is the malefactor's value as an associate: the estimated net lifetime value of maintaining interactions or a relationship with the malefactor from the point of view of the decision maker. We will refer to this as the malefactor's *Association Value* (Tooby and Cosmides 1996). Hence, we suggest that the human evolved psychological architecture contains subcomponents that are designed to spontaneously compute this index – an *Association Value index*, together with accompanying implicit representations of the degree of uncertainty about the true magnitude of the Association Value. Within a community, there will of course be a distribution of Association Values of members of the community toward an individual. In making collective decisions about the fate of a transgressor, each individual will be accessing the distinct Association Value she puts on the transgressor, with interactive negotiation within the community about a collective judgment.

Ancestrally, groups were small, and ingroup members in general could be presumed to have a positive Association Value to many co-members (family, friends, allies, etc.). We expect that this led to an elaborate evolved psychology of reparation – one that can be eclipsed in large-scale societies where strangers swamp closely networked social actors. In a small-scale society, an individual may be productive in many of his relationships, and exploitive only in a few. Under these circumstances, repair may appeal to many social actors, who would not want to lose an individual they value because of that individual's acts toward third parties they may not value as highly.

If the value of ongoing relationships with the malefactor is high enough (as it often is in small-scale societies, and sometimes is in large-scale societies), then reparative strategies may be all that are used. To the extent, however, that a transgressor's Association Value is negative, sentiments may spontaneously move toward execution or permanent ostracism (in the developed world, life in prison). Confinement is a combination of ostracism and punishment. It incapacitates by physically restraining the malefactor from having contact with potential victims, and at the same time such restrictions on movement are aversive. If the value of the malefactor to those empowered to act is low but positive, then temporary confinement (which prevents exploitation for its duration, and functions as a punishment) may be chosen. In line with this, recent research in dyadic cooperation suggests that when subjects have available as

one response the opportunity to avoid cooperative interactions with exploitive individuals, their choice to punish has a modestly hopeful meaning: They punish those who they anticipate will cooperate with them in the future – as a bargaining move – and simply avoid those they have decided not to cooperate with (Krasnow, Cosmides, and Tooby, forthcoming). Hence reparation signals the highest valuation, punishment a lower but still positive valuation, and expulsion or execution signals the lowest intrinsic valuation. The point is, that sentiments about what to do are not usually rationally arrived at, but rather are intuitively and spontaneously felt in individuals – we would suggest as the result of the interaction between individuals' evolved species-typical computational systems and the availability of relevant cues.

This constellation of selection pressures leads us to expect that reconciliation is a wary process. The offended parties should approach reconciliation with caution, and the activation of reparative motivations should coincide with the activation of attention-allocation mechanisms that motivate the offended parties to scrutinize the situation for cues of whether the strategy is indeed successful. Due to the role of guilt in the upregulation of intrinsic WTRs, these mechanisms should make us especially sensitive to the lack of remorse and repentance. In line with this, appeasement postures and expressions of apology and remorse do indeed seem to be cross-cultural elements of reconciliation rituals (Fry 2000). Reciprocally, the absence of remorse signals the ineffectiveness of reparative attempts, and intensifies punitive sentiments. Where punishment is not codified, punishers often proceed until they provoke a sufficiently intense signal of recalibration from the target of punishment. Reconciliation can be thought of as a reciprocal strategy, which starts cautiously and unfolds as coordinated signals are exchanged between the two parties. In contrast, the execution of punishment might be conceived of as a more one-sided event, in which the punisher induces recalibration in the punished without the punished necessarily acquiescing (see Tooby, Cosmides, Sell, et al. 2008).

Punishment and Reconciliation as a Sequential Process

Although punishment and reconciliation as two strategy-types exist in some tension with each other if executed at the same time, employing them in sequence – first punishment, then repair– may help each compensate for the defects of the other. With punitive sentiment,

humans have a desire to have the target experience a period of suffering. This derives from the evolutionary logic of recalibration and deterrence. An initial punishment phase creates deterrence (by setting a price)–deterrence would not be present if reparative strategies were used only by themselves. However, people may intuitively sense that punishment alone will leave the target more hostile, that is, with a lower intrinsic WTR toward social members than before. This may be exacerbated by the fact that the preferred form of punishment in developed societies – confinement – itself isolates the malefactor. Isolation, in turn, weakens or eliminates the kinds of social relationships of mutual valuation that lead individuals to harbor high intrinsic WTRs toward others – the psychological factor that functions as one primary inhibitor of exploitation.

Implicit or explicit recognition of this problem may motivate a policy of ordering punishment first, followed by repair. That is, there may be a life-cycle to recalibration because the strategies conflict with each other if carried out at the same time. Between punishment and repair, there is the "reconciliatory" signal of the termination of recalibrational efforts – we are now no longer adversarially attempting to make you suffer, and expect to return to a "normal" pro-social relationship. We mutually acknowledge that new baselines have been established, which are new starting points (i.e., you are not entitled to punish us for punishing you). People often use phrases like "paid their debt to society" – where the "debt" that is discharged is the obligation to experience an amount of recalibrational suffering commensurate with the magnitude of the crime.

When an individual leaves confinement, incapacitation is over, and repair may be seen as an important follow-up strategy as the malefactor has renewed access to potential victims. Once the malefactor is about to become free to move about in the community, and has opportunities to exploit again, the salience of adding reparative strategies should increase, with the goal of upregulating default intrinsic WTRs, and the potential for positive sum cooperative interactions (with conditional WTRs). That is, once the punitive phase is ended and the repair phase begins, the target may become the object of pro-social efforts to embed him in beneficial social relationships that will recalibrate his intrinsic WTRs, his monitored WTRs, or both. The distribution of suspended sentences or parole is expected to track this logic: It should seem intuitively appropriate that individuals who give out stronger cues that their tendency to exploit can be dealt with more easily by reparative strategies should

have shorter sentences (or suspended sentences), or should be let out on parole earlier. Factors like family, valued skills, a job, a supportive living situation, productivity, community connections, remorse, efforts at restitution, low formidability, an absence of attempts to advertise a hostile or exploitive orientation – all play into these decisions. Similarly, the intuitions underlying social work within criminal justice systems should seek to build on these factors, and establish relationships of interdependence and mutual valuation.

The Choice between Punishment and Reconciliation: An Overview

Based on the preceding arguments, Figure 5.1 provides an overview of the role of one critical factor (Association Values) in the psychological processes regulating the choice between punishment and reconciliation. These processes are initiated by an external cost-imposing event caused by an individual Y. This event activates computational programs assessing the expressed WTR and compares the costs imposed with the given baseline. If appropriate, these programs tag the act as "wrong" and the malefactor as an "exploiter." This tag sets other processes in motion. Most importantly, a suit of computational systems assesses the malefactor's Association Value and triggers the appropriate motives. In this computational process, the Association Value estimator extracts a number of cues from the environment relating to the costs of losing the individual as a social partner. A high Association Value triggers reparative motivations; a lower Association Value triggers punitive motivations, and motivations leading to expulsion or execution are triggered by a very low or negative Value. Whether these motivations in fact trigger the corresponding behaviors will depend on an assessment of the broader social situation. As argued, we expect punitive reactions to be inhibited if the malefactor Y is part of a relatively formidable coalition that is likely to retaliate with cost-infliction. In reverse, this could increase the estimated payoffs associated with a reparative strategy. If such a strategy is unsuccessful, however, the exploited individual could be forced to migrate out of the group. Hence, ancestrally, the latter strategy could have been the only option available in the face of formidable and well-connected exploiters.

To the extent that behavioral reactions are unleashed, varying degrees of attention are allocated toward scrutinizing the effects of the chosen strategy. For example, high-risk strategies such as reparation

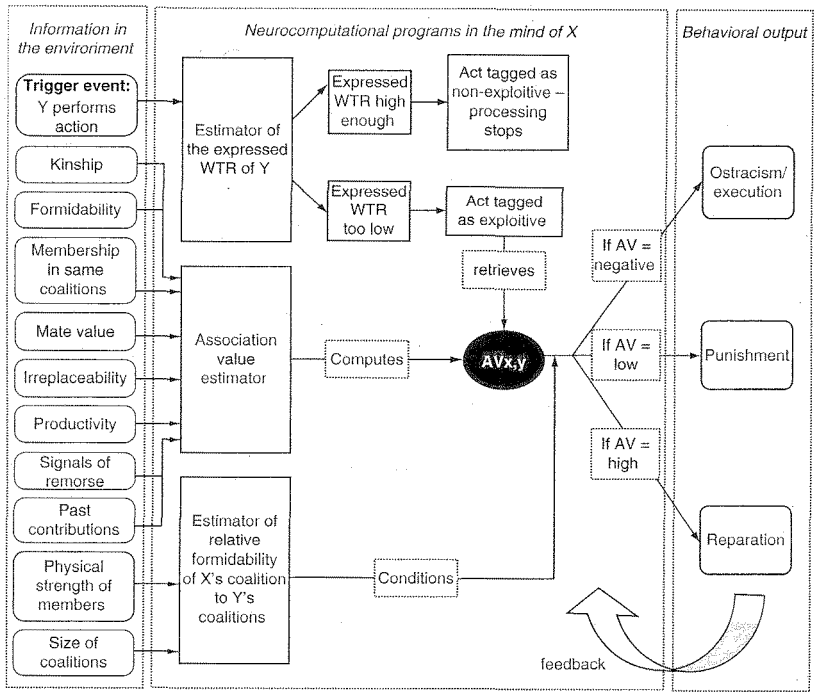

Figure 5.1 The psychological process regulating the choice between punitive and reparative responses to exploitation

should entail higher levels of attentiveness. This scrutiny creates feedback loops, where the malefactors' Association Value is recomputed and potentially new strategies are deployed. Thus, we might expect a successful use of punishment to subsequently trigger a reparative strategy. In reverse, failed reparation (i.e., Y does not show signs of remorse in response to the reparative gestures) could lead to punishment or eventually to the deployment of strategies of ostracism or execution.

Discussion about the Evidence Supporting the Argument

Several lines of evidence on human behavior suggest that Association Values play an important role in anti-exploitation decisions. First, a large-scale survey designed to test aspects of the arguments presented here suggests that humans do engage in cost-benefit analysis when choosing whether to punish or to reconcile. In the survey, attitudes

toward three specific crime cases were measured. Consistent with the arguments above, it was found that it was the perception of the specific criminal's future behavior rather than the seriousness of his act that determined whether the preferred reaction was punitive or reparative (Petersen 2007). In other words, only those who were believed to be reformable were to be rehabilitated. This effect was robust across the different crimes (rape, vandalism and assault) and across different social groups. These data also provide insights into the attention-enhancements we should expect to unlock as reparative motives arise. Those expressing general support for rehabilitation rather than punishment were also significantly more in doubt about whether this in fact constituted the right reaction, and this association was stronger the more consistently rehabilitation was supported (Petersen 2007). That is, the more firmly people support rehabilitation, the more in doubt they are. While this result may seem somewhat counter-intuitive, it is fully consistent with the view presented here.

Other types of evidence also suggest that the choice between punishment and reconciliation is based upon cost-benefit analyses. Second, in a study in cognitive neuroscience, Farrow et al. (2001) uses fMRI to analyze brain activity when people decide which crime is most forgivable. They report increased activity in the midline orbifrontal cortex, a brain area that has been linked to reward processing. Their own suggestion is that this activation represents the process of weighing the relative merits of two options against each other (presumably, to forgive one individual versus another). Third, Fry (2000) sums up the anthropological observations on reconciliation by arguing that reconciliation, cross-culturally, seems most likely to occur when relationships are important and difficult to replace – just what one would expect from the framework presented here. Finally, some evidence from studies of children's conciliatory behavior suggests that the frequency of reconciliation is higher among friends than acquaintances (Cords and Killen 1998). There is also evidence to the contrary (Butovskaya et al. 2000), but a more recent study indicates that these contrasting observations occur because friends reconcile using more implicit strategies such as simply being friendly instead of explicitly offering gifts etc. (Fujisawa, Kutsukake, and Hasegawa 2005). This last observation is not at odds with the recalibrational theory. Presumably, friends have hitherto engaged in mutual beneficial activities. An individual P's intrinsic WTR toward a friend V can probably be upregulated by V reminding P of these benefits. It might not be necessary to explicitly display the worth of the relationship by transferring new benefits.

Reconciliation has also been studied among nonhuman primates. In many ways, the argument presented here constitutes a computationally specific parallel to the way in which the primatologist de Waal (1996) makes sense of nonhuman primate conciliatory behavior (it should be noted, though, that the adaptive problem is phrased a bit differently in the primate literature, as will be discussed below). Both experimental work and observations in primate groups suggest that the likelihood of reconciliation following hostile encounters is correlated with the social value of the relationship (Aureli and de Waal 2000; although, see Silk 2002). An imaginative experimental study by Cords and Thurnheer (1993) illustrates the argument. In this study, conflicts between two long-tailed macaques were induced using food. Baseline ratings for the frequencies of reconciliation following these conflicts were obtained. Afterwards, a popcorn dispenser was installed in the macaques' cage with the unique feature that it had to be cooperatively handled to produce the reward. Conflicts were induced but this time the experimentally heightened value of the Macaques to each other increased the frequency of reconciliation by a factor of three. Furthermore, while social value seems to predict the likelihood of reconciling among nonhuman primates, data shows that the severity of an aggressive encounter has little consistent effect on the likelihood of reconciling afterwards (Silk 2002). This is consistent with the above-mentioned results on human attitudes. Finally, primate observations indicate that reconciliation is indeed a wary process, which slowly unfolds as the former antagonists exchange certain kinds of signals, presumably indicating their good faith (de Waal 2000).

Yet, the primate data needs to be viewed cautiously. In the primate literature, it is common to view reconciliation as attempts to transform adversarial relationships into pro-social relationships. Based on observations indicating that frequencies of reconciliation following aggressive encounters are correlated with degree of kinship, Silk (2000; 2002) questions this argument. Kinship-based relations are more pro-social than other types of relations, and are less likely to be enduringly disrupted by momentary conflicts (Cords 1988); in other words, they should not need reparation. Hence, another view of the function of reconciliation in nonhuman primates is as a signal that the present aggressive encounter is now terminated, and other activities can be pursued without fear of its continuation.

In the human case, however, even if reconciliation among humans is more frequent between kin than between unrelated persons

following acts of exploitation (and we expect it to be), this is not inconsistent with the WTR-derived argument. Thus, in our perspective, the adaptive problem posed by exploitation is the existence of individuals with low WTRs toward valued others (or the self), or low group-directed WTRs. These predict a likelihood of future and ongoing exploitation. This problem is real whether or not the perpetrator is related to the observer. The adaptive solution can involve upregulations of the kin's monitored or intrinsic WTR, but all else equal, the fact that kin are intrinsically valuable (Hamilton 1964) should make reparative strategies preferable (when proximity is not inherently negative sum).

Finally, it is possible to assess the predicted importance of malefactors' estimated Associational Value and related cues in the light of the historical development of modern criminal justice systems. In small-scale societies, perpetrators generally have strong ties to most members of the social group, making reparative strategies potentially effective. Moreover, the greater number of supporters an exploiter has, the more opposition there will be to punishment by the target's allies – also limiting the use of punishment. In large-scale societies, the class of potential victims falling outside of the circle of enmeshment is very large, monitoring is more difficult, and punitive agencies can become far more powerful than individual malefactors. These factors predict cross-cultural trends toward increasing use of punishment in larger scale societies. This might partly explain the rise of relatively ruthless punitive systems widely deployed in mass societies after the emergence of agriculture (Spierenburg 1984).

Yet, this development only holds until the seventeenth century, and for the last three hundred years criminal justice in the Western world has grown milder (Garland 1990).[3] This turn toward a modest use of reparative strategies can be explained by the interaction between the psychological architecture regulating the choice between punishment and reconciliation, and at least two cultural developments. First, it is possible that the rise of the printing press and, subsequently, newspapers, photography, television and film (with more direct psychophysical representations of the treatment of individuals inside criminal justice systems) causes ordinary experience in industrial societies to more closely mimic the greater engagement found among individuals in smaller scale societies. Especially, the distribution of information made possible by the printing press seems to have played an important role in establishing a sense of collective identity in large-scale societies (Anderson 1991). Second,

these processes might have been fuelled by the institutional developments of capitalist market society and later welfare state institutions. Capitalist society breeds more inclusive coalitional identities as extensive labor divisions facilitate experiences of successful social exchange with people highly dissimilar from oneself. Similarly, the establishment of social welfare schemes in the twentieth century has facilitated more equal levels of living standards, as well as a reduction in class and ethnic differentiation through clothing, and other aspects of appearance. Arguably this reinforces the perception by individuals that the nation-state is their coalition. This mental representation is no longer constantly challenged by direct observations of others in their community who appear extremely different (see, for example, Larsen 2006; Rothstein 1998). In line with this, research shows that punitiveness is lower in economically developed countries (Mayhew and van Kesteren 2002) and in countries with large welfare states (Christie 2004).

This final analysis also underscores the point that it is important not to generalize attitudes among elites in developed countries to all cultures. This is especially true in relation to the ideology that it is proper to treat all individuals "equally"; an ideology seemingly produced by perspective-taking alliance formation within democratic political processes. In general, we expect baselines with regard to acceptable levels of differential treatment to be defined by within a group, based on its internal distribution of alliances and power. In many social settings, for example, exploitive or even lethal acts against outgroup members are not viewed as crimes at all, but often as laudable. Even acts against ingroup members without social allies may be viewed similarly. Stable entrenched power differentials lead to social concepts of legitimate status-based entitlement, such as the emergence of aristocracies with prerogatives that would be viewed as criminal in democratic nations.

The Seriousness of the Act and the Quantitative Modulation of the Reaction

In the modern criminal justice system, the harmfulness and seriousness of a crime is of fundamental importance when specific sentences are measured out. One interesting feature of the argument presented here is that we should not expect our species-typical psychology to place the same weight on the seriousness of the exploitive act (i.e., the discrepancy between the revealed and the acceptable WTR)

when choosing between punishment and reconciliation. Rather, this choice is expected to be determined by a prospective estimation of the transgressor's Association Value – that is, the future value of maintaining a relationship with the exploitive person. As will be argued below, this estimate is based on numerous cues, of which the seriousness of the act is only one. Thus, in a nutshell, the argument is that we fit our reaction to the exploitive person rather than to his or her act. Even in the face of serious exploitation, we can opt for a reparative strategy. Contrasted with formalized and codified sanctioning systems found in developed countries (which are hampered by an imperative to be consistent), public opinion on criminal justice issues is expected to be more flexible, and more oriented to individualized forms of sanctioning that are specifically tailored to the criminal at hand. Studies of public opinion in different countries confirm this (Finkel et al. 1996; Petersen forthcoming).

At the same time, the recalibrational framework suggests that the perceived seriousness of the offense does play an important role in anti-exploitation decisions. While punishment and reconciliation are two qualitatively distinct reactions, both types of reactions can also be modulated quantitatively with respect to their intensity. An aggressive strategy can entail both high and low cost-infliction. Similarly, conciliatory guilt-induction and enmeshment can be aimed at inducing high or low guilt, high or low interdependence, and relationships can be curtailed or ended outright. When the evaluation of the net future value of a continued relationship with the perpetrator has inclined the actors toward either punitive or reparative strategies, the seriousness of the perpetrators' exploitation is expected to modulate the intensity of the strategy.

Within this framework, the problem posed by exploitation is that it indicates that the perpetrator does not value the victim or other members of our group sufficiently – predicting future exploitation. The costliness of the act to the victim, together with the benefit of the act to the perpetrator, gives a reliable indication of the offender's current WTR toward the target – i.e., how low the perpetrator's interest in our welfare is. Our intuitive grasp of the seriousness of a transgression reflects the discrepancy between the revealed and the acceptable WTR. Thus, the seriousness of an offense – and indirectly its costliness – tells us how much the perpetrator's WTR (whether monitored or intrinsic) toward us needs to be upregulated before it is deemed to be sufficiently high. This information should necessarily be reflected in the intensity of the reaction by which we seek to achieve

this objective. The survey data referred to above supports this role of the seriousness of the act. Thus, while the perceived seriousness of crimes does not directly influence whether punishment or restoration is preferred, it does influence the length of the preferred sentences (Petersen 2007). Similarly, other studies have consistently shown that preferred sentence lengths are determined by the seriousness of the act (Darley and Pittman 2003; Darley, Carlsmith, and Robinson 2000).

The Computational Architecture of Our Anti-Exploitation Motivational System

In deciding how to respond to exploitation, our evolved circuits need to know the net future value of remaining associated with the perpetrator. However, events in the future are inherently unobservable. Accordingly, to the extent that exploitation has been an ancestrally recurrent adaptive problem, natural selection should have selected for circuits designed to monitor cues that predict Association Value. These circuits should influence behavior by regulating the activation of motivational programs producing felt emotions (Tooby, Cosmides and Barrett 2005; Tooby and Cosmides 2008; Tooby, Cosmides, Sell et al. 2008).

Ancestral Cues to the Net Future Value of Interacting with Exploitive Persons

We propose that our evolved psychological architecture is designed to compute an Association Value index for a potentially exploitive person based on at least four analytically distinct types of cues: (1) cues relating to potential benefits of association; (2) cues relating to the likelihood of future exploitative acts; (3) cues relating to the potential harm if the act is indeed repeated; and (4) cues predicting changes in these variables if either reparative or punitive strategies are deployed.

Important cues to the potential benefits of engaging in close future social interaction with an individual would be:

- kinship (Hamilton 1964);
- the expectation of future benefits based on the history of benefit delivery from this individual in the past;
- the individual's general level of resources or productivity (status, access to resources, skills and competences, leadership abilities, positive externalities, etc.);

- whether the individual is a member of own's coalition or a rival one (because ingroup members are (all else equal) more willing cooperative partners; see Cosmides, Tooby and Kurzban 2003; Tooby, Cosmides and Price 2006);
- the attractiveness of the individual as a sexual partner;
- the general irreplaceability or value of the individual (e.g., the co-parent to one's child would serve an irreplaceable function due to his or her unique interest in the child's welfare; see Tooby and Cosmides 1996).

As matter of fact, criminological studies of the public's attitudes provide evidence that outgroup members (as defined by ethnicity, race or accent) are seen as more eligible for punishment (Dixon, Mahoney, and Cocks 2002; Hurwitz and Peffley 1997; Petersen 2007; Ugwuegbu 1979), that well-integrated ingroup members are punished less (Hembroff 1987; Goul Andersen, 1998), and that the physically attractive are treated more mildly (Mazella and Feingold 1994).

When estimating the magnitude of potential future costs of reconciling, two cues seem to be especially important. First, the costliness of the initial exploitive act (to ingroup members) should provide some estimate of how harmful potential future exploitive acts would be to decision makers. This is because it sets an upper limit on the individual's maximum intrinsic WTR toward community members. Hence, we expect the seriousness of the act to (indirectly) influence the decision about whether to punish or reconcile by having an effect on the estimate of the net prospective social value of the exploitive person to the decision makers. Second, the mind should be sensitive to cues about the exploitive person's ability to evaluate costs and benefits accurately: those who consistently misperceive the harm associated with their acts will be more prone to harm us greatly. When deciding whether to punish or conciliate, we should accordingly be motivated to look for cues that indicate whether the exploitive person indeed understands the harm he has caused.

Furthermore, we should expect our mind to monitor cues that predict how likely repeated exploitation is. We expect the following cues to be important:

- the past behavior of the exploitive person (is it a first time offense?);
- the degree to which the exploitive person expresses remorse;
- the person's degree of impulse control;
- the degree of intentionality behind the exploitive act.

At a computationally specific level, "intentional acts" can be thought of as acts resulting from decision making processes in the perpetrator that accessed information about the magnitude of the costs imposed, the magnitude of the benefits received, and on whom the costs were imposed (Sell 2006b). Compared to unintentional exploitation, intentional acts will more reliably reveal the true level of the underlying WTR and will better predict future cost-imposition.

Opinion studies have documented the cross-cultural importance of all such cues: Restorative sanctions are less preferred against repeated criminal activity (Finkel et al 1996; Petersen 2007); the more intentional exploitive acts are (e.g., from accidental to negligent to fully intentional), the more punitive sanctions are viewed as legitimate compared to restorative sanctions (Hamilton and Sanders 1992); and remorse significantly increases the perceived degree to which exploitive acts can be forgiven and the degree to which rehabilitation is viewed as the appropriate goal of the criminal sanction (Petersen 2007; Robinson, Smith-Lovin and Tsoudis 1994).

When an exploitive act is observed, our mind should be programmed to detect and process all these cues and compute an index of the net future value of the exploitive person. To the extent this Association Value is estimated as high, conciliatory motives should be activated. In contrast, punitive motives should be elicited if the value is perceived as low. Differences in this assessment to different members of the community (e.g., victim's family vs. perpetrator's family) often lead to conflicts about what course of action to take. Finally, the intensity of these motives and the behavior they give rise to should be regulated by the seriousness of the exploitive act.

The Emotional Side of Punishment and Reconciliation

It is easy to confuse scientific claims about the evolved function of a computational mechanism with claims about conscious events. However, the proposal here is about the circuit logic of evolved programs that were built into our neurocomputational architecture by natural selection, and not about conscious deliberation. This circuitry and its logic operate outside of consciousness, although it may occasionally place some of its products into conscious awareness, where we experience them as feelings, inclinations, intuitions or ways of thinking. From an evolutionary psychological perspective, emotions such as anger are simply one kind of evolved program – each with a functional problem-solving logic that deals with its respective

adaptive problem, imposed by natural selection. As discussed above, anger is the primary emotion program that evolved to deal with the recurrent adaptive problem posed by encounters with people who place too little weight on one's welfare. One of its primary outputs is the motivation to signal why the target should upregulate her WTR, through either the punitive infliction of costs, or the withdrawal of cooperative benefits, depending (speaking approximately) on the target's Association Value to the angry individual.

Since the punitive infliction of costs (driven by angry punitive sentiment) appears as a leading option when provoked by exploitation, it would be enacted as the response unless the mind's "arguments" against inflicting harm could counterbalance it. Aside from the power of the perpetrator or the community support he or she enjoys, the main factor that should diminish the impulse to harm the perpetrator is his or her Association Valuation to the person experiencing the response. If the Association Value is high enough, the prospect of inflicting costs on the perpetrator calls up a countervailing evaluative emotional subsystem – compassion. These two sets of circuits will be outputting their unique forms of value information and motivational tendencies – to injure, and to refrain from injuring, the perpetrator. Hence, the circumstances of the transgression, the characteristics of the perpetrator, and the perpetrator's relationship to supporters and detractors in the community are processed by these mechanisms, which then produce an emotional configuration representing the mind's best guess of an adaptive behavioral response. We therefore expect anger and (potentially) compassion to be key ingredients in an emotional mix elicited by exploitation (Petersen 2010).

As we sketched out, punitive and conciliatory strategies exist in tension as alternatives, because cost-infliction works against conciliatory attempts to upregulate the exploitive person's intrinsic WTR toward the punitive. Yet despite their differences, both kinds of strategies should be heavily regulated by the anger program and anger should be a key part of the emotional side of both. Anger is the data format by which the perceived seriousness of the injury is broadcast into consciousness and to other computational systems (Sell et al., forthcoming; Tooby, Cosmides, Sell et al. 2008). The felt magnitude of anger should be directly related to the displayed level of the WTR, for example, the seriousness of the exploitive act, and will play an important role in the motivation of recalibrational responses (Sell 2006b; Sell et al. 2009, forthcoming). Accordingly, we expect

anger to be a central regulator of the intensity of not only punitive strategies but also of restorative strategies. For example, someone could act in a way that did not weight another person's values sufficiently either because (1) they did not value the person's welfare (in which case punishment is a useful recalibrator), or (2) they had an incorrect model of the victim's values but a sufficiently high WTR (in which case re-education might serve as well). That is, if the perpetrator had only known what it meant to the victim, he would not have committed the act. This gives a definition of remorse: recalibrational suffering based on re-education about the victim's values. Anger not only incites behaviors designed to increase the weight the target places on the angry individual's welfare; it also motivates communication designed to educate the transgressor so that she has a correct model of the values of the angry individual – values that the exploitive act violated, and values that the malefactor should respect in the future, to the extent that he places weight on others' welfare.

In support of this claim, empirical studies show that forgiveness and reconciliation both operate within a background of anger (Averill 1982; Walker and Gorsuch 2004). Thus, while reconciliatory strategies do not impose costs on the exploitive person, they do involve an anger-motivated condemnation and denunciation, which conveys to the exploitive person the magnitude and the nature of the harm.

While anger strongly predominates when punitive strategies are induced, this is not necessarily the case in conciliatory strategies. Where the Association Value of the transgressor is high enough, signals from the compassion subsystem emerge as factors that countervail against anger. Empirical studies show that emotions of compassion or sympathy toward the exploitive person play an important motivational role when engaging in reconciliation (Gault and Sabini 2000; Petersen 2010). In general, compassion seems to be involved in abstaining from harming others and providing benefits for individuals in need (Haidt 2003; Wispé 1991).

Thus, the motivational role of compassion in reconciliation is expected within the recalibrational theory outlined here. Compassion inhibits punitive sentiments that might end the relationship, or reduce the Association Value of the transgressor by injuring or killing them. Moreover, compassion (gated by valuation) should orchestrate responses that embody the conciliatory strategy – that is, it organizes behavior that could potentially feed into the system that revises intrinsic WTRs in the target. Someone motivated by compassion acts with forbearance and kindness, advertising the degree to

which the compassionate person values the transgressor. This signal of valuation by the victim for the transgressor shows the transgressor that this is one of a limited number of persons in the world who do value them, making the victim more intrinsically valuable to the transgressor. Where Association Value is high and conciliatory strategies are emerging, the resulting conflicting emotions of anger and compassion correspond to the folk concept of feeling "hurt."

An Overview

The following is a possible interpretation of the interplay between the systems dissecting the dimensions of exploitive acts, and the motivational systems producing anger and compassion, which are triggered when harm is imposed on the self or an ingroup member. First, the act's seriousness is assessed to evaluate whether the cost-imposition reflects a welfare trade-off ratio that is too low. Second, to the extent that the act is recognized as an exploitive act, feelings of anger proportional with the seriousness are produced. Third, the release of anger – which is the data format in which the detection of a person with a low WTR is broadcasted to other mechanisms (Tooby, Cosmides, Sell et al. 2008) – serves as a vehicle for the activation of other computational systems. These systems will access the Association Value index for the transgressor – the motivational variable that reflects the mind's estimate of the net future value of associating with the exploitive person. Fourth, to the extent that this value is estimated to be high, reparative emotions toward the exploitive persons are triggered and will co-exist with anger. While anger will foster condemnation, reparative sentiments (such as compassion) should moderate cost-infliction on the exploitive person and potentially invite a restorative strategy. However, if the Association Value index is low, no checks on anger are produced and a punitive strategy will be deployed. Fifth, to the extent that reparative emotions are elicited, attention-allocating mechanisms should be activated that search for cues of remorse and guilt in the exploitive person to consolidate or abort adoption of a conciliatory strategy.

This is an individual-level description, without taking into account either how individuals and larger groups influence each other in this process, or the larger dynamics by which a community negotiates a coordinated response (if any). Obviously, there will be a distribution of different behavioral inclinations in different members of the social group based on their relationships (real or vicarious) to the

victim and the exploiter, their social distances, their vulnerability to the precedent set by the transgression, their evaluations of the relevant baselines, their formidabilities (power or lack of power) compared to the victim and malefactor, and so on. Nevertheless, based on these variables, one can use this analysis to predict systematic patterns in the responses of individuals to exploitation. Hence, we expect this architecture to regulate how we react to harmful acts at all levels of social interaction: in the family, between friends, at the workplace, and when we are confronted with crime in the media. Indeed, a number of studies of attitudes toward crime depicted in the media have concluded that, if those surveyed believe that criminals are dispositionally "good" – that they wish to behave lawfully, but are driven into criminality because of poverty or problems in their upbringing, then they favor rehabilitative strategies. If they, on the other hand, believe that criminals are dispositionally "bad," and that their criminality is the result of rational calculations or stable anti-social desires, then they support harsh punishments (see Claster 1992; Lakoff 1996; Sasson 1995; Wilson and Herrnstein 1985; the terminology of "good" and "bad: is taken from Claster [1992]). Such stereotypes seem to satisfy the input conditions of the computational systems eliciting punitive and conciliatory strategies, respectively.

Conclusion

In this chapter, we have developed an evolutionarily informed computational sketch of some of the evolved programs that are deployed when individuals deal with exploitation, crime, punishment, and reconciliation. We think that this approach might illuminate certain recurrent phenomena in formal and informal criminal justice systems, such as spontaneous political attitudes concerning crime (Petersen, forthcoming). We argue that the acts that we perceive as exploitive are acts that reveal the low value of welfare trade-off variables in the minds' of perpetrators (provided we have a pro-social orientation toward the victims). Punishment and reconciliation are two evolved strategies to remedy this adaptive problem, targeting different aspects of the exploitive person's computational architecture to upregulate the value he places on potential victims. From a more phenomenological perspective these strategies are designed to induce fear and guilt, respectively. Due to the structure of the relevant selection pressures, punitive strategies will be elicited when the net future Association Value of the criminal is estimated as low. In

contrast, when this value is estimated as high, conciliatory strategies are more likely to be favored.

This perspective, then, provides an alternative to the Freudian idea of nonpunitive orientations merely as a product of culture's sublimation of aggressive drives.[4] According to the recalibrational view, both restorative and punitive orientations emerge from the interplay between the valuation that members of the community place on the criminal or transgressor and a set of evolved programs embodying alternative defenses against exploitation. Punishment and reconciliation are both natural counterstrategies, endogenous to the human mind.

Notes

1. For example, at present there is a discourse in the United States about whether to treat certain individuals as enemy combatants or as criminals—each one involving a different set of laws and expectations. We suggest that these different sets originate in different evolved mental categories and motivational circuits corresponding to different kinds of ancestral threat. Different voices in this debate (including people of different nationalities and religious groups) might be spontaneously drawn to one side of the debate or another based on where their minds draw the boundary "ingroup member."
2. Our minds evolved to compute baselines according to this and other cognitive principles based on the net long term evolutionary payoffs of adopting one versus another. Some of the payoffs driving our species convergence include: (1) endless and inconclusive conflict emerges if different players interpret the world using different ways of establishing baselines, so there is strong selection for convergence; (2) many other ways of establishing baselines are selected against, in that they discourage convergence on benefit-benefit interactions; (3) this rule for defining baselines does not involve computing over a combinatorial explosion of counterfactual possibilities; (4) this rule is consistent with presocial ways of interpreting causation and evaluating choices when planning. These evolved principles for setting baselines underlie cross-cultural commonalities in such concepts as property, and the recognition that transgressions of commission (baseline change) are more recognizable and worse than transgressions of omission (leaving baselines unchanged).
3. At least, until quite recently. Within the last two decades countries across the Western world have thus experienced increased sentencing lengths and rising numbers of inmates (Kury and Ferdinand 1999; Prat et al., 2005).

4. Furthermore, this theory might explain why the Freudian argument might seem phenomenologically convincing. In tandem with the elicitation of conciliatory motivations, attention-allocating mechanisms will motivate ongoing scrutiny of the success of the strategy, with punitive motivations inhibited to the extent the strategy appears to be working. Accordingly, the confidence that punishment is the right choice when it is being inflicted might systematically be higher relative to the confidence that reconciliation is the proper strategy. But this typically smaller confidence in reparative strategies is not a sign that only punitiveness is endogenous. Rather this difference is the expression of functional design.

References

Anderson, B. 1991. *Imagined communities*. London and New York: Verso.

Aureli, F., and F. de Waal. (Eds.) 2000. *Natural conflict resolution*. Berkeley and Los Angeles: University of California Press.

Averill, J. 1982. *Anger and aggression*. New York: Springer-Verlag.

Boyd, R., and P. Richersen. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13(3):171–95.

Boyd, R., H. Gintis, S. Bowles, and P. Richerson. 2003. The evolution of altruistic punishment *PNAS* 100(6):3531–5.

Braithwaite, J. 2002. *Restorative justice and responsive regulation*. New York: Oxford University Press.

Buss, D. M., and J. D. Duntley. 2003. Homicide: An evolutionary perspective and implications for public policy. In *Violence and Public Policy*, edited by N. Dess, 115–28. Westport, CT: Greenwood Publishing Group, Inc.

Butovskaya, M., P. Verbeek, T. Ljungberg, and A. Lunardini. 2000. A multicultural view of peacemaking among young children. In *Natural Conflict Resolution*, edited by F. Aureli and F. de Waal, 243–62. Berkeley and Los Angeles: University of California Press.

Cameron, L. 1999. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry* 37(1):47–59.

Caprara, G. V., C. Barbaranelli, C. Pastorelli, I. Cermak, and S. Rosza. 2001. Facing guilt: Role of negative affectivity, need for reparation, and fear of punishment in leading to prosocial behaviour and aggression. *European Journal of Personality* 15(3):219–37.

Chavanne, T. J., and G. G. Gallup. 1998. Variation in risk taking behavior among female college students as a function of the menstrual cycle. *Evolution and Human Behavior* 19:27–32.

Christie, N. 2004. *A suitable amount of crime*. London and New York: Routledge.

Claster, D. S. 1992. *Bad guys and good guys: Moral polarization and crime.* Westport, CT: Greenwood Press.
Clutton-Brock, T. H., and G. A. Parker. 1995. Punishment in animal societies. *Nature* 373:209–16.
Cords, M. 1988. Resolution of aggressive conflicts by immature long-tailed macaques. *Animal Behavior* 36:1124–35.
Cords, M., and M. Killen. 1998. Conflict resolution in human and non-human primates. In *Piaget, Evolution, and Development*, edited by J. Langer and M. Killen, 193–217. Hillsdale, NJ: Lawrence Erlbaum Associates.
Cords, M., and S. Thurnheer. 1993. Reconciling with valuable partners by long-tailed macaques. *Ethology* 93(4):315–25.
Cosmides, L., and J. Tooby. 1992. Cognitive adaptations for social exchange. In *The Adapted Mind*, edited by J. H. Barkow, L. Cosmides, and J. Tooby, 163–228. Oxford: Oxford University Press.
Cosmides, L., and J. Tooby. 2005. Neurocognitive adaptations designed for social exchange. In *The Handbook of Evolutionary Psychology*, edited by D. M. Buss, 584–627. Hoboken, NJ: Wiley.
Cosmides, L., J. Tooby, and R. Kurzban. 2003. Perceptions of race. *Trends in Cognitive Sciences* 7(4):173–9.
Daly, M., and M. Wilson. 1988. *Homicide.* Hawthorne, NY: Aldine.
Darley, J. M., and T. S. Pittman. 2003. The psychology of compensatory and retributive justice. *Personality and Social Psychology Review* 7(4):324–36.
Darley, J. M., K. M. Carlsmith, and P. H. Robinson. 2000. Incapacitation and Just Deserts as Motives for Punishment. *Law and Human Behavior* 24(6):659–83.
Delton, A., D. Sznycer, T. Robertson, J. Lim, L. Cosmides, and J. Tooby. (forthcoming). *An evolved internal regulatory variable for making welfare tradeoffs.*
de Quervain, D., U. Fischbacher, V. Treyer, M. Schellhammer, S. Ulrich, A. Buck, A. and E. Fehr. 2004. The neural basis of altruistic punishment. *Science* 305:1254–8.
de Waal, F. 1992. Aggression as a well-integrated part of primate social relationships: A critique of the seville statement on violence. In *Aggression and Peacefulness in Humans and Other Primates*, edited by J. Silverberg and P. J. Gray. New York: Oxford University Press.
de Waal, F. 1996. *Good natured: The origins of right and wrong in humans and other animals.* Cambridge, MA: Harvard University Press.
de Waal, F. 2000. Primates – A natural heritage of conflict resolution. *Science* 289:586–90.
Dixon, J. A., B. Mahoney, and R. Cocks. 2002. Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology* 21(3):162–8.
Duff, A. 1996. Penal communications: Recent work in the philosophy of punishment. *Crime and Justice,* 20:1–97.

Duntley, J. D. 2005. Adaptations to dangers from humans. In *The Handbook of Evolutionary Psychology*, edited by D. M. Buss, 224–49. Hoboken, NJ: J. Wiley and Sons, Inc.
Duntley, J. D., and D. M. Buss. 2004. The evolution of evil. In *The Social Psychology of Good and Evil*, edited by A. Miller, 102–23. New York: Guilford.
Durkheim, E. 1998. Two laws of penal evolution. In *Sociology of Punishment*, edited by D. Melossi. Aldershot, UK: Ashgate.
Elias, N. 1994. *The civilizing process.* Oxford and Cambridge: Blackwell.
Ellis, L., and H. Hoffman. (Eds.) 1990. *Crime in biological, social, and moral contexts.* New York: Praeger.
Farrow, T., Y. Zheng, I. D. Wilkinson, S. A. Spence, J. F. W. Deakin, N. Tarrier, P. D. Griffiths, and P. W. R. Woodruff. 2001. Investigating the functional anatomy of empathy and forgiveness. *NeuroReport* 12(11):2433–38.
Fehr, E., and U. Fischbacher. 2004. Third-party punishment and social norms. *Evolution and Human Behavior* 25:63–87.
Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *The American Economic Review* 90:980–94.
Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415:137–40.
Finkel, N. J., S. T. Maloney, M. Z. Valbuena, and J. Groscup. 1996. Recidivism, proportionalism, and individualized punishment. *American Behavioral Scientist* 39(5):474–87.
France, A. 2002 [1894]. *The Red Lily.* McLean, VA: IndyPublish.
Frank, R. H. 1988. *Passions within reason: The strategic role of the emotions.* New York: W. W. Norton & Company.
Freud, S. 1961 [1929]. *Civilization and its discontents.* New York: W. W. Norton & Company.
Fry, D. P. 2000. Conflict management in cross-cultural perspective. In *Natural Conflict Resolution*, F. Aureli and F. de Waal, 334–51. Berkeley and Los Angeles: University of California Press.
Fujisawa, K. K., N. Kutsukake, and T. Hasegawa. 2005. Reconciliation pattern after aggression among Japanese preschool children. *Aggressive Behavior* 31(2):138–52.
Garland, D. 1990. *Punishment and modern society. A study in social theory.* Oxford: Clarendon Press.
Gault, B. A., and J. Sabini. 2000. The roles of empathy, anger, and gender in predicting attitudes toward punitive, reparative, and preventative public policies. *Cognition and Emotion* 14(4):495–520.
Goul Andersen, J. 1998. *Borgerne og lovene [Citizens and the law].* Aarhus: Aarhus University Press.
Haidt, J. 2003. The Moral Emotions. In *Handbook of Affective Sciences*, edited by R. Davidson, K. Scherer, and H. Goldsmith, 852–70. Oxford: Oxford University Press.

Hamilton, L., and J. Sanders. 1992. *Everyday justice: Responsibility and the individual in Japan and the United States.* New Haven: Yale University Press.
Hamilton, W. D. 1964. The genetic evolution of social behavior, I and II. *Journal of Theoretical Biology* 7:1–16, 17–52.
Hembroff, L. A. 1987. The seriousness of acts and social contexts: A test of Black's Theory of the Behavior of Law. *The American Journal of Sociology* 93(2):322–47.
Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (Eds.). 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Cambridge: Oxford University Press.
Hill, K. 2002. Cooperative food acquisition by Ache foragers. *Human Nature,* 13(1):105–28.
Hirsch, A., A. E. Bottoms, E. Burney, and P.-O. Wikstrom. 1999. *Criminal deterrence and sentence severity: An analysis of recent research.* Oxford: Hart Publishing.
Hoebel, E. A. 1964. *The law of primitive man.* Cambridge, MA: Harvard University Press.
Hurwitz, J., and M. Peffley. 1997. Public perceptions of race and crime: The role of racial stereotypes. *American Journal of Political Science* 41:375–401.
Jacoby, S. 1983. *Wild justice: The evolution of revenge.* New York: Harper & Row.
Kelly, R. L. 1995. *The foraging spectrum: Diversity in hunter-gatherer lifeways.* Washington: Smithsonian Institution Press.
Klepper, S., and D. Nagin. 1989. The deterrent effect of perceived certainty and severity of punishment revisited. *Criminology* 27(4):721–46.
Krasnow, M., L. Cosmides, and J. Tooby. (forthcoming). *Trust, Reciprocity and Punishment: Adaptations for Small Scale Cooperation.*
Kurzban, R., J. Tooby, and L. Cosmides. 2001. Can race be erased? Coalitional computation and social categorization. *PNAS* 98(26):15387–92.
Lakoff, G. 1987. *Women, fire, and dangerous things.* Chicago: Chicago University Press.
Landsheer, J. A., and H. Hart. 2000. Punishments adolescents find justified: An examination of attitudes toward delinquency. *Adolescence* 35(140):683–93.
Larsen, C. A. 2006. *The Institutional Logic of Welfare Attitudes.* Aldershot, UK: Ashgate.
Lee, R., and I. DeVore. 1968. Problems in the study of hunters and gatherers. In *Man the Hunter,* edited by R. Lee and I. DeVore, 3–29. Chicago: Aldine.
Lieberman, D., J. Tooby, and L. Cosmides. 2007. The architecture of human kin detection. *Nature* 445:727–31.
Ljungberg, T., L. Horowitz, L. Jansson, K. Westlund, and C. Clarke. 2005. Communicative factors, conflict progression, and use of reconciliatory

strategies in pre-school boys – A series of random events or a sequential process? *Aggressive Behavior,* 31(4):303–23.
Lowie, R. 1961. *Primitive society.* New York: Harper Torchbooks.
Mayhew, P., and J. van Kesteren. 2002. Cross-national attitudes to punishment. In *Changing Attitudes to Punishment,* edited by J. V. Roberts and M. Hough, 63–92. Portland, OR: Willan Publishing.
Maynard Smith, J. and G. Parker. 1976. The logic of asymmetric contests. *Animal Behavior* 24:169–75.
Mazella, R., and A. Feingold. 1994. The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta-analysis. *Journal of Applied Social Psychology* 24:1315–44.
Nozick, R. (1981). *Philosophical explanations.* Cambridge, MA: Harvard University Press.
Ostrom, E., R. Gardner, and J. Walker. 1994. *Rules, games and common pool resources.* Ann Arbor, MI: The University of Michigan Press.
Panchanathan, K., and R. Boyd. 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* 224(1):115–25.
Petersen, M. B. 2007. *Straf eller rehabilitering? Evolutionspsykologi, følelser og politisk holdningsdannelse [Punishment or rehabilitation? Evolutionary psychology, emotions, and political opinion formation].* Aarhus: Politica.
Petersen, M. B. 2010. Distinct Emotions, Distinct Domains: Anger, Anxiety and Perceptions of Intentionality. *Journal of Politics,* 72(2).
Petersen, M. B. (forthcoming). Public Opinion and Evolved Heuristics. Forthcoming in the *Journal of Cognition and Culture,* 9(3–4):315–37.
Petralia, S. M., and G. G. Gallup. 2002. Effects of a sexual assault scenario on handgrip strength across the menstrual cycle. *Evolution and Human Behavior* 23:3–10.
Price, M. E., L. Cosmides, and J. Tooby. 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* 23:203–31.
Robinson, D. T., L. Smith-Lovin, and O. Tsoudis. 1994. The effects of remorse on mock criminal confessions. *Social Forces* 73(1):175–90.
Rossi, P. H., J. E. Simpson, and J. L. Miller. 1985. Beyond crime seriousness: Fitting the punishment to the crime. *Journal of Quantitative Criminology* 1(1):59–90.
Rossi, P. H., E. Waite, C. E. Bose, and R. E. Berk. 1974. The seriousness of crimes: Normative structure and individual differences. *American Sociological Review* 39(2):224–37.
Rothstein, B. 1998. *Just institutions matter.* Cambridge, UK: Cambridge University Press.
Sasson, T. 1995. *Crime talk.* New York: Aldine de Gruyter.
Sell, A. 2006a. Anger expressions dissected: Why does his face look like that? Paper presented at 18th annual meeting of the Human Behavior

and Evolution Society, HBES, University of Pennsylvania, June 7–11, 2006.
Sell, A. 2006b. Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger. Doctoral Dissertation. Santa Barbara: Center for Evolutionary Psychology, University of California.
Sell, A., L. Cosmides, J. Tooby, D. Sznycer, C. von Rueden, and M. Gurven. 2009. Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B*, 276:575–84.
Sell, A., J. Tooby, and L. Cosmides. (forthcoming a) *Anger and welfare tradeoff ratios: Mapping the computational architecture of a recalibrational emotion system.*
Sell, A., J. Tooby, and L. Cosmides. 2009. Formidability and the logic of human anger. *Proceedings of the National Academy of Science,* in press.
Silk, J. B. 2000. The function of peaceful post-conflict interactions: An alternate view. In *Natural Conflict Resolution*, edited by F. Aureli, and F. de Waal, 179–81. Berkeley and Los Angeles: University of California Press.
Silk, J. B. 2002. The form and function of reconciliation in primates. *Annual Review of Anthropology* 31:21–44.
Spierenburg, P. 1984. *The spectacle of suffering.* Cambridge: Cambridge University Press.
Stylianou, S. 2003. Measuring crime seriousness perceptions: What have we learned and what else do we want to know. *Journal of Criminal Justice* 31:37–56.
Sznycer, D., J. Price, J. Tooby, and L. Cosmides. 2007. Recalibrational emotions and welfare trade-off ratios: Cooperation in anger, guilt, gratitude, pride, and shame. Paper presented at 19th annual meeting of the Human Behavior and Evolution Society, HBES, College of William & Mary, Virginia, May 30–June 3, 2007.
The Bible, 21st Century King James Version, www.biblegateway.com
Thornhill, R., and C. Palmer. 2000. *A natural history of rape: Biological bases of sexual coercion.* Cambridge: MIT Press.
Tooby, J., and L. Cosmides. 1992. The psychological foundations of culture. In *The Adapted Mind*, edited by J. H. Barkow, L. Cosmides, and J. Tooby, 19–135. Oxford: Oxford University Press.
Tooby, J., and L. Cosmides. 1996. Friendship and the bankers paradox: Other pathways to the evolution of adaptations for altruism. *Proceedings of the British Academy* 88:119–43.
Tooby, J., and L. Cosmides. 2005. Conceptual foundations of evolutionary psychology. In *The Handbook of Evolutionary Psychology*, edited by D. M. Buss, 5–67. Hoboken, NJ: J. Wiley & Sons, Inc.
Tooby, J., and L. Cosmides. 2008. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In

Handbook of Emotions, edited by M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, 3rd edition, 114–37. New York: Guilford.
Tooby, J., L. Cosmides, and H. C. Barrett. 2005. Resolving the debate on innate ideas. In *The Innate Mind: Structure and Content*, edited by P. Carruthers, S. Laurence, and S. Stich, 305–37. New York: Oxford University Press.
Tooby, J., L. Cosmides, and M. E. Price. 2006. Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics* 27:103–29.
Tooby, J., L. Cosmides, A. Sell, D. Lieberman, and D. Sznycer. 2008. Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In *Handbook of approach and avoidance motivation*, edited by A. J. Elliot, 251–71. Mahwah, NJ: Lawrence Erlbaum Associates.
Tooby, J., N. Thrall, and L. Cosmides. 2006. The role of 'outrages' in the evolved psychology of intergroup conflict. Paper presented at 18th annual meeting of the Human Behavior and Evolution Society, HBES, University of Pennsylvania, June 7–11, 2006.
Trivers, R. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46:35–57.
Ugwuegbu, D. 1979. Racial and evidential factors in juror attribution of legal responsibility. *Journal of Experimental Social Psychology* 15:133–46.
Vangelisti, A. L., J. A. Daly, and J. R. Rudnick. 1991. Making people feel guilty in conversations: Techniques and correlates. *Human Communication Research*, 18:3–39.
Walker, D., and R. Gorsuch. 2004. Dimensions underlying sixteen models of forgiveness and reconciliation. *Journal of Psychology and Theology* 32:21–5.
Warr, M. 1989. What is the perceived seriousness of crime? *Criminology* 27(4):795–822.
Williams, G. C., and D. Williams. 1957. Natural selection of individually harmful social adaptations among sibs with special reference to social insects. *Evolution* 11(1):32–9.
Wilson, J. Q., and R. J. Herrnstein. 1985. *Crime and human nature.* New York: Simon and Schuster.
Wispé, L. 1991. *The psychology of sympathy.* New York: Plenum Press.