19

Institutional Design Capitalizing on the Intuitive Nature of Decision Making

Mark Lubell, Rapporteur

Christoph Engel, Paul W. Glimcher, Reid Hastie, Jeffrey J. Rachlinski, Bettina Rockenbach, Reinhard Selten, Tania Singer, and Elke U. Weber

Neural and Psychological Processes as the Substrate of Institutional Design

What insights about institutional design can be gained by considering that the processing of information is not always under conscious control? At first, it might seem that the very purpose of most institutional efforts to control decision making is to produce decisions that are conscious, slow, deliberative, effortful, rational, and unemotional. Institutional design would thus seem targeted to produce what psychologists sometimes refer to as "system 2" thinking (Kahneman 2003; Kahneman and Frederick 2005). From this perspective, institutional interventions should drive out cognitive processes that are unconscious, rapid, automatic, associative, holistic, and emotional; that is, to reduce to "system 1" thinking. If there is reason to believe that system 1 processes will be at work, additional institutions would aim at bringing cognition under conscious control, or at least at limiting the undesirable influence of system 1 thinking on decisions and behavior.

In this report we demonstrate that institutional efforts to improve decision making are not, and should not be, directed solely at producing system 2 thinking. No cognitive process is superior in all settings. Consequently, good institutional design ought to aim at the best match between cognitive processes and tasks. In appropriate contexts, this includes capitalizing on the power of intuition.

As other chapters in this volume reveal, the distinction between conscious and unconscious is a somewhat artificial construct (Platt et al., this volume). So, too, is the supposed "system 1" versus "system 2" distinction (Keysers et al., this volume). Human behavior is certainly the product of a combination of processes that implicate concepts of consciousness and intentionality in different ways. Since the cognitive revolution, psychologists have worked to identify and define cognitive processes, without regard to the level of awareness that they implicate. The picture becomes exponentially more complicated if one also takes the underlying neural structure into account.

Despite the artificial nature of the distinction between conscious and unconscious processing, many social institutions assume that such a distinction exists and is important. For example, the urge to impose a sanction on another person if one has been hurt "intentionally" is felt more intensely than if the same mischief has been caused by this person's negligence (Falk et al. 2000). In addition, institutional settings often provide instructions to decision makers to make careful, deliberative choices, as in the jury system (discussed below). The use of the conscious/unconscious distinction as a useful shorthand at the institutional level, however, produces methodological and practical challenges for social scientists, which we illustrate in Figure 19.1.

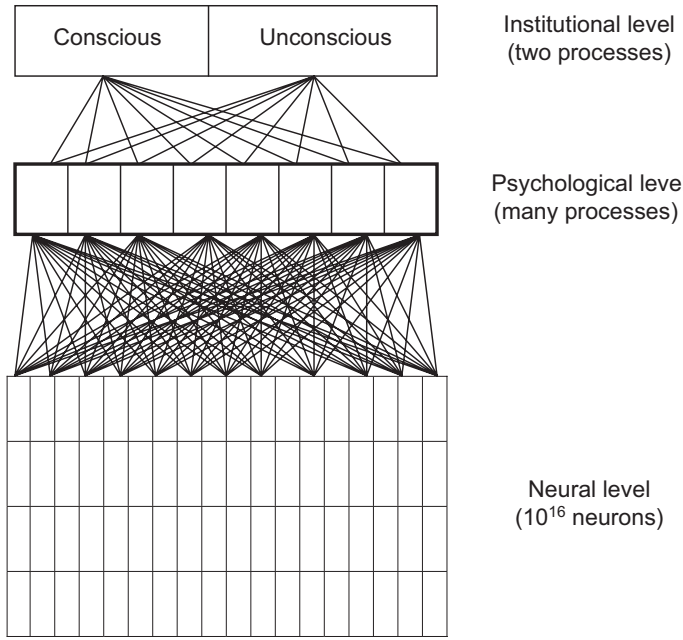

Figure 19.1 A mapping of institutional, psychological, and neurological processes.

Figure 19.1 reveals the impediment to converting an understanding of neural processes into institutional design. This might seem obvious to many; after all, understanding a tiny neuron could hardly be translated easily into understanding a complex social institution. It is tempting to think of Figure 19.1 as an inverted pachinko machine: a neural firing at the bottom level could translate into any number of different, unpredictable cognitive processes, which in turn translate into either a conscious or unconscious cognitive mechanism, thereby producing almost any conceivable behavioral output. The case studies, below, suggest that the analogy is not quite apt. We believe that, if used appropriately, psychological and even neurological knowledge can indeed help improve institutional design (although, we admit, pachinko players also harbor such beliefs, and they are certainly mistaken).

In this chapter, we identify three cases in which consciously inaccessible mental processes matter for institutional design. In principle, the distinction between the conscious and unconscious affects the analysis of all institutions. This distinction affects the constellations of formal rules and informal social norms that structure human behavior and thereby define social institutions (North 1990). It should be relevant, whether the institution has been deliberately designed or emerged through some evolutionary process, and whatever its stated or constructed goal. In our case studies, however, we rely predominantly on legal institutions: judicial decision making by judges or jurors, the interaction between legal contracts and interpersonal trust in shaping cooperative relationships, and standards of criminal responsibility. Our three case studies track the different types of interactions between the conscious/unconscious distinction and the legal system. This distinction influences the development of rules and structures that the legal system uses to advance the quality of decisions it makes; it influences how the system attempts to affect the behavior of citizens; and it guides how the system judges the conduct of its citizens. For the first case, we focus on the mental activity of the people who make decisions within the legal system; that is, judges or jurors. In the second and third cases, we investigate the efforts of the legal system to shape behavior. The first case may spur us to formulate questions that neuroscience might be able to answer in the future. For the other two cases, neuroscience data exists upon which legal institutions might build. In the case of legal contracts and trust, a better understanding of psychological and neural processes would without a doubt be helpful for institutional analysis and even design. In criminal responsibility, however, understanding the underlying mental forces may make it more difficult for the legal system to serve its historical function.

Judge and Jury Decision Making

Reid Hastie (this volume) describes a simulated trial in which the defendant Frank Johnson was charged with the first-degree murder of his drinking buddy,

Alan Caldwell. Although many of the events in the trial are undisputed (e.g., Johnson definitely stabbed Caldwell), a series of disputed facts exist such as whether or not Johnson acted in self-defense. The disputed facts all bear on legal standards that define guilt and type of offense, as well as punishment severity.

In this context, the goal of the jury system as an institution is to determine whether or not Caldwell is guilty according to the legal standards. The task of the judge or jury in such cases is "to get the story right" in order to apply the relevant standards correctly. The legal system desires to avoid both wrongful convictions and false acquittals, but it is weighted more heavily toward avoiding wrongful convictions, as most legal systems treat these as much more destructive than false acquittals.

Designing an institution to maximize correct verdicts while satisfying the criteria of weighing wrongful convictions more heavily than false acquittals would be simple enough, if the decision maker relies solely on conscious, system 2 processes. The judge or jury would observe the evidence presented by both sides, weigh the evidence appropriately in reference to the legal standards, and then integrate the available evidence into an unbiased summary judgment. The procedural rules for such a decision maker would need to be designed to ensure an efficient presentation of the evidence. Live presentation, for example, would hardly be meaningful, unless it enhanced the decision maker's memory of the relevant facts. Likewise, the order of the presentation would matter little. Oral arguments by attorneys would be unnecessary, except insomuch as they would remind the decision maker of the task or evidence. The emotional state of the witnesses would be of little worth, unless it provided a clear indication of the witnesses' mental state. Likewise, invidious influences (e.g., race, income, or gender) of the parties or witnesses would matter only to the extent of their strict Bayesian relationship to the underlying facts. Finally, any decision maker with the cognitive skills necessary to master the rules of inference and process the evidence would suffice.

Of course, this scenario depicts no process in use by any country. Trials are messy everywhere, and decision makers less predictable. The prosecution and the defense bring forth pieces of evidence in a scrambled order, based on witness availability and attorney strategy. Both sides are represented by professionals, who know they get a high premium if they successfully shift how the decision maker views the case. Legal systems adopt rules that carefully prescribe the mode, order, and content of evidence presentation as if all of these seemingly extraneous rational factors affect judgment. Research supports the influence of such factors.

How does the decision maker react to all of this? Hastie presents a "story model" of juror and judge decision making that is described in terms of general models of explanation-based judgment (Hastie 1999, 2001, this volume; Hastie and Pennington 2000). The basic intuition of this model is that legal decision makers reason about the evidence to form a mental model of events and causal relations; thereafter they compare that abstract mental model to relevant legal

classifications to arrive at a decision. The legal decision makers will consider multiple possible models and choose the model with the highest level of confidence, as determined by criteria such as coverage (taking into account all the facts), consistency, plausibility, and completeness (Hastie, this volume).

Hastie's story model, however, does not fit easily into the simple distinction between conscious and unconscious. To refer back to Figure 19.1, the story model is a construct that operates at the psychological level. Nevertheless, it seems to describe the juror decision-making process accurately. Hence, institutional-level approaches that embrace a binary distinction between rational or conscious cognitive processes and the irrational or unconscious cognitive process will necessarily miss the mark in terms of their value to accurate fact-finding. The story model would generally encourage procedural rules that facilitate jurors' natural tendencies to try to construct an accurate account of the events. This might include such reforms as allowing jurors to ask questions or take notes. Judges might also provide instructions at the beginning of the case, rather than at the end, to provide jurors with a cognitive template to organize the evidence that they will receive. Furthermore, a court might attend more carefully to the order in which the evidence is presented, perhaps ensuring a chronological presentation, to facilitate the construction of a stable story.

In the United States, at least, evidentiary rules do not embrace the psychological-level concept of the story model; instead, they are attuned to concerns about the distinction between conscious and unconscious processing. The system distrusts unconscious influence, in some instances, and embraces rules to constrain the influence of unconscious processes. In other contexts, the system attempts to take advantage of the strengths of unconscious inference processes. Finally, legal systems rely sometimes on the intuition of the decision maker assessing the appropriate level of confidence in their judgment. Below, we review briefly each of these three approaches to the distinction between conscious processing.

In the U.S., the legal system has adopted numerous procedural rules to limit the influence of seemingly undesirable unconscious influences. For example, the Federal Rule of Evidence (Rule 403) precludes the introduction of evidence that presents a risk of "unfair prejudice." The Advisory Committee to the Federal Rules of Evidence explains the rule specifically as forbidding evidence that has "an undue tendency to suggest decision on an improper basis, commonly, though not necessarily, an emotional one." A common application of this rule restricts the presentation of particularly graphic evidence, such as bloodied bodies or other evidence of physical pain (Bright and Goodman-Delahunty 2006). Graphic evidence might initiate automatic processing to evoke a desire for vengeance against the defendant, regardless of whether the evidence supports a conviction. The rule assumes that a juror exposed to such evidence will attend excessively to the single piece of graphic evidence, thereby biasing jurors' assessment of the rest of the evidence. Restrictions like this on evidence

are thought to be essential to restricting emotional processes that jurors cannot control, even if they are instructed to do so.

In other cases, the legal system takes full advantage of intuitive processing and does not attempt to restrict its influence. For example, judgments of culpability and liability frequently require inferences about the motives, goals, intentions, and emotions of the actors involved. It is easy to underestimate how difficult such inferences are to make. Mental states must be reliably inferred from behavior, otherwise the system will be too erratic. Furthermore, the members of the society in which the legal system operates must have a shared system of understanding intentionality. The human brain appears to be quite adept at such inferences. Social neuroscience has started to understand the neural networks underlying such inferences, for example, through cognitive (mentalizing) and emotional perspective taking (empathizing). Mentalizing refers to our ability to represent intentions and beliefs of others, whereas empathy refers to our capacity to share feelings and motivational states with others. Empathy, in particular, has been shown to occur on an implicit level of processing; that is, people seem to represent the emotional states of others quite automatically without having to engage in deliberative thinking (Singer 2006; see also McCabe and Singer, this volume). The same networks that process the sensations of touch, pain, and disgust in one person are reactivated when that person perceives others experiencing similar emotions. These automatic empathic responses can be modulated by perceived fairness, similarity, or affective link to the other (for a more detailed account, see de Vignemont and Singer 2006; Singer 2006; Singer et al. 2004, 2006).

A legal system that bases important judgments on various mental states must depend on these automatic, neural processes. Lawyers recognize clearly the existence of such processes in their strategic presentation of evidence and jury selection. Even more significantly, the system itself depends on these processes. Basically, legal systems assume that judges and juries will be able to infer mental states reliably. Jury instructions, for example, instruct jurors to assign verdicts on the basis of mental states, but provide little or no guidance as to how they are to ascertain mental states.

The case that Hastie used to illustrate the story model is typical in the instructions it provides to jurors on how they are to infer the mental states of the actors. The instructions advise the jurors to convict the defendant of murder if the defendant intended the death of the victim when he acted. They further instruct the jurors to convict of a more serious form of murder if the defendant formed this intent in advance and planned the actions in some way. They also require that the jurors acquit the defendant if they conclude that the defendant acted out of a reasonable belief that his own life was in danger. Jurors who have the cognitive mechanisms for understanding purposive action, planning, and fear can make such judgments without further elaboration. The system provides no clear rules for making such inferences, and it need not do so. The evidence from neuroscience suggests that the human brain can perform this

function to a reasonable degree of reliability on its own, without invoking an elaborate set of deductive rules. Legal systems seem to depend on this ability.

In addition, legal systems appear to rely on intuitive processes in a more indirect way. Recall that most criminal justice systems are more concerned with avoiding wrongful convictions than false acquittals. Operationalizing this concern can be challenging. The English jurist, William Blackstone, asserted that he would rather let ten guilty men go free than convict one innocent man. This aphorism provides a sense of the magnitude of the weighing, but can hardly be translated into a handy rule of inference. The American system demands that a trier of fact in a criminal case (usually a jury) not convict unless they are certain of the defendant's guilt "beyond a reasonable doubt." This phrase admits of no easy definition, but it seems to convey a sense of the mental state required, as this passage from the well-known jurist Justice Lemuel Shaw (*Commonwealth* v. *Webster* 1850) illustrates:

> [W]hat is reasonable doubt? It is a term often used, probably pretty well understood, but not easily defined. It is not mere possible doubt; because every thing relating to human affairs, and depending on moral evidence, is open to some possible or imaginary doubt. It is that state of the case, which, after the entire comparison and consideration of all the evidence, leaves the minds of jurors in that condition that they cannot say they feel an abiding conviction, to a moral certainty, of the truth of the charge.

The concept of reasonable doubt refers then to a state of mental certainty, not to a probability judgment (*Victor* v. *Nebraska* 1994).

The concept of *conviction intime*, which governs how judges report their decisions in French and German courts, presents similar issues, albeit with a somewhat different resolution. The principle requires that judges state they are 100% convinced of their reasoning in a particular decision, even if the evidence suggests a more probabilistic or uncertain conclusion. Judges are not allowed to make any quantitative probabilistic statements that quantify their uncertainty about their judgment (e.g., I am 95% certain that the defendant did it). Essentially, *conviction intime* asks judges to rely on their expert intuition to persuade themselves fully about the correct story. Especially in uncertain cases, this process calls on a judge to make a variety of inferences on the basis of contradictory or uncertain evidence, so as to facilitate a kind of overconfident certainty in the decision maker. As with the reasonable doubt standard, *conviction intime* may be best understood as inducing the decision maker to reach a particular mental state. Judges must all share the same sense of what constitutes this mental state or else the principle cannot produce reliable judgments. *Conviction intime* appears to demand a somewhat more rigorous, almost deductive certainty, thereby distinguishing it from the reasonable doubt standard. However, like the reasonable doubt standard, it takes advantage of a shared understanding of how mental states interact with judgment.

These three examples of how procedural rules interact with intuition suggest several important puzzles about legal systems. First, why do legal systems sometimes attempt to suppress intuition and at other times they rely on it? The answer might be mundane: some intuitions are thought to be impediments to the system's goals whereas others are seen to further its goals. Second (and relatedly), do legal systems purport to suppress unconscious processes in an effort to maintain the appearance of deliberative processing? *Conviction intime* seems to have that character, given its insistence on an expression of utmost certainty. Social order might depend on the perception that the legal system inflicts harsh punishment only in the right cases, and only for acceptable reasons. We certainly cannot rule out the alternative hypothesis; namely, that rules such as *conviction intime* do what a cursory inspection might suggest—they increase deliberative processing. If this is true, then either deliberative processing produces better judge and jury decisions, or there is an intriguing trade-off between maintaining the legitimacy of the political system and making correct legal decisions.

Third, why does the legal system seem to divide cognitive processes into conscious and unconscious, instead of sorting them at the level of psychological processes? This distinction certainly has its costs, as it impedes efforts to assess reforms that might otherwise take advantage of the subtle way that fact finders approach their decisions. The notion that some jurors process some evidence consciously and other evidence unconsciously fails to describe what it is that jurors are doing accurately. In reality, conscious and unconscious or semiconscious processes all interact to facilitate the production of an account of the facts that judges or juries can map onto various legal categories. The legal system's folk theorizing about human cognition might be a bit too wooden.

Finally, an assessment of the cognitive processes that influence judgment raises important puzzles as to the sources of difference in legal doctrines between countries. For example, the legal system in the U.S. does not have the *conviction intime* principle. Although the reasonable doubt standard is somewhat similar, as described, it tolerates doubt explicitly. Furthermore, the U.S. legal system relies heavily on juries. How these variations developed can be traced historically in each country. The observable differences seem to have come about because (a) the goals of the systems were different; (b) the nature of the decision process varies cross-culturally; or (c) one of the two systems is more effective.

To summarize, we view legal systems as relying on a distinction between conscious and unconscious processes in crafting procedural rules that govern trials. Legal systems sometimes suppress unconscious inferences, while relying on them in other contexts. As Hastie's research has shown, however, this distinction is both artificial and somewhat detrimental, as it emphasizes a particular set of questions that do not address how decision makers actually assess the facts.

Trust and Social Cooperation

Cooperation When Contracts Are Incomplete

In many social situations, individual self-interest impedes a mutually beneficial outcome that requires cooperative effort among individuals. In such situations, individuals face incentives to free ride; that is, they make no contribution to the costs for preservation while benefitting from the efforts of others. If each member acts as a free-rider, social benefits will not be realized. Game theorists formalize this situation as a public goods game, the multiplayer version of the popular Prisoner's Dilemma.

In the Prisoner's Dilemma, a row and a column player simultaneously choose to cooperate or to defect (see Figure 19.2). If both cooperate, both achieve their second best outcome. If one of them cooperates while the other defects, the defector attains her best outcome, while the cooperator gets her worst outcome. Thus, both players have an incentive to defect, producing their second-worst outcomes. A combination of greed and fear leads to the socially detrimental result (Macy and Flache 2002). Greedy individuals attempt to obtain their individual best outcome: the free lunch. Fearful individuals anticipate this and defect because they do not want to be the loser. This distribution of payoffs and the underlying preference structure are specific to the Prisoner's Dilemma. However, there are many more positive sum games that are not games of harmony (Rapoport et al. 1976).

The dilemma would not exist, if the parties involved could write a complete contract prior to playing. This would mean that each player would commit to a strategy (preferably cooperate), all subsequent actions in the game are verifiable (i.e., whether a player cooperated or defected can be confirmed beyond doubt), and the actions are enforceable by a third party. Writing such a complete contract is frequently not possible under real-life conditions. Anticipating all of the possible means of defection is rarely possible, and attempting to do so can be expensive.

Rather than draft a complete contract demanding cooperation in all circumstances, parties can rely instead on trust. In a trust relationship, a trustor deliberately takes the risk of being exploited by the trustee (Hardin 2001). How these trust relationships develop in everyday life have been experimentally analyzed in two classes of games: the trust game (Berg et al. 1995) and the gift exchange game (Fehr et al. 1993).

	Contribution	Defection
Contribution	2, 2	0, <u>3</u>
Defection	<u>3</u>, 0	<u>1</u>, <u>1</u>

Figure 19.2 Prisoner's Dilemma: the left-hand payoff is for the row player while the right-hand payoff is for the column player. Best results are underlined.

The basic setup of the trust game involves a sender, who is endowed with some amount of money (e.g., ten tokens) and a receiver. The sender may send any fraction of their endowment to the receiver, which is augmented en route, say tripled. The receiver then has the option of returning any amount of the received money back to the sender. Obviously, the joint profit of both players will be maximized if the sender transfers the entire amount to the receiver. There have been literally hundreds of experiments on the trust game conducted under very different conditions which, roughly summarized, show that the more senders give, the more is returned. A recent study has shown that this phenomenon occurs as well in market-like situations (Brown et al. 2004).

Although the evidence of widespread trust is impressive, these experiments also reveal that trust represents a fragile solution for solving the cooperation problem. Some senders do not send and some receivers do not return. In the absence of knowing whether one is facing someone who is apt to be trustworthy, one is well advised to spend the extra cost to enter into enforceable contractual protections. Any institutional intervention that increases levels of trust would therefore reduce the need to undertake such costs. By identifying the mechanisms that inspire trust, neurobiological research suggests interventions that might induce greater levels of trust and increase the probability of obtaining cooperative solutions to public goods problems.

Research in neurobiology has revealed a widespread set of mechanisms that operate in an unconscious fashion to improve trust. For example, a recent study (Kosfeld et al. 2005) has shown that the level of trust increases in subjects if oxytocin[1] has been sprayed into their nose. In the trust game subjects exposed to oxytocin sent more money. Preliminary results from imaging studies (fMRI) suggest that this effect is mediated through the amygdala, a brain structure crucial to fast emotional processing, fear conditioning, and reward learning. Oxytocin reduces activation in the amygdala, thereby muting an alarm signal that would otherwise induce avoidance tendencies. This study demonstrates a deeply embedded neurological pathway that induces trust.

In addition, preliminary studies from Tania Singer's lab suggest that subjects exhibiting mistrust do not differ with respect to their empathic brain responses elicited when others suffer pain (ACC, insula), but rather with respect to their amygdala activation elicited when they anticipate receiving painful stimulation themselves. Such subjects seem to have enhanced amygala-mediated fear responses to aversive events (Singer et al., in preparation). Going back to the Prisoner's Dilemma, these results would suggest that cooperation fails not because of greed, but rather fear. Hence, institutional mechanisms that

[1] Oxytocin is a peptide produced primarily in the hypothalamus; it works both as a hormone and a neuromodulator. In nonhuman mammals, oxytocin has a central role in behavioral regulation, particularly in positive social interactions ranging from social attachment such as pair bonding, to maternal care and sexual bonding (Carter 1998; Carter et al. 2001; Insel and Young 2001; Insel and Shapiro 1992; Pedersen 1997; Uvnas-Moberg 1998). Generally it is thought to facilitate approach behavior and to overcome natural avoidance tendencies.

reduce the perceived risk of victimization would be the key to increasing trust, and hence cooperation.

Evidence from experimental economics invites a further hypothesis at the border between psychology and the neurosciences. If true, the effect would not modulate trust, but trustworthiness. In oligopoly experiments, a number of experimental subjects compete. The experimenter is interested in exploring deviations from the Walrasian equilibrium (i.e., price equal marginal cost). In principle, demand is not part of the research question. To economize on the subject pool, many oligopoly experiments have replaced demand by a computer that is programmed to buy according to a predetermined demand function. Some experimenters wondered whether this manipulation is as innocent as it seems. This turns out not to be true. Collusion plummets if the computer is replaced by human buyers (Engel 2006). One plausible interpretation is that experimental subjects are hesitant to inflict harm on their peers.

To the best of our knowledge, this hypothesis has not yet been tested experimentally. It has, however, been shown that experimental subjects exhibit empathy when they see pain inflicted on others (Jackson et al. 2005; Singer et al. 2004, 2006). This is also the case if, in a coordination game, those who have behaved fairly nonetheless get punished (Singer et al. 2006). What happens if subjects have a chance to inflict harm on their peers? This remains to be tested. Arguably, empathy results in sympathy and should make them shy away from this. If true, the definition of peers and the perception of whether a subject feels treated fairly by them becomes important.

Therefore, being vulnerable in a positive sum game appears to be beneficial for all players. If this interpretation holds, then a person might be motivated to exploit the neurological mechanism that brings about trust by projecting an image of oneself as being vulnerable. One way to do this would be to expose oneself to a serious (but not vital) risk.

Although trust may help parties secure gains from cooperation, it remains a risky choice. A solid body of experimental evidence demonstrates that actors are heterogeneous. A rough estimate has it that Western societies typically are composed of about 60% of conditional cooperators, 30% of utility maximizers, and 10% of altruists. In such populations, blind trust is not sustainable, as can be shown by modeling the situation in terms of evolutionary game theory (Boyd et al. 2003; Gintis et al. 2003). In the spirit of rational choice, in such a setting actors will watch out for information that signals the type of their interaction partners.

This insight might help explain a curious feature of both common and continental law. In both traditions, a contract represents a promise that is enforceable through the power of the courts. To be legally enforceable under the common law system, however, a contract requires an offer and acceptance of a valuable consideration; a mere promise is not enough. For example, if someone promises to give you their laptop computer tomorrow and asks for nothing in return, when this person fails to deliver the laptop, you would not be able to enforce

the promise through the courts. By contrast, if a person promises to sell you their laptop for $1 and you agree, both of you have exchanged consideration, and the promise becomes enforceable. An exchange does not need to be made up of goods of equal value; they must simply have some value. Furthermore, consideration is said sometimes to have "failed" when the consideration is too vaguely described. Hence, if I promise to give you my laptop and ask that you will "consider it as a favor," the arrangement still lacks valuable consideration. Why should the law insist on a concrete exchange? Perhaps because the exchange of concrete consideration triggers the neurological mechanisms described above: it induces people into mutual exchange, and thereby encourages the kind of cooperative thinking that will reduce the chances that the parties will exploit each other later as the conditions of the contract unfold.

Interestingly, not all legal systems require a consideration. Under many continental legal orders, a mere promise is binding, but only if it is affixed in writing (for a discussion on the actual degree of convergence, see Rothoeft 1968). The act of writing, however, might serve the same psychological function as the mutual exchange in the common law system, as the construction of a written agreement requires the type of cooperative thinking that also underlies the exchange of consideration.

Although such interventions reduce the risk that a trust relationship breaks down, or is not initiated, they are not, in and of themselves, sufficient to sustain trust. A sustainable strategy combines three elements: signal your willingness to cooperate at the outset; commit to punish defectors; in punishment, signal that you want to continue the relationship (Selten et al. 1997).

Having addressed the first element, let us now turn to the second. There is good evidence that, in a social dilemma, the chance to pay for inflicting harm on defectors improves cooperation (Fehr and Gächter 2002). From a rational actor perspective, this is surprising. Why should an individual be willing to give up some of their own utility for doing harm to cheaters? The original Prisoner's Dilemma repeats: each member of the community is best off if all others punish while the individual member does not (and keeps all her money). This is anticipated. No one ever punishes (Heckathorn 1989).

Actually, experimental subjects have a preference for punishment as an institution. If given a choice, despite a strong initial reluctance, subjects migrate rapidly to a sanctioning institution which "produces" almost full contributions and thus no punishment (Gürerk et al. 2006). This even holds if they could alternatively choose reputation to back up cooperation (Milinski and Rockenbach 2007).

Recent studies in neuroscience have started to investigate the neural foundation of such punishment sentiments. De Quervain et al. (2004) scanned subjects while they were deciding whether and how much punishment points they wanted to distribute to people who had behaved unfairly in a previous game. Activation in the dorsal striatum, a brain structure involved in processing stimulus-reward-response contingencies, was activated when subjects decided

to punish, and these activations were positively correlated with the amount of punishment delivered. A recent study by Singer et al. (2006) suggests that punishment is not only rewarding when given the opportunity to execute the act, but also when subjects merely watch other defectors being punished (i.e., without executing the punishment themselves). In this experiment, subjects were scanned (and their empathic brain responses measured) as they observed people, who had either behaved fairly or unfairly in a sequential trust game, receiving painful stimulation to their hand. Both women and men showed empathic brain responses toward the suffering of fair players. Men, however, showed an absence of empathic brain responses when the unfair player got "punished," revealing instead an increased activation in the ventral striatum, an area involved in primary reward processing. The higher these men expressed subjective desires for revenge in later questionnaires, the higher their response in these reward areas. These results suggest that seeing defectors punished is rewarding and may help explain the motivation for altruistic, that is, costly punishment.

Thus prepared, we return to the greed versus fear explanation of mistrust. If indeed fear is key, society might be better off insuring trustors for the (arguably small) risk of being exploited by their counterparts. In an experimental trust game, this could be implemented the following way: at the outset, the experimenter imposes a tax. Out of her endowment, each subject has to pay a fraction into a pool. The pool compensates trusting subjects who have been exploited. Of course, punishment and insurance could also be combined. In such a scheme, not only would the defector's utility be reduced by (costly) punishment; the money thus collected by the bank would be used to (partly) compensate victims for their losses.

In the literature on punishment, this is usually seen as unilateral. The victim inflicts harm on the perpetrator. Institutional practice exhibits an alternative that engages the perpetrator. U.S. legislation, for example, propagates apologies. It does so mainly by prohibiting the apology to be introduced as evidence in a civil trial, thereby protecting the defendant from liability and reducing the cost of apology (Robbenolt 2006). Anecdotally, people injured through the negligence of others often tell their lawyers that all they really want is an apology. Furthermore, some defendants seem to want to apologize (O'Hara and Yarn 2002). One interpretation of this is that despite the defection, which led to litigation, an apology signals that one is planning to maintain a cooperative relationship. Apologies are probably effective because of the emotional content they often convey, which may be subject to automatic processing. The neuroscientific foundation of the effect of apologies has not yet been investigated. It may be hypothesized that apologies reduce reward-related signals motivating punishment behavior and, through this, reconstituting empathy. Alternatively, apologies may also have a similar effect to oxytocin in that they reduce fear signals of the amygdala and recover trust.

We now turn to the third element in the Selten et al. (1997) argument. Punishment is foolproof only in very simple settings. Whenever there is noise, the risk exists that the person on which punishment is inflicted sees this as unfair and counter-punishes (Konow 2005; Nikiforakis 2008). From this, a war of revenge can easily result, which brings cooperation to an abrupt end.

We hypothesize that "pedagogical" punishment (as found to perform best in Selten et al. 1997) and "anger" punishment are different neurologically. If institutions are able to make anger punishment unlikely (e.g. by imposing social norms), the risk of the first punishment resulting in a war of revenge is substantially reduced. Neuroscientific studies may shed some light on this suggestion. However, clever experiments will need to be designed to contrast brain signals underlying punishment motivated by retributive motives and punishment motivated purely by consequentialist motives to see whether reward-related processing is found only in the former decision. Another interesting study would be to train people in empathy-enhancing techniques and to see whether, using a similar paradigm to that used by Singer et al. (2006), subjects are able to overcome retributive motivation toward defectors after empathic induction.

Thus far, the only complications we have introduced are actor heterogeneity and noise. We have suggested minimal institutional interventions to stabilize trust. However, in many real-world interactions, a higher degree of formality is hard to avoid. Division of labor must be organized in greater detail. Some risks are so characteristic to the types of interaction that a responsible partner will specify future outcomes, should these risks materialize. In short: formal contract is advisable.

In such contexts, the contracting parties face a serious challenge, provided making the contract complete still remains impossible or inappropriate. Fehr and Rockenbach (2003) have shown that subjects who deliberately dispense the possibility of punishing undesired behavior, although it is contractually possible, reach higher efficiency than those who incorporate a punishment clause in the contract. Along similar lines, Falk and Kosfeld (2006) show that it is not beneficial for the proposer to reduce the reaction space of the opponent by withdrawing the actions that are "unfair" with respect to the proposer, because this signals distrust. Asking for a certain clause may be a strong signal of distrust. More importantly even, the parties must manage to negotiate in a way that leaves room for building trust.

Default contracts stipulated in a statute provide an alternative to this dilemma, as do standard form contracts, which are widely applied. Either way, the parties are not forced to make unlikely risks explicit. Further research is needed on the conditions under which negotiations do build, instead of erode, trust. Anecdotal evidence seems to point to the possibility that trustworthiness is signalled by the way in which negotiations are conducted.

Therefore, legal institutions governing contracts seem to take limited advantage of the kinds of mechanisms that neuroscientists studying the biological basis of trust have uncovered. Considering that common law and continental

legal systems have evolved over centuries, it would perhaps be surprising if these systems took no account of the role of the human brain, as well as its strengths and weaknesses. However, once again, the law as an institution emphasizes primarily the first-level analysis of conscious versus unconscious. We suspect that the examples we have presented, in which the law incorporates some insights from neuroscience, are the exception rather than the rule. Most contract law, like procedure, treats people as rational, conscious agents who are held responsible for their actions, unless these actions can be shown to be the product of significant unconscious influences, akin to duress and deception. Hence, we anticipate that the law has much to gain from the insights of neuroscience, in particular from the light it sheds on how trust is inspired.

Standards of Criminal Responsibility

Our last case study builds on an analysis of the idea of "criminal intent" conducted by Paul Glimcher (this volume). Glimcher discusses a U.S. Supreme Court case from the early 1990s concerning a murder suspect named Solomon Hotema. In this case, the Court established the standards of criminal intent as a voluntary action and stated that subjects who are absolutely controlled by the delusions of a "diseased" brain are entitled to an acquittal. Glimcher's main argument is that this notion of voluntary action is inconsistent with the materialist view of neurobiology. Thus, in this section, we will utilize a somewhat reverse logic from that of the previous two, and begin by examining this institution with an eye toward assessing the implications of neurobiology. We discuss how the institution might change if neurobiological evidence were taken seriously.

Perhaps more so than in our previous discussions, criminal law categorizes human actions into conscious and unconscious. The law holds people responsible for their actions that it deems to be the product of conscious thought—equating conscious thought with volition. It exempts from liability people who commit a wrongful act, but who can be shown to lack the same conscious mental state as most wrongdoers. For example, courts would not treat a person who throws a knife into another as a result of an epileptic fit as they would an assassin who commits the same act as part of a contract murder. Courts approach cases of severe mental illness that lead to claims of insanity in much the same way; they try to identify whether a wrongdoer's mental state most closely approximates the cool deliberation of the assassin or the uncontrollable gyrations of the epileptic. Sufficient evidence of irrational thought leads courts to treat the mentally ill offender as the law would treat the epileptic. In cases of severe mental illness, the courts conclude that the crime is attributable to the person's (irrational, unconscious) brain, not their (rational, conscious) mind.

Neurobiology challenges this dichotomy. For example, serotonin levels in the brain have a close relationship with the likelihood of violent crime. Decreases in serotonin level are known to increase the rate of violent acts and

depression in many individuals. (Rossby 2003). Imagine that the relationship was perfectly linear: more serotonin, more violent crime. It would be hard to maintain the dichotomy of criminal law in the face of such evidence. How could one say that a person is responsible for their serotonin level? Still, serotonin bears only indirectly on the distinction between conscious and unconscious processes that the courts rely on in cases of claimed insanity. Although neat linear relationships between neurological activity and behavior have yet to be identified, neurologists believe that neuroscience will uncover findings that will undermine the basic dichotomy of criminal law.

With this background, one can make a distinction between a retributive model of a criminal punishment and a consequentialist model. Most modern legal systems are built strongly on a retributive model, which asserts that punishment should be meted out to exact retribution on the criminal, punishing him for his defection from society. The retributive model is supported by experimental social dilemmas with punishment institutions, where subjects often express anger at defectors and exhibit a desire for vengeance. Retribution is critical to the justification of criminal law. A moral society must have a reason to inflict hardship on one of its members. The attribution of deliberate, intentional misconduct provides the basis for retribution against offenders. Punishing those whose conduct was the product of something other than rational, volitional thought would be no different than punishing an innocent.

The retributive model, however, rests on the assumption of voluntary action: punishing a criminal who did not voluntarily commit the crime is a morally bankrupt undertaking if done solely for retributive ends. Criminals with "diseased minds" that impair voluntary action do not deserve punishment.

Retributivism, however, is not the only motive for a criminal justice system. Punishment also serves to decrease the incidence of crime in society, either by removing violent individuals or deterring other potential criminals. This deterrent motive is a form of consequentialist justification for inflicting hardship on those who commit crime. The consequentialist model is concerned solely with improving the overall welfare of society by discouraging defection. Because retribution is irrelevant to the consequentialist view, it does not require any notion of voluntary action or freewill; its desired effect is achieved by changing the expected costs of being caught. A materialistic view of neurobiology rejects the notion of voluntary action and thus implies logically that only a consequentialist model of criminal justice is morally defensible. A consequentialist view of criminal justice would have a punishment severity function to track directly the likelihood of violent crime.

Although both retributivist and consequentialist goals animate the existence and function of most systems of criminal law, retributivism predominates today. In many instances, retributivism and consequentialism do not produce different policies. For example, neither would advocate punishing the epileptic in our earlier example, and both would assign harsh punishment to the contract killer. When the policies diverge, however, people reject consequentialism and

embrace retributivism (Robinson and Darley 2003). For example, consequentialism demands that people administer punishment in ways that are sensitive to the probability that a wrongful act is detected, and yet people do not behave this way (Sunstein et al. 2000). Similarly, even though roughly two-thirds of the citizens in the U.S. support the use of the death penalty, less than one-third believe that it provides a useful deterrent[2]—retributive ends provide the support for this punishment.

Consider the implications of the relationship between serotonin and crime control. Retributivism and consequentialism implicate different policies in response to the relationship between serotonin levels and the propensity to commit crime. A retributivist would see the role of serotonin as an involuntary cause of crime that is not properly attributed to the perpetrator. Hence, for a retributivist, increased serotonin levels would justify less harsh punishment. A consequentialist, however, would view the increased propensity to commit murder as justifying an increase in the penalty for crimes committed by those with high serotonin levels. To deter crime effectively, society has to send a stronger signal to those with high serotonin levels. Although we have not done a survey, we suspect that most people would see a high serotonin level as justifying less punishment. As neuroscientists identify the precursors of crime ever more clearly, more factors like serotonin may come to affect the criminal law. Neuroscience is thus on a collision course with the intuition behind criminal law (Greene and Cohen 2004).

It is not clear whether abandoning the notion of voluntary action (favored by a retributive system) and adopting a consequentialist institution would have positive benefits for society. Research on social dilemmas suggests that the opportunity for retributive punishment is an important basis for maintaining cooperation. Thus, if the opportunity for punishment were to be removed, other people in society may be less likely to cooperate. In many ways, the legal system is designed to make punishment more efficient by providing economies of scale and by removing the need for individuals to mete out their own vengeance. Most importantly, legalizing punishment is a technology for containing the risk of counter-punishment, which we have analyzed in the previous section. This, however, is not likely to work if acts of the criminal justice system are perceived as patently unfair. This is precisely what happens if the courts punish defendants with "a broken brain." Thus adopting a consequentialist system might result in reducing a citizen's belief in the justice of the legal system, thereby diminishing cooperation on the part of an individual citizen and increasing the frequency of vigilante style justice. This potential for overall decay in social order is one reason for maintaining a retributive model of criminal justice, favored by many legal scholars. In effect, the dichotomy between conscious, rational thought and unconscious, irrational thought provides the illusion of freewill, which might be an essential intuition to a stable society.

---

2 ABC News/Washington Post Poll, June 22–25, 2006.

Conclusions

This report offers an analytical framework for understanding how integrating notions of deliberative and automatic processing as well as other neurobiological processes can change how we analyze institutions. Several plausible signatures of how institutions have adapted to automatic processing were presented. Most of these institutional adaptations increase the performance of decision making either by capitalizing on or contolling automatic processes in some way. The trial process relies on the intuition of its decision makers in assessing mental states of actors and in using intuitive senses of certainty of the decision makers. The contract system seems also to prod unconscious processes that produce greater degrees of trust. The criminal justice system depends on a classification of behavior as the product of conscious, rational thought or unconscious, irrational thought.

The three examples that we discussed underscore the premise that social institutions (most notably legal systems) seem to need to categorize cognitive processes into conscious and unconscious. This ignores the psychological approach of identifying many types of cognitive processes and certainly ignores the neurological level. This failure leads to difficulties as well as opportunities. In the trial process, it leads the system to underrate the value of reforms that would produce more accurate judgment. The contract system has only begun to appreciate the depth of neurological triggers to trust. The criminal justice system seems on a collision course with neurological understanding.

Still, the distinction between conscious and unconscious seems pervasive at the institutional level. Even though we can readily identify the shortcomings of this distinction, it has proven to be a valuable, workable way of categorizing human thought and action and thus is not discarded easily. Indeed, the criminal justice system's deep embrace of the dichotomy might itself motivate an enormous amount of law-abiding behavior, which otherwise might be lost were the illusion it produces to be shattered. Given the great utility of the distinction between conscious and unconscious thought at the institutional level, reformers who would rely instead on different levels of analysis bear the burden of proof that their way is better. In view of the advances in psychology and neuroscience that we might expect in the future, this burden might, on occasion, be met.

References

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity and social history. *Games Econ. Behav.* **10**:122–142.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson. 2003. The Evolution of Altruistic Punishment. *Proc. Natl. Acad. Sci.* **100**:3531–3535.

Bright, D. A., and J. Goodman-Delahunty. 2006. Gruesome evidence and emotion: Anger, blame and jury decision-making. *Law Hum. Behav.* **30**:183–202.

Brown, M., A. Falk, and E. Fehr. 2004. Relational contracts and the nature of market interactions, *Econometrica* **72(3)**:747–780.

Carter, C. S. 1998. Neuroendocrine perspectives on social attachment and love. Psychoneuroend. **23(8)**:779–818

Carter, C. S., M. Altemus, and G. P. Chrousos. 2001. Neuroendocrine and emotional changes in the post-partum period. *Prog. Brain Res*. **133**:241–249.

Commonwealth v. *Webster.* 1850. 59 Mass. 295, 320.

De Quervain, D. J. F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder et al. 2004. The neural basis of altruistic punishment. *Science* **305**:1254–1258.

de Vignemont, F., and T. Singer. 2006. The empathic brain: How, when and why? *Trends Cogn. Sci.* **10(10)**:435–441.

Engel, C. 2006. How Much Collusion? A Meta-Analysis on Oligopoly Experiments. MPI Collective Goods Preprint No. 2006/27. http://ssrn.com/abstract=951160.

Falk, A., and M. Kosfeld. 2006. Distrust: The hidden costs of control. *Am. Econ. Rev*. **96**:1611–1630.

Falk, A., E. Fehr, and U. Fischbacher. 2000. Testing Theories of Fairness – Intentions Matter, Institute for Empirical Research in Economics, Working Paper No. 63. University of Zurich.

Fehr, E., and B. Rockenbach. 2003. Detrimental effects of sanctions on human altruism. *Nature* **422(6928)**: 137–140.

Fehr, E., and S. Gächter. 2002. Altruistic Punishment in Humans. *Nature* **415**: 137–140.

Fehr, E., G. Kirchsteiger, and A. Riedl. 1993. Does Fairness prevent Market Clearing? An Experimental Investigation. *Qtly. J. Econ*. **108**:437–460.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2003. Explaining altruistic behavior in humans. *Evol. Hum. Behav*. **24**:153–172.

Greene, J. D., and J. D. Cohen. 2004. For the law, neuroscience changes nothing and everything. *Philos. Trans. R. Soc. Lond. B* **359**:1775–1785.

Gürerk, O., B. Irlenbusch, and B. Rockenbach 2006. The competitive advantage of sanctioning institutions. *Science* **312(5770)**:108–111.

Hardin, R. 2001. Conceptions and explanations of trust. In: Trust in Society, ed. Karen S. Cook, pp. 3–39. New York: Russell Sage Foundation.

Hastie, R. 1999. The role of "stories" in civil jury judgments. *Michigan J. Law Ref.* **32(2)**:1–13.

Hastie, R. 2001. Emotions in jurors' decisions. *Brooklyn Law Rev.* **66**:991–1009.

Hastie, R., and N. Pennington. 2000. Explanation-based decision making. In: Judgment and Decision Making: An Interdisciplinary Reader, ed. T. Connolly, H.R. Arkes, and K.R. Hammond, 2nd ed., pp. 212–228. New York: Cambridge Univ. Press.

Heckathorn, D. D. 1989. Collective action and the second-order free-rider problem. *Ration. Soc.* **1**:78–100.

Insel, T. R., and Shapiro, L. E. 1992. Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles. *Proc. Natl. Acad. Sci*. **89**:5981–5985.

Insel, T. R., and L. J. Young. 2001. The neurobiology of attachment. *Nature Neurosci.* **2**:129−136.

Jackson P. L., A. N. Meltzoff, and J. Decety. 2005. How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage* **24**:771–779.

Kahneman, D. 2003. A perspective on judgment and choice: Mapping bounded rationality. *Am. Psychol.* **58**:697–720.

Kahneman, D., and Frederick, S. 2005. A model of heuristic judgment. In: The Cambridge Handbook of Thinking and Reasoning, ed. K.J. Holyoak, and R.G. Morrison, pp. 267–293. Cambridge Univ. Press.

Konow, J. 2005. Blind spots: The effects of information and stakes on fairness bias and dispersion. *Social Justice Res.* **18**:349–390.

Kosfeld M., M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr. 2005. *Nature* **435**:673–676.

Macy, M. M., and A. Flache. 2002. Learning dynamics in social dilemmas. *Proc. Natl. Acad. Sci.* **99**:7229–7236.

Milinski, M., and B. Rockenbach. 2007. Spying on others evolves. *Science* **317(5837)**:108–111, 464–465.

Nikiforakis, N. S. 2008 . Punishment and Counter-Punishment in Public Good Games. Can We Really Govern Ourselves? *J. Public Econ*., in press.

North, D. C. 1990. Institutions, Institutional Change, and Economic Performance. New York: Cambridge Univ. Press.

O'Hara, E., and D. Yarn. 2002. On apology and consilience. *Washington Law Rev*. **77**:1121–1192.

Pedersen, C. A. 1997. Oxytocin control of maternal behavior. Regulation by sex steroids and offspring stimuli. *Ann. NY Acad. Sci*. **807**:126–145.

Preston, S. D., and F.B.M. de Waal. 2002. Empathy: Its ultimate and proximate bases. *Behav. Brain Sci*. **25**:1–72.

Rapoport, A., M. J. Guyer, and D. Gordon. 1976. The 2 x 2 Game. Ann Arbor: Univ. of Michigan Press.

Robbennolt, J. K. Apologies and Settlement Levers. 2006. *J. Emp. Legal Stud.* **3**:333–373.

Robinson, P. H., and J. M. Darley. 2003. The role of deterrence in the formulation of criminal law rules: at its worst when doing its best. *Georgetown Law J*. **91**:949–1002.

Rossby, P. 2003. Serotonin Deficit and Impulsive Violence: Does Your Case Fit?, National Legal Aid & Defender Assoc. *Cornerstone:* www.nlada.org/DMS/Documents/1066920620.52/ serotonin.pdf

Rothoeft, D. 1968. System der Irrtumslehre als Methodenfrage der Rechtsvergleichung. Dargestellt am deutschen und englischen Vertragsrecht. Tübingen: Mohr.

Selten, R., M. Mitzkewitz and G. R. Uhlich. 1997. Duopoly strategies programmed by experienced players. *Econometrica* **65**: 517–555.

Singer, T. 2006. The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neurosci. Biobehav. Rev.* **30**:855–863

Singer, T., B. Seymour, J. O'Doherty, H. Kaube, R. J. Dolan, and C. D. Frith. 2004. Empathy for pain involves the affective but not sensory components of pain. *Science* **303(5661)**:1157–1162.

Singer, T., B. Seymour, J. P. O'Doherty, K. E. Stephan, R. J. Dolan, and C. D. Frith. 2006. Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439(7075)**:466–469.

Sunstein, C. R., D. Schkade, and D. Kahneman. 2000. Do people want optimal deterrence? *J. Legal Stud.* **29**:237–253.

Uvnas-Moberg, K. 1998. Oxytocin may mediate the benefits of positive social interaction and emotions. *Psychoneuroend.* **23(8)**:819–835.

Victor v. Nebraska, 1994. 511 U.S. 1.