6-14-2017

Sentiments, Conduct, and Trust in the Laboratory

Vernon Smith
Chapman University, vsmith@chapman.edu

Bart J. Wilson
Chapman University, bjwilson@chapman.edu

Recommended Citation

Smith, V. L., & Wilson, B. J. (2017). *Sentiments*, conduct, and trust in the laboratory. *Social Philosophy & Policy, 34*(1), 25-55. doi: 10.1017/S0265052517000024

Sentiments, Conduct, and Trust in the Laboratory

Sentiments, Conduct, and Trust in the Laboratory[*]

Vernon L. Smith[†] and Bart J. Wilson[‡]
Economic Science Institute
Chapman University

17 JANUARY 2014

> 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

Background and Motivation

Current interest in trust games by experimental economists originated in the 1990's (Berg et al., 1995) following upon earlier studies of simple two-person ultimatum and dictator games (Guth et al., 1982; Kahneman et al., 1986). The finding that decisions in these games collided with the predictions of game theory subsequently ignited a large literature on trust games.[1,2] This literature has extensively replicated and explored the robustness of the original

[*] 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 paper has benefitted from seminar presentations at George Mason University, the University of Alaska Anchorage, and Wofford College.

[†] Email: *vsmith@chapman.edu*

[‡] Email: *bartwilson@gmail.com*

[1] Yet equilibrium theory had performed well in a great variety of experimental markets in which subjects traded using the institutions (language) of message exchange and contracting that had been observed in business and financial practice such as bid-ask double auctions, posted pricing, and call markets (Smith, 1982). It is important to observe, however, that when trust games are conducted under the same conditions as the experimental markets—repeat play and private information on payoffs—convergence is to equilibrium predicted outcomes. (McCabe et al., 1998, pp. 16-19). Since private information forecloses any prospect of the players reading or signaling conduct in their actions, this condition removes all social content from trust games much as in our representations of the extended order of markets. But in market experiments that provide complete information, convergence to equilibrium is still observed, even if that convergence is less rapid than with private information (Smith, 1982).

[2] Experimentalists long have asked whether their replicable findings in student subject population are special to that group or can be extended to other populations. One answer is to go to the field for subjects; examples include the bilateral bargaining experiments reported in Siegel and Harnett (1964) comparing undergraduates with General Electric executives. In this tradition, forty years later, Fehr and List (2004) report a comparison of CEOs and university students, both in Costa Rica, using the Berg et al. (1995) trust game. In this particular comparison the "CEOs exhibit considerably more trustful and trustworthy behavior than students; as a consequence, CEOs reach substantially higher efficiency levels" (Fehr and List, 2004, p. 764).

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 empirical findings.

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 (Berg et al., 1995, pp. 138-9):

> 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.[3]

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' (Hoffman et al., 1998). Following upon Berg et al. (1995) 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.[4] In this paper we return to Adam Smith (1759) 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.[5] 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*

[3] Before his death, John Dickhaut helped to instigate an extension of these original experiments to the study of three-person trust games in which person A could transfer money which was tripled, to person B, who could transfer money that was tripled again to person C. Person C could then return money to B, and B could return money to A. The original qualitative patterns of trust and trustworthiness continued to be represented in the three-person case (see Rietz et al., 2013).

[4] For discussions of stakes and context, see Camerer (2003, pp. 60-61) and Smith (2008, Chapter 10); for intentions see McCabe et al. (2000) and Fehr and Rockenbach (2003).

[5] 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 Smith (2010, pp. 5-9).

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 et al. (2007) supply a concise summary of models that enrich utility by the inclusion of 'other' rewards. 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 paper.

In his first book, *The Theory of Moral Sentiments*, Adam Smith (1759; hereafter *Sentiments*, or *TMS* for specific reference citation) articulated a theory of human sociality devoted to understanding moral human action; i.e., the "practice of the duties of life" (C.J. Smith, 1894, p. 574). 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.[6] In this paper we (1) provide a brief account and interpretation of *Sentiments* showing that it (2) departs fundamentally from contemporary patterns of thought in economics[7] 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 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." (Smith, 1759, p. 219) 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

[6] For a broader perspective on Adam Smith and his cohorts in Scotland's intellectual community, see Buchan (2003) and Phillipson (2010).

[7] We recognize that *Sentiments* may have important pre-visionary connections to psychology, social psychology, philosophy, sociology, and anthropology, but any such discussion is well beyond the scope of what we attempt here.

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.

'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).[8] 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 (e.g., trust) games or in small groups (public good and common property games).[9]

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 actions as signals of rule-following conduct, it seems particularly appropriate to probe the substance and implications of that model.[10] 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.

[8] So abbreviated and further discussed by McCloskey (2006).

[9] Kahneman et al. (1997) provide a particularly clarifying and thoughtful distinction between the concept of Bentham's (also Jevons' and Edgeworth's) intensity of experienced utility, and the writings of later and contemporary economists based on decision utility. These utilitarian concepts, however, deflect modeling attention away from the foundation of social action in conduct as we find it developed in *Sentiments*.

[10] Recall that the motivation for new models, beginning in the 1980's and 1990's, grew out of a body of replicable and replicated experimental data that contradicted the traditional game-theoretic (Max-U) framework for the individual who was postulated to expect that others would do the same. That failure experience jump-started the search for alternatives, and one consequence was that the enterprise settled upon modifying the arguments of the utility function to which the maximization calculus was applied.

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* (Smith, 1776; hereafter *Wealth*). 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" (*TMS*, p. 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. (*TMS*, p. 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. (*TMS*, p. 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" (*TMS*, p. 38). 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. The first 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".[11] For Smith, behavior is about rules all the way down (see Samuel Johnson's definitions at the head of the paper). 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 ill 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.[12] 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 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.[13] Smith clearly saw the latter as efficient, but hastened to make plain that their efficiency is not what explains why people follow them.

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 "behaviour", with no modifier, that he references already regards others because those regards are embedded in the rules that govern human intercourse (*TMS*, p. 225). As Charles John Smith in his *Synonyms Discriminated* (1894) 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" (p. 159).

By using both CONDUCT and BEHAVIOUR, the meticulous Adam Smith intends to place the confidence of men of virtue in two distinct concepts. Chiefly though, his project throughout

[11] 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.

[12] See Kurzban (2001) for an evolutionary psychologist's critique of experimental economics as an essentially behaviorist enterprise.

[13] 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 *TMS*, pp. 82-5 and Smith (2013).

Sentiments is about conduct.[14] [15] 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; i.e., rule-following) human action (*TMS*, p.110):

> 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 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;[16] 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" (*TMS*, p. 175).[17]

Sense Perceptions and the Non-Specifiability of Human Conduct

What then is the disconnect between a utilitarian model of behavior (in the modern social science sense) and human conduct? The deficiency of an egoist utilitarian approach, as Pettit (1995) explains, stems from equating the outcome of acting with the motivation for acting in a social situation (pp. 311-12):

[14] Adam Smith uses CONDUCT 309 times in the 338-page *Sentiments*, 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). BEHAVIOUR, on the other hand, never appears in a title and is used only 80 times. Moreover, the conjunction X AND BEHAVIOUR is used 17 times where X is CONDUCT, CHARACTER, SENTIMENTS or COUNTENANCE; again, he refers five times to WHOLE BEHAVIOUR. The substance of Smith's thought process—one to which we are not accustomed—is revealed in his careful diction.

[15] Charles John Smith discriminates the synonyms for us (p. 159):

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.

[16] Notice that the mirror of society is *in* the behavior of those with whom one lives, i.e., behavior regards others. Observe also the inference from Adam Smith's thought experiment that the concept of the individual, of one's own character, of self-knowledge, is ultimately derived from the idea of social mind or social psychology.

[17] 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 (2006) who treats the rules of all virtues as rules of grammar. 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" (*TMS*, pp. 175-176).

> When I act on a desire to help an elderly person across the road, I act so as to satisfy that desire but I do not act for the sake of such satisfaction; I act for the sake of helping the elderly person. To think otherwise would be to confuse the sense in which I seek desire-satisfaction in an ordinary case like this and the sense in which I seek it when I relieve the longing for a cigarette by smoking or the yearning for a drink by going to the pub.

The difference lies in the hypothesis that action maps into outcome and thence preference; rather, 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 matter. Such considerations, however, are at the core of *Sentiments*, which articulates principles and propositions that unpack the meaning conveyed in the social context. Even 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. That my action can be "explained" by a preference function defined on my experience of "smoking" and my having learned that it is in my interest to honor my wife's concerns, does not provide insight into the principles that govern 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. Such prescience, also predating modern psychology, deserves a sympathetic retrospective hearing.

Hayek (1963) explains that our ability to perceive a pattern in human conduct does not necessarily imply an ability to completely specify it. 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 mathematician Barbara Shipman can completely specify it as a flag manifold projected onto a perceivable 2-D plane (Frank, 1997).

Human conduct falls into the third category. We can obviously sense a pattern in conduct, but it is a non-specifiable pattern. 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 $U_i(s_1,...,s_n)$'s for $i = 1,...,n$, which is why Frank Knight says "[i]t is not enlightening to be told that conduct consists in choosing between possible alternatives" (1922, p. 467). 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 our own perceptions of a one-to-one correspondence between action and preference as elements of a scientific explanation.[18]

Denying that our own perceptions of conduct can be used 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 Smith's *The Theory of Moral 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 motivation:

  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" (*TMS*, pp. 113-114).
- On conduct and self-command (*TMS*, p. 83; italics added):

  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" (*TMS*, p. 114).

  "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, by sympathy with the approbation of this supposed equitable judge. If otherwise, we enter into his disapprobation, and condemn it" (*TMS*, p. 110).

[18] See also Knight (1924) who noted this problem with behavioral representations in economics 90 years ago.

- On rules being derived from experience, not reason (*TMS*, p. 159):

> 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.[19]

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 consensus, 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.

A Formal Representation of Smith's Model of Action[20]

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

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

where *PR* and *PW* are (0, 1) indicators, respectively, that action deserves social praise (1), or not (0), and is praise-worthy (1), or not (0); and α_i, β_i, γ_i and δ_i are nonnegative functions weighting PR and PW in determining its propriety.[21] 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.

[19] 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 in Smith (2008).

[20] We develop here the implications of our brief footnote 5 in Smith and Wilson (2014).

[21] We assume it is stoically obvious, as it is in *Sentiments,* that an appraisable 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, etc. Self-respect implies that I also see it as praiseworthy even where 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. (Fehr and Gachter, 2000). In *Sentiments,* 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!

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 satisfying or pleasing 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 "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 $\alpha_i(C) = 0$, 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; viz, 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 evolution. *Ex post* social preference and reciprocity explanations of unpredicted forms of other regarding behavior are similarly still incomplete. Smith's model was predictive of the results in trust and other such early games, but for our ignorance of it.

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 a familiar concept to the reader,[22] BENEFICENCE has an archaic ring to it, sounding more like an 18th century word than a 21st century one (when did you last use the word in conversation?). Beneficence is, literally from Latin, *well doing*, and according to the *OED* only entered the lexicon in the 16th 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

[22] "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" (*TMS*, p. 86).

receiver of the signal require the whole context, $C(m_1,m_2)$, to interpret the action. Charles John Smith (1894) distinguishes BENEVOLENCE from BENIGNITY (another 18th century-sounding word), HUMANITY, and KINDNESS (pp. 165-166):

> 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" (pp. 175-176).[23] 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 (Wilson, 2010 & 2012).[24] "Disagreement" implies disequilibrium 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 'relational propositions' from *Sentiments* (Part II, Section ii, Chapter I):[25]

- *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" (*TMS*, p. 78).
- *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" (*TMS*, p. 78).[26]

[23] Cf. Fn 16.

[24] 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 $U_j(\cdot)$.

[25] We name and number them as propositions; *Sentiments* does not.

[26] In Smith and Wilson (2014), we 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. Under this interpretation the proposition denies that ultimatum offers

- *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" (*TMS*, p. 78).

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'.[27] Let's see how well *Sentiments* predicts conduct over 250 years after it was first offered.

Trust Game: Designs for the Reward and Punishment of Actions

Our starting point, displayed in Figure 1, is a slightly modified two-person trust game of McCabe and Smith (2000); the latter has been replicated by Cox and Deck (2005) and Gillies and Rigdon (2008), and we want to test for significant deviation in our new experiments. The first modification is that the payoffs are increased by 20%. If, as the first-mover, Person 1 ends the game by playing right, each person receives $12 instead of the original $10. Similarly, if Person 1 plays down and Person 2 plays right, Person 1 receives $18 and Person 2 $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). Notice that this game, like that in McCabe and Smith (2000), 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 predicted), 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%) 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

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.

[27] The conditionals in these propositions are all subject to error: e.g., 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).

the first mover playing down, or are simply willing to exploit it self-interestedly under anonymity.[28] 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.

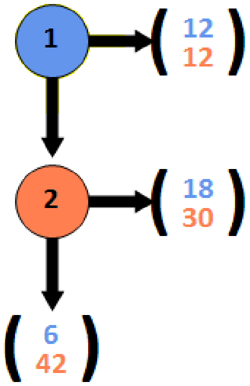


Figure 1. ***No Punish (NP) Game***

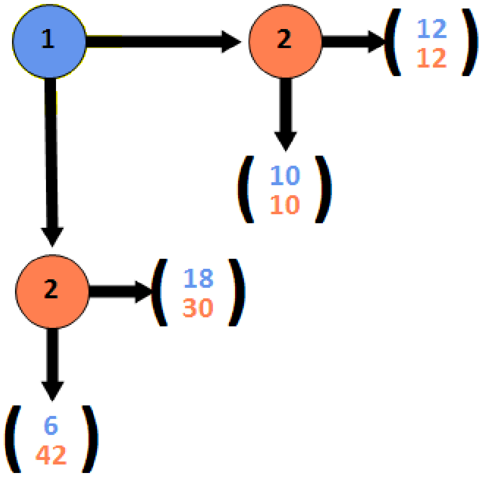


Figure 2. ***Punish Want of Beneficence (PWB) Game***

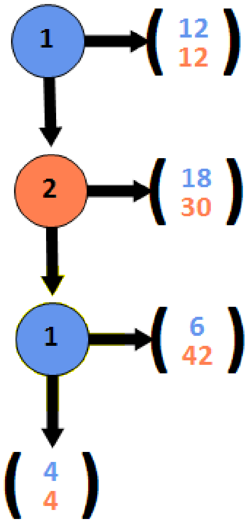


Figure 3. ***Punish Hurt (PH) Game***

One take-away point—before we extend this basic extensive form game[29]—is that this problem is not nearly as simple as traditional game theory presumes it to be for its human

[28] See also Smith and Wilson (2014).

[29] 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

subjects. All the subjects have to cleave to are their past experiences in life and the bare structure of an unfamiliar, single play, extensive form game tree.

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% or 350%.[30]

The *Punish Hurt* (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.

Procedures

We originally recruited 150 students with a variety of majors from the undergraduate population at 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

former (see, e.g., McCabe et al., 2000). 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 *Sentiments* 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.

[30] We note that because of these inequitable outcomes, models of inequity aversion (e.g., Fehr and Schmidt, 1999) also predict that Person 1 would not play down. But inequity aversion fails to predict that when Person 1 has no right play option, must play down, and Person 2 sees that nothing is given up, the results flip to strong support of the self-interested outcome. In the language of *Sentiments* no benefit can be intended by Person 1, and no gratitude felt by Person 2 (Smith and Wilson, 2014).

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 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 B. 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.[31] The experiment lasted well under the 60 minutes for which they were recruited.

<u>*Results*</u>

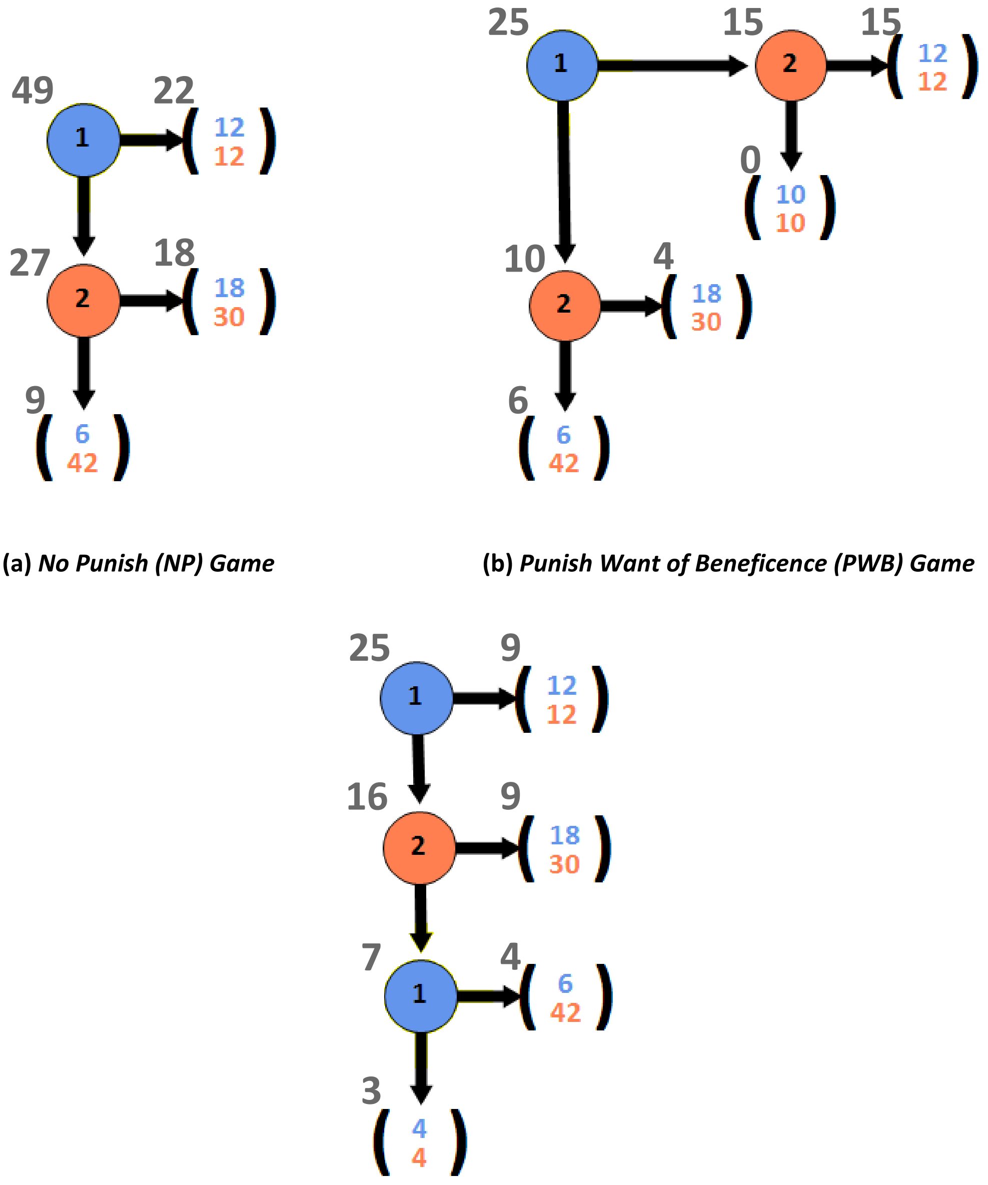
Figure 4 reports the number of decisions at each node and the number of outcomes reached in each of the three games. Our first finding establishes a baseline by comparing our *No Punish Game* to the previous trust game studied by McCabe and Smith (2000) and replicated by Cox and Deck (2005) and Gillies and Rigdon (2008).

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%) move down. Previous studies combined have found that 46 out of 98 first movers (47%) 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%) play right. Previous studies have found that 31 out of 46 (67%) 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*.

[31] Thus, the 296 subjects were paid a sum total of $7,432 for participating in this experiment.



(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

Finding 2: *Fifteen out of 15* (100%) *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.) Instead of 33% of the Person 2's playing down after Person

1 plays down, 6 out of 10 (60%) 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), but statistically insignificant in a logit model for panel data of the 10 sessions conducted in this experiment (see Appendix A, Table A2: marginal effect is -0.3288 with a *p*-value of 0.1821.) 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%). In the logit model for panel data, this effect is marginally significant (Appendix A, Table A1: marginal effect = -0.2753, *p*-value = 0.0563). Even though Adam Smith successfully predicts that Person 2 will not punish the want of beneficence 100% 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 unintended 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%*) *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).[32]

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 (E.g., McCabe et al., 2002, Figures 3 and 4, also reported in Smith, 2008, pp. 269-270; Fehr and Rockenbach, 2003; Fehr and List, 2004). 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* (Appendix A, Table A2: marginal effect = -0.4087, *p*-value = 0.0760) In effect the offer no longer is an unambiguous signal of trust and hence this circumstance undermines trustworthiness.[33] 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.*

[32] Recall that the subjects are randomly assigned to one of the three treatments in each of 5 sessions.

[33] In Fehr and Rochenbach (2002) when the threat of punishment is conveyed in advance of the offer to cooperate, cooperation is reduced. Similarly, in Fehr and List (2004) the punishment option reduces the trustworthiness of both CEO and student subjects.

At this point it could be said that these findings can be rationalized with nearly every model of social preferences (as well as by simple payoff maximization). But every one of those 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 (Wilson 2008, 2010, 2012):

> 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.[34]

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 (*TMS*, p. 78).

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.

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 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.

[34] Furthermore, if Person 1 is inequality averse at the third node, why didn't he or she choose ($12, $12) to begin with?

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*? Notice that there are four possible combinations of outcomes for the game in Figure 5:

(i) Person 1's may act beneficently with the same frequency that they do in Figure 4(a) and Person 2's may reward that beneficence with the same frequency as they do in Figure 4(a);

(ii) Person 1's may act (less) beneficently with the same frequency that they do in Figure 4(b) and Person 2's may (fail to) reward that beneficence with the same frequency as they do in Figure 4(b);

(iii) Person 1's may act beneficently with the same frequency that they do in Figure 4(a) and Person 2's may (fail to) reward that beneficence with the same frequency as they do in Figure 4(b); and

(iv) Person 1's may act (less) beneficently with the same frequency that they do in Figure 4(b) and Person 2's may reward that beneficence with the same frequency as they do in Figure 4(a).

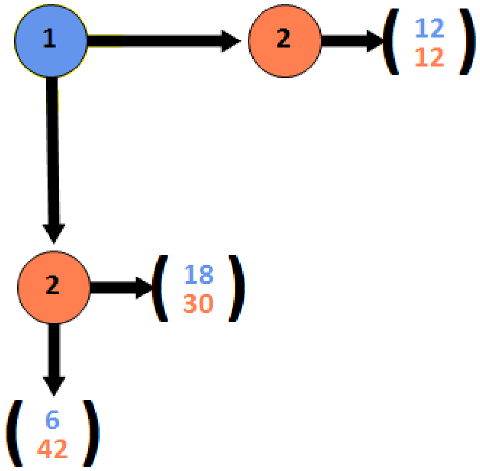


Figure 5. ***No Punish Pass (NPP) Game***

Which set of results would you, the reader, predict? And how confident are you in that prediction? Out interpretation of *Sentiments* having failed to anticipate the results in Figure

4(b), we had no clearly reliable insight, which is why we conducted the explicating treatment; without it we could not have claimed that it was only the prospect of punishment that mattered.

We found in Figure 4 (b) and (c) that the set of sub-games available in the decision tree affect choice; these results are congruent with *Sentiments* although they do not change the analysis by self-loving players. 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.

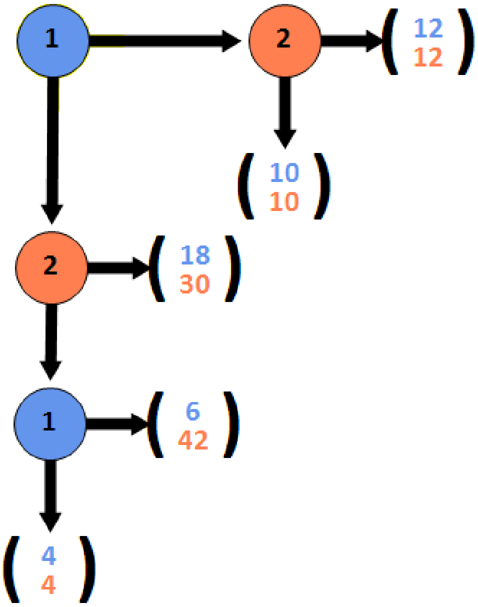


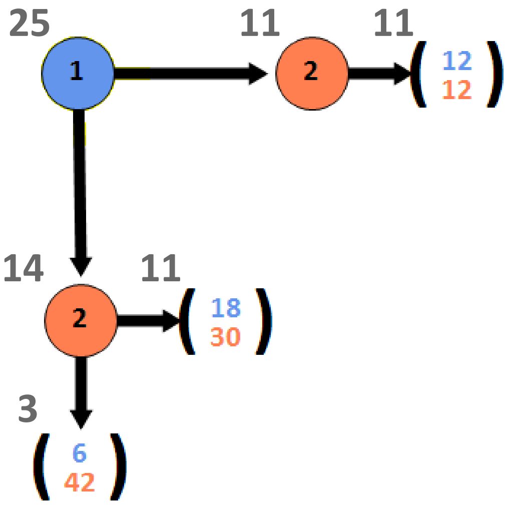
Figure 6. ***Punish Either (PE) Game***

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*.

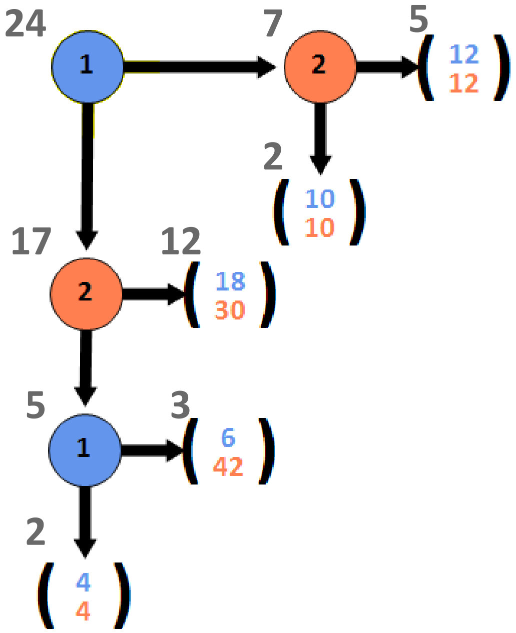
Finding 4: *In the* NPP Game *Person 2's reward the beneficence of Person 1 at (a) a higher frequency than in the* PWB Game *and (b) the same frequency as in the* NP Game*.*

Fourteen out of 25 Person 2's in the *NPP Game* have the opportunity to reward the beneficence of Person 1, and 11 (79%) 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 (Appendix A, Table A2: marginal effect = 0.4374, *p*-value = 0.0768).

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%) Person 1's act beneficently toward Person 2, which is closer to the observed frequency in the *NP Game* (55%) than the observed frequency of the *PWB Game* (40%). Having observed the combination (i) above, 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.



(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

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"* (Hayek, 1973, p. 17). It does not emerge as a community convention because, as explained by Adam Smith, the want of beneficence does no real positive harm.[35] 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, 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 motivations, 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

[35] 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.

adding the opportunity to punish hurt. Seventeen of 24 (71%) Person 1's act beneficently, which is quite close to the 16 of 25 (64%) who do so in the *PH Game*. Twelve of 17 (71%) Person 2's reward the act of beneficence, which is in line with the 67% who do so in the *NP Game*. For the 5 Person 1's who were hurt by Person 2, two of them (40%) punish that hurt, which is consistent with the 43% 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. As reported above in Finding 2, we previously found that zero out of 15 Person 2's were willing to punish the want of beneficence.

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. But the latter, in effect, are reduced to games against an inert nature. 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" (1925, p. 374), echoed *Sentiments* in which individual actions are signals of rule-governed conduct. Context matters because it gives meaning to outcomes.

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 predictions. *Sentiments* is about the ethical rules that constitute the character of an inherently sociable person who strives for a better life, and it 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" (Knight, 1922, p. 459).

References

Ashraf, Nava, Colin Camerer, and George Lowenstein. 2005. "Adam Smith, Behavioral Economist," *Journal of Economic Perspectives*, 19(3): 131-145.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 10: 122–142.

Bicchieri, Cristina. 2006. *The Grammar of Society*. New York: Cambridge University Press.

Buchan, James. 2003. *Crowded with Genius*: *The Scottish Enlightenment*: *Edinburgh's Moment of the Mind*. New York: Harper.

Camerer, Colin F. 2003. *Behavioral Game Theory*. Princeton, NJ: Princeton University Press.

Cox, James C. and Cary A. Deck. 2005. "On the Nature of Reciprocal Motives," *Economic Inquiry*, 43: 623–635.

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

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159-181.

Fehr, Ernst and John List. 2004. "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOs" *Journal of the European Economic Association,* 2: 743–771.

Fehr, Ernst and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 422: 137-140.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114(3): 817-868.

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

Kahneman, Daniel, Peter Wakker, and Rakesh Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics*, 112: 375-405.

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

Hayek, F. A. 1973. *Law Legislation and Liberty, Vol. I Rules and Order.* Chicago: University of Chicago Press.

Hayek, F. A. 1963. "Rules, Perception and Intelligibility," *Proceedings of the British Academy*, 48: pp. 321-344.

Knight, Frank H. 1922. "Ethics and the Economic Interpretation," *Quarterly Journal of Economics*, 36(3): 454-481.

Knight, Frank. H. 1924. "The Limitations of Scientific Method in Economics," in *The Trends of Economics*, R.G. Tugwell (ed.), New York: Alfred A. Knopf, pp. 229-268.

Knight, Frank H. 1925. "Economic Psychology and the Value Problem," *Quarterly Journal of Economics*, 39(3): 372-409.

Kurzban, Robert. 2001. "Are Experimental Economists Behaviorists and is Behaviorism for the Birds?" *Behavior and Brain Sciences*, 24(3): 420-421.

McCabe, K., Stephen Rassenti, and Vernon L. Smith. 1996. Game Theory and Reciprocity in Some Extensive Form Experimental Games. *Proceedings of National Academy of Sciences*, 93: 13421–13428.

McCabe, Kevin, Stephen Rassenti and Vernon L. Smith . 1998. "Reciprocity, Trust and Payoff Privacy in Extensive Form Bargaining," *Games and Economic Behavior*, 24: 10-24.

McCabe, Kevin and Vernon L. Smith. 2000. "A Comparison of Naïve and Sophisticated Subject Behavior with Game Theoretic Predictions," *Proceedings of the National Academy of Arts and Sciences*, 97: 3777-81.

McCabe, Kevin, Vernon L. Smith, and Michael LePore. 2000. "Intentionality Detection and 'Mindreading': Why Does Game Form Matter," *Proceedings of the National Academy of Sciences*, 97(8): 4404-4409.

McCloskey, Deirdre N. 2006. *The Bourgeois Virtues: Ethics for an Age of Commerce*. Chicago: University of Chicago Press.

Pettit, Philip. 1995. "The Virtual Reality of '*Homo Economicus*'," *The Monist*, 78(3): 308-329.

Phillipson, Nicholas. 2010. *Adam Smith: An Enlightened Life.* New Haven: Yale University Press.

Rietz, Thomas A., Roman M. Sheremeta, Timothy W. Shields, and Vernon L. Smith. 2013. "Transparency, Efficiency and the Distribution of Economic Welfare in Pass-Through Investment Trust Games," *Journal of Economic Behavior and Organization*, forthcoming.

Siegel, Sidney and Donald Harnett. 1964. "Bargaining Behavior: A Comparison between Mature Industrial Personnel and College Students," *Operations Research*, 12:300-304.

Smith, Adam. 1759. *The Theory of Moral Sentiments*. Indianapolis: Liberty Fund (1982).

Smith, Charles John. 1894. *Synonyms Discriminated: A Dictionary of Synonymous words in the English Language*. New York: Henry Holt and Company.

Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science," *American Economic Review*, 72(5): 923-955.

Smith, Vernon L. 2008. *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge: Cambridge University Press.

Smith, Vernon L. 2010. "Theory and Experiment: What are the Questions?" *Journal of Economic Behavior and Organization*, 73(1): 3-15.

Smith, Vernon L. 2013. "Adam Smith: From Propriety and Sentiments to Property and Wealth," *Forum for Social Economics,* 43 (1): 1-15.

Smith, Vernon L. and Bart J. Wilson. 2014. "Fair and Impartial Spectators in Experimental Economic Behavior," *Review of Behavioral Economics*, forthcoming.

Wilson, Bart J. 2008. "Language Games of Reciprocity," *Journal of Economic Behavior and Organization*, 68(2): 365-377.

Wilson, Bart J. 2010. "Social Preferences aren't Preferences," *Journal of Economic Behavior and Organization*, 73(1): 77-82.

Wilson, Bart J. 2012. "Contra Private Fairness," *American Journal of Economics and Sociology*, 71(2): 407-435.

Appendix A. Logit Model Maximum Likelihood Estimates for Panel Data

Table A1. Person 1's First Decision to Play Down (Y_1 = 1)

Variable[†]	Marginal Effect	*p*-value
Punish Want of Beneficence	-0.2753	0.0563
Punish Hurt	-0.0426	0.7707
No Punish Pass	0.1301	0.3585
Punish Either	0.2827	0.0585
	n	148

[†] Fixed effects by session (*s* = 1, 2, …, 10) omitted.
No Punish is the baseline conducted in each session.

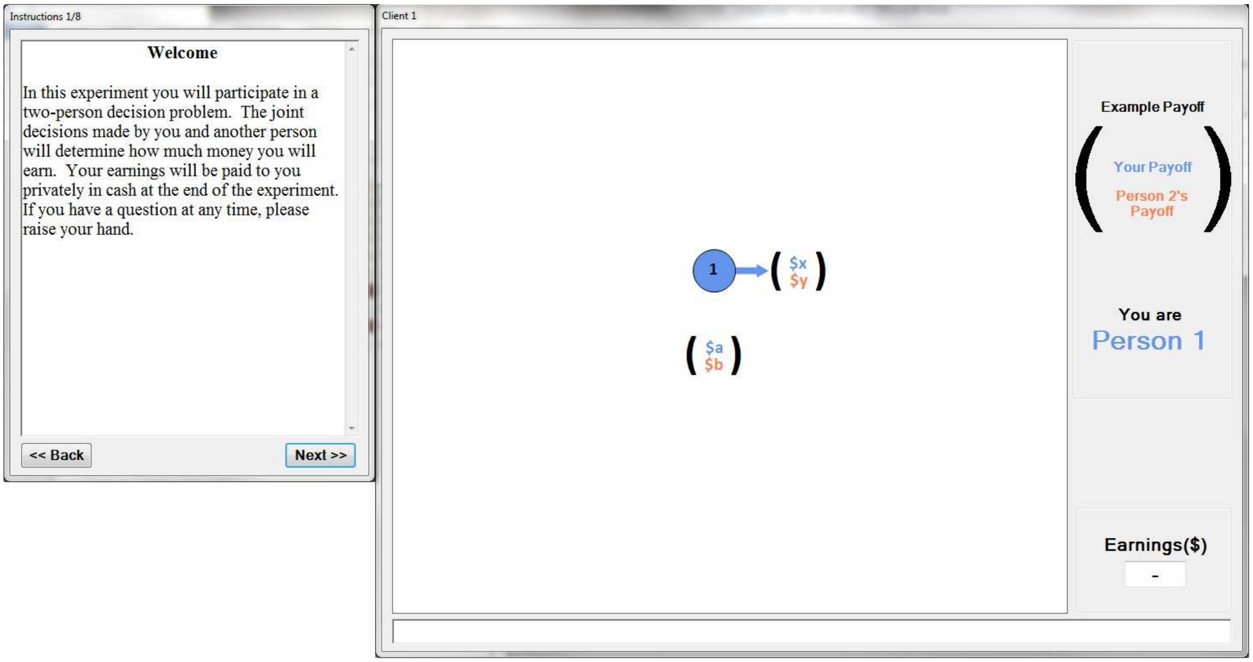
Table A2. Person 2's First Decision to Play Right (Y_2 = 1) Conditional on Person 1 Playing Down

Variable[†]	Marginal Effect	*p*-value
Punish Want of Beneficence	-0.3288	0.1821
Punish Hurt	-0.4087	0.0760
No Punish Pass	0.4374	0.0768
Punish Either	0.2858	0.1645
	n	84

[†] Fixed effects by session (*s* = 1, 2, …, 10) omitted.
No Punish is the baseline conducted in each session.

Appendix B. Experiment Instructions

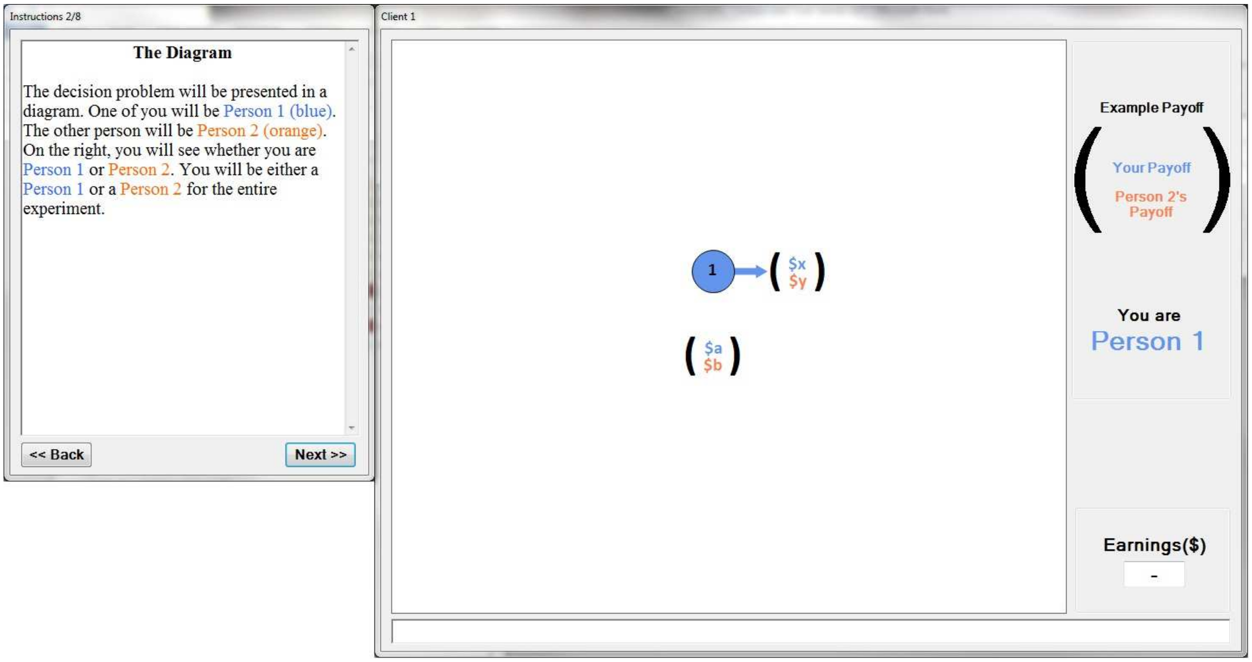
Please note that prior to making a decision, the colored arrows at the active decision node were flashing, alternating between the choices.



Instructions 1/8

Welcome

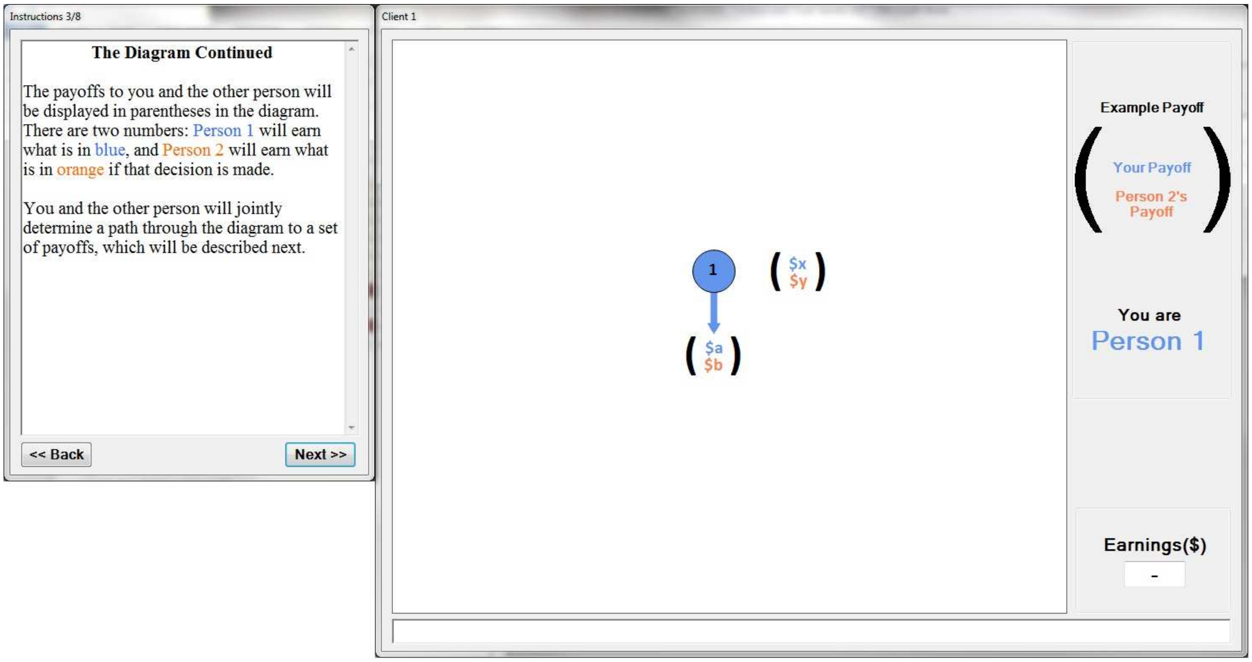
In this experiment you will participate in a two-person decision problem. The joint decisions made by you and another person will determine how much money you will earn. Your earnings will be paid to you privately in cash at the end of the experiment. If you have a question at any time, please raise your hand.



Instructions 2/8

The Diagram

The decision problem will be presented in a diagram. One of you will be Person 1 (blue). The other person will be Person 2 (orange). On the right, you will see whether you are Person 1 or Person 2. You will be either a Person 1 or a Person 2 for the entire experiment.

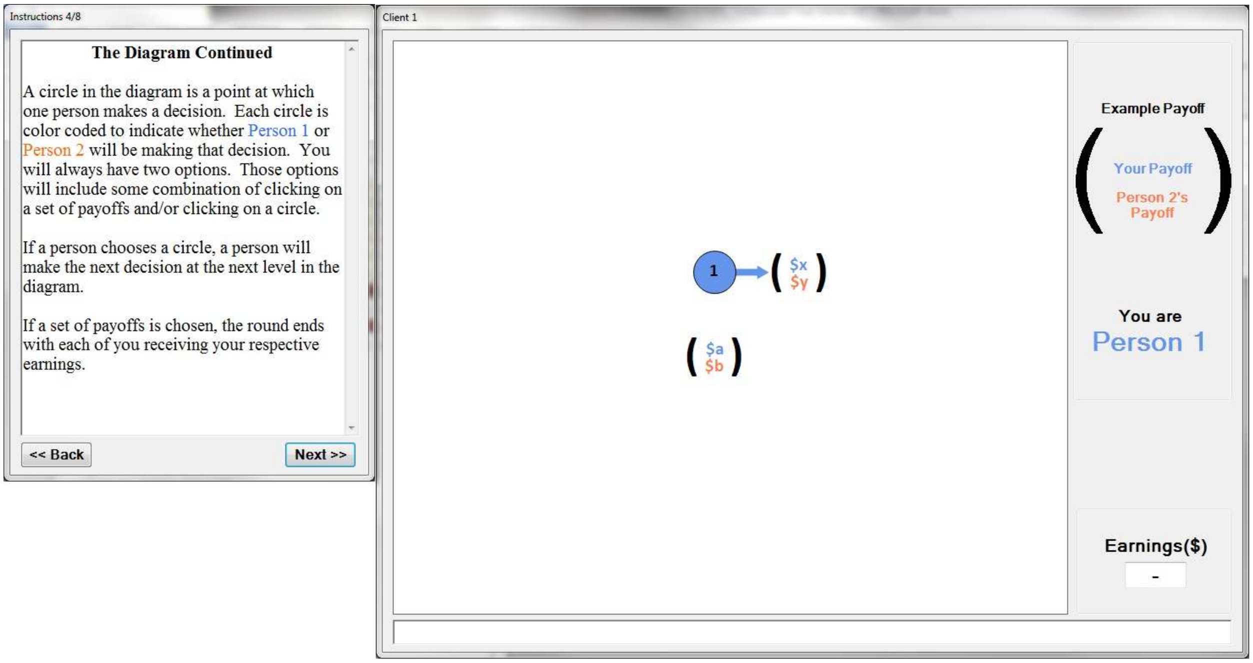


Instructions 3/8

The Diagram Continued

The payoffs to you and the other person will be displayed in parentheses in the diagram. There are two numbers: Person 1 will earn what is in blue, and Person 2 will earn what is in orange if that decision is made.

You and the other person will jointly determine a path through the diagram to a set of payoffs, which will be described next.



Instructions 4/8

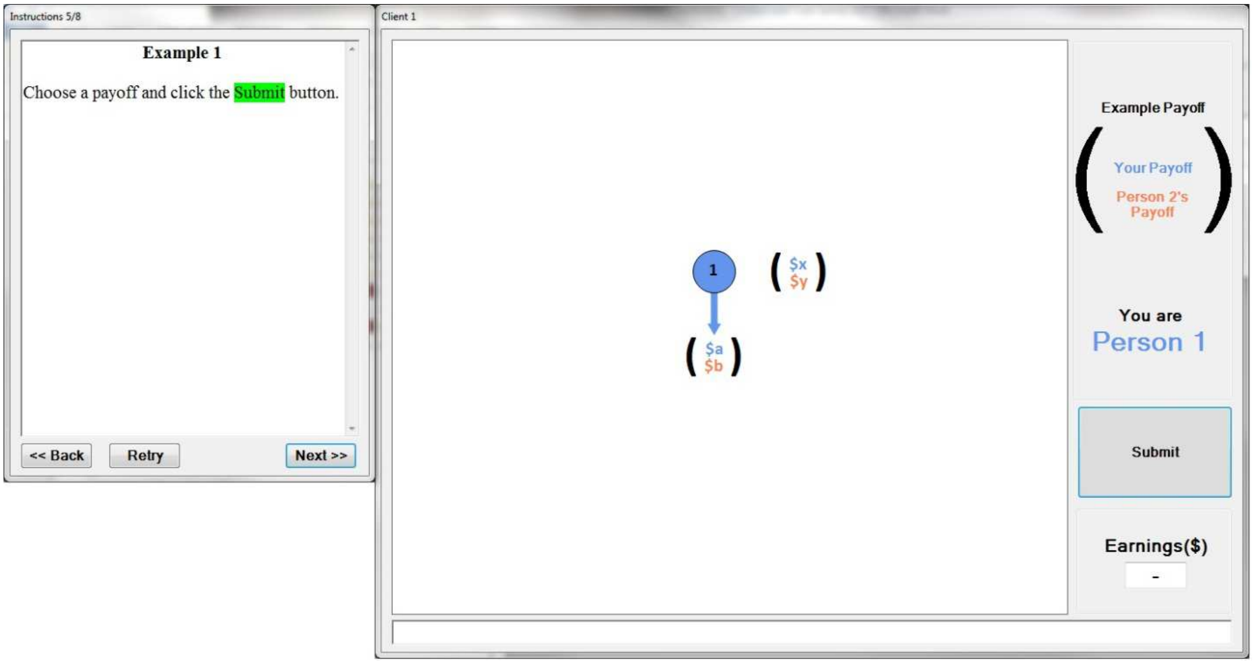
The Diagram Continued

A circle in the diagram is a point at which one person makes a decision. Each circle is color coded to indicate whether Person 1 or Person 2 will be making that decision. You will always have two options. Those options will include some combination of clicking on a set of payoffs and/or clicking on a circle.

If a person chooses a circle, a person will make the next decision at the next level in the diagram.

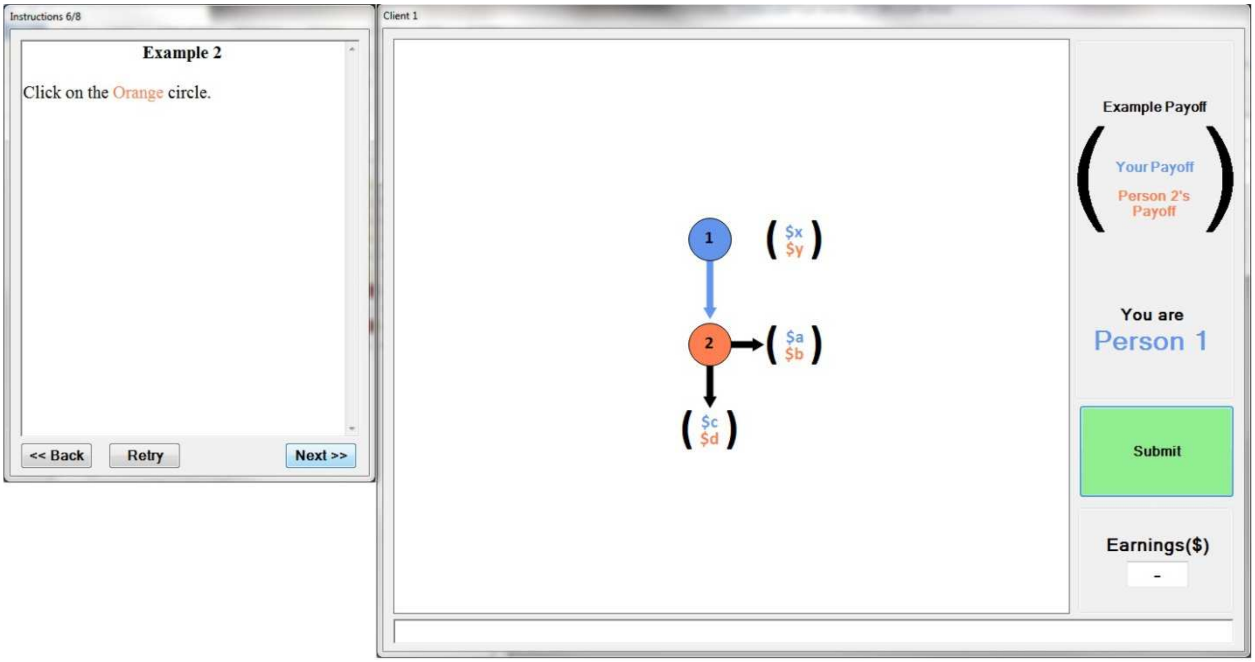
If a set of payoffs is chosen, the round ends with each of you receiving your respective earnings.

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



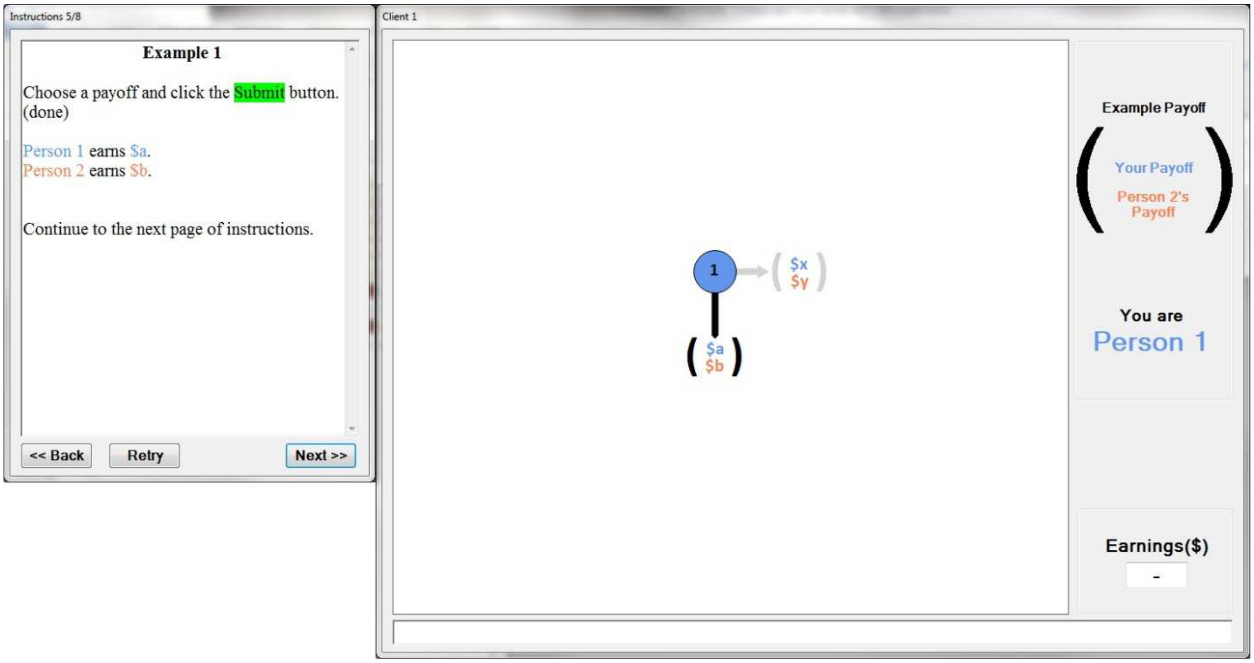
Example 1

Choose a payoff and click the Submit button.



Example 2

Click on the Orange circle.



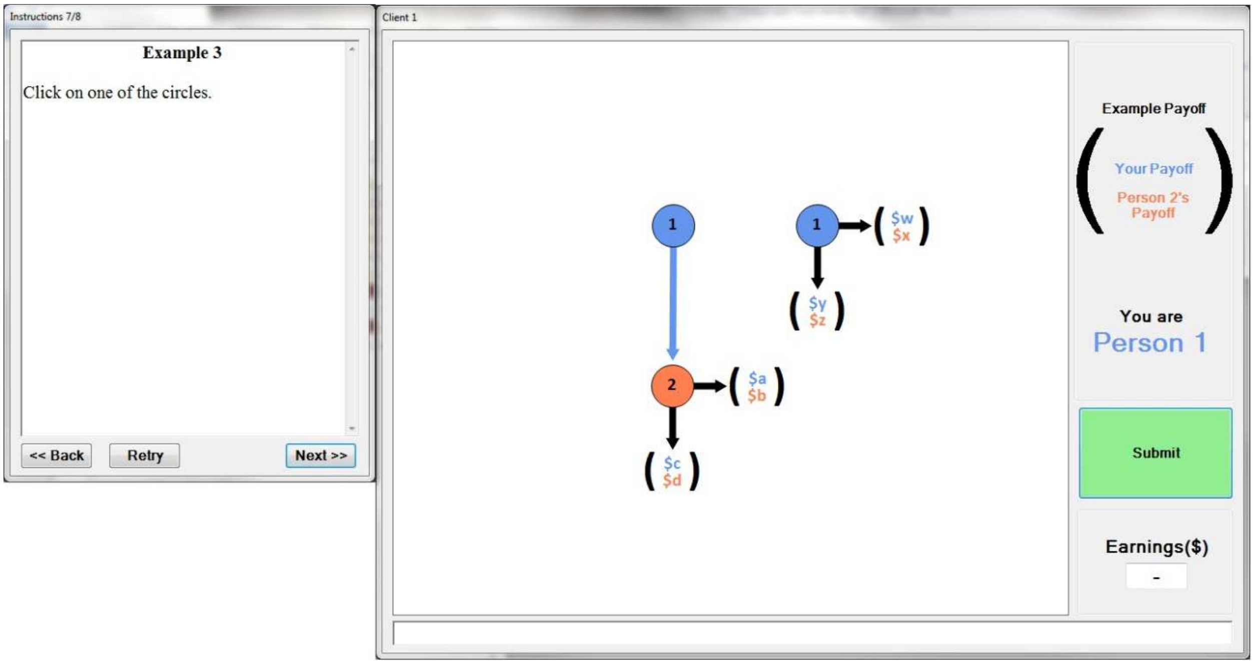
Instructions 5/8

Example 1

Choose a payoff and click the Submit button. (done)

Person 1 earns $a.
Person 2 earns $b.

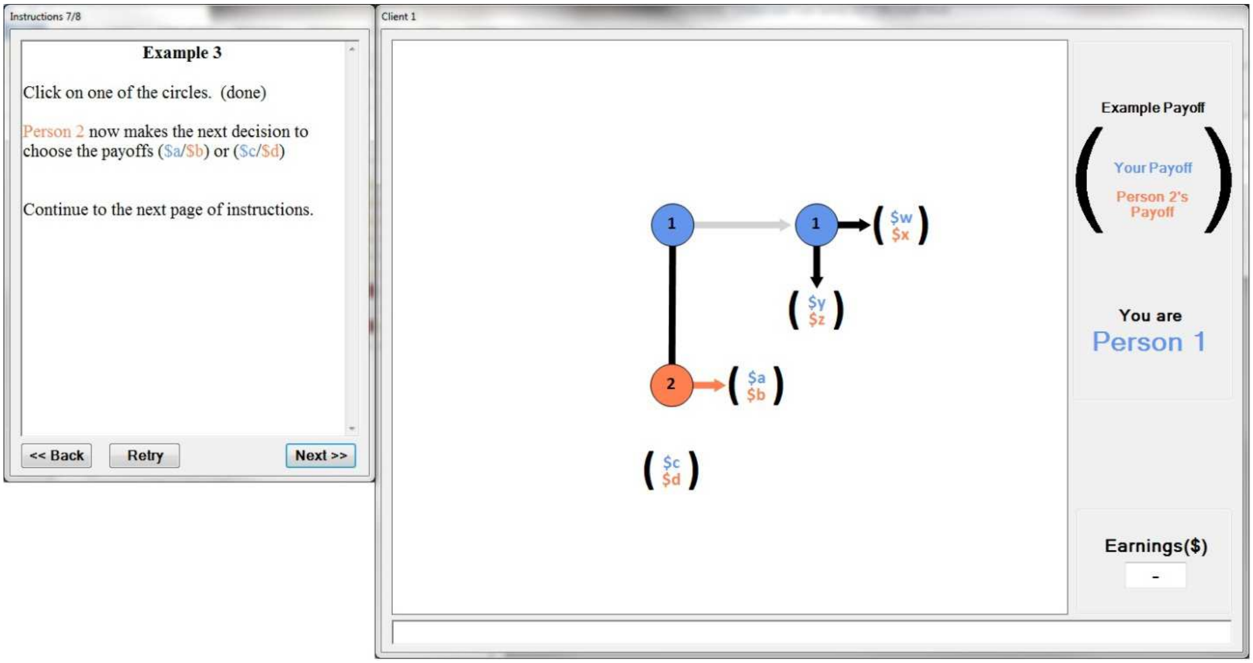
Continue to the next page of instructions.



Instructions 7/8

Example 3

Click on one of the circles.



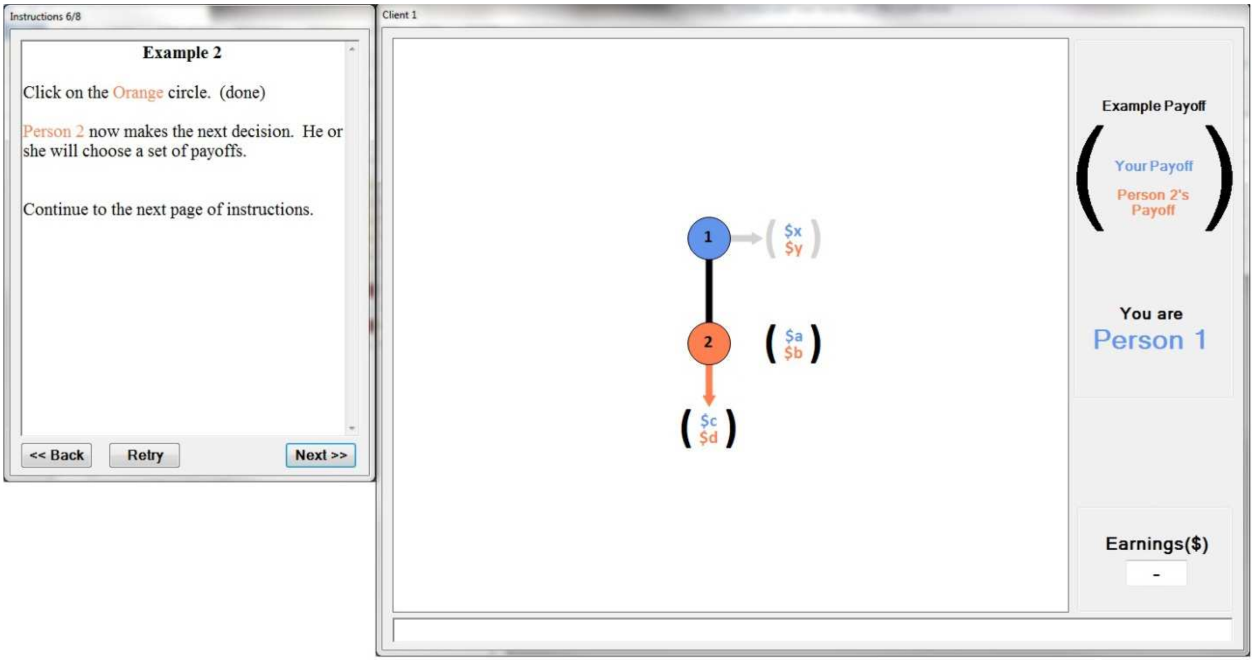
Instructions 7/8

Example 3

Click on one of the circles. (done)

Person 2 now makes the next decision to choose the payoffs ($a/$b) or ($c/$d)

Continue to the next page of instructions.



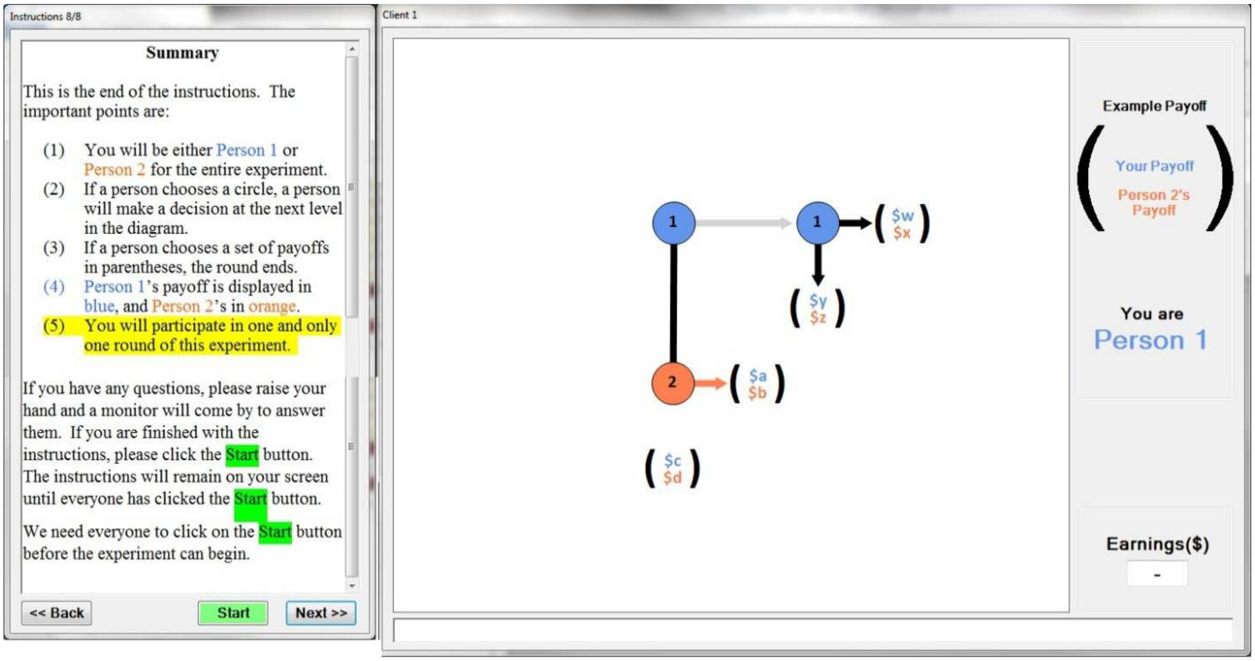
Instructions 6/8

Example 2

Click on the Orange circle. (done)

Person 2 now makes the next decision. He or she will choose a set of payoffs.

Continue to the next page of instructions.



Instructions 8/8

Summary

This is the end of the instructions. The important points are:

(1) You will be either Person 1 or Person 2 for the entire experiment.

(2) If a person chooses a circle, a person will make a decision at the next level in the diagram.

(3) If a person chooses a set of payoffs in parentheses, the round ends.

(4) Person 1's payoff is displayed in blue, and Person 2's in orange.

(5) You will participate in one and only one round of this experiment.

If you have any questions, please raise your hand and a monitor will come by to answer them. If you are finished with the instructions, please click the Start button. The instructions will remain on your screen until everyone has clicked the Start button.

We need everyone to click on the Start button before the experiment can begin.