https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

SENTIMENTS, CONDUCT, AND TRUST IN THE LABORATORY*

By Vernon L. Smith and Bart J. Wilson

Abstract: *In this essay we provide a brief account and interpretation of* The Theory of Moral Sentiments *showing that it departs fundamentally from contemporary patterns of thought in economics that are believed to govern individual behavior in small groups, and contains strong testable propositions governing the expression of that behavior. We also state a formal representation of the model for individual choice of action, apply the propositions to the prediction of actions in trust games, report two experiments testing these predictions, and interpret the results in terms directly related to the model. In short, we argue that the system of sociability developed by Adam Smith provides a coherent non-utilitarian model that is consistent with the pattern of results in trust games, and leads to testable new predictions, some of which we test.*

KEY WORDS: Adam Smith, experimental economics, trust, conduct, beneficence, justice

> Behaviour.
>
> 1. Manner of behaving one's self, whether good or bad; manners. . . .
> 5. Conduct; general practice; course of life.
>
> Conduct.
>
> 6. Behaviour; regular life.
>
> Samuel Johnson, *Dictionary of the English Language* (1755)

I. Background and Motivation

Current interest in trust games by experimental economists originated in the 1990s following upon earlier studies of two-person ultimatum and

* We thank Jeffrey Kirchner for his software programming *par excellence*; Jennifer Cunningham for diligently recruiting our subjects; Chapman University for financial support; Sean Crockett, Ryan Hanley, and Sigve Tjøtta for constructive comments; and finally the many students with whom we have read and discussed the ideas in Adam Smith's two great books. The essay has benefitted from seminar presentations at five universities (George Mason, Alaska Anchorage, Simon Fraser, Baylor, and Wofford College), and from interaction with the other contributors to the present volume.

doi:10.1017/S0265052517000024
© 2017 Social Philosophy & Policy Foundation. Printed in the USA.

dictator games.[1,2,3] The finding that decisions in these games collided with the predictions of game theory subsequently ignited a large literature on trust games. This literature has extensively replicated and explored the robustness of the original findings and launched a search for explanations and models, and the testing of their predictive implications in an effort to better account for the original discordant results.

Much of the subsequent research was motivated by the original reciprocity or exchange interpretation of these results and the costly punishment and reward strategies that characterized subject behavior:[4]

> In conclusion, experiments on ultimatum game, repeated prisoners' dilemma games, and other extensive form games provide strong evidence that people do punish inappropriate behavior even though this is personally costly. Furthermore, subjects take this into account when they make their decisions. The investment game provides evidence that people are also willing to reward appropriate behavior and this too is taken into account. Taken together these results suggest that both positive and negative forms of reciprocity exist and must be taken into account in order to explain the development of institutional forms which reinforce the propensity to reciprocate.

The reciprocity narrative as an explanation of trust/trustworthiness derived much of its weight from concepts in evolutionary theory and in particular the developing field of evolutionary psychology theory that involved social exchange algorithms for "mind reading," "intentionality," and "cheater detection."[5] Following Berg, Dickhaut, and McCabe, many experiments established that intentions ("appropriate behavior") mattered; moreover, treatments that manipulated intentions or context had a greater impact on choices than treatments that varied payoffs.[6]

[1] See Joyce Berg, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior* 10 (1995): 122–42.

[2] Werner Güth, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3, no. 4 (1982): 367–88.

[3] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler, "Fairness And the Assumptions of Economics," *Journal of Business* 59, no. 4 (1986): S285–S300.

[4] Berg et al., "Trust, Reciprocity, and Social History," 138–39.

[5] Elizabeth Hoffman, Kevin A. McCabe, and Vernon L. Smith, "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry* 36, no. 3 (1998): 335–52.

[6] For discussions of stakes and context, see Colin F. Camerer, *Behavioral Games Theory* (Princeton, NJ: Princeton University Press, 2003), 60–61; and Vernon L. Smith, *Rationality in Economics: Constructivist and Ecological Forms* (Cambridge: Cambridge University Press, 2008), chap. 10; for intentions, see Kevin McCabe, Vernon L. Smith, and Michael LePore, eds., "Intentionality Detection and 'Mindreading': Why Does Game Form Matter," *Proceedings of the National Academy of Sciences* 97, no. 8 (2000): 4404–4409; and Ernst Fehr and Bettina Rockenbach, "Detrimental Effects of Sanctions on Human Altruism," *Nature* 422 (2003): 137–140.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

In this essay, we return to Adam Smith who provided a rich non-utilitarian model of conduct in human intercourse.

The growing empirical evidence in support of the original findings led to a second, more formal, response in which the traditional game-theoretic assumption of strictly self-interested agents was replaced by a utility function defined over both "own" and "other" reward payoffs, while retaining all the other assumptions.[7] Reciprocity was thereby interpreted as a form of revealed other-regarding behavior, and this could be rationalized within the game-theoretic framework by simply postulating that agents were driven by an "other-regarding" utility criterion. We wish to emphasize, however, that *when a key prediction of a theory fails, all of its assumptions must be on the table for reconsideration, and the search for a resolution must not exclude consideration of entirely different ways of thinking, representing, and modeling the phenomena.*

Cox, Friedman, and Gjerstadt supply a concise summary of models that enrich utility by the inclusion of "other" rewards.[8] Their model is particularly noteworthy in prefacing the experiments we report below because they parameterize utility to include postulated emotional responses — such as status, gratitude, and resentment — to the intentions conveyed by the first mover in two person games. They model only second-mover responses, but that is the obvious first step in a program to reform and redirect the theory exercise, and in itself is not the source of the problem with this approach as we shall address it in this essay.

In his first book, *The Theory of Moral Sentiments* (hereafter, *Sentiments* or *TMS*), *Adam Smith* articulated a theory of human sociality devoted to understanding moral human action, that is, the "practice of the duties of life."[9] He and his intellectual cohorts in the Scottish enlightenment were astute observers of their respective worlds of primary interest as they searched for the hidden rules that ordered the complex phenomena they studied. In this essay we (1) provide a brief account and interpretation of *Sentiments* showing that it (2) departs fundamentally from contemporary patterns of thought in economics that are believed to govern individual behavior in small groups, and (3) contains strong testable propositions governing the expression of that behavior; also we (4) state a formal representation of the model for individual choice of action, (5) apply the propositions

[7] These assumptions were: backward induction (players look ahead and apply reason to the analysis of other and own decisions); decisions are independent of the players' history or future (the game is played exactly once by anonymously paired players) and complete information on payoffs (fully displayed to both players). For further discussion, see Vernon L. Smith, "Theory and Experiment: What Are the Questions?" *Journal of Economic Behavior and Organization* 73, no. 1 (2010): 5–9.

[8] James C. Cox, Daniel Friedman, and Steven Gjerstad, "A Tractable Model of Reciprocity," *Games and Economic Behavior* 59 (2007): 17–45.

[9] Charles John Smith, *Synonyms Discriminated: A Dictionary of Synonymous Words in the English Language* (New York: Henry Holt and Company, 1894), 574.

to the prediction of actions in trust games, (6) report experiments testing these predictions and (7) interpret the results in terms directly related to the model. In short, we argue that the system of sociability developed in *Sentiments* provides a coherent non-utilitarian model that is consistent with the pattern of results in trust games, and leads to testable new predictions, some of which we test in what follows.

Although the model in *Sentiments* is neither outcome-based nor utilitarian, this does not mean that people were unconcerned with self: "Every man . . . is first and principally recommended to his own care; and . . . fitter and abler to take care of himself than any other person."[10] But Smith saw the socialization process as modifying action in the self-interest to bring it down to what one's contemporaries could go along with. Hence, both the social preferences and reciprocity models of action in the trust game are seen as flawed. Remarkably, his model anticipated many of the resolutions and interpretations of the experimental community in coming to terms with the striking predictive failure of traditional economic analysis in small group interactions.

II. "Pleasure" and the Mainsprings of Human Action in Sentiments

In contemporary representations by economists and cognitive psychologists, "pleasure" gives rise to "utility" whose measure is related functionally to a desirable (or undesirable) outcome resulting from the action. Given a choice among alternatives, an individual is postulated to choose the action that maximizes preference (hereafter, Max-U).[11] Utilitarian preference functions perform heavy duty work in modeling a vast range of human decisions: isolated individuals in psychophysical measurements, individuals choosing among uncertain probabilistic prospects, interactive agents in supply-and-demand, auction and asset markets, and individuals interacting through choices in two-person (for example, trust) games or in small groups (public good and common property games).

Max-U applied to small groups constitutes a fundamental departure from the intellectual modeling framework of *Sentiments*, but Adam Smith's systematic account of human action illuminates the processes that govern action in small groupings, of which trust games are an excellent example. Experimentalists have already designed and reported various experiments showing that intentions matter, and that the focus on outcomes needs to be re-examined. Since Adam Smith's model eschews outcomes and their utility (including social preference) and begins with

[10] Adam Smith, *The Theory of Moral Sentiments* (*Sentiments* or *TMS*) (Indianapolis, IN: Liberty Fund, [1759] 1982), 219.

[11] So abbreviated and further discussed by Deirdre N. McCloskey, *The Bourgeois Virtues: Ethics for an Age of Commerce* (Chicago: University of Chicago Press, 2006).

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

actions as signals of rule-following conduct, it seems particularly appropriate to probe the substance and implications of that model. In *Sentiments*, the individual is painted as inseparably connected from birth with overlapping social groupings based in family, extended family, friends and neighbors; these groupings in turn prepare and enable the individual to reach much beyond these narrow circles into daily life experiences. As Smith saw it, this is the world that first and originally defines the content and meaning of sociability, defines the individual within that social context, and out of which the civil order of society emerged based on property, defined as rights to undertake (or not) certain actions, conditional on circumstances.

The world of *Sentiments* envisions a pre-civil law community as a proving ground for fashioning the rules of social order in an environment disciplined by propriety, and bereft of any external enforcement of property. But of course it was a world in which individuals continued to engage and thrive long after the emergence of civil government, national economies, and the extended order of specialization and markets; the latter is the world Smith sought to understand in his better-known and phenomenally successful second book, *An Inquiry Into the Nature and Causes of the Wealth of Nations* (hereafter, *Wealth*).[12] But *Wealth* was a systematic treatment of economic development in civil society based on third party enforcement of property right rules — justice. The two worlds were distinct but complementary, and *Sentiments* articulated the critical preconditions for the emergence of justice and the enabling of civil society.

In *Sentiments* Adam Smith frequently makes reference to the "pleasure" associated with an action chosen by an individual. What did "pleasure" mean in *Sentiments*? The title of Part I, Section I, Chapter II, provides the key definition: "Of the Pleasure of mutual Sympathy."[13] It refers to the fellow-feeling which Smith saw as the critical common feature of human sociability that governs individual conduct:

> But whatever may be the cause of sympathy, or however it may be excited, nothing pleases us more than to observe in other men a fellow-feeling with all the emotions of our own breast; nor are we ever so much shocked as by the appearance of the contrary. Those who are fond of deducing all our sentiments from certain refinements of self-love, think themselves at no loss to account, according to their own principles, both for this pleasure and this pain.[14]

[12] Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (Indianapolis, IN: Liberty Fund, [1776] 1981), Vols. 1–2.
[13] Adam Smith, *TMS*, 13.
[14] Ibid., 13.

> As the person who is principally interested in any event is pleased with our sympathy, and hurt by the want of it, so we, too, seem to be pleased when we are able to sympathize with him, and to be hurt when we are unable.[15]

Since the modern reader may think that this framework surely must only be about intimate friends, we hasten to add Smith's dictum that such sentiments, ". . . when expressed in the countenance or behavior, even towards those not peculiarly connected with ourselves, please the indifferent spectator upon almost every occasion."[16] Provisionally, therefore, *Sentiments* should be viewed as articulating a theory for all social occasions.

Smith's system was not about outcomes, nor about equilibrium in outcomes, nor especially about "behavior" in its ordinary usage in the standard social science model. Samuel Johnson's dictionary of the English language in 1755 did not even include the modern social scientific meaning of behavior, which is definition 5 in the *Oxford English Dictionary* (*OED*): "the manner in which a thing acts under specified conditions or circumstances, or in relation to other things."[17] For Smith, behavior is about rules all the way down (see Samuel Johnson's definitions at the head of the essay). A person interfacing with others either acts within the rules, for which he is deemed to be "good" or at least "not bad," or he acts outside the rules, for which he is deemed to be "bad." Thus, in the course of life a person as a general practice either conducts himself well in a morally upstanding way, or regularly or on occasion falls short.

In contrast, modern positivistic economic interpretations of observations in the laboratory abstain from attributing moral judgments as a mainspring to human action, treating people as things acting in relation to other things under specified conditions, like molecules in an Erlenmeyer flask. Translated into modern language, "moral judgment" means assessing the fitness between actions conditional on context and the norms/conventions that have emerged by consent. The appropriate equilibrium concept is in rule space. The individual judges the propriety of his or her own and other action given the context (including all outcome values in the game). But over time the normative rules of propriety change by group consent.[18] Smith clearly saw the latter as efficient, but hastened to make plain that their efficiency is not what explains why people follow them.

[15] Ibid., 15.

[16] Ibid., 38.

[17] The *OED* only traces this definition back to 1674, whereas definition 1, the manner of conducting oneself in the external relations of life, goes back to 1490. Examples for definition 2, which is word for word the same as Samuel Johnson's in 1755, are dated 1521 (perhaps) and 1535.

[18] The rules of propriety governing pre-civil order in small groups evolved into the rules of property in the civil order of law — a topic beyond our reach in this paper. See Adam Smith, *TMS*, 82–85 and see Vernon L. Smith, "Adam Smith: From Propriety and Sentiments to Property and Wealth," *Forum for Social Economics* 43, no. 1 (2013): 1–15.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

Smith's model applied to the trust games of the 1980s and 1990s yields a much different resolution of why Max-U would fail predictively and why it was inappropriate to replace self-loving preferences with social preferences, and hence self-regarding behavior with other-regarding behavior. For Smith, "self-regarding behavior" is an oxymoron and "other-regarding behavior" a pleonasm. When he says that "[m]en of virtue only can feel that entire confidence in the conduct and behaviour of one another, which can, at all times, assure them that they can never either offend or be offended by one another," the "behavior," with no modifier, that he references already regards others because those regards are embedded in the rules that govern human intercourse.[19] As Charles John Smith in his *Synonyms Discriminated* explains, "BEHAVIOUR . . . refers to all those actions which are open to the observation of others as well as those which are specifically directed to others. As BEHAVIOUR refers more especially to actions, so DEMEANOUR . . . refers more directly to manners; or in other words, DEMEANOUR regards one's self, BEHAVIOUR regards others."[20]

By using both "conduct" and "behavior," the meticulous Adam Smith intends to place the confidence of men of virtue in two distinct concepts. Chiefly though, his project throughout *Sentiments* is about conduct.[21,22] An isolated individual *j* abstracted from society is but a counterfactual thought experiment to impress upon the reader the central role of sociability in (moral; that is, rule-following) human action:[23]

> Were it possible that a human creature could grow up to manhood in some solitary place, without any communication with his own species, he could no more think of his own character, of the propriety or demerit of his own sentiments and conduct, of the beauty or deformity of his own mind, than of the beauty or deformity of his own face . . . Bring him into society, and he is immediately provided with the mirror he

[19] Adam Smith, *TMS*, 225.

[20] Charles John Smith, *Synonyms Discriminated*, 159.

[21] Adam Smith uses "Conduct" three hundred nine times in the 338-page *TMS*, twice in a chapter title and once in the title of the very important Part III ("Of the Foundation of our Judgments concerning our own Sentiments and Conduct, and of the Sense of Duty"). "Behavior," on the other hand, never appears in a title and is used only eighty times. Moreover, the conjunction *X* and "Behavior" is used seventeen times where *X* is Conduct, Character, Sentiments, or Countenance; again, he refers five times to Whole Behavior. The substance of Smith's thought process — one to which we are not accustomed — is revealed in his careful diction.

[22] Charles John Smith discriminates the synonyms for us: "As BEHAVIOUR belongs to the minor morals of society, so CONDUCT to the graver questions of personal life . . . We speak of a man's behaviour in the social circle, of his conduct in his family, as a citizen, or in life. Good conduct is meritorious and virtuous. Good behaviour may be natural or artificial. The conduct has relation to the station of men's lives, or the circumstances in which they are placed. Good conduct will include right behaviour as part of it, and a proper demeanour will flow necessarily out of it" (Charles John Smith, *Synonyms Discriminated*, 159).

[23] Adam Smith, *TMS*, 110.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

> wanted before. It is placed in the countenance and behavior of those he lives with, which always mark when they enter into, and when they disapprove of his sentiments;[24] and it is here that he first views the propriety and impropriety of his passions, the beauty and deformity of his own mind.

In ordinary human intercourse, we *feel* when we *experience* the mirror of life, the *sentiments* of which then lead us to *conduct* ourselves accordingly. Moreover, in the practice of virtues we direct our conduct in the circumstances in which we find ourselves "by a certain idea of propriety, by a certain taste for a particular tenor of conduct, [rather] than by any regard to a precise maxim or rule."[24,25]

III. Sense Perceptions and the Non-Specifiability of Human Conduct

What then is the disconnection between a utilitarian model of behavior (in the modern social science sense) and human conduct? The difference lies in the modern hypothesis that action maps into outcome and thence preference; rather, in *Sentiments* the mapping is from the social situation, *including outcome as one of its components*, into action. Experimentalists discovered this when Max-U failed in trust games; they explored why, and found that intentions mattered. Such considerations, however, were already at the core of *Sentiments*, which articulates principles and propositions for unpacking the meaning conveyed in the social context. In order to read the meaning of actions in their social context, individuals require common knowledge that more of a good (or less of a bad) is better for each, for then each knows which action is beneficial or hurtful to whom. In the absence of this common knowledge of individual non-satiation ("egoism") *no one can judge and select socially agreeable actions*. Consequently, when I smoke, I might respect my wife's unwillingness to "go along" by leaving the house to light up. My self-command action already reflects its rule-governed social context, as well as our knowledge of each other's preferences.

That my action can be "explained" by an appended second preference function defined on my experience of "smoking" and my having learned

[24] Ibid., 175.

[25] The exception is the practice of the virtue of justice. This nontrivial distinction between the rules of justice and the rules of all other virtues separates Adam Smith from Bicchieri, who treats the rules of all virtues as rules of grammar in Cristina Bicchieri, *The Grammar of Society* (New York: Cambridge University Press, 2006). For Adam Smith, "the rules of justice may be compared to the rules for grammar; the rules of the other virtues [however], to the rules which critics lay down for the attainment of what is sublime and elegant in composition. The one, are precise, accurate, and indispensable. The other, are loose, vague and indeterminate, and present us rather with a general idea of the perfection we ought to aim at, than afford us any certain and infallible directions for acquiring it" (Smith, *TMS*, 175–76).

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

that it is in my social interest to honor my wife's concerns, incoherently fails to provide insight into the principles that govern the emergence of rule-following conduct. *Sentiments* provides such a system. If it had been part of the tradition in economics in the 1980s and 1990s the results from trust and other such games would have been anticipated, as well as the subsequent demonstrations that intentions matter. As we shall see below, adding or subtracting nodes to a trust game changes actions because it changes the meaning people read into actions, but individual preferences — more reward is better, less is worse — do not change. Such prescience, in Smith, also predating modern psychology, deserves a sympathetic retrospective hearing.

In most social situations we do not know others as well as I know my wife, but that does not mean I am oblivious to the sensibility of others. Hayek explains that our ability to perceive a pattern in human conduct does not necessarily imply an ability to completely specify it.[26] Our sensory perceptions of patterns fall into three categories: (1) those that we can sense and can explicitly describe, (2) those that we cannot sense but can explicitly specify, and (3) those that we can sense but not explicitly specify. For example, in the first category we can sense a pentagon, like this one ⌂, and discursively describe it as such. The description fully describes the perception of a shape. But in the case of the second category, we cannot sense the 6-D pattern of the bee waggle dance, though the mathematician Barbara Shipman can completely specify it as a flag manifold projected onto a perceivable 2-D plane.[27]

Human conduct largely falls into the third category. We can obviously sense a pattern in conduct, but it is a non-specifiable pattern subject to error. We can recognize the actions and associated motivations of someone as being just, fair, or equitable, or beneficent, kind, or humane, but we cannot specify all of the perceptual elements that we treat as part of the same rule pattern (the sense of just as JUST but not FAIR, the sense of beneficent as BENEFICENT but not HUMANE). Our perception of conduct contains shades and subtleties of ethics and aesthetics that cannot be precisely specified by a set of s_i's and concomitant $U_i(s_1, \ldots, s_n)$'s for $i = 1, \ldots, n$, which is why Frank Knight says "[i]t is not enlightening to be told that conduct consists in choosing between possible alternatives."[28] Human conduct is not explicitly specifiable like a perceivable pentagon or a bee waggle dance in unperceivable six dimensions. It is an entirely different kind of sensible pattern. As experimentalists it is important to avoid the error of seeing patterns of human conduct in subjects' actions that use

[26] Friedrich A. Hayek, "Rules, Perception and Intelligibility," *Proceedings of the British Academy* 48 (1963): 321–44.

[27] Adam Frank, "Quantum Honeybees," *Discover Magazine*, 1997, http://discovermagazine.com/1997/nov/quantumhoneybees1263#.UXA6f0raix0.

[28] Frank H. Knight, "Ethics and the Economics Interpretation," *Quarterly Journal of Economics* 36, no. 3 (1922): 454–81.

our own perceptions of a one-to-one correspondence between action and preference as elements of a scientific explanation.[29]

Denying that our own perceptions of conduct can be explicitly known as legitimate elements of scientific explanation, however, does not entail denying that our subjects are perceiving non-specifiable patterns in each other's conduct and acting upon them. That our subjects do so must form a datum for analysis, and moreover, it is the foundation of *Sentiments*. Adam Smith is concerned with understanding conduct, the fair-play rules governing that conduct, and the trial-and-error processes through which those rules might have emerged. Thus,

- *On the propriety of action:*
  Man has a "love of praise and of praise-worthiness" and a "dread of blame and blame-worthiness", and "[t]he love of praise-worthiness is by no means derived altogether from the love of praise . . . though they resemble one another . . . [and] . . . are connected . . . , [they] are yet, in many respects, distinct and independent of one another."[30]
- *On conduct and self-command:*[31]
  If he would act so as that the impartial spectator may enter into the principles of his conduct . . . he must . . . upon all . . . occasions, humble the arrogance of his self-love, and bring it down to something which other men can *go along with* . . . In the race for wealth, and honours, and preferments, he may run as hard as he can, and strain every nerve and every muscle, in order to outstrip all his competitors. But if he should justle, or throw down any of them, the indulgence of the spectators is entirely at an end. It is a violation of fair play, which they cannot admit of. This man is to them, in every respect, as good as he: they do not enter into that self-love by which he prefers himself so much to this other, and cannot go along with the motive from which he hurt him.
- *On process:*
  " . . . to attain this satisfaction, we must become the impartial spectators of our own character and conduct."[32]
  "We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it. If, upon placing ourselves in his situation, we thoroughly enter into all the passions and motives which influenced it, we approve of it,

[29] See also Frank H. Knight, "The Limitations of Scientific Method in Economics," Rexford G. Tugwell, ed., *The Trends of Economics* (New York: Alfred A. Knopf, 1924), 229–68, who noted this problem with behavioral representations in economics over ninety years ago.
[30] Adam Smith, *TMS*, 113–14.
[31] Ibid., 83, italics added.
[32] Ibid., 114.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

by sympathy with the approbation of this supposed equitable judge. If otherwise, we enter into his disapprobation, and condemn it."[33]

- *On rules being derived from experience, not reason:*[34]
  Our continual observations upon the conduct of others, insensibly lead us to form to ourselves certain general rules concerning what is fit and proper either to be done or to be avoided . . . . They are ultimately founded upon experience of what, in particular instances, our moral faculties, our natural sense of merit and propriety, approve, or disapprove of. We do not originally approve or condemn particular actions; because, upon examination, they appear to be agreeable or inconsistent with a certain general rule. The general rule, on the contrary, is formed, by finding from experience, that all actions of a certain kind, or circumstanced in a certain manner, are approved or disapproved of.[35]

Consequently, actions signal conduct or responses to the conduct inferred from the actions of others; and the general rules governing conduct become fixed through the discipline of their propriety; what others will "go along with" (by agreement from circumstanced experience) shapes the rule and determines its fitness. Only within the self-governing discipline of these general rules, is there scope for the individual to seek self-loving personal gain.

IV. A Formal Representation of Smith's Model of Action[36]

Think of an action, a_i, by individual i as depending on i's judgment of its propriety, given the action's contextual circumstances:

$$a_i(\text{Propriety}|C) = \alpha_i(C) \cdot PR + \beta_i(C) \cdot PR \cdot PW + \gamma_i(C) \cdot PW + \delta_i(C), \qquad (1)$$

where PR and PW are (0, 1) indicators, respectively, that action is praised by others (1), or not (0), and is praiseworthy (1), or not (0); and α_i, β_i, γ_i and δ_i are nonnegative functions weighting PR and PW in determining

[33] Ibid., 110.
[34] Ibid., 159.
[35] Hence general rules are not a product of reason, or rational construction; they are formed "insensibly" out of experience and, if efficient, are ecologically rational, as noted in Vernon L. Smith, *Rationality in Economics: Constructivist and Ecological Forms* (Cambridge: Cambridge University Press, 2008), 269–70.
[36] We develop here the implications of our brief footnote 5 in Vernon L. Smith and Bart J. Wilson, "Fair and Impartial Spectators in Experimental Economic Behavior," *Review of Behavioral Economics* 1, no. 1 (2014): 1–26.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

its propriety.[37] In the second term, *PW* leverages *PR*, while the third term expresses the sentiment that *PW* may yield stand-alone value (self-command), "distinct" from *PR*, even where no praise is possible. For example, in double anonymity experiments weight is still given to *PW* even where no one can know your decision.

Let $C(m_1,m_2)$ be the circumstances including game structure, choice alternatives and the vector of all nodal payoffs (m_1,m_2) allowing the signal to be deciphered. An action is based on conduct that is more or less appropriate conditional on circumstances, where the action chosen best satisfies or "fits" a socially mediated criterion. The additive function $\delta_i(C)$, independent of the social indicators, allows unmitigated "self-love" to be expressed — in *Sentiments* this motive always has an important stoic role. Where *i* cannot infer the intent of other and reward beneficence, then $\alpha_i(C) = \beta_i(C) = 0$, and $\delta_i(C)$ looms larger than otherwise in determining the action chosen, but this is still mediated by a positive weight for *PW*.

An expression similar to (1) applies to a hurtful action subject to blame/blameworthiness.

Equation (1) defines a rule for *i*; namely, *i*'s choice of an action given $C(m_1,m_2)$ and the self-command weights (judgments) *i* places on *PR* and *PW*. Simultaneously, there exists in *i*'s social world a convention ("customs") defining what "people will go along with" for the choice of α_i, subject to the same conditionals. If *i*'s choice is out of order, he or she will receive corrective feedback ("disapprobation"). Hence, *Sentiments* is primarily about the adaptation of individuals to what is "fit and proper." The model is open ("vague") concerning the inertial processes whereby the demands of social conventions emerge and change through time, but there is an implication of evolutionary change and adaptation.

V. Propositions on Beneficence and Justice

Within this framework, Smith states three relational propositions on beneficence and justice. While the indispensable virtue of justice may be

[37] We assume it is stoically obvious, as it is in *TMS,* that a praisable action by me is one that raises your payoff (more is better); a blamable action reduces your payoff, but in the model, context must make it transparent that the action was extortion-free, intentionally motivated, not just accidental, and so on. Self-respect implies that I also see it as praiseworthy even when this cannot be known by others. Each of these elements (weights) is subject to error, to misreading, and to disagreement. Notice that there may be contexts in which explicit incentives (tipping, bonuses, and wage differentials) are inconsistent with the praise/praiseworthiness of an action, leading to a "crowding out" of their effort-enhancing purpose. See Ernst Fehr and Simon Gachter, "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* 14, no. 3 (2000): 159–81. In *TMS,* however, this is due not to outcome fairness (social preference), but to a foul — rule (norm) violation. At three universities we have had to justify paying experimental subjects their participation "earnings" because members of the Institutional Review Board thought such payments were "coercive" (or in violation of psychology practice). Remarkably, however, it is often thought that our strict ban on the use of subject deception is an unreasonable restraint on science!

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

a familiar concept to the reader, "beneficence" has an archaic eighteenth century ring to it.[38] Beneficence is, literally from Latin, *well doing*, and according to the *OED* only entered the lexicon in the sixteenth century. Its older Latin relation, "benevolence," is used by Chaucer (that would be in 1384). Benevolence is, literally, *well willing*. Thus, benevolence consists of the intention to do good for another, beneficence the action that does good for another. A niggardly, selfish, or mischievous man, as determined by the circumstances, cannot be beneficent even if what he does is good. Thus, beneficence always presupposes benevolence; specific beneficent actions signal intent, but both sender and receiver of the signal require the whole context, $C(m_1,m_2)$, (including egoistic non-satiation) to interpret the action. Charles John Smith distinguishes "benevolence" from "benignity" (another eighteenth-century-sounding word), humanity, and kindness:[39]

> Benignity is, as it were, dormant, or passive benevolence. It is a matter more of temperament than will . . . As benevolence is inherent, so benignity may be shown on special occasions only…Humanity expresses an impulse rather than a quality . . . [and] is not so much a virtue when exhibited as something the absence of which is positively disgraceful and evil . . . Kindness is very like benevolence, but is rather a social than a moral virtue. It applies to minor acts of courtesy and good will, for which benevolence would be too serious a term.

The conceptual distinction between kind intentions as applicable to minor acts and benevolence as applicable to more serious, and hence beneficent, acts reinforces Adam Smith's claim that the general rules of conduct are "loose, vague and indeterminate, and present us rather with a general idea of the perfection we ought to aim at, than afford us any certain and infallible directions for acquiring it."[40,41] In other words, as a non-specifiable pattern, there is room for disagreement within general rules of conduct for interpreting an act in context as connoting minor or major intentions of doing good.[42,43,44] "Disagreement" implies disequilibrium

[38] "Society . . . cannot subsist among those who are at all times ready to hurt and injure one another. The moment that injury begins, the moment that mutual resentment and animosity take place, all the bands of it are broke asunder . . . Beneficence, therefore, is less essential to the existence of society than justice. Society may subsist, though not in the most comfortable state, without beneficence; but the prevalence of injustice must utterly destroy it" (Adam Smith, *TMS*, 86).

[39] Charles John Smith, *Synonyms Discriminated*, 165–66.

[40] Adam Smith, *TMS*, 175–76.

[41] Cf. Hayek, "Rules, Perception and Intelligibility."

[42] Bart J. Wilson, "Social Preferences Aren't Preferences," *Journal of Economic Behavior and Organization* 73, no. 1 (2010): 77–82.

[43] Bart J. Wilson, "Contra Private Fairness," *American Journal of Economics and Sociology* 71, no. 2 (2012): 407–435.

[44] Where in modern game theory is the assumption of agreement on the interpretation of the act? Hidden obscurely in the assumption that every individual j always chooses the largest possible pot of utilitarian pleasure $U_j(\cdot)$.

and impels change; either individuals must adapt the rules they follow, or the normative rules of propriety must adapt. The latter is a long-run adaptation: if individuals continue to resist compliance with the norms, then the norms adapt to the popular will.

With this background, here are three (of many) "relational propositions" from *Sentiments* (Part II, Section ii, Chapter I):[45]

- *Beneficence Proposition 1*: *Properly motivated beneficent actions alone require reward.*
  Why? " . . . because such alone are the approved objects of gratitude, or excite the sympathetic gratitude of the spectator."[46]
- *Beneficence Proposition 2*: *The want of beneficence cannot provoke resentment and punishment.*
  Why not? "Beneficence is always free, it cannot be extorted by force, the mere want of it exposes to no punishment; because the mere want of beneficence tends to do no real positive evil."[47,48]
- *Injustice Proposition*: *Improperly motivated hurtful actions alone deserve punishment.*
  Why? " . . . because such alone are the approved objects of resentment, or excite the sympathetic resentment of the spectator."[49]

These are strong falsifiable propositions applying to choice behavior [*sic*, conduct; hereafter we will use the language of *Sentiments*] in a wide range of games in which actions are voluntary, extortion-free, and convey intentions likely to be unambiguously interpreted as "beneficial" or "hurtful."[50] Let's see how well *Sentiments* predicts conduct in trust games, over two hundred fifty years after it was first offered.

[45] We name and number them as propositions; *TMS* does not.
[46] Adam Smith, *TMS*, 78.
[47] Ibid.
[48] Smith and Wilson use this proposition to interpret the standard ultimatum game context as projecting a form of involuntary extortion: the first player's choice is subject to veto by the second, the players' roles having been determined at random (Smith and Wilson, "Fair and Impartial Spectators in Experimental Economic Behavior"). Under this interpretation, the proposition denies that ultimatum offers can be described as involving "beneficence," or that the responses involve "gratitude" or "punishment" independent of their perception as extortionist. It also suggests that the ultimatum game outcomes will be sensitive to changes in the context wherein procedures or narratives rationalize a process whereby subjects have reached the ultimatum stage game.
[49] Adam Smith, *TMS*, 78.
[50] The conditionals in these propositions are all subject to error: for example, if the game context allows me to punish your defection on a trusting action by me, you may not find my trusting action credible (it is coercive), whereas my action was not dependent on my having the option to punish you (I would have trusted you in any case).

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

VI. Trust Game: Designs for the Reward and Punishment of Human Actions

Our starting point, displayed in Figure 1, is a slightly modified two-person trust game of McCabe and Smith;[51] the latter has been replicated by Cox and Deck[52] and Gillies and Rigdon,[53] and we want to test for significant deviation in our new experiments. The first modification is that the payoffs are increased by 20 percent. If, as the first-mover, Person 1 ends the game by playing right, each person receives $12 instead of the original $10. [The payoff for Person 1 (Person 2) is displayed on top (bottom) in Figure 1.] Similarly, if Person 1 plays down and Person 2 plays right, Person 1 receives $18 and Person 2 receives $30 instead of the original $15 and 25, respectively. The second modification is that if Person 2 plays down, Person 1 receives, instead of nothing, the modest non-zero amount of $6, and Person 2 receives $42, as opposed to $40 in the original game. The non-zero amount for Person 1 is necessary to implement the test of two of our propositions in our two other initial treatments (with all payoffs positive). This game, like that in McCabe and Smith, seems quite hazardous for a Person 1 who plays down.

Previous results combined from the three studies listed above indicate that 46 out of 98 first movers pass the play to the second mover, and consistent with *Beneficence Proposition 1* (back predicts), 31 second movers reward the beneficent actions of the trusting first-mover by playing right. In what we will call the *No Punish Game* (hereafter *NP* Game) we will test whether our slight modification of Person 2's choices affects the conduct of our Person 2's. Notice that 15 of the 46 second movers (33 percent) do not agree with their corresponding 15 first movers (who did not judge this situation as one in which the second mover recognizes and rewards the first mover for wishing well and doing well). There is something about how this action is *re*-presented to the minds of these second-movers that they do not see beneficence in the first mover playing down, or are simply willing to exploit it self-interestedly under anonymity.[54] Perhaps they perceive mere kindness or not even that. Only further experimental designs could flesh out the motivations of these second-movers and how they differ from the other 31.

[51] Kevin McCabe and Vernon L. Smith, "A Comparison of Naïve and Sophisticated Subject Behavior with Game Theoretic Predictions," *Proceedings of the National Academy of Arts and Sciences* 97 (2000): 37770–81.

[52] James C. Cox and Cary A. Deck, "On the Nature of Reciprocal Motives," *Economic Inquiry* 43 (2005): 623–35.

[53] Anthony S. Gillies and Mary L. Rigdon, "Epistemic Conditions and Social Preferences in Trust Games," (Working paper, University of Michigan, 2008).

[54] See also Smith and Wilson, "Fair and Impartial Spectators in Experimental Economic Behavior."

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

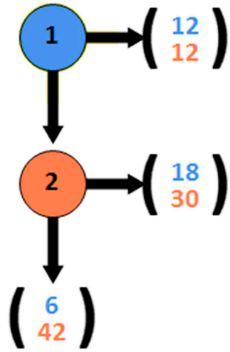


FIGURE 1. No Punish (NP) Game

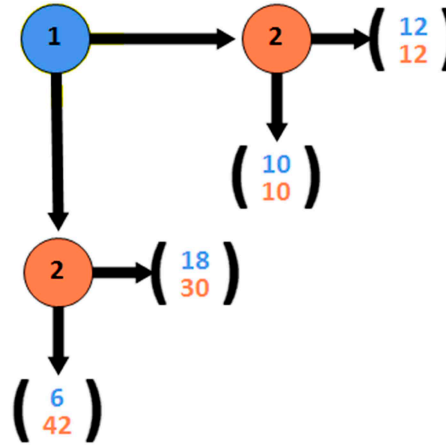


FIGURE 2. Punish Want of Beneficence (PWB) Game

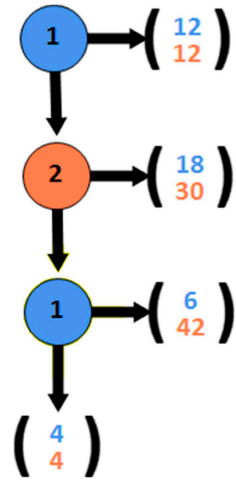


FIGURE 3. Punish Hurt (PH) Game

Before extending this basic extensive form game, we note that this problem is not nearly as simple as traditional game theory presumes it to be for its human subjects.[55] The subjects have only their past experiences in life to cleave to, and the bare structure of an unfamiliar, single play, extensive form game tree.

[55] Extensive form trust games yield results quite different than their strategic or normal game-theoretic form representations; in particular people are less trusting and less trustworthy in the latter, and we study only the former noted in Kevin McCabe, Vernon L. Smith, and Michael LePore, "Intentionality Detection and 'Mindreading': Why Does Game Form Matter,"

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

To test *Beneficence Proposition 2*, we construct the extensive form game in Figure 2, which we will call the *Punish Want of Beneficence Game* (hereafter, *PWB Game*). The difference between this game and the *NP Game* is that if Person 1 plays right, Person 2 can either punish the want of beneficence of Person 1 by playing down, yielding $10 to each person, or Person 2 can play right yielding $12 to each person. Playing down by Person 1 is an unambiguously beneficent action (at least we hypothesize so) towards Person 2 as Person 2's payoff increases by 250 percent or 350 percent.

The *Punish Hurt Game* (hereafter, *PH Game*) in Figure 3 is designed to test the *Injustice Proposition*. Compared to the *NP Game* in Figure 1, if Person 2 plays down, Person 1 can either accept the ($6, $42) outcome as in Figure 1, or choose to punish hurt for not receiving $18. Punishing the hurt of Person 2 by playing down comes at the cost of $2 for Person 1, but it also reduces the payoff of Person 2 from $42 to $4.

Notice that the personal cost of punishment by Person 2 in the *PWB Game* and by Person 1 in the *PH Game* is $2 in both. *Sentiments* predicts that (a) Person 2 will not punish want of beneficence by playing down if Person 1 plays right in Figure 2, but that (b) Person 1 will punish hurt by playing down in response if Person 2 plays down in Figure 3.

A. Procedures

We originally recruited 150 students with a variety of majors from the undergraduate population at Chapman University, a private university with approximately 5,000 undergraduates. Each of 5 sessions conducted over 3 weeks consisted of 30 students equally and randomly assigned to each of the three games described above. No student had any prior experience in an extensive or normal form game though many had experience in a prior experiment interacting with more than one other person. To explicate these results, we recruited another 146 students from the same population and assigned them (nearly) equally and randomly to each of three games, one of which was the *NP Game* in Figure 1. The two new additional games in the second series in this study will be presented and discussed below. But first we will report the results for 49 total pairs in the *NP Game*, 25 total pairs in the *PWB Game*, and 25 total pairs in the *PH Game*.

Each subject was paid $7 for showing up on time and was seated in one of two adjacent computer laboratories. One laboratory contained 14 people

Proceedings of the National Academy of Sciences 97, no. 8 (2000): 4404–4409. You get choice data at every decision node from the strategic form given 2N subjects, but — once having established these findings — from the perspective of *TMS* the strategic form represents a context that increases error in the interpretation of actions as signals that convey the meaning in outcomes. We choose the extensive form to reduce that noise.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

seated in 14 carrels and the other 16 people were seated in the front portion of a 24-carrel room. The roles of Person 1 and Person 2 in all three games were distributed nearly proportionately between the two rooms. The subjects read at their own pace the interactive computerized instructions contained in the Appendix. After playing one and only one of the extensive form games, the subjects were privately paid their earnings, which averaged $18.11, excluding the show up payment. The experiment lasted well under the 60 minutes for which they were recruited.

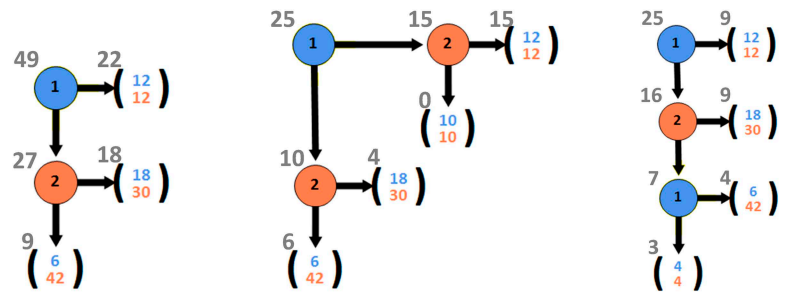
B. *Results*

Figure 4 reports the number of decisions at each node and the number of outcomes reached in each of the three games. Our first finding establishes a baseline by comparing our *No Punish Game* to the previous trust game.

Finding 1: *First movers in the* NP Game *beneficently play down, and second movers reward that action at frequencies statistically consistent with those observed in the previous trust games.*

Of the 49 Person 1's in our sample, 27 (55 percent) move down. Previous studies combined have found that 46 out of 98 first movers (47 percent) trust their anonymous second mover. Using a two-sided two-proportion z-test, we fail to reject the null hypothesis of equal proportions ($z = 0.93$, p-value = 0.3507). Of the 27 Person 2's who have the opportunity to move, 18 (67 percent) play right. Previous studies have found that 31 out of 46 (67 percent) second movers honor the trust of the first mover. Again, we fail to reject the null hypothesis of equal proportions with a two-sided test ($z = 0.06$, p-value = 0.9493).

Having replicated the previous rate at which Person 2's reward the beneficent actions of Person 1, our next finding reports whether Adam Smith's *Beneficence Proposition 2* holds in the *PWB Game*.



(a) *No Punish (NP) Game* **(b)** *Punish Want of Beneficence (PWB) Game* **(c)** *Punish Hurt (PH) Game*

FIGURE 4. Number of Decisions by Node and Outcome

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

Finding 2: *Fifteen out of 15* (100 percent) *Person 2's do not punish Person 1 for failing to beneficently play down in the* PWB Game.

Empirically, comparing Figure 4 (b) with Figure 4 (a), the "simple" addition of a decision node for Person 2 when Person 1 plays right changes the frequency at which Person 2's play right to reward Person 1 when the latter beneficently plays down. (Consistent with *Sentiments*, circumstances, $C(m_1,m_2)$, matter whereas a social preference function explaining cooperation in the first game fails to predict action in the second.) Instead of 33 percent of the Person 2's playing down after Person 1 plays down, 6 out of 10 (60 percent) fail to reward the beneficent action of Person 1. Pooling the second mover decisions across the four studies with the basic trust game, this difference is marginally significant with a two-sided test ($z = 1.67$, p-value = 0.0941, $n_1 = 73$, $n_2 = 10$). Interestingly, Person 1's appear to anticipate that Person 2's may be untrustworthy by only playing down in 10 out of 25 pairs (40 percent). Even though Adam Smith successfully predicts that Person 2 will not punish the want of beneficence 100 percent of the time when Person 1 fails to act beneficently, it appears that giving Person 2 the option to punish want of beneficence has the unanticipated consequence, in the counterfactual treatment, of changing the response to Person 1 beneficently playing down. We will flesh out this observation in the next section with a new game designed to explicate this result.

Our next finding reports a test of the *Injustice Proposition*:

Finding 3: *Seven Person 1's in the* PH Game *faced the decision of whether or not to punish the perfidy of Person 2 and 3* (43 percent) *do.*

Comparing the first-move decisions of Person 1, in the *PWB Game* (Figure 4b) with those in the *PH Game* (Figure 4c) the option to punish conditional on the hurt of Person 2 increases the frequency at which Person 1 plays down (10/25 = 40% in the *PWB Game* versus 16/25 = 64% in the *PH Game*, $z = 1.70$, p-value = 0.0894).

Interestingly, the frequency at which Person 2's follow *Beneficence Proposition 1* in the *PH Game* falls slightly relative to our baseline *NP Game* (9/16 vs. 18/27). Similar declines in the cooperative response when Person 1's have the option to punish defection have been noted often.[56] The interpretation was that "trust" as a signal of intentions by Person 1's loses credibility in the presence of the punishment option. Person 2's play right in the *PH Game* much less than in the baseline *NP Game*. In effect, the offer no longer is an unambiguous signal of trust and hence this circumstance undermines trustworthiness. All these follow as applications of Smith's general and deeply engrained principle that *beneficent actions must be properly motivated, be freely given and cannot be extorted by force.*

At this point it could be said that these findings can be rationalized with nearly every model of social preferences. But every one of those

[56] See, for example, Vernon L. Smith, *Rationality in Economics: Constructivist and Ecological Forms* (Cambridge: Cambridge University Press, 2008), 269–70.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

social preference models was itself an *ex post* rationalization of observing outcomes in the laboratory that had falsified Max-U. This contrasts with the prior prediction of Smith's *Beneficence Proposition 1 and 2*. Because the motivations of the subjects cannot be *directly* observed (the pattern of actions is recognizable but not explicitly specifiable), reasoning by *ex post* models of social preferences is inescapably circular:[57]

A: What is the *ex post* inferred inequality-averse outcome when Person 1 moves at the third node of the *PH Game*?
B: Person 1 chooses ($4, $4) over ($6, $42).
A: Why did Person 1 choose ($4, $4) over ($6, $42)?
B: Because he or she is postulated from the observations to be averse to inequality.

Adam Smith's theory of moral sentiments, however, is not circular. His *ex ante* theory *explains* the conduct of Person 1 by separately specifying the motivations of both Person 1 and Person 2:

A: What is the predicted outcome by which Person 1 punishes Person 2 at the third node of the *PH Game*?
B: Person 1 chooses ($4, $4) over ($6, $42).
A: Why did Person 1 choose ($4, $4) over ($6, $42)?
B: Because Person 2, with the improper motivation of an even higher payoff, took the hurtful action of playing down. Such an action by Person 2 is an approved object of resentment by Person 1.

Smith's model did not predict the proportion of subjects that would invoke the costly punishment option, but Max-U predicted that no one would invoke it.

C. *Two Additional Games*

The results of the *PWB Game* indicate that the opportunity to punish want of beneficence simultaneously (a) reduces the frequency of Person 1's beneficence to Person 2 and (b) reduces the frequency at which Person 2 rewards the properly motivated beneficent actions of Person 1. We did not predict this from our *ex ante* interpretation of Smith's propositions and will investigate it with new tests. Is this because, vis-à-vis the *NP Game*, Person 1 does not directly choose ($12, $12) — a perceived "loss of control?" — or because Person 2 has recourse to punishing the want of beneficence by choosing ($10, $10)? That is, the *PWB Game* introduces two alterations in the *NP Game*: (1) generically, regardless of the choice

[57] Bart J. Wilson, "Language Games of Reciprocity" *Journal of Economic Behavior and Organization* 68, no. 2 (2008): 365–77.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

by Person 1, Person 2 is the controlling person who determines the final outcome and (2) Person 2 can now punish the want of beneficence by Person 1. To test the effects of the first change only, we conducted what we will call the *No Punish Pass Game* (hereafter the *NPP Game*) in Figure 5.

The *NPP Game,* like the *NP Game*, provides no opportunity for Person 2 to punish the want of beneficence by choosing ($10, $10). The difference is that, if Person 1 fails to act beneficently toward Person 2 by playing down, Person 2 has the involuntary option of ending the game with payoffs ($12, $12). Like the *PWB Game*, Person 2 must take an action, albeit one inconsequential to the payoffs, if Person 1 fails to act beneficently. The question is, is the frequency of Person 1's who act beneficently *and* the frequency of Person 2's who reward that beneficence closer to what we observe in the *NP Game* or in the *PWB Game*?

We found in Figure 4 (b) and (c) that the set of subgames available in the decision tree affect choice; these results are congruent with *Sentiments*, although they do not change the traditional game-theoretic analysis. In Figure 6 we combine the opportunities to punish in each of these into a single game in which either hurt or want of beneficence can be punished: the *Punish Either* (hereafter, *PE*) *Game*. In this form there is no imbalance in punishment opportunities as occurs in Figure 4 (b) and 4 (c). Again, we recruited subjects to be randomly assigned within each session to the games in Figure 1, Figure 5, and Figure 6 in proportions nearly equal.

Figure 7 reports the number of decisions at each node and the number of outcomes reached in the *NPP Game* and the *PE Game*. We report first the results of the *NPP Game*.

Finding 4: *In the* NPP Game *Person 2's reward the beneficence of Person 1 at (a) a higher frequency than in the* PWB Game *and (b) the same frequency as in the* NP Game.

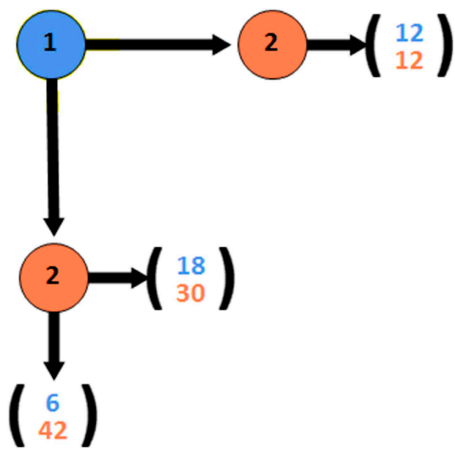


FIGURE 5. No Punish Pass (NPP) Game

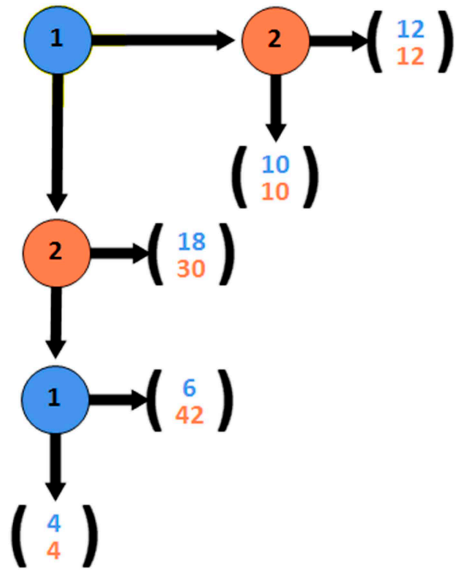


FIGURE 6. Punish Either (PE) Game

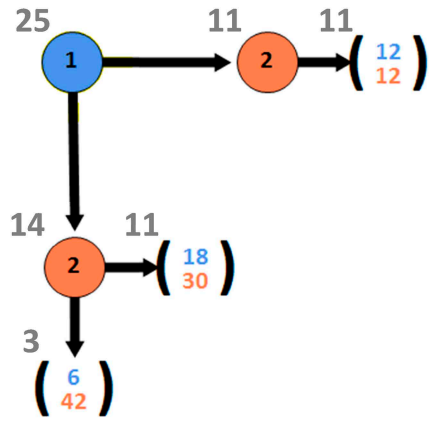
Fourteen out of 25 Person 2's in the *NPP Game* have the opportunity to reward the beneficence of Person 1, and 11 (79 percent) support Adam Smith's *Beneficence Proposition 1*. This observed frequency is nearly double that of what we observed in the *PWB Game*. Using a two-sided test, we reject the null hypothesis of equal proportions in the *NPP Game* and the *PWB Game* ($z = 1.92$, p-value = 0.0272). If anything, the involuntary option of Person 2 when Person 1 fails to act beneficently leads Person 2's to be *more* inclined to reward the beneficence of Person 1.

Thus, Person 1's in the *NPP Game* appear to correctly anticipate that Person 2's will properly reward their beneficence. Fourteen of 25 (56 percent) Person 1's act beneficently toward Person 2, which is closer to the observed frequency in the *NP Game* (55 percent) than the observed frequency of the *PWB Game* (40 percent). The results confirm that our *NPP Game* replicates the *NP Game*. Thus, *we conclude that it is the opportunity to punish the want of beneficence by Person 2 that leads Person 2's to be less likely to reward the realized beneficence of Person 1*.

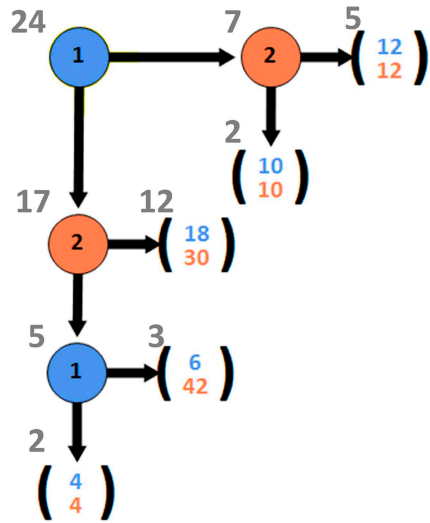
As a counterfactual experimental treatment, the *PWB Game* allows us to measure the effect of a hypothetical rule on observed conduct. Punishing the want of beneficence is a rule of "what is *not*."[58] It does not emerge as a community convention because, as explained by Adam Smith, the want of

[58] F. A. Hayek, *Law, Legislation, and Liberty, Rules and Order* Volume I (Chicago: University of Chicago Press, 1973), 17.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press



(a) ***No Punish Pass (NPP) Game***



(b) ***Punish Either (PE) Game***

FIGURE 7. Number of Decisions by Node and Outcome for the Two Additional Games

beneficence does no real positive harm.[59] Thus, if *Beneficence Proposition 2* holds, introducing the opportunity to punish the want of beneficence interferes with, or "distorts," reading the motives of one's counterpart,

[59] We can imagine all sorts of beneficence in our favor, but punishing the want of it invites resentment from those whose circumstances we do not fully know. They might not agree that their beneficence to us is appropriate given the context, or we might not know that they are incapable of being beneficent to us.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

particularly in a situation stripped of the normal contextual cues that we rely upon to make such assessments. Keep in mind that in *Sentiments* actions signal intentions and propriety, and are not just inert if-and-only-if determiners of outcomes. So perhaps we should not have been surprised at the sensitivity of conduct by both Person 1's and Person 2's in the *PWB Game*. (Except to parse the cause we needed the *NPP* game results). This observation reminds and humbles the hypothesizing social scientist that the human taste for a particular tenor of conduct is a rather sensitive and complicated palate; it also demonstrates how experimental tests can provide more specific content for the model in *Sentiments*.

Finally, the results of the *PE Game* in Figure 7(b) indicate that any differences induced by including (asymmetrically) the option to punish the want of beneficence are offset by also adding the opportunity to punish hurt. Seventeen of 24 (71 percent) Person 1's act beneficently, which is quite close to the 16 of 25 (64 percent) who do so in the *PH Game*. Twelve of 17 (71 percent) Person 2's reward the act of beneficence, which is in line with the 67 percent who do so in the *NP Game*. For the 5 Person 1's who were hurt by Person 2, two of them (40 percent) punish that hurt, which is consistent with the 43 percent that do so in the *PH Game*. Symmetry (*PE*) restores the original order (*NP*)! Lastly, we observe our first two (out of 7) Person 2's who punish a want of beneficence, compared with zero out of 15 in Finding 2.

VII. Conclusion

With the predictive failure of Max-U, as it applied to trust games in the 1980s and 1990s, experimentalists offered social preference and reciprocity explanations of the results. Neither of these adaptations positioned us to think outside the bounds of our accustomed, preference-and-exchange framework of thought. Max-U had served well-enough the observational demands of decision in games against nature and in market supply and demand experiments under perfect enforcement of property. The intermediate interactive world of personal social exchange required a plethora of new experiments designed to understand why the postulated mapping from action to outcome to utility preference was so sensitive to the particular context. Frank Knight's dictum, "the economist meets the problem of conduct and motive at every point and stage of his work," echoes *Sentiments* in which individual actions are signals of rule-governed conduct. Context matters because it gives meaning to outcomes.[60]

We show how propositions from *Sentiments* anticipated both the observations in trust games and the many subsequent explorations of the role of intentions. Moreover, we show that these propositions make new testable

[60] Frank H. Knight, "Economic Psychology and the Value Problem," *Quarterly Journal of Economics* 39, no. 3 (1925): 372–409.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

predictions. *Sentiments* is about the ethical rules that constitute the character of an inherently sociable person who strives for a better life, and it furthermore has predictive power where Max-U had failed. This great book is the foundation for lost insights into a quintessentially humanistic science of economics. "Life is not fundamentally a striving for ends, for satisfactions, but rather for bases for further striving."[61]

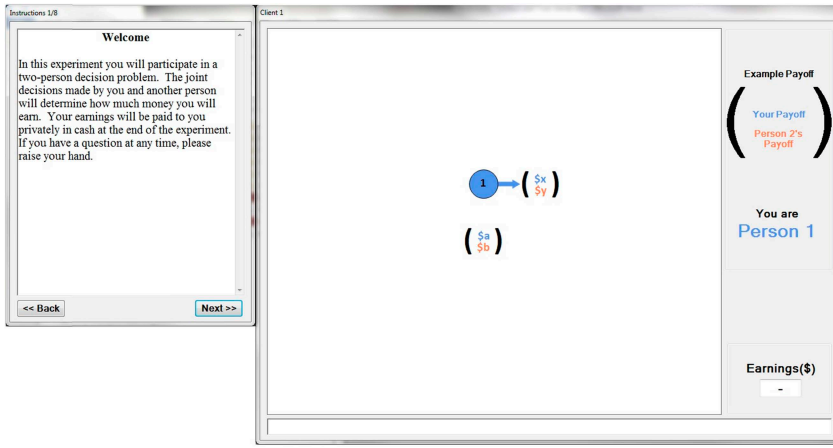
Economics and Law, Chapman University

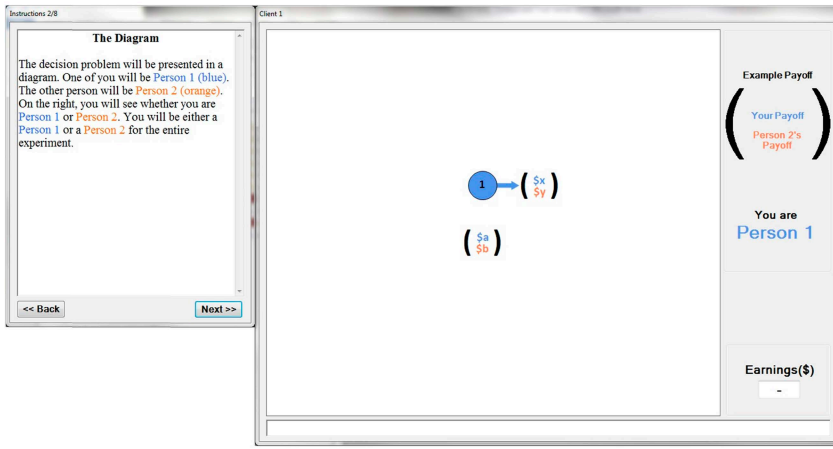
[61] Frank H. Knight, "Ethics and the Economics Interpretation," 459.

https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press

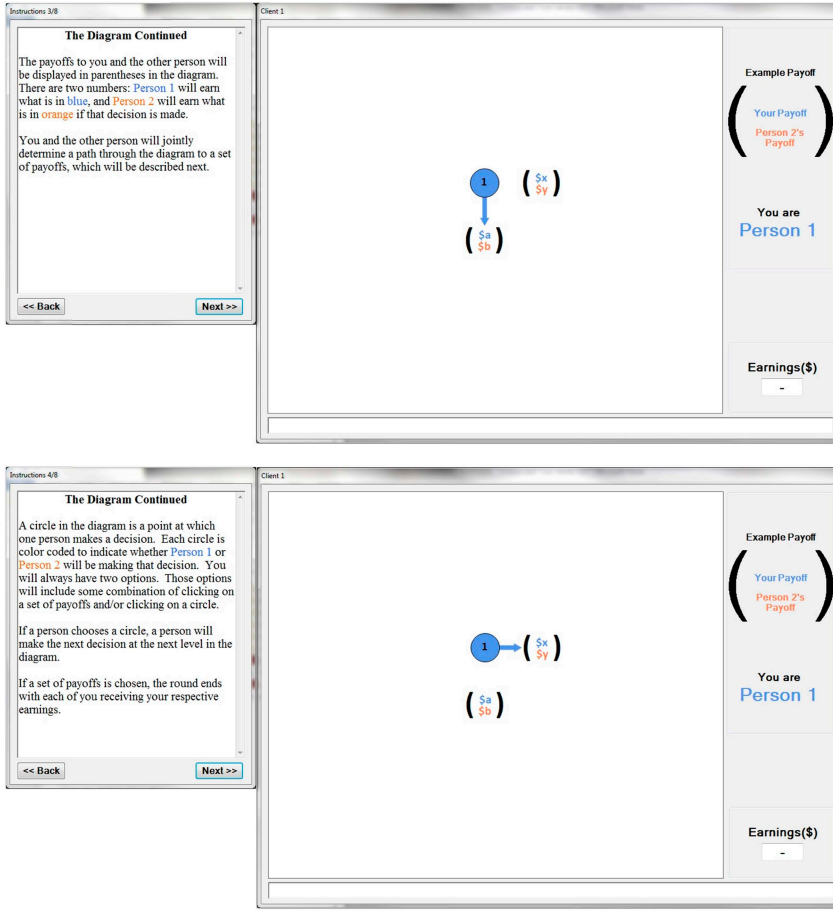
Appendix
Experiment Instructions

Please note that prior to making a decision, the colored arrows at the active decision node were flashing, alternating between the choices.



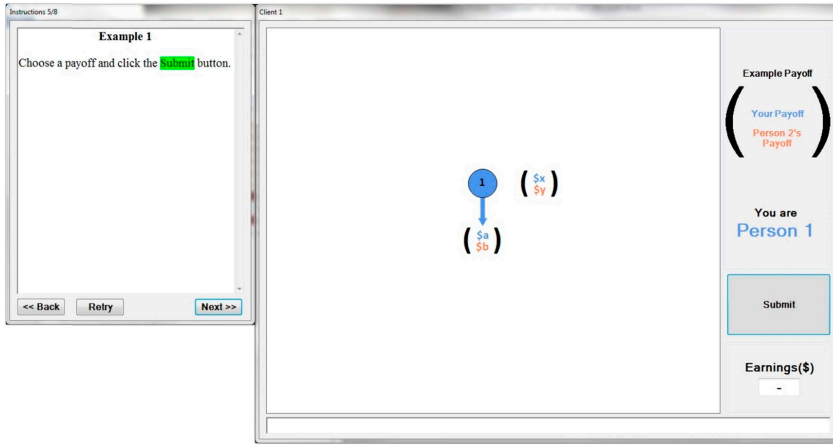


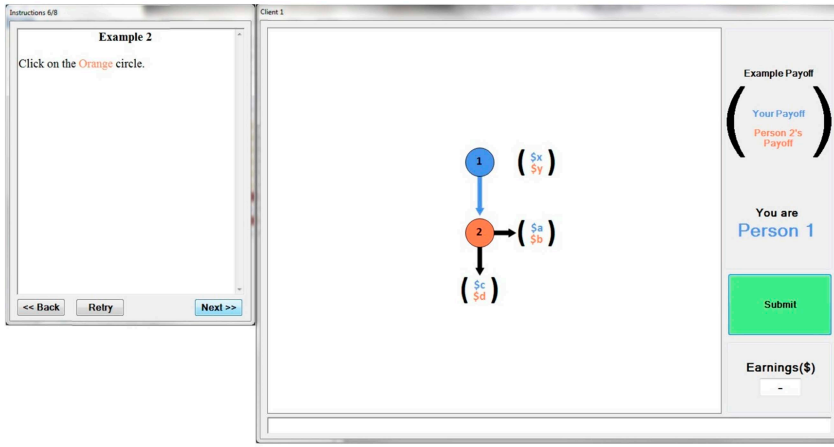
https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press



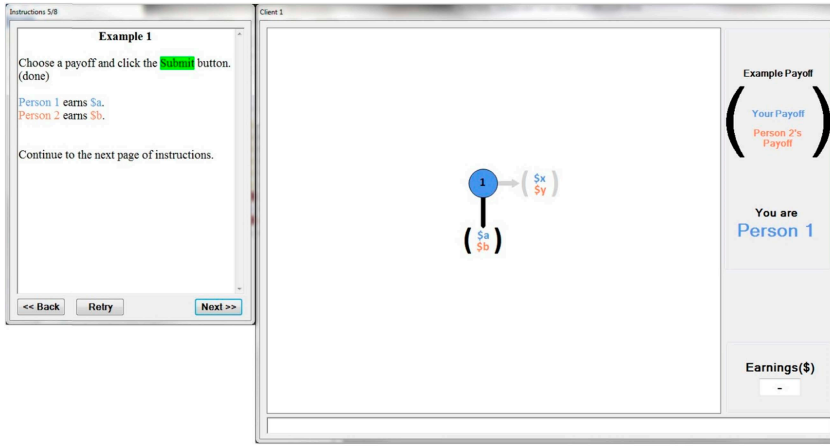
Please note that prior to making a decision, the Submit button flashed green until it was clicked.

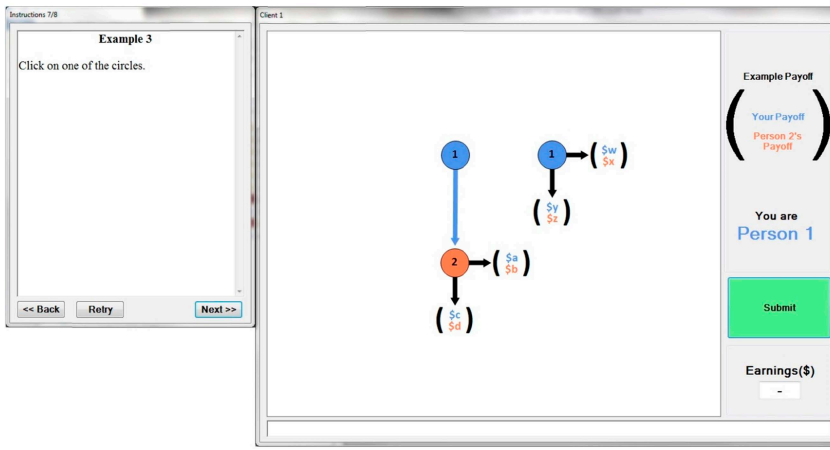
https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press



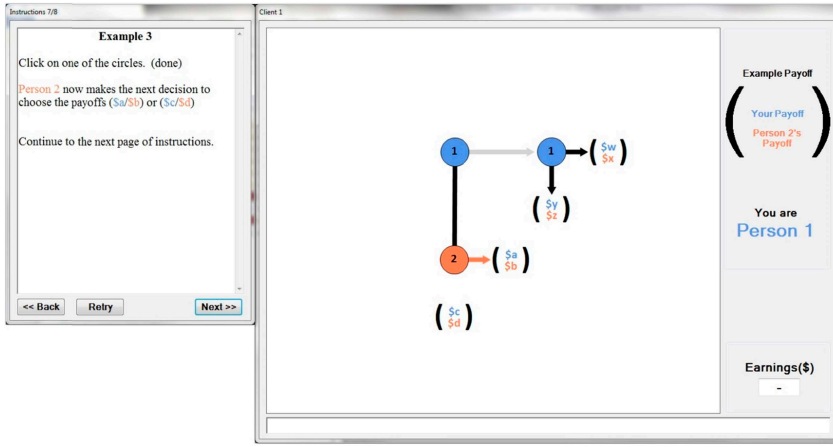


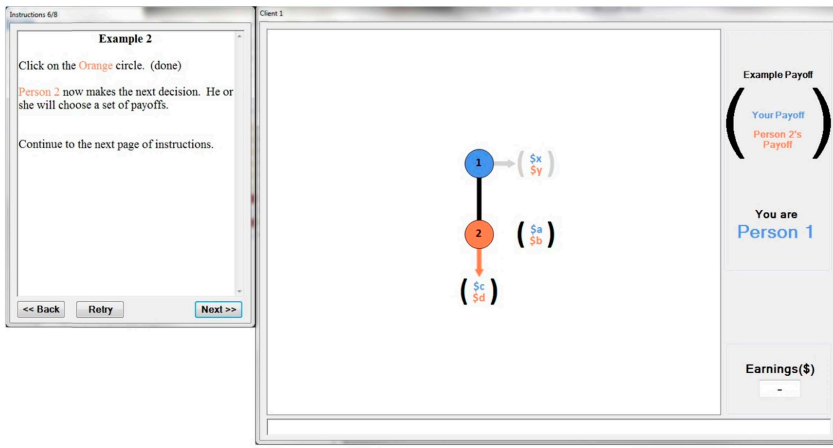
https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press



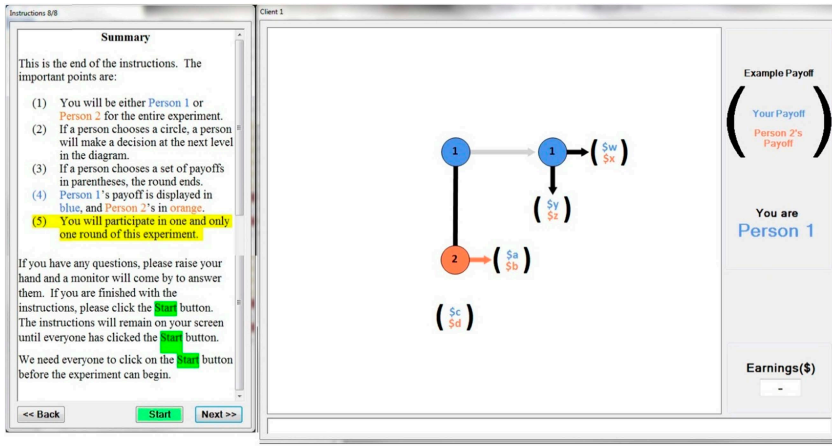


https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press





https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press



https://doi.org/10.1017/S0265052517000024 Published online by Cambridge University Press