THE ROYAL SOCIETY

Published online 26 November 2004

The neuroeconomic path of the law

Morris B. Hoffman

Second Judicial District (Denver), State of Colorado, 1437 Bannock Street, Courtroom 9, Denver, CO 80203, USA (*morris.hoffman@judicial.state.co.us*)

Advances in evolutionary biology, experimental economics and neuroscience are shedding new light on age-old questions about right and wrong, justice, freedom, the rule of law and the relationship between the individual and the state. Evidence is beginning to accumulate suggesting that humans evolved certain fundamental behavioural predispositions grounded in our intense social natures, that those predispositions are encoded in our brains as a distribution of probable behaviours, and therefore that there may be a core of universal human law.

Keywords: neuroeconomics; law; human evolution; behaviour; brain

1. INTRODUCTION

Developments at the intersection of evolutionary biology and neuroscience are beginning to approach a powerful resonance with the foundations of law. In this essay, I argue that significant advances in our understanding of evolution, the brain and behaviour presage a similar revolution in our understanding of the roots of law. The old paradigm of law as a purely cultural construct to repress our natural aggressions is giving way to a deeper understanding of the law's adaptive value and its role as an institutional expression of evolved social behaviours.

It is becoming clearer, with each advance in evolutionary biology, that the distinctions between animal morphology and animal behaviour are arbitrary, and that evolution is a powerful tool to explain aspects of both. At the same time, advances in neuroscience are suggesting a brain-to-behaviour mechanism that may supply what has been the missing link in the notion that complex behaviours have a significant evolutionary component. Together, these disciplines are not only completing evolution's tale, they are also shedding significant light on the age-old question of human nature, and therefore on the foundations of human nature's institutional analogue, the law.

2. THE DARWINIAN PRAGMATISM OF HOLMES

Law's first significant and quite unsatisfactory encounter with evolutionary biology began in the late 1800s as a kind of legal version of Social Darwinism. Its champion, Oliver Wendell Holmes Jr, constructed an elegant foundation built on what, at that time, was thought to be evolution's relentless drive towards self-interest through aggression.

In 1897, Holmes published a law review article in the *Harvard Law Review* called 'The path of the law' (Holmes 1897). It was profoundly influential, and set the stage for a revolution in twentieth century jurisprudence. Richard Posner has called it 'the best article-length work on law ever written' (Posner 1992, p. x). The revolution it began, and the one in whose midst we still find ourselves today, has had enormous consequences, both good and bad.

Holmes's jurisprudential model was based on a rather startling (for its time) synthesis of law and biology. The core idea, which Holmes first published 16 years earlier in his small book *The common law* (Holmes 1881), was that the march of the common law is like the march of evolution—not guided by any external goals (i.e. 'natural law') but rather shaped by the interaction of individual judges' proximate decisions and the ultimate judgement of precedent. Judges push the boundaries of the law just enough to accommodate what their experience tells them should be an acceptable result in a single case. Rules that work (that is, rules that are accepted over time) survive as precedent. Rules that do not work, die.

In 'The path of the law', Holmes (1897) polished and tightened these insights into four forceful and striking axioms, which he argued completely described and informed all of jurisprudence.

(i) If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict . . . . (p. 459).
(ii) The prophecies of what courts will in fact do, and nothing more pretentious, are what I mean by the law (p. 461).
(iii) The duty to keep a contract at common law means a prediction that you must pay damages if you do not keep it—and nothing else (p. 462).
(iv) [I]t would . . . be a gain if every word of moral significance could be banished from the law altogether (p. 464).

These insights were a breath of fresh air to a jurisprudence suffocating from a stilted kind of formalism, in which legal thinking was almost entirely limited to the mundane acts of identifying, classifying and labelling legal principles. In a very real sense, Holmes had begun to do for jurisprudence what Darwin had done for biology: insights untethered to any grand external assumptions were making it

One contribution of 16 to a Theme Issue 'Law and the brain'.

© 2004 The Royal Society

possible to see both disciplines on a very large scale, and therefore to see deep connections between what had before seemed to be isolated observations.

But Holmes's insights came at a heavy price, in part because they were based on a primitive understanding of evolution, not unlike Herbert Spencer's primitive understanding. The revolution Holmes began was not just a revolution against legal formalism; it was, as Professor Albert Alschuler has so aptly described it, a revolution against the very idea of right and wrong (Alschuler 2000, p. 10). In the law, as in the natural world, Holmes insisted that there is no right or wrong, only a relentless struggle for survival.

As Professor Alschuler also recognized, Holmes's valueless philosophy is the grandfather of the principal school of modern American jurisprudence: law and economics (Alschuler 2000, pp. 2–8). Although the law and economics movement has contributed greatly to a deep understanding of the law and its applications (e.g. Posner 1983), it has its limitations, especially to those of us who suspect that there may be more to the rule of law than setting arbitrarily deterrent levels of game theoretic payoffs. Classical economics can predict certain human behaviours and can also tell us whether behaviours are efficient or inefficient, but it cannot tell us whether behaviours are right or wrong.

To a great extent, the law and economics vision sees law simply as the expression of a relentlessly hedonistic, and therefore quite mutable, marketplace. In Posner's extreme world, we do not ask whether it is right or wrong for one person to torture another, we insist only that the torturing decision be made in a free market; that is, no one can be physically forced to submit, and the torturer and the tortured must be free to agree to a price that reflects the intersection of their mutual preferences. Because most people prefer not to be tortured, the price for torture will skyrocket, and torturing behaviours will be driven out by a price that reflects most people's disinclination to be tortured (and their disinclination to pay high prices to torture), not by laws prohibiting the torture itself (Posner 1983, p. 82). According to this view, our traditional labelling of torture as 'wrong' is an unnecessarily normative-laden synonym for 'unpopular'.

This analysis can be quite enlightening, especially to those of us with libertarian inclinations, but it is far from a complete description of the foundations of law. As I will discuss in more detail below, biology is now making it clear that humans are not relentlessly hedonistic, at least not in the classic economic sense. Moreover, just like Holmes's model, the assumptions of law and economics beg the deepest questions of all: *why* do people have the preferences they have, and should they really be free to express any preference as long as the marketplace can absorb it? In fact, Holmes himself expressed exasperation about universal preferences he recognized but could not explain, calling them 'can't helps' (Alschuler 2000, p. 24).

Other attempts to describe the axioms of law—by legal philosophers such as John Rawls (1971) and Robert Nozick (1974)—are also incomplete for the same reason: they presuppose that people, and therefore the law, must act in certain ways. For example, Nozick posits three sets of rules from which he argues all our notions of distributive justice emanate: (i) rules governing how un-owned property is acquired; (ii) rules governing how owned property may be transferred; and (iii) rules governing how violations of the acquisition and transfer rules should be rectified (Nozick 1974, pp. 150–153). But why must we have rules governing the acquisition and transfer of property, and what should those rules say? Why must we have rules to rectify violations of the acquisition and transfer rules, and how exactly should those violations be rectified?

At the bottom of all these taxonomies in the legal subfloor lies the real foundation: why do people behave the way they do, and how should society react to those behaviours?

3. THE ANCIENT DEBATE ABOUT JUSTICE

Of course, the debate about human nature, and therefore about justice, has been going on since the dawn of the human race. The Greek version of the debate, as retold by Plato, pitted Thrasymachus ('Justice is nothing else than the interest of the stronger') (Plato *ca.* 360 B.C., 1901 edition, p. 19) against Socrates ('Justice is the excellence of the soul', p. 43). Hobbs and Locke tangled over the same question. Indeed, the Enlightenment debate about the essence of human nature was very much a part of the compromises that made up the American constitution (McGinnis 1996).

All of these efforts to describe and justify the rule of law—Socrates's, Locke's and Jefferson's divine man, Thrasymachus's, Hobbes's, Alexander Hamilton's and Spencer's selfish man, Holmes's bad man, Posner's rational man and Nozick's free man—depend on unstated assumptions about why people behave the way they do. Those assumptions tend to coalesce around two quite unsatisfactory poles. The Socratic side of the debate generally attributes its assumptions about why people behave the way they do, and especially why people *should* behave in particular ways, to the divine, and for that reason the Socratic approach has, quite unfortunately, gone out of favour in our post-modern world. The Holmesian side of the debate, buoyed by its misinterpretation of Darwin, assumes that all behaviours are ultimately expressions of raw self-interest.

David Hume added an important ingredient to the debate—and the twinkle of a synthesis—by recognizing the emotional component to human behaviour. He was among the first post-Renaissance philosophers to observe that humans do not always act 'rationally' in the sense of making conscious, calculated choices. Instead, our behaviours are often driven by emotion, and our sense of deliberation is often a mere artefact of conflicting emotions. Hume also recognized, long before Darwin, that some aspects of our emotion-driven behaviours—both for good and evil—seem to be 'kneaded into our frames' (Hume 1748, p. 271).[1]

Research in evolutionary biology and neuroeconomics is suggesting that Socrates and Hume may have had it right all along, if we replace Socrates's 'divine' with 'evolved reciprocity', and Hume's 'kneaded into our frames' with 'inherited'. It appears the Holmesian assumptions about why people behave the way they do are not at all accurate, and may be misinterpretations of even deeper truths about human behaviour, truths that much more completely describe, and may even justify, the rule of law. When we add evolutionary and neuroeconomic insights to the way we frame these foundational questions, and specifically the insight that we are often driven by our evolved neuroarchitectures to act in a very complicated, yet often predictable, kind of socially modulated self-interest, a case can be made, and I will try to make it here, that there is indeed a relatively fixed and immutable set of right and wrong human behaviours. In this neuroeconomic model, the law is neither an inexplicable divine good, nor a cultural veneer against inexplicable original sin, nor a mere lubricant for arbitrary market preferences. It is an expression of our evolved natures as a profoundly social species.[2]

Before I expand on neuroeconomics and its effects on the foundations of law, let me refer to some more general observations about the relationship between evolution and behaviour.

4. EVOLUTION AND BEHAVIOUR

Evolutionary biology has undergone a revolution of Copernican proportions since Darwin's and Holmes's time, and may be about to undergo another. The first revolution was about evolutionary morphology: how exactly do genes express themselves in physical traits? The rediscovery of Mendel, integrated with the discovery of DNA and the mechanics of replication and protein production, began to create a powerful picture of exactly how genetic information is transmitted across generations, how mutations can arise in that transmitted information, how, once transmitted, the genetic information is expressed in physical traits, and how an individual animal's interaction with its environment can make an inherited trait more, or less, likely to appear in subsequent generations. Adding game theory into the mix has given evolutionary biologists a powerful insight into the ways adaptation works within and between populations.

A similar synthesis is now taking place with respect to evolved *behaviours*. Anthropologists, biologists, economists and linguists are coming to understand that the evolutionary forces that shaped all animal bodies, including ours, have also shaped the menu of behaviours that animals, including us, entertain in response to certain stimuli, as well as the probabilities associated with each particular choice in that menu (see Wilson 1975; Dawkins 1989). Indeed, because natural selection is the evolutionary motor, it is incorrect to think of morphology and behaviour as separate animal functions; everything an animal is—the package it comes in and the way that package interacts with its environment—is the ultimate, though of course not proximate, product of evolutionary pressures. That is, opposable thumbs make evolutionary sense only because they came with the behaviours to use them. Function and form—and therefore behaviour and morphology—are inseparable in evolution's long march (see Dawkins 1999).

E. O. Wilson has even suggested that this evolutionary synthesis between mind and body presages a fusion between the social and natural sciences (Wilson 1998). But even on a less grand scale, the idea that complex animal behaviours, and especially human behaviours, might have a significant evolutionary component has been the object of intense criticism. There have been two primary challenges: (i) the puzzle of altruism; and (ii) the question of how it is that genes act on brains to produce behaviours that can be inherited.

5. ALTRUISM

The initial reaction, by biologists and social scientists alike, to the realization that Earth's creatures are the product of an evolutionary process that naturally weeds out the weak and infertile, was the assumption that every animal would therefore tend to become, as one modern philosopher of science has put it, a 'powerful loner, who knows when to cheat and can do it well; he is a kind of Nietzschean *übermensch* who breaks the conventions of sociability and morality with one powerful swipe of his well-oiled and bloodied fighting appendage' (Casebeer 2003). Indeed, Spencer's social philosophy and Holmes's legal philosophy are rather straightforward extensions of this primitive view of evolution.

It turns out, of course, that nature herself is much more complex. We have known for some time that individuals across animal populations do not exhibit anything like the uniform level of self-interested sexual and male–male aggression that natural selection seems to predict. In social and not-so-social species alike, even the most aggressive individuals seldom behave like Nietzschean über-animals, practising instead all kinds of more modulated behaviours. *Displays* of aggression, for example, often substitute for actual aggression. Cooperation, not competition, seems to rule. In fact, at the extreme end of this selfish/selfless continuum of behaviour, individual animals have been known to sacrifice their own fitness, even their own lives, for the benefit of other related and even non-related individuals. How can this kind of altruistic behaviour be explained by natural selection?

The puzzle was neatly solved in its weakest form (kin altruism) by W. D. Hamilton, who demonstrated that seemingly altruistic behaviours between related individuals make perfect evolutionary sense if we refocus the fitness inquiry from individual animal fitness to individual gene fitness (Hamilton 1964). From the gene's point of view, parents should always sacrifice themselves for three or more children (or five or more grandchildren) since, on average each child carries one-half of the parent's genes (and one-quarter of the grandparent's genes). Parental sacrifice is not about an individual acting altruistically; it is about parental genes acting quite selfishly.

But how does evolution explain altruistic behaviours between non-related individuals? One explanation is that natural selection simply is not perfect, and behaviours that are adaptive on average can go terribly wrong on occasion in individual circumstances. That is, we all have powerful and perfectly adaptive urges to save our kin, graded to relatedness, and perhaps those urges occasionally get misapplied to non-kin. Versions of this argument were often made in biology texts in the 1960s, to explain, for example, the evolution of warning cries in birds (Williams 1966, p. 206).

That explanation is not terribly satisfying, for several reasons. First, non-kin altruism seems to be much more widespread than this explanation would predict. Moreover, one would expect that the profound genetic cost of misdirected kin altruism would have led to highly evolved and widespread mechanisms to recognize one's own kin, yet it is an almost universal problem in the animal kingdom that half of all parents (males) are, by the very nature of sexual reproduction, unsure of the paternity of their offspring.

Evolutionary theorists have discovered a deeper, simpler explanation for non-kin altruism: just like the illusion of kin 'altruism', non-kin 'altruism' is not altruism at all if what is going on is the payment of a direct cost in exchange for a less direct, but nonetheless palpable, benefit. Biologists have long known of a phenomenon, later dubbed by evolutionary theorists as 'return effect altruism', in which individual animals act in ways that appear in isolation to be altruistic, but which, when viewed in a larger, often social context, clearly confer an adaptive advantage on the allegedly altruistic actor. Bird warning cries are a good example. Even a purely 'altruistic' cry—that is, one whose frequency does not depend on the proximity of kin—can convey a net adaptive advantage. Although giving such a cry can be extremely costly to the individual in the short-run (and, in fact, so costly that there is no long-run left), it also can trigger responsive cries in other nearby prey animals. Because many predators have coevolved strategies to give up the hunt if they hear too many cries, '[w]arning your [unrelated] neighbour that a predator is nearby may be the quickest way to get the predator to move on elsewhere' (Trivers 1971).

In addition to return-effect altruism, biologists have also long recognized that apparently altruistic behaviours are sometimes just one side of symbiotic relationships, and that when one measures the net costs and benefits of the symbiotic whole the behaviours can make perfect adaptive sense, for both participants. For example, symbiotic cleaning behaviours are fairly widespread in the ocean—more than 45 species of fishes and six species of shrimp are known to be cleaners, and innumerable species of fishes serve as hosts (Feder 1966). For both the cleaner (who spends enormous energy benefiting the host) and the cleaned (who resists the temptation to swallow the cleaner), the individual behaviours seem maladaptive when viewed in isolation. However, when considered together, and taking into account the costs of not being cleaned (that is, the damage done to the host fish by ectoparasites), the difficulty and danger of finding a cleaner, the site specificity of cleaners, the lifespans of cleaners, and the ability of hosts to find the same cleaner repeatedly, theorists have demonstrated that cleaning behaviours can confer a net adaptive advantage on both the cleaner and the cleaned (Trivers 1971).

However, traditional symbiosis is not the only kind of 'reciprocal altruism', as theorists have labelled this sort of mutually beneficial interaction. It is a much more generalized, and indeed central, aspect of evolution itself. Robert Trivers was among the first evolutionary biologists to recognize that when one considers not just the survival behaviours of a single individual in isolation, but also that that individual must make guesses about the survival behaviours of his competitors and prospective mates, optimum solutions exist that involve a mix of aggressive and cooperative strategies, with the cooperative strategies often taking the form of a time-delayed exchange, or what evolutionary theorists have come to call reciprocal exchanges.

Trivers showed that animals engage in a whole host of complex reciprocal exchanges, even with (and, indeed, especially with) unrelated individuals that can accrue to their net long-term individual benefit, even if they seem to be to their net short-term individual disadvantage. If the net long-term advantage sufficiently outweighs the net short-term cost, the behaviour can become adaptive and animals can inherit a tendency to act in ways that appear 'altruistic'.

Displays of aggression also have obvious adaptive value, when viewed not in isolation but as an exchange between individuals. John Maynard Smith demonstrated that the kinds of 'graded signals', as he called them, seen in displays of aggression, make perfect evolutionary sense when one considers not just the obvious benefits of aggression but also its obvious and not-so-obvious costs—being killed or injured, or having one's mate lured away during combat by a 'sneaky male'. Perhaps more importantly, Smith demonstrated that both within and between species, when both the costs and benefits of behaviours are carefully considered, including the delayed costs and benefits in reciprocating, natural selection operates to drive individuals (and populations) to a relatively stable distribution of behavioural strategies, strategies he called 'evolutionarily stable strategies' (Maynard Smith 1982).

Of course, aggression, displays of aggression, cleaning symbioses and warning cries are just a tiny part of the behavioural toolbox. In recent years, evolutionary biologists have examined a whole host of others kinds of reciprocal behaviours, both in humans and non-humans, including foraging, communication, nepotism, sibling rivalry, parent–child conflict, habitat selection, predator–prey interaction and even learning (Dugatin 2001). Game theorists have added a mathematical rigour to the analysis of reciprocal exchanges, treating them as a set of multiple non-zero-sum games. Evolutionarily stable strategies are the Nash equilibria (Nash 1950) for the particular 'game' in question, under the constraints biologists believe were operating when a particular strategy evolved.

In intensely social species like humans, reciprocity seems to play an especially important role. After all, our evolved behavioural tendencies are the product of a long and complex interaction between our individual ancestors and the small groups in which they evolved, and living in groups requires rather sophisticated mechanisms to regulate relationships between members. The social behaviours we evolved in that environment are central to what we now view as human nature.

Any group needs rules for admission and exclusion. Thanks to the combination of our large brains, our ability to speak and our intense social natures, it seems we literally evolved an instinct for rules. In fact, our brains and our language structures are themselves most probably tools we evolved to increase the efficiency of our reciprocal social exchanges (Pinker 1994). Being in a group at all—following rules, enforcing violations, accepting punishment—is a constant series of reciprocal trade-offs between short-term individual gains/losses ('Should I steal that pile of food our hunters brought in today?') and long-term individual gains/losses ('Do I want to stay in this group and benefit from mate availability and economies of scale in food gathering and defense?') (Axelrod & Hamilton 1981). The scientific literature is now flush with studies of reciprocity, or 'cooperation' as some biologists now prefer (Burnstein *et al.* 1994; Dugatin 2001; Brown & Moore 2000; Gintis *et al.* 2003).

It now appears beyond cavil that all animals, including humans, have evolved tendencies to behave in certain predictable manners under certain conditions. Of course, the extent to which these ultimate tendencies actually express themselves, how they interact with an individual's proximate environment and, in the case of higher-order animals, their learned behaviour, continues to be open to great debate.

But how do brains turn evolutionarily stable strategies into behaviours? Before turning to that profoundly important question, I consider three general classes of human behaviours, which I propose are genetically based and which may inform much of the foundations of law.

6. THREE CORE HUMAN PRINCIPLES

It appears that humans, and indeed all intensely social animals, have a predisposition to follow three central behavioural rules: (i) promises to reciprocate must be kept (contract); (ii) reciprocal exchanges must be relatively equal (tort and criminal); and (iii) serious violations of the first two principles must be punished (enforcement).[3] These three rules form the nucleus of a kind of neo-natural law that I suspect is part of our inherited natures, and therefore is both universal and relatively invariant.

This is a profoundly different view of justice than the incomplete evolutionary views of people like Holmes and Spencer. Under this construct, we have an instinct for justice not because justice is 'a brooding omnipresence in the sky', as Holmes once derisively described efforts to externalize the law (Holmes 1917), but because we are the complex products of an evolution that made such social behaviours adaptively beneficial to our individual survival.

There is compelling game-theoretic evidence that all humans are indeed armed with versions of these three internal principles, which classical economics cannot explain but which become entirely rational if we look at our behaviours as a constellation of evolved reciprocal exchanges. For example, in the so-called 'ultimatum game', involving two players, A and B, player A is given money (or useful goods) and both players are told that A must choose a fraction to offer to B. Both players know the total amount available for division. They are also told that if B accepts the offer, the money will be divided as A has proposed, but that if B rejects, neither player gets anything. Classical economics, which assumes that A and B will act in unbounded 'self-interest', predicts that A will offer next to nothing, and that no matter how small the offer, B will accept it. Neither prediction holds true.

In industrial societies, A offers an astonishingly 'altruistic' average of *ca*. 40%, simultaneously acknowledging that this exchange should be roughly equal, that A should probably get a little more because he is the one who started out with the money,[4] and that if he offers much less B will reject and neither will get anything. In fact, offers of less than 30% are frequently rejected (Gintis 2000). These same general results occur in pre-industrial cultures (Henrich *et al.* 2001).[5]

Just as with classically 'altruistic' behaviours, this kind of behaviour is not altruistic at all. Humans have built-in regulators, evolved over aeons of intense social interaction, that tell us not to be unfair to each other, lest today's player A will become tomorrow's player B. That these preferences for a generalized kind of fairness are the adaptive product of evolutionary pressures is clear not only from their human universality, but also from the fact that other intensely social primates exhibit similar test behaviours (Brosnan & de Waal 2003).[6] That these preferences are bound up with evolved *social* behaviours is clear because when the other player is a stranger (or worse still, a computer), people tend to revert to more classically self-interested behaviours (Cook & Hegtvedt 1992). However, our hunches about a universal natural justice are still far ahead of the supporting science. The most difficult gap is in the neurology of behaviour.

The challenge is that brains, not genes, generate behaviour. Although our brains are themselves the ultimate product of evolution, neuroscientists have yet to discover the behavioural analogue to DNA—the mechanism by which our brains transmit behavioural predispositions to us, but there is encouraging and exciting evidence from the science of neuroeconomics.

7. THE PROMISE OF NEUROECONOMICS

As the mechanisms of the brain are being uncovered in both humans and non-humans, neuroscientists are doing two important things: (i) they are isolating and studying the actual brain structures involved in decision-making; and (ii) from these discoveries they are piecing together a new probabilistic paradigm of how brains make decisions, a paradigm that may go a long way toward explaining how adaptive behaviours are expressed in individuals and then transmitted across generations. A remarkably comprehensive and clear exposition of these developments can be found in Paul Glimcher's new book (Glimcher 2003).

As Glimcher reminds us, Rene Descartes believed that all human behaviour could be divided into two fundamentally different categories: simple behaviours, which were the deterministic motor responses of given sensory inputs; and complex behaviours (or what we would call 'cognitive' behaviours), which Descartes saw as the indeterminate product of unknown and unknowable forces, and which he simply called 'the soul' (Descartes 1649). For centuries, this Cartesian dualism has both enlightened and burdened neuroscientists examining the etiology of behaviour.

It enlightened the investigation because, by focusing on simple motor reflexes, it allowed investigators to discover quite a lot about the neural pathways between stimulus and response. However, it also profoundly burdened the investigation because it presumed that what was going on in the brain in these so-called simple motor reflexes was the same sort of deterministic and essentially linear transmissions that were observed elsewhere along the neural path. In this reflex model of simple behaviours, the brain does very little more than reflect the stimulus signal back to the appropriate response path. However, as discussed in more detail below, evidence is accumulating that when the brain 'decides' what the 'appropriate' response path should be, something is going on that is much more complicated than the reflex model predicts.

Of course, Cartesian dualism also had the effect, and indeed the intended effect, of cleaving the behavioural world into two mutually exclusive pieces: the scientifically accessible reflex piece and the mysterious, religious, scientifically inaccessible cognitive piece. Neurological research on non-cognitive behaviours proceeded apace, but research on cognitive behaviours was, at least early on, condemned to the realm of the metaphysical. Advances in neuroscience are driving a fundamental synthesis between these two realms.

Many neuroscientists believe that all behaviour can be explained by combining a sufficient number of reflexive units in a sufficiently complex way, and that the very notion of a separate and distinct kind of 'cognitive' mechanism is a false dichotomy created simply to avoid touchy issues about consciousness and free will. For example, as early as 1950, the renowned German physiologist Erik Von Holtz and a colleague wrote:

> The sooner we recognize that the [higher functions] which leave the reflex physiologist dumbfounded in fact send roots down to the simplest basal functions of the CNS, the sooner we shall see that the previously terminologically insurmountable barrier between the lower levels of neurophysiology and higher behavioural theory simply dissolves away.
>
> (Von Holtz & Middelstaedt 1950)

At the time of Von Holtz's predictions, Cartesian dualism still dominated the study of brain anatomy and function. Most 1950s textbooks divided the cortex into three functional parts: sensory, motor and what was dubbed 'association', which is a word that had its origins in Pavlov's conditioned associations, but which came to be a kind of catch-all all category corresponding to Descartes' 'complex' (Glimcher 2003, p. 233).

The association parts of the cortex remained shrouded in mystery primarily for technical reasons. Researchers could measure the activity of single sensory and motor neurons by exposing those neurons in anaesthetized subjects. But association neurons could not be explored in this manner because anaesthetized subjects are unconscious, and therefore cannot 'associate'—that is, do any of the cognitive activities presumed being done in these areas of the cortex.

This technical limitation was overcome in the late 1950s, when researchers perfected a technique of inserting tiny wires into the cortices of conscious subjects (Glimcher 2003, pp. 233–234). From that moment on, and accelerated by many other technological advances, including the functional magnetic resonance imaging, a flood of data began to be gathered about the association areas of the cortex, and the data started to suggest that the essentially Cartesian distinctions between sensory, motor and association were not at all accurate. Perhaps Von Holtz was right. Perhaps the 'cognitive' function occurring in the association part of the cortex was merely a complicated arrangement of reflexes.

But it turns out that the classical reflex model is a very poor predictor of whole categories of determinate behaviours, no matter how many simple reflex circuits we postulate, and no matter how complex the connections. For example, rhythmic behaviours—like stepping—have simply not lent themselves to any kind of modelling based on combinations of reflexive pathways (Glimcher 2003, p. 111). Moreover, electrode studies of many of these rhythmic behaviours show neural activity that is quite different, both in kind and intensity, than one would expect from a reflex model (Glimcher 2003, pp. 111–112).

As result of these difficulties, some neuroscientists are suggesting that Descartes' dualism should be closed in the other direction—by positing that there are really no reflexes, and that all behaviours are the product of a much more complicated process. In an amazing experiment in 1987, James Gnadt and Richard Andersen discovered that a portion of the parietal cortex associated with saccadic eye movements—gaze-aligning movements that rotate the eye at high speeds when an animal switches its gaze from one object to another—and which had been categorized as sensory, exhibited a suspiciously cognitive kind of pre-movement memory. When saccadic neurons in the parietal cortex of macaques trained to stare at a primary visual stimulus were activated by a secondary visual stimulus, those neurons not only remained active after the visual stimulus was removed, but remained active until the macaques moved their eye to gaze at the remembered location of the secondary stimulus. As Gnadt and Andersen put it, these saccadic neurons 'appeared to be related to the pre-movement planning of saccades in a manner which we have chosen to describe as *motor intention*' (Gnadt & Andersen 1988).

That very phrase—motor intention—sounds quite strange to ears conditioned to 400 years of Cartesian dualism, and perhaps even stranger to more modern adherents of the all-reflex approach. Neurons whose function was presumably sensory (we might even say 'autonomic') turn out to behave in a surprisingly cognitive way. What we thought was a simple sensory/motor reflex—detect a new object and move the eye to it—turns out to involve a cognitive delay—detect a new object, remember its location, and decide later whether to move the eye to the remembered location.

It seems Von Holtz's synthesis is being realized, but in the opposite direction. Even the simplest 'reflexes' may involve neural activities once thought to be associated with 'cognition'. However, where does all this leave us in our hunt for a connection between genes and behaviour?

In recent years, Paul Glimcher and other neuroscientists have suggested an answer. Because it is becoming apparent that the distinction between 'reflex' and 'cognition' is artificial, perhaps the distinction between 'determinate' and 'indeterminate' is also artificial. Perhaps the brain—both in its 'simple' and 'complex' activities—is a probability machine rather than some contraption that inexplicably switches back and forth between reflexive/determinate outcomes (burn your hand, pull it back) and cognitive/indeterminate outcomes (you decide you will walk home today rather than take the bus). Perhaps all behaviours are represented in the brain by a set of probability distributions, which are then continuously influenced by the interaction between ultimate causes (the initial probabilities that evolution built into brains) and proximate causes (the particular environmental challenges brains are called upon to solve).

In this model, the 'reflex' is just an extreme kind of probability distribution—one with a very high probability bunched near a single action, the response. When you burn your hand, it is extremely likely (but not determinate) that your brain will decide to pull the hand back. The high probability of that particular action masks its inherently indeterminate (and cognitive) character. Likewise, when you decide to walk home rather than take the bus, you are also engaging a distribution of probable behaviours, though the probabilities are more evenly distributed over a wider range, leaving you with the conscious sense that you 'decided' what to do. But in both cases, according to the probabilistic model, the particular 'decision' is an indeterminate outcome bounded only by the probability distribution of all outcomes.

Glimcher and his colleagues have performed a series of spectacular neurological experiments strongly suggesting that the brain works exactly in this probabilistic way. Using a variety of neurobiological techniques to study the neural firings in the brains of monkeys and humans as they make decisions during various kinds of games, the experimenters found that when the strength and frequency of those firings are accumulated and plotted over time, they look virtually identical to the probabilistic outcomes in decision-making by individuals over time—the so-called 'utility function' of modern economics (Glimcher 2003, pp. 322–336; Glimcher *et al.* 2004). This is a remarkable result, suggesting an essential unity between the way a single brain makes a single decision, and the patterns that emerge when brains make many decisions over time.

Thus, although we cannot predict whether the brain of any particular person will offer 40% in the ultimatum game, we can surmise that the population-wide average of 40% reflects the fact that the brains in that population have a probability distribution for this behaviour that peaks near 40%. The genius of the probabilistic model is that it preserves the indeterminacy (free will?) of a particular individual's behaviour, while explaining the perfectly determinate behaviour of large groups of individuals, or of a single individual over many trials. It also suggests the real possibility that some behaviours are heritable, because brains have of course been inherited.

To complete this second post-Darwinian revolution, neuroscientists will need to discover exactly how behavioural probability distributions are encoded in the brain. When and if that happens, neuroeconomics may do for the evolution of behaviour what Watson and Crick did for the evolution of physical traits.

8. LAW AS AN EXPRESSION OF EVOLVED PROBABILISTIC BEHAVIOURS

Of course, law is a special kind of game, a sort of metagame, where the thing in play is not a direct payoff, but the very question of what should be the rules of the game. Law may well 'evolve' in the short run in the way Holmes suggested. But if it is true that humans have evolved a set of basic social behaviours to navigate our way through the social world, and that those basic social rules form the core of a kind of natural justice, then law may well have evolved in a much deeper and profound way.

At those critical points where judges, juries or legislatures have to make decisions, those decisions may not be the valueless preferences Holmes presumed. They may not be the arbitrary behavioural cousins of the mutation, waiting for time and survival pressures to sort the useful from the useless. They may instead be preferences that reflect the interaction between the case at hand and neuroeconomically evolved, probabilistic, norms that all judges, jurors and legislators carry inside their brains. It seems likely that we cannot help but to give some distributed weight to the core principles that promises should be kept, that social exchanges should be roughly equal and that serious violations of these two aspects of our imbedded social contract should be punished.

We might therefore reformulate Holmes's axioms this way, at least with regard to that portion of the body of law we believe is encompassed by our core of natural justice:

(i) If you want to know what the law is, look at it as would a 'good man', someone who is not interested in the outcome of a dispute but recognizes that one day he may be subject (in either direction) to the rule derived from the dispute.
(ii) Law is nothing more than a prophecy of what rules good people are likely to agree are well settled.
(iii) 'Duty', whether in contract or tort, is a prediction of how we would agree in advance to treat one another, without knowing ahead of time whether we will be the breacher or breachee, the tortfeasor or the injured, the criminal or the victim, the defendant or the sentencing judge.
(iv) The law should not be embarrassed to label some of its most basic rules in moral and ethical terms, because that is exactly what they are.[7]

9. THE RELATIONSHIP BETWEEN FREEDOM AND JUSTICE

The economist Paul Rubin has written a provocative book arguing that the freedom to leave one group and join another, and thus avoid coercion by dominants, is a deep part of our evolved natures as humans (Rubin 2002). Rubin argues that our profound sense of individuality, which has survived in tandem with our profound social natures, was a kind of ultimate veto over both dominant and collectivist excess. Exit freedom had the effect of imposing constraints on dominant individuals in the group: if a few powerful individuals got too powerful, they risked loss of members, and thus loss of some of the net advantage of living in groups. Likewise, even the majority in any group had to keep a keen eye on majoritarian excess.

Justice is what happens when our deepest social axioms—which, as I have suggested, themselves contain an embedded core of justice—are given efficient expression. The key to these social axioms is that they are the evolved product of *reciprocal* social exchanges. That is, the small groups in which we evolved contained an important element of freedom—the freedom to enter into mutually beneficial social interactions, the freedom to decline to do so, and, as Rubin points out, the freedom to leave the group and go join another. Laws enacted or developed without these complimentary forces in play will themselves tend to be unjust.

Thus, a dictator is inclined to write laws that are not just, both because the dictator is unlikely to become an enforcement object of his own laws and because he may have the power to limit his subjects' exit. Those laws will tend to reflect only the dictator's unconstrained self-interest, not the social connections out of which the dictator, and all of us, evolved.

By contrast, the deepest social connections that bind us bind us only because, in the end, we are free to disregard them. They have become powerful precisely because they must have had enough long-term utility to overcome their short-term costs, and to keep us from exercising our freedom to exit the group. They do not achieve that status if they are forced upon all group members by some *a priori* collective will. Individuals, not groups, are the functional units through which genes act, and social norms become adaptive only because they confer a net benefit to individuals. Thus, collectivist regimes are also inclined to write laws that are unjust.

Ultimately, as Rubin so powerfully argues, there is an inextricable evolutionary link between justice and democracy. The ability of any justice system to accommodate the biological tension between individual freedom and social norms depends to a great extent on its own ability to develop those norms as a free expression of social consensus. The best laws work because they efficiently confer, and express, enough long-term benefits on enough individuals that those individuals are willing to remain in the group and pay the short-term price of compliance. The genius of democracy is that it provides a continuous feedback mechanism on these social norms, constantly recalibrating them to current individual preferences.

In effect, democracy creates a market for the governed, in which conflicting preferences for individual freedom and social restraint compete freely to obtain optimal results. This is hardly a new insight, but what may be new is the evolutionary vision that suggests democratic institutions are not artificial constructs, but rather are expressions of our own evolved, and complimentary, desires for freedom and social stability. If we say that 'justice' represents our biologically embedded tendency to accommodate the tension between self and others, a tension that presupposes we have the freedom to act selfishly or selflessly, then our best institutions are those that most efficiently express that accommodation.

Rubin argues that a democratic nation with a free market economy is the highest expression of the human spirit simply because humans are built for freely entering into mutually beneficial reciprocal exchanges with other humans and because democracy is the most efficient accommodation between social constraint and individual freedom. Admittedly, Rubin's thesis, and indeed mine in this essay, both depend on many assumptions about the conditions under which humans evolved. Although palaeontologists and anthropologists are learning more and more about the ecological details of the so-called 'era of evolutionary adaptivity'—that portion of the Palaeolithic 50 000 to 100 000 years ago when the current human genome is thought to have emerged—much remains unknown.

For example, we do not know much about the ecological conditions that caused humans to stop living in small mobile groups of mostly related hunter–gatherers and start living in larger sedentary groups of mostly unrelated hunter–gatherers, although some anthropologists have speculated that this change was driven by population pressures (e.g. Tudge 1998; Carniero 2000). That transition, which preceded the transition to horticulture then agriculture, is a key to understanding the extent to which nonkin reciprocity may have shaped our genome. Regardless of when and how the sedentary transition happened, it seems highly unlikely that groups as large as 'nations' have existed long enough to have had any adaptive impact, except possibly as misinterpreted cues for social behaviour grounded on a much smaller scale.

Nevertheless, Rubin's core insights about the evolutionary relationship between economic and political freedom are tantalizing, and are precisely the kind of interdisciplinary approach that is beginning to shed light on the nature of humans and their institutions, including law.

10. CONCLUSION

Advances in neuroeconomics are suggesting the physical mechanisms by which animal behaviours are inherited. All behaviours—whether simple motor reflexes or high-order cognition—may be generated not by a collection of determinate stimulus/response pathways, but rather by the indeterminate triggering of a particular behaviour from a probabilistic distribution of possible behaviours. These insights have profound implications, not only for age-old paradoxes of free will and the interaction of the mind and body, but also for the foundations of law.

The idea that some behaviours are heritable as an array of probabilities meshes quite nicely with what evolutionary theory and game theory have been teaching us about human behaviour. We are the products of evolutionary forces that shaped us to survive as individuals, but in small intensely social groups. Our very being is about accommodating the ancient tension between self and others by passing all our decisions through a distribution of probabilities that has a built-in shape, and that peaks at three Nash equilibria: (i) do not break promises; (ii) be fair; and (iii) punish serious violations of (i) and (ii).

Because these core principles are only peaks in a distribution of probable behaviours, there is no doubt that there will be great individual variation in behaviours. Indeed, the variations between individuals, and in individuals over time seem to take on the same distribution shape as the internal brain distributions. These core principles are a kind of behavioural fractal. The paradox of predictable macroeconomic patterns appearing out of unpredictable microeconomic choices is neatly solved if the machines we use to make our choices are themselves probability machines.

Of course, the majesty of the brain, and what makes it the king of adaptive tools, is that brains can learn—they can, over time, change the shape of their decision curves. Experience, whether gained from actual encounters between a brain and the world, or through the accumulated and communicated wisdom of other brains, can alter the initial probability distributions with which our models were originally equipped. In fact, this constant feedback loop between the outside world and the brain's representation of the outside world is precisely what brains are all about.

The ultimate nurture/nature debate may thus collapse into a quantitative debate about the malleability of our initially set decision curves. Although the proximate effects of culture can hardly be overestimated when we are talking about a machine built to soak up experience, it is equally clear that the initial settings matter too.

I suspect we will discover that our deepest social instincts—of the kind I postulate in this essay—operate like our deep language structures. They form a template upon which the syntax of human interaction can unfold. There are endless variations in behaviour among individuals and among cultures (just as there are language variations), but, in the end, my guess is that we will discover that the syntax of social interaction is universal and invariant. The deepest roots of law express those universal and invariant rules of social syntax.

Holmes's insights about the relationship between individual decisions and the false majesty of the law were profound and powerful applications of what little was then known about evolution and human nature. However, his resonating logic has led us to the ironic precipice of denying our own humanity. Holmes understood only half of the engine of human evolution—that we, as individuals, are the product of a relentless struggle to survive. He did not realize that in the course of that struggle, the path upon which evolution took us was a path of intense social cooperation.

As a judge, who every day imposes drastic penalties on the free-riders we manage to detect and capture, I must confess that it is comforting to contemplate that the law is not merely a lubricant of market preferences or a collection of arbitrary predilections of the ruling class. It may well reflect our deepest commitments to each other, commitments that are at the heart of our evolved natures as social animals.

I thank Albert Alschuler, Paul Glimcher, Oliver Goodenough, Richard Posner, Paul Rubin and the anonymous peer reviewers, all of whom provided comments on earlier drafts of this essay. I also thank the late Margaret Gruter and all my friends at the Gruter Institute for Law and Behavioral Research for their contributions and inspiration.

ENDNOTES

[1] Two of most insightful modern heirs to Hume are probably Antonio Damasio (1994) and Robert Frank (1988).

[2] I am hardly the first person to propose a biologically driven return to a natural sense of justice. See, for example, Masters & Gruter (1992). See also the groundbreaking work of Owen Jones (2001*a*,*b*) on the relationship between law and biology.

[3] There is, of course, a fourth fundamental behaviour, shared by many social and non-social species alike: the recognition of property rights. The notion that things possessed by one individual cannot be taken by others may well be the most fundamental of all evolved behaviours, because survival is itself bound up with the use of things (food and shelter, for example). Indeed, the three core human principles that I posit in this essay are arguably derivative of the notion of property: it may be that living in groups, and its attendant reciprocity, is simply a strategy to deal with the problem of scarce resources. But I leave a discussion of property to my colleague Jeffrey Stake.

[4] This is a version of the so-called 'endowment effect' (Kahneman 1991), and it seems to operate even where, as in the ultimatum game, the property 'owner' had no original claim to the money. Possession, it turns out, may be nine-tenths of the law because even fleeting possession evokes powerful feelings of entitlement.

[5] There were some interesting differences between industrial and pre-industrial societies. A's offer tends to be lower in pre-industrial societies (26% mean) than in industrial societies (40% mean). There was also more variation in A's offer in pre-industrial societies than in industrial cultures. The lowest offers occurred in societies where the incidence of cooperation and market practices was low, and here rejection was rare. Offers were higher where exchange was more frequent. However, where local custom imposed on B a future obligation to reciprocate at a time to be determined by A, even offers greater than 50% were sometimes refused (Henrich *et al.* 2001).

[6] There are, nevertheless, important differences between the 'economies' of humans and other primates. For example, it appears that although capuchins and chimpanzees have a primitive ability to monetize goods, they are unable to recognize different denominations of money (Brosnan 2004).

[7] When I say 'good man' in these neo-Holmesian formulations, I simply mean people whose ordinary social constraints—that is, their evolved accommodations between short-term self-interest and long-term self-interest—have not been disabled.

REFERENCES

Alschuler, A. W. 2000 *Law without values: the life, work and legacy of Justice Holmes*. University of Chicago Press.

Axelrod, R. & Hamilton, W. D. 1981 The evolution of cooperation. *Science* **211**, 1390–1396.

Brosnan, S. F. 2004 Primate economics, presented at the Gruter Institute for Law and Behavioral Research, Squaw Valley, CA, 22 May 2004.

Brosnan, S. F. & de Waal, F. 2003 Monkeys reject unequal pay. *Nature* **425**, 297–299.

Brown, W. M. & Moore, C. 2000 Is prospective altruist-detection an evolved solution to the adaptive problem of subtle cheating in cooperative ventures? *Evol. Hum. Behav.* **21**, 25–37.

Burnstein, E., Crandall, C. & Kitayama, S. 1994 Some Neo-Darwinian rules for altruism: weighing cues for inclusive fitness as a function of the biological importance of the decision. *J. Person. Soc. Psychol.* **67**, 773–789.

Carniero, R. L. 2000 The transition from quantity to quality: a neglected causal mechanism in accounting for social evolution. *Proc. Natl Acad. Sci. USA* **97**, 12 926–12 931.

Casebeer, W. D. 2003 Book review: evolution and the capacity for commitment. *Hum. Nat. Rev.* **3**, 12–14.

Cook, K. & Hegtvedt, K. 1992 Empirical evidence of the sense of justice. In *The sense of justice: biological foundations of law* (ed. R. Masters & M. Gruter), pp. 187–210. Newbury Park, CA: Sage Publications.

Damasio, A. R. 1994 *Descartes' error: emotion, reason, and the human brain*. New York: Putnam.

Dawkins, R. 1989 *The selfish gene*, 2nd edn. Oxford University Press.

Dawkins, R. 1999 *The extended phenotype: the long reach of the gene*. Oxford University Press.

Descartes, R. 1649 *L'homme*. Cambridge, MA: Harvard University Press. [*Treatise on man*, 1972 translation by T. S. Hall.]

Dugatin, L. A. 2001 Subjective commitment in nonhumans: what should we be looking for, and where should we be looking?. In *Evolution and the capacity for commitment* (ed. R. Neese), pp. 120–137. New York: Russell Sage.

Frank, R. H. 1988 *Passions within reason: the strategic role of the emotions*. New York: W.W. Norton.

Feder, H. M. 1966 Cleaning symbioscs in the marine environment. *Symbiosis* **1**, 327–380.

Gintis, H. 2000 *Game theory evolving*. Princeton University Press.

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

Glimcher, P. W. 2003 *Decisions, uncertainty, and the brain: the science of neuroeconomics*. Cambridge, MA: MIT Press.

Glimcher, P. W., Dorris, M. C., Bayer, H. M., & Lau, B. 2004 Physiologic utility theory and the neuroeconomics of choice. *Games Econ. Behav*. (In the press.)

Gnadt, J. W. & Andersen, R. A. 1988 Memory related motor planning activity in posterior parietal cortex of macaque. *Exp. Brain Res.* **70**, 216–220.

Hamilton, W. D. 1964 The genetical evolution of social behavior. *J. Theor. Biol.* **7**, 1–52.

Henrich, J., Boyd, J., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElreath, R. 2001 In search of homo economicus: behavioral experiments in fifteen small-scale societies. *Am. Econ. Rev.* **91**, 73–78.

Holmes, O. W. Jr 1881 *The common law*. Cambridge, MA: Harvard University Press.

Holmes, O. W. Jr 1897 The path of the law. *Harv. Law Rev.* **10**, 457–478.

Holmes O.W., Jr 1917 Southern Pac. Co. v. Jensen, 244 US 205, 222.
Hume, D. 1748 *Enquiries concerning human understanding and concerning principles of morals*, 3rd edn. Oxford: Clarendon. [1975 edition, ed. L.A. Selby-Bigge.]
Jones, O. D. 2001*a* Time-shifted rationality and the law of law's leverage: behavioral economics meets behavioral biology. *Northwestern Univ. Law Rev.* **95**, 1141–1205.
Jones, O. D. 2001*b* Proprioception, non-law and bio-legal history (2001 Dunwody Distinguished Lecture in Law). *Florida Law Rev.* **53**, 831–874.
Kahneman, D. 1991 The endowment effect, loss aversion, and status quo bias. *J. Econ. Perspect.* **5**, 193–194.
McGinnis, J. O. 1996 Original constitution and our origins. *Harv. J. Law Pub. Pol.* **19**, 251–261.
Masters, R. D., Gruter, M. (eds) 1992 *The sense of justice*. Newbury Park, CA:Sage
Maynard Smith, J. 1982 *Evolution and the theory of games*. New York: Cambridge University Press.
Nash, J. F. 1950 Equilibrium points in n-person games. *Proc. Natl Acad. Sci. USA* **36**, 48–49.
Nozick, R. 1974 *Anarchy, state and utopia*. New York: Basic Books.
Pinker, S. 1994 *The language instinct*. New York: W. Morrow.
Plato *ca*. 360 B.C. *The republic*. New York: Colonial Press. [1901 edition, translated by B. Jowett.]
Posner, R. A. 1983 *The economics of justice*. Cambridge, MA: Harvard University Press.
Posner, R. A. 1992 *Essential Holmes: selections from the letters, speeches, judicial opinions and other writings of Oliver Wendell Holmes Jr*. University of Chicago Press.
Rawls, J. 1971 *A theory of justice*. Cambridge, MA: Harvard University Press.
Rubin, P. H. 2002 *Darwinian politics: the evolutionary origin of freedom*. New York: Rutgers.
Trivers, R. L. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57.
Tudge, C. 1998 *Neanderthals, bandits, and farmers: how agriculture really started*. New Haven, CT: Yale University Press.
Von Holtz, E. & Middelstaedt, H. 1950 Das Reafferenzprinzip: Wechselwirkung zwischen Zentralnervensystem und Perepherie [The reafference principle: interaction between the central nervous system and the periphery.] *Naturwissenschaften* **37**, 464–476. In *The behavioral physiology of animals and man: the selected papers of E. Von Holtz* (ed. R. D. Martin). Miami, FL: University of Miami.
Williams, G. 1966 *Adaptation and natural selection*. Princeton University Press.
Wilson, E. O. 1975 *Sociobiology: the new synthesis*. Cambridge, MA: Harvard University Press.
Wilson, E. O. 1998 *Consilience*. New York: Knopf.