Intertemporal bargaining predicts moral behavior, even in anonymous, one-shot economic games

Commentary on:
Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: the evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59-78.

George Ainslie
School of Economics, University of Cape Town, Rondebosch 7701, South Africa; and Department of Veteran Affairs, 151 VA Medical Center, Coatesville, PA 19320.

George.Ainslie@va.gov **http://www.Picoeconomics.org**

Abstract
To the extent that acting fairly is in an individual's long-term interest, short-term impulses to cheat present a self-control problem. The only effective solution is to interpret the problem as a variant of repeated prisoner's dilemma, with each choice as a test case predicting future choices. Moral choice appears to be the product of a contract because it comes from self-enforcing intertemporal cooperation.

Text
The target article by Baumard et al. argues that an intrinsic motive for fairness has been socially selected and has thus evolved as one of the "mental and social mechanisms that produce moral judgments and interactions" (Abstract). Alternatively (it seems), the authors suggest that people may feel like selfishly free-riding, but are restrained by "a prudence which . . . is built into our evolved moral disposition" (sect. 2.2.2, para. 3). Either way, an innate moral preference is said to account for three otherwise anomalous kinds of self-depriving behavior: where a subject (1) helps strangers without expectation of return, (2) cooperates in anonymous, one-shot games, and (3) pays to punish others for their moves in public goods games (sect. 2.2). The argument for social selection is well thought out. However, before we add either special motive to the long list of elementary needs, drives, and other incentives that have been discerned in human choice (e.g., Atkinson & Raynor 1975), we should examine whether known properties of reward might not explain a preference for fairness, or for the very similar traits of inequity aversion (Frohlich et al. 2004) and game-theoretic choreography (Gintis 2009, pp. 41–44).

Much of the target article discusses how people arrive at cognitive judgments of fairness, but the tough problem is motivational. It may be that "competition among cooperative partners leads to the selection of a disposition to be intrinsically motivated to be fair" (sect. 2.2.1, para. 12), but people continue to have a disposition to be selfish as well, and perhaps also a disposition to be altruistic and leave themselves open to exploitation. Among these dispositions, morality does not compete like just another taste, but leads people to "behave *as if* they had passed a contract" (sect. 3.2.2, para 1, italics in

the original; see also sects. 1 and 2.2.2). The article's central problem is, "since [people] didn't, why should it be so?" (sect. 1, para. 2). The authors' proposal of an innate moral preference to solve this "puzzle of the missing contract" (sect. 1, para. 3) just names the phenomenon, rather than supplying a proximate mechanism for the contract-like faculty.

Rather, we should look at the purpose of the contract. The payoffs for selfish choices are almost always faster than the payoffs for moral ones. If I fake fairness like an intelligent sociopath, I may eventually be found out, but I will reap rewards in the short run; and the likelihood that I will get away with any given deception increases my temptation to try it. Thus, even if I realize that fairness serves my own long-term interests, I face ongoing pressure from my short-term interests to cheat. There is still controversy over whether people overvalue imminent rewards generally (hyperbolic discounting; see Ainslie 2010, 2012 in press) or only when we are emotionally aroused (hyperboloid discounting; see McClure et al. 2007), but in either case I will often have the impulse to cheat when it is against my long-term interest. Since faking my motives is an entirely intrapsychic process, the only way I can commit myself not to do it is to interpret my current choice as a test case for how I am apt to choose in the future: "If I am hypocritical [or biased, or selfish . . .] this time, why wouldn't I expect to be next time?" Thus bundled together, a series of impulses loses leverage against a series of better, later alternatives – greatly if the discounting is hyperbolic, less so but still possibly if the discounting is hyperboloid (Ainslie, 2012). Then, to the extent that I am aware of my temptation problem, I will have an incentive to make *personal rules* against deciding unfairly – that is, to interpret each choice where I might be unfair as a test case of whether I can expect to resist this kind of temptation in the future. I draw the line between fair and unfair by the kind of reasoning that Baumard et al. describe, and then face reward contingencies that will be similar to those of a repeated prisoner's dilemma. Whatever my reputation is with other people, I will have a reputation with myself that is at stake in each choice, and which, like my social reputation, is disproportionately vulnerable to lapses (Monterosso et al. 2002).

This dynamic can account for two of the three phenomena that the authors highlight as seeming anomalies for mutualism:

1. Although helping strangers without expectation of return can be rewarding in its own right, I may also help them because of a personal rule for fairness at times when I would rather cheat and could do so without social consequences. Then I do behave as if I had made a social contract. The contract is real, but exists between my present self and my expected future selves. Like the oral contracts among traders that Baumard et al. list (sect. 2.1.3, para. 1), my contract is self-enforcing. I may still get away with cheating, by means of the casuistry with personal rules called rationalization; or I may instead become hyper-moral, if I am especially fearful of giving myself an unfavorable self-signal (Bodner & Prelec 2001). Either deviation moves me away from optimal social desirability, but my central anchor is just where Baumard et al. say it should be.
2. To the extent that my reputation with myself feels vulnerable, I may reject an experimenter's instruction to maximize my personal payoff in a one-shot Prisoner's Dilemma or Dictator game, and instead regard the game as another test case of my

character (Ainslie 2005). Such an interpretation makes it "not that easy . . . to shed off one's intuitive social and moral dispositions when participating in such a game" (sect. 3.3.2, para. 1).

3. No further explanation seems necessary for the punishment phenomenon. It is not remarkable that subjects become angry at either cheating or moralizing stances by other subjects, and pay to indulge this anger. As with problem (2), the seeming anomaly arises from experimenters' assumptions that the reward contingencies they set up for a game are the only ones in subjects' minds.

As for the cognitive criteria for partners' value, talent and effort probably do not exhaust the qualities that are rationally weighed in social choice. Wealth or status conveyed by inheritance or the happenstance of history have always been factors, and transparency itself – how easy it is to be evaluated – must be one. But the authors' proposal of social selection will work perfectly well with other criteria for estimation. The hard part of their goal ("to contribute . . . proximate and ultimate explanations of human morality"; target article, Abstract) has been to explain the semblance of bargaining when counterparties are apparently absent. This can be accomplished by the logic of internal intertemporal bargaining, without positing a specially evolved motive.

ACKNOWLEDGMENT
This material is the result of work supported with resources and the use of facilities at the Department of Veterans Affairs Medical Center, Coatesville, PA. The opinions expressed are not those of the Department of Veterans Affairs or of the US Government.

References

Ainslie, G. (2005) You can't give permission to be a bastard: Empathy and self-signaling as uncontrollable independent variables in bargaining games. *Behavioral and Brain Sciences* 28:815–16.

Ainslie, G. (2010) The core process in addictions and other impulses: Hyperbolic discounting versus conditioning and framing. In: *What is addiction?* ed. D. Ross, H. Kincaid, D. Spurrett & P. Collins, pp. 211–45. MIT Press.

Ainslie, G. (2012) Pure hyperbolic discount curves predict "eyes open" self-control. *Theory and Decision* 73:3-34 . doi:10.1007/s11238-011-9272-5.

Atkinson, J. W. & Raynor, J. O. (1975) *Motivation and achievement*. Winston.

Bodner, R. & Prelec, D. (2001) The diagnostic value of actions in a self-signaling model. In: *Collected essays in psychology and economics*, ed. I. Brocas & J. D. Carillo, pp. 105–23. Oxford University Press.

Frohlich, N., Oppenheimer, J. & Kurki, A. (2004) Modeling other-regarding preferences and an experimental test. *Public Choice* 119:91–117.

Gintis, H. (2009) *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton University Press.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. (2007) Time discounting for primary rewards. *Journal of Neuroscience* 27:5796–804.

Monterosso, J. R., Ainslie, G., Toppi-Mullen, P. & Gault, B. (2002) The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:437–48.