

Hume's Natural History of Justice: Social Exchange and Self-Control

Mark Collier
University of Minnesota, Morris

In Book III, Part 2 of the *Treatise*, Hume presents a natural history of justice conventions. Clearly, prudential considerations play a central role in his account. After all, Hume maintains that our ancestors established justice conventions for the sake of reciprocal advantage. But this is not what makes his approach so novel and attractive. Hume recognizes that strategic rationality cannot fully explain how human beings – with our propensity to defect for the sake of short-term gains – could establish conventions for delayed social exchange in large-scale societies. This leads him to propose an innovative account of the role that sympathy plays in establishing trust between impulsive and shortsighted agents. One might object that his sentimentalist proposal amounts to wishful thinking, but it receives a good deal of support from emerging research in neuroeconomics and experimental game theory.

Nature, according to Hume, appears to have exercised particular “cruelty” towards human beings (T 3.2.2.2; SBN 484). When one surveys the rest of the animal kingdom, one observes a delicate balance between *what creatures want* and *what they can accomplish*. Lions have voracious appetites, but they have the means to satisfy them; sheep have simple desires, but these are easily fulfilled. It is in man alone that we find an “unnatural conjunction of infirmity, and of necessity” (T 3.2.2.2; SBN 485). Upon further reflection, however, it becomes apparent that nature has provided us with a remedy for this unfortunate predicament: social cooperation (T 3.2.2.3; SBN 485).

‘Tis by society alone he is able to supply his defects, and raise himself up to an equality with his fellow creatures... By the conjunction of forces, our power is augmented: By the partition of employments, our ability increases: And by mutual succour we are less expos’d to fortune and accidents. (T 3.2.2.3; SBN 485)

Social cooperation allows us to compensate for our feeble frames. We are not the fastest, strongest, or sturdiest creatures, but we make up for these shortcomings by joining forces with one another.

This raises an important question: what is it about human beings that allow us, unlike other animals, to cooperate on such a vast scale? Hume recognizes that the origin of human cooperation is a puzzle. One would expect that conflict, rather than cooperation, would have characterized the lives of our primitive ancestors. After all, resources in this ancestral environment would have been *scarce*; there were simply not enough goods to go around. Moreover, these goods would have been *unstable*, since individuals competing for them would have been driven by insatiable greed for possessions.

This avidity... of acquiring goods and possessions for ourselves and our nearest friends, is insatiable, perpetual, universal, and directly destructive of society. There scarce is any one, who is not actuated by it; and there is no one, who has not reason to fear from it, when it acts without any restraint, and gives way to its first and most natural movements. (T 3.2.2.12; SBN 491-2)

When creatures with unlimited appetites strive for limited resources, they would inevitably come to blows. Thus, it is hard to understand how our ancestors could have cooperated with one another in these circumstances.

Hume maintains that one cannot simply appeal to generosity or benevolence in order to solve this puzzle. He does not deny that we are capable of acting unselfishly; the problem is that human kindness is narrowly tailored (T 3.2.2.5; SBN 487). We naturally care about our friends and relatives, but we are also largely indifferent towards those outside our tribe (T 3.2.2.6; SBN 487). Thus, it is easy to understand why we refrain

from the possessions of those in our close circle, but it is difficult to see why we would do the same with regard to strangers, whose welfare is of little concern to us.

The solution to this puzzle, according to Hume, lies with our capacity for strategic reasoning.

Men being naturally selfish, or endow'd only with a confin'd generosity, they are not easily induc'd to perform any action for the interest of strangers, except with a view to some reciprocal advantage, which they had no hope of obtaining but by such a performance. (T 3.2.5.8; SBN 519)

Our ancestors began to cooperate with strangers, on this account, because they recognized that doing so would further their interests. Hume maintains that this lesson would not have been difficult to learn. After all, our ancestors would have had “repeated experience” of the “inconveniences” associated with the uninhibited pursuit of scarce resources (T 3.2.2.10; SBN 490). These inconveniences would have stood in stark contrast with the benefits of cooperation exhibited within family life (T 3.2.2.4; SBN 486). Thus, it would have required only minimal intelligence to recognize the advantages of a policy of mutual restraint from the possessions of others. As Hume puts it, this would have been “palpable and evident, even to the most rude and uncultivated of human race” (T 3.2.7.1; SBN 534).

It would not have been enough, of course, to believe that there is reciprocal advantage in abstention from the possessions of others. There is no profit in adopting such a policy unless others are willing to do the same. Thus, one must also believe *that others believe* that they would benefit if they restrained their natural appropriative impulses. Moreover, one must believe *that others are willing to put this policy into practice*. How could our ancestors have assured one another that this was the case? Hume maintains that they must have *expressed* their interests to one another. Property

conventions would have been established when these expressions gave rise to patterns of mutual restraint.

This convention... is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behavior. (T 3.2.2.10; SBN 490)

In other words, our ancestors managed to abstain from each other's possessions because they expected that others would reciprocate; and they expected that other would reciprocate because they acknowledged and expressed their interests in doing so.

Humean conventions involve agreements where “the actions of each of us have a reference to those of the other” (T 3.2.2.10; SBN 490). Let us call individuals *strategically rational agents* if they recognize that the outcomes of their decisions can depend upon the choices of others. We can then say that two persons enter into *conventions for the stable possession of property* if and only if:

- (A) Each person believes that the other is a strategically rational agent.
- (B) Each person believes that mutual abstention from the possessions of others would promote their individual interests.
- (C) Each person believes that the other person believes that mutual abstention from the possessions of others would promote their individual interests.
- (D) The beliefs in (A)-(C) produce a suitable regularity in their behavior.

This analysis makes it clear that property conventions require a good deal of cognitive sophistication. Agents must not only be capable of strategic reasoning (B), but they must also be able to attribute this capacity to others (A). Indeed, agents must attribute quite sophisticated mental states to one another. We must not only believe *that others believe that mutual abstention promotes their interests*, but we must also believe *that others*

believe that we believe this, and that *they believe that we believe that they believe this*, etc. In other words, the common knowledge condition (C) requires that we can form recursive, higher-order expectations (Lewis 1969, p.56).

Conventions for the stable possession of property promote our interests, but they offer little help when we find ourselves with a *surplus of goods*. In order to reap the full benefits of social cooperation, according to Hume, our ancestors must have settled upon conventions for the exchange of property.

Different parts of the earth produce different commodities; and not only so, but different men both are by nature fitted for different employments, and attain to greater perfection in any one, when they confine themselves to it alone. All this requires a mutual exchange and commerce.... (T 3.2.4.1; SBN 514; cf. T 3.2.2.3; SBN 485)

It would have occurred to our ancestors by “plain utility and interest” that they would each benefit from the reciprocal exchange of surplus goods and services (T 3.2.4.2; SBN 515). Thus, we can say that two agents enter into *conventions for social exchange* if and only if:

- (A) Each person believes that the other is a strategically rational agent.
- (B) Each person believes that social exchange would promote their individual interests.
- (C) Each person believes that the other person believes that social exchange would promote their individual interests.
- (D) The beliefs in (A)-(C) produce a suitable regularity in their behavior.

Hume maintains that conventions for social exchange would have gradually evolved as strategically rational agents learned to regulate their behavior on the expectation that others would reciprocate in turn.

Hume recognizes that social exchange conventions are difficult to explain, however, in the context of large-scale societies. The problem is that the transfer of goods and services in economies of any significant size will involve *temporally deferred*

exchanges. Suppose that I promise to reimburse you next week if you provide me with your goods or services today. In such circumstances, you would not agree to the exchange unless you were assured that I will keep my word when the appropriate time comes. If we are strangers to each other, however, you would have no reason to believe that I will do so. Consider Hume's "farmer's dilemma" (Skyrms and Vanderschraaf 1997, pp. 4-6).

Your corn is ripe today; mine will be so to-morrow. 'This profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and should I labour with you upon my account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (T 3.2.5.8; SBN 520-521)

We can trust that our friends and family will keep up their end of the bargain, since they are concerned about our welfare. But the problem is that such assurance is much harder to come by when we are dealing with strangers.

Conventions for social exchange could not be established unless strategically rational agents were assured that their partners will reciprocate. But how could they trust one another to do so? Hume maintains that strategically rational agents would recognize that it is in their *long-term* interests to cooperate in social exchanges. They would understand, as he puts it, that anyone caught cheating "must never expect to be trusted any more". (T 3.2.5.10; SBN 522)

Hence I learn to do a service to another, without bearing him any real kindness; because I foresee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising

from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal. (T 3.2.5.9; SBN 521)

The crucial point is that agents would *stake their reputations* on the exchange. If one of them fails to reciprocate, therefore, he would deprive himself of the benefits of future commerce. Even though unilateral defection is a dominant strategy in a *one-shot* farmer's dilemma, mutual cooperation would promote their interests in an *iterated* sequence of games (Vanderschraaf 1998, pp. 223-225). Thus, conventions for social exchange could be established because strategically rational agents would expect each other to act on their long-term interests.

Hume recognizes that this *rational solution* to the farmer's dilemma is not sufficient as it stands. The problem is that human beings have a strong propensity to *discount the future*.

[E]very thing, that is contiguous to us... commonly operates with more force than any object, that lies in a more distant and obscure light. Tho' we may be fully convinc'd, that the latter object excels the former, we are not able to regulate our actions by this judgment; but yield to the solicitations of our passions, which always plead in favour of whatever is near and contiguous. This is the reason why men so often act in contradiction to their known interest; and in particular why they prefer any trivial advantage, that is present, to the maintenance of order in society, which so much depends upon the observance of justice. (T 3.2.7.2-3; SBN 535)

The rational solution to the farmer's dilemma is that prudence counsels us to reciprocate for the sake of our long-term agents interests. But it seems that human beings would be incapable of taking such advice, because of our tendency to prefer *small-rewards now* over *large-rewards later*.¹

¹ This psychological propensity to prefer the contiguous over the remote has been confirmed in a wide variety of experimental studies (Ainslie 2001, Ch.3). Indeed, researchers have demonstrated that this propensity is strongly correlated with defection in iterated Prisoner's Dilemmas (Harris and Madden 2002; Yi et al. 2005).

We seem to have fallen right back into the jaws of the farmer's dilemma. Neither farmer would have reason to cooperate in these circumstances because they would not expect their partner to reciprocate.

You have the same propension, that I have, in favour of what is contiguous above what is remote. You are, therefore, naturally committed to commit acts of injustice as well as me. Your example both pushes me forward in this way by imitation, and also affords me a new reason for any breach of equity, by shewing me, that I should be the cully of my integrity, if I alone shou'd impose on myself a severe restraint amidst the licentiousness of others. (T 3.2.7.3; SBN 535)

It seems that our previous analysis of social exchange conventions was incomplete. Agents must not only believe that their partners understand the reciprocal advantage of social exchange; they must also believe *that their partners are self-controlled enough to act on their acknowledged interests*. If human beings are incapable of delayed gratification, however, this condition could never be satisfied. Strategically rational agents would not agree to sequential, deferred exchanges with shortsighted and impulsive partners.

Hume recognizes that conventions for social exchange could never have been established if our propensity to discount the future was left unchecked. In order to complete his natural history of justice conventions, he attempts to explain how our ancestors overcame this obstacle to cooperation. Hume rejects any solution that appeals to the strenuous effort of individuals; it is simply naive to think that agents could conquer their impulsivity through "frequent meditation" or a "repeated resolution" to be strong-willed (T 3.2.7.5; SBN 536). The remedy must be significantly harsher. Indeed, he maintains that weak-willed agents could never achieve self-control unless they submit themselves to the coercive power of institutional sanctions.

Men are not able to radically cure, either in themselves or others that narrowness of soul, which makes them prefer the present to the remote. They cannot change their natures. All they can do is change their situation, and render the observance of justice the immediate interest of some particular persons, and its violation their most remote. These persons, then, are not only induc'd to observe those rules in their own conduct, but also to constrain others to a like regularity, and enforce the dictates of equity thro' the whole society. (T 3.2.7.6; SBN 537)

In contemporary terms, these institutions serve as “commitment devices” (Schelling 1960). Like alcoholics who voluntarily check themselves into a clinic, we become aware of our propensity to discount the future, and thus we commit ourselves to a system of sanctions that will force us to pursue our long-term interests. According to Hume, we do not create the rule of law in order to safeguard us from others; rather, it is invented to protect us from ourselves (Harrison 1981, p. 172).

Hume’s institutional solution, however, is flawed in several important respects. For one thing, sanctions would prove to be ineffective deterrents for shortsighted subjects; if agents have a propensity to discount future consequences, threats of *future punishment* would provide little help. Moreover, it is unclear why those who run these institutions would not succumb to the very same propensity. It might very well be the case that enforcing reciprocal exchanges promotes their long-term interests; but the only way that they could avoid discounting these interests would be to submit themselves to a higher coercive institution, and so on *ad infinitum*.² In short, the institutional solution creates more problems than it solves: a system of sanctions could never be established in

² Hume attempts to forestall this objection by stipulating that it is in the *immediate interests* of the magistrates to serve as a commitment device. But it is hard to see why this would be the case. There would presumably be a wide range of private commercial exchanges towards which the magistrates would be entirely indifferent. They would only be concerned with defections in private exchanges because they might lead to the *eventual* collapse of social cooperation. But the problem is that shortsighted creatures would discount these future threats.

the first place, and even if it could, it would hardly be useful when it comes to creatures with a propensity to discount the future.

Nevertheless, Hume provides a more promising solution to the problem of self-control. Interestingly, we can find important clues about this alternative account in his often overlooked discussion of chastity norms. Hume's genealogy of chastity begins with a number of plausible assumptions:

- (i.) Men and women have common interest in raising children to maturity.
- (ii.) Children require lengthy and costly parental care.
- (iii.) Men are unwilling to contribute to these costs unless they are assured that the child is related to them.
- (iv.) Men cannot be assured of paternity unless they can trust that their mate is faithful.

Given these assumptions, we can see that men face an *assurance problem* when it comes to raising children: they will not contribute to the costs of child care unless they are confident that they have not been cuckolded. But how can this condition ever be satisfied? That is, how can they possibly trust that their mates are faithful?

It is obvious that *vows of fidelity* cannot solve this assurance problem. The question at hand is whether someone is trustworthy. Thus, one cannot simply take them on their word. (In the language of game theory, this is "cheap talk"). Hume also considers, and quickly rejects, the notion that *institutional sanctions* could provide men with the requisite assurance. The problem is that "legal proof", as he puts it, would be "difficult to meet with in this subject" (T 3.2.12.4; SBN 571). This is not the case in the court of public opinion, however, where the standards for indictment are notoriously low. Reputations are lost upon the slightest presumption of impropriety. Perhaps this simple fact is all we need in order to solve the problem.

What restraint, therefore, shall we impose on women, in order to counter-balance so strong a temptation as they have to infidelity? There seems to

be no restraint possible, but in the punishment of bad fame or reputation; a punishment, which has a mighty influence on the human mind, and at the same time is inflicted by the world upon surmises, and conjectures, and proofs, that would never be receiv'd in any court of judicature. (T 3.2.12.4; SBN 571)

Women have an acknowledged interest in maintaining their reputations in society. They will carefully avoid any situation, therefore, that might provoke indignation from others. Thus, men can be assured that their mates will avoid even the slightest suggestion of infidelity; the risks would simply be too great.

This solution is plausible as far as it goes; but it does not go far enough. The problem is that women, like everyone else, have a propensity to discount the future. As a result, they will have a tendency to overlook the damage that might be done to their reputations.

All human creatures, especially of the female sex, are apt to over-look remote motives in favour of any present temptation: The temptation is here the strongest imaginable: Its approaches are insensible and seducing: And woman easily finds, or flatters herself she shall find, certain means of securing her reputation, and preventing all the pernicious consequences of her pleasures. (T 3.2.12.5; SBN 572-3)

It is not enough that that men believe *that their mates believe that infidelity goes against their long-term interests*. Even if women acknowledge their long-term interests in fidelity, they might be unable to resist the allure of the present moment. Thus, we find ourselves back where we started. Men would not contribute to the costs of child care unless they believe that their mates are self-controlled enough to act on their acknowledged interests. But if it is common knowledge that human beings are impulsive, this condition would never be satisfied.

Hume maintains that trust could never be established in these circumstances unless men believe *that their partners are disposed to feel repugnance at the very thought of infidelity.*

‘Tis necessary, therefore, that, beside the infamy attending such licenses, there shou’d be some preceding backwardness or dread, which may prevent their first approaches, and may give the female sex a repugnance to all expressions, and postures, and liberties, that have an immediate relation to that enjoyment (T 3.2.12.5; SBN 572).

The crucial insight is that these feelings of repugnance occur at the *moment of deliberation*, and thus will not be discounted along with the future consequences (cf. Frank, 1988, p. 82). Agents act impulsively because they overestimate the value of present goods; thus, the only way this propensity could be corrected is if these rewards were rendered less attractive. And this is precisely what is accomplished by the feeling of repugnance: it serves as a contrary hedonic impulse.

Hume recognizes that he must account for these contrary impulses. After all, it seems *prima facie* implausible to maintain that women would feel repugnance at “the approaches of a pleasure, to which nature has inspir’d so strong a propensity” (T 3.2.12.6; SBN 572). His psychological explanation appeals to the mechanisms of sympathy. Our capacity to sympathize with the welfare of others leads us to disapprove of conjugal infidelity and its socially destructive consequences. The mechanisms of sympathy also communicate this disapproval through a process of emotional contagion; as a result, even those without any particular interests in chastity are “carried along with the stream” and feel repugnance at the very thought of infidelity (T 3.2.12.7; SBN 572). The association between infidelity and shame is reinforced by education and the rhetoric of politicians. Women are conditioned, from a very early age, to look upon infidelity with a weary eye.

Thus, Hume's official position is that sympathy and socialization work together to insure that women acquire an aversion to an otherwise natural impulse.

Hume's natural history of chastity norms suggests an alternative solution to the problem of how impulsive agents could establish conventions for delayed exchanges. The challenge is to explain why strategically rational agents would trust their shortsighted partners to reciprocate once they have received the benefits of the exchange. Hume recognizes that prudential considerations are insufficient to solve this assurance problem. Agents would not participate in delayed social exchanges unless they had reason to believe that their partners were *disposed to feel repugnance at the very thought of injustice*. Mutual trust would be impossible between agents who lack what he calls an "abhorrence of injustice" (T 3.2.2.25; SBN 501). It is the fact that you detest cheating which assures me that you are not only prudent enough to recognize the benefits of reciprocal exchange, but that you are also self-controlled enough to keep up your side of the bargain.

As with the case of chastity, Hume recognizes that he is obligated to account for these feelings of repugnance towards injustice. It seems implausible to maintain that predominantly selfish agents would be averse to cheating strangers, especially when unilateral defection offers the highest material payoff. His psychological explanation once again appeals to the mechanisms of sympathy: agents initially favor cheating whenever they stand to benefit, but they come to disapprove of this behavior when they consider its negative consequences on human commerce. It is our sympathy with the interests of the public, as he puts it, which "gives us an uneasiness" towards "all kinds of private injustice" (T 3.2.8.7; SBN 545). And it is this psychological aversion to cheating

which enables shortsighted agents to overcome their propensity to defect for the sake of short-term gains.

Hume anticipates that one will object to his sentimental solution on the grounds that it rests upon wishful thinking and “chimerical speculation” (T 3.2.12.6; SBN 572). According to his proposal, we reciprocate in social exchanges because we detest cheating. But is there any empirical evidence that human beings acquire these psychological aversions to injustice? Hume was not in a position, of course, to test his account. It is perhaps for this reason that scholars have largely refrained from endorsing it (Mackie 1980, p. 90). But it is no longer the case that we must remain neutral on this issue: researchers in the emerging field of neuroeconomics have developed an experimental paradigm which allows us to put these speculative claims to the test.

In the neuroeconomic experiments, researchers measure the brain activity of participants as they make decisions in games involving social exchange and collective action. This experimental approach is admittedly in its early stages. But the results are coming in, and they provide a good deal of support for Hume’s sentimentalist solution to the puzzle of cooperation. In one groundbreaking experiment, researchers demonstrate that areas of the brain which process *negative hedonic rewards* become activated when players make uncooperative moves in iterated Prisoner’s Dilemmas (Rilling et al. 2002, p. 399). Indeed, these findings suggest that this neural reward circuitry enables participants to cooperate in these tasks by inhibiting the impulsive temptation to defect for the sake of short-term gains (ibid., p. 403).

There is also neuroeconomic evidence in favor of Hume’s proposal that *shared expectations* play a crucial role in establishing conventions for social exchange.

Researchers measured the brain activity of subjects as they participated in two-person reciprocal exchanges. The results indicate that areas of the frontal cortex associated with *joint attention* become activate when subjects cooperate with each other in these tasks. Indeed, these experiments suggest that this ability to form convergent expectations about mutual gain promotes cooperative behavior by helping to inhibit the impulsive desire for immediate gratification (McCabe et. al 2001, p. 11834).

Previous commentators have correctly emphasized the informal game-theoretic notions in Hume's theory of justice (Lewis 1969, Gauthier 1979, Mackie 1980, Vanderschraaf 1998, Hardin 2007). Clearly, self-interest and strategic rationality play a pivotal role in his account. Indeed, Hume maintains that these considerations are sufficient to explain the development of justice conventions in small-scale societies. But Hume makes it clear that prudential rationality cannot fully explain how imperfect creatures such as ourselves – with propensities to discount the future – could establish these conventions in large-scale societies. This leads him to present an innovative account of the psychological prerequisites of justice conventions: human beings cooperate with one another to the unique extent that we do because we are strategically rational creatures with a heart.

REFERENCES

Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press.

Frank, R. (1988). *Passions within reason: the strategic role of the emotions*. W.W. Norton & Co.

Gauthier, D. (1979). David Hume: contractarian. *Philosophical Review*, 88, 3-38.

- Hardin, R. (1982). *Collective action*. Johns Hopkins University Press.
- Hardin, R. (1993). From power to order, from Hobbes to Hume. *The Journal of Political Philosophy*, 1, 69-81.
- Hardin, R. (2007). *David Hume: Moral and Political Philosophy*. Oxford University Press.
- Harris, A. and Madden, G. (2002). Delay discounting and performance on the prisoner's dilemma game. *The Psychological Record*, 52, 429-440.
- Harrison, J. (1981). *Hume's theory of justice*. Oxford University Press.
- Hume, D. (1998). *An enquiry concerning the principles of morals*. Oxford University Press.
- Hume, D. (2000). *A treatise of human nature*. Oxford University Press.
- Lewis (1969). *Convention: a philosophical study*. Harvard University Press.
- Mackie, J. (1980). *Hume's moral theory*. Routledge, Kegan, and Paul.
- McCabe et al. (2001). A functional imagining study of cooperation in two-person reciprocal exchange.
- Rilling, J., Gutman, D. et al. (2002). A neural basis for social cooperation. *Neuron*, 35, 395-405.
- Schelling (1960). *The strategy of conflict*. Harvard University Press.
- Skyrms, B. and Vanderschraaf, P. (1997). Game theory. In P. Smets (Ed.) *The Handbook of Practical Logic*. Kluwer Press.
- Taylor, M. (1987). *Anarchy and Cooperation*. John Wiley and Sons.
- Vanderschraaf, P. (1998). The informal game theory in Hume's account of convention. *Economics & Philosophy*, 14, 215-47.
- Yi, R., Johnson, M. W., & Bickel, W. K. (2005). Relationship between cooperation in an iterated prisoner's dilemma game and the discounting of hypothetical outcomes. *Learning and Behavior*, 33, 324-336.