# Voluntary Interaction and the Principle of Mutual Benefit

Andrea Isoni[*]

Robert Sugden[†]

Jiwei Zheng[‡]

29 July 2022

[*] Behavioural Science Group, Warwick Business School, Coventry CV4 7AL, UK. University of Cagliari, Italy. Email: a.isoni@warwick.ac.uk.

[†] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia, Norwich NR4 7TJ, UK. Email: r.sugden@uea.ac.uk.

[‡] Department of Economics, Lancaster University Management School, Lancaster LA1 4YX, UK. Email: j.zheng18@lancaster.ac.uk.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 670103. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper. We thank the Editor and two anonymous reviewers for useful comments.

## Voluntary Interaction and the Principle of Mutual Benefit

### Abstract

Most social preference theories are based on observations of non-voluntary interactions. Non-selfish behaviour may take fundamentally different forms in voluntary interactions, such as market transactions. We investigate the *Principle of Mutual Benefit* – an injunctive norm requiring individuals who enter interactions voluntarily to conform to common expectations about behaviour within them. This norm induces patterns of behaviour inconsistent with existing social preference theories, and allows extrinsic incentives to crowd in trustworthiness. We embed this norm in a model consistent with evidence about promise-keeping, gift exchange, and 'avoiding the ask'. We present new experimental evidence that people adhere to it.

**Keywords:** voluntary interaction; Principle of Mutual Benefit; social norm; crowding in; non-selfish behaviour

**JEL Classification:** D01, D63, D91.

In one of the most famous sentences in economics, Adam Smith (1776/ 1976: 26–27) tells us that it is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. His hypothesis is that behaviour in markets is predominantly motivated by self-interest, whatever may be the case in other domains of social life.[1] This hypothesis has been maintained by generations of economists, and is currently endorsed even by many of those who study non-selfish motivations.

Behavioural economists sometimes comment on an apparent contrast between market experiments, in which subjects' behaviour is often well-explained in terms of self-interest, and experiments on Dictator, Ultimatum, Trust and Public Good Games, which have generated a large body of evidence of other-regarding behaviour. Social preference theories explain the latter evidence by assuming that individuals are willing to make trade-offs between personal material benefits and various kinds of 'social' benefit, such as material gains to other people (e.g., Becker, 1974: 1083–1085), reductions in inequality (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), increases in economic efficiency (e.g., Charness and Rabin, 2002), confirmations of other people's expectations of benefit (e.g., Battigalli and Dufwenberg, 2007), and the rewarding or punishing of other people for their kindness or unkindness (e.g., Rabin, 1993). But why is there not more evidence of social preferences in market experiments?

A common explanation is that the rules by which markets operate do not give individuals the power to implement social preferences (e.g., Levine, 1998: 605–606; Fehr and Schmidt, 1999: 830; Falk and Fischbacher, 2006: 307): markets *frustrate the exercise of* pro-social motivations. A stronger claim, that markets *repress* pro-social motivations, has been made on the basis of experiments that elicit individuals' willingness to forgo material payoffs to act 'morally' towards animals (Falk and Szech, 2013) or with 'social responsibility' when their actions may have negative externalities (Bartling et al., 2015). These experiments find that implicit valuations of morality and social responsibility are lower when subjects interact in bilateral bargaining or experimental markets than when they make decisions as individuals.[2] Falk and Szech (p. 707) and Bartling et al. (p. 225) interpret their results as

[1] Smith's (1759/ 1976) theory of moral sentiments gives an important role to benevolence and 'fellow-feeling' in non-market social life. But his explanation of trust *in business* relies on each merchant's self-interest in maintaining a reputation for trustworthiness (Smith, 1763/1978: 538–539).

[2] Lower but not zero: Bartling et al. (2015) find a price premium for goods embodying social responsibility. Sutter et al. (2020) find that trading volumes in experimental markets are reduced if exchanges have socially irresponsible externalities.

supporting philosophical arguments that markets can corrode morality (Anderson, 1993; Sandel, 2012). A similar conclusion is often reached in the literature on *intrinsic motivation*, by invoking the hypothesis that material incentives for behaviour that would otherwise be motivated only by altruism or public spirit can inhibit (or 'crowd out') that behaviour (Titmuss, 1970; Frey, 1994; Katz and Handy, 1998; Gneezy and Rustichini, 2000; Hayes, 2005).

Our paper offers a fundamentally different perspective on the relationship between non-self-interested motivations and behaviour in markets. We argue that existing discussions about markets and morality overlook a crucial difference between, on the one hand, the relationship between participants in the experiments that produce evidence in support of theories of social preference and, on the other, the relationship between the parties to a market transaction. This difference is that market transactions, in common with many other interpersonal relationships in civil society, are *voluntary*: they take place only with the prior consent of all parties.

We characterise a type of motivation that is neither benevolent nor self-interested, and that applies to voluntary interactions. It expresses a moral attitude to the relationship between co-participants in activities that are directed at mutual benefit. This motivation comes into play only after a voluntary interaction has been initiated, but can induce non-self-interested behaviour *within* such an interaction (for example, returning another participant's trust). We argue that markets tend to support this motivation rather than erode it.

Our paper builds on the work of Sugden (2018), who characterises a concept of voluntary interaction and a norm of behaviour – the 'Principle of Mutual Benefit' (PMB) – that applies to such interactions.[3] This principle requires that, if there are common expectations about how people behave within some type of voluntary interaction, then any individual who enters such an interaction should conform to those expectations for as long as other participants do the same. Sugden (2018) focuses on the moral status of PMB, arguing that it can be recommended to citizens who are seeking to agree on the moral rules that they will uphold. Our paper investigates the implications of the hypothesis that there is an empirical tendency for people to conform to PMB. We compare this hypothesis with social

[3] There are some parallels between Sugden's approach and that of Smith and Wilson (2019). Both emphasize the significance of voluntariness, and question the applicability of social preference theories to market interactions. The similarities and differences between the two approaches will be discussed in Sections 3.5 and 6.3.

preference theories as explanations of decisions about whether or not to enter voluntary interactions and about behaviour within such interactions.

We focus on a class of two-person interactions that take place recurrently within a large population of potential participants. These interactions have three key features that are characteristic of trading relationships when contracts are not fully enforceable. First, they are voluntary: an interaction takes place if and only if both potential participants choose to enter it. Second, within each interaction, there is a commonly known sequence of moves by the participants than can be interpreted as 'honest play'. Each participant's net gain from honest play (relative to non-entry) is private knowledge, and it may be positive or negative; with positive probability, net gain is positive for both. Third, one of the participants may have an opportunity to 'cheat' (i.e., deviate from honest play) in a way that benefits them but harms the other participant.

In the theoretical part of our paper, we use these three properties to show that PMB makes a distinctive set of predictions that differ from those of existing theories of social preferences. If there is a sufficiently strong motivation to conform to PMB, individuals whose self-interest dictates entering to cheat may instead enter and play honestly, or they may choose non-entry; but they do not enter unless that is in their self-interest. Both forms of non-self-interested behaviour can be crowded *in* by material incentives. These predictions should be read as hypotheses about general tendencies in actual behaviour in the real-world settings that the model represents in a necessarily stylised form – i.e., the model's *target domain*.

In parallel experimental work, we investigate behaviour in a laboratory environment designed to capture those same three key properties in a context that mimics real-world cases of markets with imperfectly enforceable contracts.[4] We find strong evidence of the differences between behaviour before and after entry that are implied by PMB, but no evidence either of crowding in (as implied by both PMB and guilt aversion) or of crowding out (as suggested in some contributions to the 'markets and morals' literature).

[4] These two lines of work were chronologically as well as conceptually parallel, with a common origin in Sugden's (2018) philosophically-oriented analysis. The theoretical model has been progressively refined since the experiment was run, partly in response to suggestions from referees. The experiment is therefore best interpreted as providing relevant evidence rather than a direct test of the model.

Our paper is organised as follows. In Sections 1 and 2, we formalise the concept of a voluntary interaction and define PMB in relation to this. In Section 3, we compare PMB with accounts of pro-sociality offered by other theories. In Sections 4 and 5, we present our theoretical model and derive implications about its equilibrium properties. In Section 6, we review evidence from relevant previous laboratory and field experiments. In Section 7, we report our experiment and discuss the implications of its results. Section 8 concludes.

## 1. Voluntary interactions

As a preliminary to developing our game-theoretic model of a voluntary interaction, it is useful to have an intuitive illustration of the type of real-world situation that the model is intended to represent. This interaction – a familiar case of imperfectly enforceable contracts – is between Passenger, who is standing on a city street, and Driver, who is driving a vacant taxi along the same street. Passenger can choose whether to signal that she wants a ride. If she signals, Driver can choose whether to drive on or stop and pick her up. If the ride is initiated, Driver's behaviour during the ride can be either courteous or rude. At the end of the ride, Passenger chooses whether or not to add a conventional percentage to the metered fare.

Formally, we define a voluntary interaction as a component of an *extensive game form*. The concept of a *game form*, introduced by Gibbard (1973), is now a standard tool for the analysis of voting systems, opportunities and rights, and is a basic building block of implementation theory (Maskin and Sjöström, 2002: 245–247). A game form differs from a game by using 'physical outcomes' rather than utilities as the primitive descriptions of what happens if a given combination of strategies is played, and hence by eliminating all information about players' preferences. The concept of an *extensive* game form (i.e., a sequential game, but with physical outcomes in place of utilities) is due to Moore and Repullo (1988). If game theory is being used to derive predictions about the behaviour of rational players, the utility information included in a game is crucial. However, game forms can be more useful when the theory is being used to analyse the *rules* ('mechanisms' in implementation theory) that govern interactions between people.

An extensive game form has a set $N$ of *players*, where $|N| = n \geq 2$.[5] Each player $i \in N$ is identified by a distinct description of their role in the game, not as a specific person. The game form is represented by a *tree*, made up of *nodes* and *actions* (i.e., directed links between

[5] In referring to properties of extensive game forms, we use the terminology of Hart (1992).

nodes). The tree is 'rooted' at a unique *initial node*. Every node is *either* (i) a *terminal* node (i.e., a node with no successors) *or* (ii) a *(decision) node of* a specific player who chooses between the available actions at that node *or* (iii) a *chance node* at which 'nature' randomises between two or more actions according to some probability distribution. An extensive form game with no chance nodes is *deterministic*. *Information sets* (i.e., sets of nodes of a given player, such that that player cannot distinguish between them) are defined in the usual way. Every directed path from the initial node to a terminal node is a *path of play*. We use the notation $\langle a_1, \ldots, a_k\rangle$ to represent a path of play (in a game or subgame) that consists of the consecutive actions $a_1, \ldots, a_k$.

For any given extensive game form, for each player $i \in N$, there is a set $C_i$ of possible (physical) *outcomes* for that player, interpreted as mutually exclusive descriptions of things that might happen to $i$. At this stage, we impose no particular structure on possible outcomes except for the requirement that the description of an outcome does not refer to any specific person (as opposed to player role) or to any player's preferences. For each path of play $P$ (and thereby for each terminal node), there is an outcome $c(i, P) \in C_i$ for each player $i$; the $n$-tuple of these outcomes (the *collective outcome* of $P$) will be denoted by $\mathbf{c}(P)$. Notice that two or more distinct paths of play may have the same collective outcome.

In representing voluntariness, we start from the intuitive idea that if an (as yet, formally undefined) 'interaction' $V$ involving a given set of players $N$ is to count as 'voluntary', each member of $N$ should previously have faced a choice between participation in $V$ and some outside option, and should have chosen participation. That choice should have been made in circumstances in which other members of $N$ were not able to use threats or promises to induce that player to participate. Thus, the description of each player's outside option should be independent of the decisions of other players.

Our formal representation of voluntariness uses a specific type of extensive game form, an *IVI* ('initiation and voluntary interaction') game. Intuitively, an IVI game has two parts – an 'initiation procedure' and a 'voluntary interaction'. The initiation procedure determines whether the voluntary interaction takes place. Formally, an IVI game is a deterministic[6] extensive game form for which there exists an $n$-tuple of players' *outside*

[6] In the Supplemental Appendix, we show how our analysis can be extended to game forms with chance nodes.

*options*, $\mathbf{O} = (O_1, \ldots, O_n)$ with $O_i \in C_i$ for all $i$, and that satisfies two conditions defined in the following paragraph.

As a first step, we define a special type of decision node, a *reset node*. A reset node is a node of a specific player $j$ who, at that node, chooses between exactly two actions. One of these actions, the *reset action*, is such that, at all succeeding terminal nodes, each player $i \in N$ gets the outcome $O_i$. Thus, a player who chooses a reset action nullifies the effects of all players' previous actions, restoring the status quo ante. The other action is the *continuation action*. A path of play is a *consent path* if it includes at least $n$ decision nodes, the first $n$ of which are reset nodes, one for each player in $N$, at which the continuation action is chosen. A node that is reached by the $n$th continuation action of a consent path is a *transition node*. (Intuitively, a transition node is reached immediately after all players have rejected their outside options.) A path of play is a *veto path* if it includes at least one reset node at which the reset action is chosen prior to any transition node. The two conditions that define an IVI game are:

*Voluntariness*. Every path of play is either a consent path or a veto path.

*Common Knowledge of Transitions*. Every transition node is the initial node of a subgame.

Since there must be some path on which no reset action is chosen, Voluntariness implies the existence of at least one consent path. Since there are no chance nodes at which consent paths could diverge, these conditions imply that every IVI game has a unique consent path, and hence a unique transition node; if and when that transition node is reached, that fact is common knowledge. On every other path, at least one player chooses a reset action, and the outcome for each player $i$ is $O_i$. The subgame that begins at the transition node is the *voluntary interaction*, denoted by $V$. The tree structure that contains the transition node and all nodes not in $V$ is the *initiation procedure*. Because the only decision nodes in the initiation procedure are reset nodes, the initiation procedure provides no opportunities for threats or promises.

Returning to the opening example, Figure 1 shows how the interaction between Passenger and Driver can be represented as an IVI game – the *Taxi Ride Game*. Actions in the voluntary interaction are shown as solid lines; actions in the initiation procedure are shown as dashed lines. (For the moment, ignore the distinction between light and heavy solid lines.) The players' outside options are $O_P$ and $O_D$. The possible outcomes of the voluntary interaction are represented in an obvious notation. For example, the outcome for Passenger of experiencing rudeness and not giving a tip is $RD_{\mathrm{P}}$; the outcome for Driver of being courteous and receiving a tip is $CT_{\mathrm{D}}$.

*[Insert Figure 1 here]*

## 2. The Principle of Mutual Benefit

We use the concept of an IVI game as a generic representation of a class of similar but not identical *episodes* that occur within some large population of individuals who interact recurrently. For example, a specific episode of the Taxi Ride Game would occur at a specific time and place; it would involve two specific *persons* (i.e., members of a population of taxi drivers and taxi users) occupying the two player roles in the game; there would be a specific destination intended by the customer; and so on. Over an extended period of time, each person is likely to be involved in many such episodes, typically with different persons as co-players. (We do not require that every member of the population is a potential occupant of every role.) The IVI representation is *generic* in the sense that it describes salient features that are common to all episodes; episode-specific features are suppressed.

In focusing on such a representation, we are assuming that each person, through their membership of the relevant population, has access to information and experiences that generate common knowledge of those population-level features of the game that are properties of the IVI description.[7] We do not assume common knowledge of features that are episode-specific, and do not include such features in the generic descriptions of outcomes. (For example, in an episode of the Taxi Ride Game, Driver may not know that Passenger is late for a flight that will take her to an important meeting.)

Within a population, there may be common knowledge not only about the properties of an IVI game, but also about how that game is usually played, conditional on the initiation of the voluntary interaction. Referring to the latter knowledge as ‘normal expectations’, Sugden (2018: 256–281) expresses PMB informally as the following ethical prescription:

[7] Lewis (1969) analyses how common knowledge can be generated within a population of individuals who recurrently play games that are similar, but not necessarily identical, to one another. Cubitt and Sugden (2003) present a reconstruction of Lewis’s theory. A key component of this theory is that individuals can make inductive inferences between situations that are perceived as similar (compare Gilboa and Schmeidler, 1995).

> When participating with others in a voluntary interaction, and for as long as others' behaviour in that interaction is consistent with this very principle, behave in such a way that the other participants are able to satisfy normal expectations about the consequences of the interaction for them.

For example, suppose there is a normal expectation that taxi drivers act courteously and that passengers give tips to courteous drivers. Then PMB prescribes that a person who has chosen to take a taxi ride and has been treated courteously *ought* to give a tip.

We now formalise this principle. Consider any $n$-player IVI game $I$ that is played recurrently in some large population $M$. Let $V$ be the voluntary interaction embedded in $I$. Let $P^*$ be some path of play in the subgame $V$ (i.e., a directed path from the transition node to a terminal node). Suppose it is common knowledge in $M$ that, in episodes in which $V$ is initiated, what happens in $V$ is 'almost always' described by $P^*$, with the implication that the collective outcome is 'almost always' $\mathbf{c}(P^*)$. Then $P^*$ is the *practice* for the voluntary interaction. In any given episode, a player *$i$ conforms to* a practice $P^*$ if, at every decision node for $i$ in $P^*$ that is in fact reached, $i$ chooses the action in $P^*$. For the present, we do not formalise the concept of 'almost always'; we merely assume that, at any given time, an IVI game has no more than one practice (but may have none).[8] We will refine the definition of a practice when, in Sections 4 and 5, we analyse a specific class of IVI games. The concept of a practice is related to, but not quite the same as a 'convention', as the latter is normally understood in game theory. For our purposes, the most important difference is that a practice describes a regularity in behaviour without saying anything about *why* people behave in this way, while a convention is a regularity that is a Nash equilibrium in relation to individuals' preferences and beliefs.

We can now state the normative principle that is the subject of our paper:

> *Principle of Mutual Benefit*. Consider any IVI game $I$ with voluntary interaction $V$ played in some population $M$. Suppose that some path of play $P^*$ in $V$ is the practice for $I$. Then in any episode in which the game is played by members of $M$ and in which its voluntary interaction is initiated, each player is required to conform to $P^*$.

[8] Imprecise concepts analogous with 'almost always' are used in many theories of conventions and norms. For example, Bicchieri's (2006: 11–12) definition of a 'social norm' includes clauses about the behaviour and beliefs of 'sufficiently large subsets' of the population.

Here, 'required' is to be interpreted normatively. An individual who *endorses* the principle thereby accepts that, under the stated conditions, she ought to conform to $P^*$; an individual *adheres* to the principle by in fact conforming to $P^*$ whenever those conditions apply.

To get a sense of the normative appeal of PMB, consider the Taxi Ride Game when the practice is the path of play ⟨*courteous*, *tip*⟩ represented by the heavy solid line in Figure 1. Consider a specific episode of the game: Jane is on the kerb, signalling for a ride, and Joe is driving the taxi. A decision by either player $i$ to enter the voluntary interaction is a signal to the other player $j$ that, conditional on $j$ choosing to enter too, all of the following is very probably true: $i$ will conform to the practice; $i$ will do so in the expectation that $j$ will conform too; and their joint conformity to the practice will bring about the collective outcome that $i$ expected when he chose to enter. So, if Jane were to take the ride, be treated with courtesy but not give a tip, she would be acting contrary to the commonly understood meaning of a signal that she had given voluntarily. The same would be true of Joe if he gave Jane the ride and then failed to treat her with courtesy. Notice that, whatever regularities of behaviour there may be in the population, PMB does not require Jane to signal for a ride, and if she does, it does not require Joe to pick her up. And if the voluntary interaction is initiated but Joe is discourteous, there is no requirement for Jane to give a tip: as a result of Joe's action, it is not possible for Jane to stay on the path of the practice, and so PMB imposes no further requirements on her.[9]

Why do we use the term 'mutual benefit'? Formally, the concept of benefit plays no part in our analysis, because PMB makes no reference to preferences; this concept features only in our interpretation of that analysis. Informally, we treat a player as benefiting from their participation in a voluntary interaction if that interaction has a practice and if they and their coplayers conform to it. Under these circumstances, a player $i$ who chooses to enter a voluntary interaction $V$ with practice $P^*$ knows that, if $V$ is initiated and if they conform to $P^*$, they will (almost certainly) get the outcome $c(i, P^*)$ rather than their outside option $O_i$. If, knowing this, player $i$ chooses entry and gets $c(i, P^*)$, we say that they 'benefit' relative to getting $O_i$. That use of words is normal in economics: think, for example, of revealed preference theory and its application in cost-benefit analysis in the form of hedonic pricing.

[9] As formulated above, PMB does not permit players to deviate from practices in ways that would impose no cost on, or even benefit, their coplayers (for example, by giving larger-than-customary tips). A natural amendment to PMB would be to permit deviations that do not impose restrictions on other players' choice sets.

## 3. Mutual benefit and other forms of non-self-interested behaviour

We now compare the obligations deriving from endorsing PMB with the various forms of non-self-interested behaviour implied by other theories of pro-sociality.

### *3.1. Mutual benefit and reciprocity*

The hypothesis that individuals have preferences for reciprocity has been proposed by Rabin (1993), Charness and Rabin (2002) and Dufwenberg and Kirchsteiger (2004). In Rabin's (1993) seminal theory, the *kindness* of one player *i* to the other player *j* is assessed by taking *i*'s beliefs about *j*'s strategy as given and then considering the decision problem faced by *i* as if it were a non-strategic choice among the alternative Pareto-efficient distributions of material payoffs between the players that are feasible for *i*. Player *i* shows kindness (unkindness) towards *j* by choosing a distribution that is relatively favourable (unfavourable) to *j*. Each player derives utility from their own material payoff and from rewarding the other player for acting on kind intentions (a preference for positive reciprocity) or punishing them for acting on unkind intentions (a preference for negative reciprocity).

This form of reciprocal self-sacrifice is fundamentally different from the reciprocity that is expressed by PMB. This difference is exhibited in the 'Paradox of Trust' – that Rabin's theory cannot explain the persistence of a practice of mutually beneficial trust and trustworthiness in a recurrent Trust Game (Isoni and Sugden, 2019). If such a practice were mutually beneficial, trust by the first mover would not be self-sacrificing, and therefore neither kind nor unkind. Thus, a preference for reciprocal kindness would not motivate the second mover to be trustworthy if that was contrary to their self-interest.[10] In contrast, self-interested adherence to PMB can activate non-self-interested trustworthiness in voluntary interactions (e.g., the combination of courtesy and tipping in the Taxi Ride Game).

### *3.2. Mutual benefit and guilt aversion*

Another recurring theme in the social preference literature is the *guilt-aversion* hypothesis that people are averse to taking actions that disconfirm other people's expectations of benefit (Pelligra, 2005; Bacharach et al., 2007; Battigalli and Dufwenberg, 2007). With respect to

[10] Rabin (1993: 1296–1297) draws a similar conclusion in relation to a non-voluntary Trust Game represented in strategic form. Dufwenberg and Kirchsteiger (2004) amend Rabin's definitions of kindness and unkindness in a way that allows trust by the first mover to be classified as kind. Isoni and Sugden (2019) argue that this amendment lacks a convincing psychological rationale and has counter-intuitive implications.

behaviour *within* voluntary interactions, adherence to PMB can induce effects similar to those of guilt aversion (e.g., costly conformity to a practice of promise-keeping). However, the two approaches have divergent implications for decisions about *entering* voluntary interactions. Suppose it is common knowledge between individuals *i* and *j* that *i* has an empirically well-grounded expectation that *j* will participate with him in some voluntary interaction, and that this would be to *i*'s benefit. A guilt-averse *j* might choose to enter, even if that was contrary to her self-interest. PMB imposes no such obligation.

### *3.3. Mutual benefit and norms*

The hypothesis that non-self-interested behaviour is explained by adherence to social norms is often presented as an alternative to theories of social preference. Following Cialdini et al. (1990), we distinguish between *descriptive* norms (common expectations about how people in fact behave) and *injunctive* norms (common perceptions about the appropriateness or inappropriateness of types of behaviour). Our paper investigates the hypothesis that PMB is a norm in both senses: in the relevant population, people both endorse it and adhere to it. Theories of social norms often assume that an individual's motivation to comply with an injunctive norm is conditional on their belief that the norm is also descriptive (e.g., Bicchieri, 2006). This conditionality is built into the specification of PMB, since PMB never requires anyone to do anything that does not involve conforming to a descriptive norm.

The norms that are discussed in relation to (or in contrast to) social preferences are typically less general in scope than PMB. There has been particular interest in norms of *promise-keeping* (e.g., Ellingsen and Johannesson, 2004; Vanberg, 2008) and *truth-telling* (e.g., Gneezy, 2005; Lundquist et al., 2009). These norms are clearly important in facilitating the realisation of mutual benefit in market transactions and other voluntary interactions. PMB can be interpreted as unifying a large class of norms of voluntary interaction.

Undoubtedly, social norms have a dark side: they can support unthinking conformism to pointless or even cruel practices, and can perpetuate arbitrary inequalities in the distribution of the surplus created by social cooperation. In the case of PMB, however, these effects are mitigated by the fact that PMB requires individuals to conform only to the practices of those interactions that they choose to enter.

### *3.4. Mutual benefit and crowding out*

The *crowding-out* hypothesis proposes that socially-oriented motivations are inhibited by decision environments that give extrinsic incentives for behaviour that such motivations might otherwise have induced. A possible underlying mechanism, suggested for example by Bénabou and Tirole (2003), is that individuals form beliefs about other people's motivations by drawing inferences from those people's behaviour, and that, as proposed by Bem's (1967) theory of self-perception, this form of reasoning extends to inferring one's own motivations from one's own behaviour. If a person chooses some action knowing there was an extrinsic incentive to do so, that knowledge can induce the belief that the incentive was their reason for performing it, crowding out the thought that the action has social or moral value. For this mechanism to work, the scrutiny of one's own and other people's intentions must play a significant role in moral reasoning. In contrast, PMB takes no account of the intentions that lie behind behaviour. As we will show later, extrinsic incentives which directly induce some people to conform to a practice out of self-interest can indirectly induce others to conform when this is *not* in their self-interest – a *crowding-in* mechanism.

### *3.5 Mutual benefit and 'humanomics'*

'Humanomics' is the neologism coined by Smith and Wilson (2019) to describe an analysis of pro-sociality that is heavily indebted to Adam Smith's *Theory of Moral Sentiments* (*TMS*: 1759/1976). They advocate a 'fundamental rethinking of human sociability' building on Smith's idea that rules of conduct in human societies are the products of context-dependent social learning (pp. xiv, 10–11).

Although this approach is in the same spirit as PMB,[11] the theory of other-regarding behaviour that Smith and Wilson (2019: 81–94) reconstruct from *TMS* has rather different foundations, in some respects resembling reciprocal kindness. The central concepts in this theory are *beneficence*, *gratitude* and *reward* and their negative correlates *injustice*, *resentment* and *punishment*. A person's action is beneficent (respectively: unjust) if it intentionally confers a benefit (harm) on another person. 'Benefit' and 'harm' are measured relative to a 'normal baseline condition' (p. 73). Beneficent (unjust) actions induce the sentiment of gratitude (resentment) in the person who is benefited (harmed), and are judged

[11] These similarities are rooted in Sugden's own engagement with *TMS* (e.g., Sugden, 2002). In presenting PMB, Sugden (2018: 272–277) argues that its psychological substrate is the mechanism of 'correspondence of sentiments' first proposed by Adam Smith in *TMS*.

by observers to deserve reward (punishment).[12] Smith and Wilson characterise all voluntary market transactions as 'exchanges of gifts, in the sense that each has to give in order to receive' (pp. 70–71).

### *3.6 Mutual benefit and team reasoning*

The fundamental hypothesis of team reasoning is that individuals who have a shared sense of group identity are motivated to play their respective parts in combinations of strategies that further the collective interest of the group (Sugden, 1993; Bacharach, 2006; Karpus and Radzvilas, 2018). Although theories of team reasoning differ about how 'group identity' and 'collective interest' should be interpreted, theorists generally agree that group identity can be created by mutual consent, and that if one possible outcome for a group is preferred to another by all group members, then the group has a collective interest in reaching the former outcome rather than the latter. In some versions of the theory of team reasoning, each individual's motivation to play their part is conditional on their expectation that the other individuals will play theirs (Gold and Sugden, 2007). Thus, given the revealed preference definition of 'mutual benefit', adherence to PMB can interpreted as 'a form of team reasoning' (Sugden, 2018: 232–235).

## 4. The Mutual Benefit Game

In this Section, we describe the *Mutual Benefit Game*. This is a type of IVI game with the three features we described in the introduction as characteristic of market transactions when contracts are imperfectly enforceable. Its simple structure – that of a Dictator Game that is played only if both players agree – is particularly useful for identifying and understanding differences between alternative theories of non-self-interested behaviour.

The Mutual Benefit Game is a two-player IVI game with the extensive game form shown in Figure 2. For the moment, ignore the entries enclosed by square brackets. This game form should be interpreted as a generic representation of a class of similar but not identical episodes that recur in a large population. For concreteness, we have specified the initiation procedure so that Player 2 (she, the final mover in the game if the voluntary interaction is initiated) moves before Player 1 (he), but this has no significance for our

[12] *TMS* gives Smith and Wilson the philosophical resources to avoid the Paradox of Trust. Adam Smith (1759/ 1976: Part VII, Section ii, Chapter 3) argues that an action that benefits both the actor and another person can be praiseworthy by simultaneously expressing self-love *and* benevolence.

analysis; we could have reversed the order of the players' moves, or made them simultaneous. The outcomes $O_1$, $O_2$, $X_1$, $X_2$, $Y_1$ and $Y_2$ are generic descriptions of what happens if the relevant terminal node is reached.

*[Insert Figure 2 here]*

Since we want to compare the implications of PMB with those of social preference theories, our model of the Mutual Benefit Game includes information about players' (real-valued and finite) material payoffs. These are shown in square brackets in Figure 2, normalised so that each player's outside option gives a payoff of zero. The *payoff parameters* $x_1$, $x_2$, $y_1$ and $y_2$ can take different values in different episodes.

In each episode, the values of Player 1's payoff parameters $x_1$ and $y_1$ are drawn at random from a non-degenerate joint distribution, constant across episodes, with continuous *parameter density function* $g(x_1, y_1)$. Similarly, Player 2's payoff parameters are drawn, independently of Player 1's, from a joint distribution with continuous density function $h(x_2, y_2)$. These distributions are common knowledge in the population, and hence between the players in any episode. In any given episode, each player knows the realisations of their own parameters, but not those of their coplayer, before any actions are taken.

Since we will be interpreting the path of play $\langle X \rangle$ as 'honest play' in a market transaction, we assume (i) $\text{pr}(x_1 > 0) > 0$, (ii) $\text{pr}(x_2 > 0) > 0$, and (iii) $\text{pr}(x_1 > y_1) = 1$. Assumptions (i) and (ii) imply there is non-zero probability that the path of play $\langle in, in, X \rangle$ is *mutually beneficial* (relative to outside options), i.e., it gives positive payoffs to both players. Assumption (iii) implies that if $\langle X \rangle$ were the practice and if the voluntary interaction were initiated, Player 1 would be certain to be harmed if Player 2 failed to conform to that practice.[13] In any particular episode, whether (and how far) it would be in Player 2's self-interest to 'cheat' in this way depends on the value of $y_2 - x_2$. During the initiation procedure, this value is known to Player 2 but not to Player 1.

The asymmetry between the two players serves an important modelling purpose by separating two sets of factors that can influence entry decisions. For Player 1, entry decisions involve issues of trust: he has no opportunity to cheat and has to weigh up the risk of being

[13] Assumption (iii) also implies that, if $\langle Y \rangle$ were the practice and if the voluntary interaction were initiated, a failure to conform to that practice by Player 2 would unambiguously *benefit* Player 1. In the spirit of footnote 9, we interpret PMB as imposing no obligations on Player 2 in this case.

cheated. For Player 2, they involve issues of trustworthiness: she has no risk of being cheated, and knows the costs and benefits of cheating.

It may aid intuition to think of the Mutual Benefit Game as representing a non-synchronised exchange transaction. By both choosing *in*, the players agree that Player 1 will send some specific good to Player 2 and that Player 2 will then send some specific good in return. If Player 2 breaches this contract, she may incur some penalty; the size of this penalty is a random variable, the realisation of which is known only to Player 2. The payoffs $x_1$, $x_2$, $y_1$ and $y_2$ represent the normalised values of the goods to the relevant players, minus any penalty in the case of $y_2$.

For our purposes, the interest of the game centres on whether $\langle X \rangle$ can be the practice. In answering this question, it is convenient to divide Player 2's parameter space into the regions A, B, C and D, as defined in Table 1, and to define $\pi_A$, $\pi_B$, $\pi_C$ and $\pi_D$ as the probabilities that the realised parameter pair ($x_2$, $y_2$) lies strictly within the respective region, i.e., the probabilities of cases A, B, C and D. Notice that, because of the continuity of $g(.,.)$ and $h(.,.)$, there is zero probability that a parameter pair lies on the boundary between any two regions. Throughout our analysis, we will ignore behaviour that is conditional on such zero-probability events. By virtue of the assumption that pr($x_2 > 0$) > 0, $\pi_B + \pi_C > 0$.

For each case, Table 1 shows Player 2's best response if she acts on unconditional self-interest, using the notation *in/X* (respectively: *in/Y*) to denote choosing *in* at her first node and *X* (*Y*) at her second if this is reached. The table also shows Player 2's best response if she acts on self-interest subject to the constraint of not choosing *in/Y*. Constrained and unconstrained best responses differ in two cases, C and D. In both cases, *in/Y* is the self-interested best response. In case C, the constrained best response is *in/X*; in case D, it is *out*.

*[Insert Table 1 here]*

The strategic structure of a Mutual Benefit Game depends on the value of $\pi_B/(\pi_B + \pi_C + \pi_D)$. If $\pi_B/(\pi_B + \pi_C + \pi_D) = 1$, there is no conflict between PMB and self-interest. Thus, there is a subgame-perfect equilibrium in self-interest in which $\langle X \rangle$ is the practice. This can be interpreted as the special case of *fully enforceable* contracts. If $\pi_B/(\pi_B + \pi_C + \pi_D) = 0$ (*non-enforceable contracts*), a self-interested Player 2 would never conform to $\langle X \rangle$; the practice can be sustained only by non-self-interested behaviour. The main interest of the Mutual

Benefit game lies in the spectrum of *imperfect enforceability* – cases in which $0 < \pi_B/(\pi_B + \pi_C + \pi_D) < 1$.

## 5. Mutual benefit equilibrium

We now present a model of how the Mutual Benefit Game is played in populations in which people are motivated to adhere to PMB. Since our aim is to explore the behavioural implications of PMB, we will assume that this norm is the *only* source of non-self-interested motivations.

For any given Mutual Benefit Game (characterised by its specific parameter density functions $g(., .)$ and $h(., .)$), we define the population-level rate of *conformity* to $\langle X\rangle$, denoted by the endogenous variable $\rho$, as the ex ante probability that, in an episode involving two randomly selected members of the population and conditional on both of them having chosen *in*, Player 2 chooses $X$.[14] We assume that the population is sufficiently large that, in the analysis of any individual's behaviour, their own impact on population-level expectations can be ignored.

Recall that PMB imposes no moral requirements: (i) on Player 1, or (ii) on Player 2 if she or Player 1 chooses *out*, or (iii) on Player 2 unless $\langle X\rangle$ is the practice, for which (since there are only two paths of play in the voluntary interaction), $\rho > 0.5$ is a minimal condition. Since our aim is to compare PMB with theories of social preference, we conform with other models of norm-following (e.g., Brekke et al., 2003; Bicchieri, 2006) in treating norm violations as 'morally costly' for the violator, thus allowing for possible trade-offs between material and moral costs. To model the moral cost that Player 2 incurs by choosing $Y$ at her second node, we define a continuous[15] and weakly increasing *moral cost function* $m_i(\rho)$ for each person $i$ (i.e., each member of the population), such that $m_i(\rho) = 0$ if $\rho \leq 0.5$. This formulation eliminates the imprecision in the concept of 'almost always', as used in Section 1, and allows us to represent the possibility that individuals are more averse to contravening a

[14] To deal with the possibility that Player 2's second node is reached with zero probability, our definition of equilibrium (explained later) incorporates a form of subgame perfection.

[15] Formally, this continuity assumption excludes the limiting case in which $i$ treats PMB as an absolute constraint and interprets 'almost always' as a rate of conformity no less than some exact threshold value. However, provided that individuals' threshold values are drawn from a continuous distribution, our results can be extended to this case.

given practice, the greater the population-level rate of conformity to it. We assume that each player maximises expected material payoff minus the expected moral cost of violations of PMB.[16] We will use the term ‘self-interest’ to refer only to material payoffs.

Notice that the behaviour of each player is fully determined by the realised values of their payoff parameters, their moral cost function (in the case of Player 2), and ρ. Since ρ is the only endogenous variable, we can define an equilibrium state of the model as a value of ρ that reproduces itself when players’ strategies are best responses to that value. We interpret such a *mutual-benefit equilibrium* as a rest point in a dynamic process of reinforcement learning (Erev and Roth, 1998). As a first step in analysing equilibrium, we characterise each player’s best responses to ρ.

First consider Player 1. His decision has no effect on his payoff unless Player 2 chooses *in*. Conditional on her doing so, Player 1’s expected payoff is $\rho x_1 + (1 - \rho)y_1$ if he chooses *in* and zero otherwise; since $x_1 > y_1$, this expectation increases in ρ. So, Player 1 chooses *in* if this expectation is positive and *out* if it is negative.

Now consider a given person *i* in the role of Player 2, facing a given conformity rate ρ. In Table 1, we used two provisional definitions of ‘best response’ for Player 2 – one based on unconditional self-interest, the other based on self-interest conditional on not choosing *in*/*Y*. We are now able to use a single definition in terms of material payoff minus moral cost. In cases A and B, Player 2’s best response (*out* and *in*/*X* respectively) is independent of moral cost. In case C ($y_2 > x_2 > 0$), the best response for any individual *i* is *in/X* if $y_2 - m_i(\rho) < x_2$ (case C′) and *in/Y* if that inequality is reversed (case C″). In Case D, *i*’s best response is *out* if $y_2 - m_i(\rho) < 0$ (case D′) and *in/Y* if that inequality is reversed (case D″). Thus, Player 2’s parameter space can be divided into six regions, A, B, C′, C″, D′ and D″, with the best responses shown in Figure 3. We use $\pi_{C'}(i, \rho)$ to denote the probability that *i*’s payoff parameter pair lies in region C′; $\pi_{C''}(i, \rho)$, $\pi_{D'}(i, \rho)$ and $\pi_{D''}(i, \rho)$ are defined similarly. We will say that a proposition holds ‘if moral costs are sufficiently high’ if it holds when there is some $\rho' < 1$ such that, for each relevant individual *i*, $m_i(\rho) > \max(y_2)$ at all $\rho \geq \rho'$. (Notice that $m_i(\rho) > \max(y_2)$ implies $\pi_{C''}(i, \rho) = \pi_{D''}(i, \rho) = 0$.)

*[Insert Figure 3 here]*

[16] Risk neutrality is assumed for reasons of simplicity, and is not crucial for our main results.

Let $\chi_i(\rho)$ be the ex ante probability that, for a given value of $\rho$, a given person *i* (in the role of Player 2, with a given parameter density function *h*(., .)) chooses *X* at her second node, conditional on both players having chosen *in* (either as a best-response strategy, or as the result of a 'tremble' in the sense used in the standard analysis of subgame-perfect equilibrium[17]). Let $\chi(\rho)$ be the corresponding ex ante probability defined for a randomly selected person; $\chi(.)$ is the *best response function*.

Using the pattern of best responses shown in Figure 3:

(1) $\chi_i(\rho) = [\pi_B + \pi_{C'}(i, \rho)] / [\pi_B + \pi_C + \pi_{D''}(i, \rho)]$.

This expression is well defined because, by the definition of a Mutual Benefit Game, $\pi_B + \pi_C > 0$. If $\rho \leq 0.5$, $\pi_{C'}(i, \rho) = \pi_{D'}(i, \rho) = 0$ because $m_i(\rho) = 0$, and so $\chi_i(\rho) = \pi_B / [\pi_B + \pi_C + \pi_D]$, which we denote by $\pi_U$ (the 'unconstrained' conformity rate that would result from self-interested best responses to any $\rho$). If $\rho \geq 0.5$, $\pi_{C'}(i, \rho)$ is continuous and weakly increasing in $\rho$, $\pi_{D''}(i, \rho)$ is continuous and weakly decreasing in $\rho$, and hence $\chi_i(\rho)$ is continuous and weakly increasing in $\rho$. Since these results hold for all $\chi_i(\rho)$, they also hold for $\chi(\rho)$. Hence:

> **Proposition 1:** $\chi(\rho) = \pi_U$ for all $\rho \leq 0.5$; and $\chi(\rho)$ is continuous and weakly increasing in the interval [0.5, 1].

Formally, a mutual-benefit equilibrium is a rate of conformity to *X*, $\rho^*$, such that $\chi(\rho^*) = \rho^*$. Such an equilibrium is *unconstrained* if $\rho^* = \pi_U$ and *constrained* if $\rho^* > \pi_U$. We can now state an existence result:

> **Proposition 2 (Existence of mutual-benefit equilibrium):** At least one equilibrium exists. If moral costs are sufficiently high for all individuals, there is at least one constrained equilibrium.

The first part of Proposition 2 follows immediately from the fact that $\chi(\rho)$ is continuous with $\chi(0) \geq 0$ and $\chi(1) \leq 1$. To prove the second part, it is sufficient to notice that, if moral costs are sufficiently high, $\rho^* = 1$ is an equilibrium.

Figure 4 illustrates possible equilibria by showing three alternative best response functions, $\chi^{I}(\rho)$, $\chi^{II}(\rho)$ and $\chi^{III}(\rho)$. We will interpret these functions as induced by three alternative parameter density functions $h^{I}(x_2, y_2)$, $h^{II}(x_2, y_2)$ and $h^{III}(x_2, y_2)$. (Alternatively, one

[17] The clause about trembles covers the case, mentioned in footnote 14, in which Player 1's best response selects *out* with probability 1.

might interpret them as induced by different assumptions about individuals' moral cost functions.) The dashed 45-degree line $\chi(\rho) = \rho$ represents the equilibrium condition. $E^{\mathrm{I}}$ is an unconstrained equilibrium; $E_1^{\mathrm{II}}$, $E_2^{\mathrm{II}}$, $E_3^{\mathrm{II}}$ and $E^{\mathrm{III}}$ are constrained equilibria.[18]

*[Insert Figure 4 here]*

We now state six qualitative implications that follow immediately from the preceding analysis. For this purpose, we classify players according to whether, given their realised payoff parameters, their payoffs from the path $\langle in, in, X\rangle$ would be positive or negative. Thus, P1$^+$ (P1$^-$) denotes a Player 1 for whom $x_1 > 0$ ($x_1 < 0$); P2$^+$ (P2$^-$) denotes a Player 2 for whom $x_2 > 0$ ($x_2 < 0$). Each implication is a proposition about the ex ante probability of a specified strategy choice by a randomly selected player, conditional on that player's payoff from $\langle in, in, X\rangle$ being positive or negative.[19]

**Implication 1:** In all equilibria, the probability that P1$^+$ chooses *in* is weakly increasing in $x_1$ and $y_1$.

**Implication 2:** In all equilibria with $y_1 < 0$, P1$^-$ chooses *in* with probability zero.

According to Implications 1 and 2, people do not enter voluntary interactions unless it is in their self-interest to do so.

**Implication 3:** In all equilibria, P2$^+$ chooses *in* with probability one.

**Implication 4:** In all equilibria, the probability that P2$^-$ chooses *in* is weakly increasing in $y_2$. If $\pi_D > 0$ and if moral costs are sufficiently high, there is an equilibrium in which the probability that P2$^-$ chooses *in* conditional on $y_2 > 0$ is strictly less than one.

Implications 3 and 4 demonstrate a first distinctive form of non-self-interested behaviour predicted by PMB. People will always enter voluntary interactions from which they can

[18] Given our interpretation of equilibrium as a rest point of a process of reinforcement learning, it is natural to define an equilibrium as stable (unstable) if $\chi(\rho)$ cuts the 45-degree line from above (below). Using this definition, Proposition 2 can be strengthened to: At least one *stable* equilibrium exists; and if moral costs are sufficiently high, at least one *stable* constrained equilibrium exists. The 'existence of equilibrium' statements in Implications 4, 5 and 7 below can be strengthened in a similar way.

[19] The second parts of Implications 4 and 5 are true by virtue of the fact that, if moral costs are sufficiently high, $\rho^* = 1$ is an equilibrium.

benefit by following a practice. However, they may refrain from entering if they can benefit only by deviating from the practice.

> **Implication 5:** In all equilibria, the probability that P2$^+$ chooses *in/X* is weakly increasing in ($x_2 - y_2$). If $\pi_C > 0$ and if moral costs are sufficiently high, there is an equilibrium in which the probability that P2$^+$ chooses *in/X* conditional on $x_2 < y_2$ is strictly positive.
>
> **Implication 6:** In all equilibria, P2$^-$ chooses *in/X* with probability zero.

Implications 5 and 6 demonstrate a second distinctive form of non-self-interested predicted by PMB. Within an interaction they have initiated voluntarily, people may act against their self-interest by conforming to a practice.

So far, we have treated the parameter density functions $g(.,.)$ and $h(.,.)$ as fixed. We now consider two Mutual Benefit Games $I^0$ and $I^1$ which differ only in terms of Player 2's parameter density functions, respectively $h^0(x_2, y_2)$ and $h^1(x_2, y_2)$. We will say that $h^1(x_2, y_2)$ differs from $h^0(x_2, y_2)$ by an *increasing shift in* $x_2$ if $h^0(x_2, y_2)$ can be transformed to $h^1(x_2, y_2)$ by shifting probability mass from left (i.e., lower values of $x_2$) to right (i.e., higher values of $x_2$) in the payoff space represented in Figure 3. Similarly, $h^1(x_2, y_2)$ differs from $h^0(x_2, y_2)$ by a *decreasing shift in* $y_2$ if $h^0(x_2, y_2)$ can be transformed to $h^1(x_2, y_2)$ by shifting probability mass from top (i.e., higher values of $y_2$) to bottom (i.e., lower values of $y_2$) in that space. Using Figure 3, it is easy to see that any increasing shift in $x_2$ or any decreasing shift in y2 implies that $\pi_D + \pi_{C''}(i, \rho)$ weakly decreases and $\pi_B + \pi_{C'}(i, \rho)$ weakly increases. Thus, using equation 1, any such shift weakly increases the value of $\chi_i(\rho)$ at all $\rho$. Since this holds for all $\chi_i(\rho)$, it holds for $\chi(\rho)$. This is the key to our final implication:

> **Implication 7:** Consider any Mutual Benefit Games $I^0$ and $I^1$ which differ only in terms of Player 2's parameter density functions, respectively $h^0(x_2, y_2)$ and $h^1(x_2, y_2)$, and hold constant the moral cost function $m_i(.)$ of each person $i$. Let $\rho^0_{\text{low}}$ and $\rho^0_{\text{high}}$ (respectively: $\rho^1_{\text{low}}$ and $\rho^1_{\text{high}}$) be the lowest and highest equilibrium values of $\rho$ in game $I^0$ (respectively: $I^1$). If the shift from $h^0(x_2, y_2)$ to $h^1(x_2, y_2)$ is either an increasing shift in $x_2$ or a decreasing shift in $y_2$, then $\rho^1_{\text{low}} \geq \rho^0_{\text{low}}$ and $\rho^1_{\text{high}} \geq \rho^0_{\text{high}}$.

To illustrate Implication 7, refer back to Figure 4. In this figure, $\chi^{\text{I}}(\rho)$, $\chi^{\text{II}}(\rho)$ and $\chi^{\text{III}}(\rho)$ are induced by parameter density functions $h^{\text{I}}(x_2, y_2)$, $h^{\text{II}}(x_2, y_2)$ and $h^{\text{III}}(x_2, y_2)$; the shifts from $h^{\text{I}}(x_2, y_2)$ to $h^{\text{II}}(x_2, y_2)$ and from $h^{\text{II}}(x_2, y_2)$ to $h^{\text{III}}(x_2, y_2)$ are increasing shifts in $x_2$ and/or decreasing shifts in $y_2$. The lowest equilibrium shifts from $E^I$ to $E^{II}_1$ and then to $E^{III}$; the highest equilibrium shifts from $E^I$ to $E^{II}_3$ and then to $E^{III}$.

Implication 7 is the crowding-in property of our model. Increasing shifts in $x_2$ and decreasing shifts in $y_2$ can be interpreted as increases in Player 2's material incentive to conform to the practice $\langle X \rangle$, conditional on the initiation of the interaction. (In terms of the example used in Section 4, an increase in the expected penalty incurred by Player 2 for breach of a contract would be a decreasing shift in $y_2$.) Such an increase in material incentives, other things being equal, tends to induce a higher rate of conformity to $\langle X \rangle$. This increase in conformity can be decomposed into two complementary effects. The *payoff effect* is the increase in the value of $\chi(\rho)$ induced by the shift in $h(., .)$, $\rho$ being held constant. The *moral cost effect* is the increase in the value of $\chi(\rho)$ induced by the change in $\rho$, $h(., .)$ being held constant. Implication 7 demonstrates another distinctive prediction of PMB: increases in material incentives to conform to a practice tend to crowd in moral motivations to conform.

## 6. Existing experimental evidence

The voluntary interaction in a Mutual Benefit Game is a Dictator Game. However, the Mutual Benefit Game has two features which rarely appear in combination in Dictator Game experiments, namely that the game is voluntary for both players, and that it is played recurrently and anonymously by pairs of individuals drawn from a larger population. In this section, we review evidence from experiments that use one-shot or recurrent Dictator Games that are at voluntary for at least one of the players. In certain respects, these experimental environments are similar to the target domain of our model. By virtue of these similarities, the findings of these experiments can provide preliminary indications of the explanatory potential of the PMB approach.

### *6.1. Promise-keeping*

There is strong experimental evidence that people incur costs to keep cheap-talk promises (e.g. Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2007). Charness and Dufwenberg's design is effectively a one-shot Dictator Game that is voluntary for Player 1 but not for Player 2 (i.e., it is the Mutual Benefit Game as shown in Figure 2 but without

Player 2's reset node); the payoffs satisfy $x_1 > 0 > y_1$ and $y_2 > x_2 > 0$.[20] In one treatment, Player 2 was allowed to send a free-form message to Player 1 before the game began. This increased the frequency of the $\langle in, X\rangle$ path of play: Player 2s sent promises to choose *X*, Player 1s believed these promises, and Player 2s kept them, contrary to self-interest.

Charness and Dufwenberg explain this as an effect of guilt aversion: Player 2 has a preference for confirming Player 1's expectation of benefit. However, later experiments that have replicated Charness and Dufwneberg's results have shown that people do not have a similar willingness to confirm expectations induced by other people's promises (Vanberg, 2008) or by random devices (Ederer and Stremitzer, 2017). A possible explanation, aligned with Implication 5 of our model, is that promise-keeping is a practice in the population from which subjects are drawn, and that Player 2's action of promising to choose *X* is equivalent to an *in* action in the initiation procedure of an IVI game. If the obligation to choose *X* derives from PMB, it arises only if the Dictator Game is voluntary.[21]

### *6.2. Avoiding the ask*

The phenomenon of 'avoiding the ask' has been observed in the laboratory by Dana et al. (2006) and Lazear et al. (2012) and in the field by DellaVigna et al. (2012) and Andreoni et al. (2017). We focus on Dana et al.'s seminal experiment. This is a one-shot Dictator Game that is voluntary for Player 2 (the dictator) but not for Player 1. If Player 2 chooses *in*, she can choose any integer division between herself and Player 1 of a total material payoff of 10. In the *standard* treatment, Player 1 then learns that this game is being played; in the *private* treatment, he does not. If Player 2 chooses *out*, the outcome is (0, 9), and Player 1 never learns that he has been involved in a game. In the standard treatment, 20 out of 61 Player 2s chose *out*; 16 of these subjects (the 'reluctant sharers') would have shared if the outside option had not been available. In the private treatment, in contrast, only one out of 14 Player 2s chose *out*. Dana et al. interpret their results as showing that 'a concern for not appearing selfish can sometimes motivate generosity, but a kind of generosity that participants would

[20] This equivalence holds after normalising Charness and Dufwenberg's payoffs and substituting expected values for risky prospects.

[21] Ederer and Stremitzer (2017) amend Charness and Dufwenberg's guilt-aversion hypothesis, proposing that an individual feels guilt only if the expectation they disconfirm was a reasonable inference from an explicit or implicit promise that they had intentionally made. Our model generalises this idea.

rather have avoided' (p. 195). Lazear et al. use a similar design, find similar results, and draw similar conclusions.

Implication 3 of our model suggests an alternative explanation, namely that Player 2 is *following* a norm that prescribes sharing in Dictator Games that are entered voluntarily. As an analogy, consider the Mutual Benefit Game, interpreted in terms of taxi rides and tips. The practice $\langle X \rangle$ is that the passenger (Player 2) tips the driver (Player 1). Consider an episode in which the benefit that Player 2 would get from the ride is more than the metered fare but less than the fare plus the tip (i.e., $x_2 < 0 < y_2$). If Player 2 adheres to PMB, she will not take the ride. In doing this she is acting against her self-interest, just as reluctant sharers do. But she is not secretly violating a norm that tells her to share her wealth with taxi drivers; she is choosing *not* to violate the norm of paying the customary amount for a service.

### *6.3. Gift exchange*

Building on Akerlof's (1982) model of 'partial gift exchange' in labour markets in which employers cannot fully monitor workers' effort, experimental economists have investigated Gift Exchange Games in which a subject in the role of 'employer' sets a wage rate and then a 'worker' who knows she will be paid this wage chooses an effort level. A common finding is that employers offer wages above workers' reservation levels and that workers respond by exerting more effort than self-interest requires. For example, Fehr and Falk (1999) implement a sequence of anonymous double-auction markets in which employers make bids to buy labour and workers make offers to sell it; after a bilateral contract has been made, the worker chooses how much (costly) effort to perform. In this design, the worker's post-contract role corresponds with that of Player 2 in our model, i.e., the active player in a fully voluntary Dictator (sub)Game. Fehr and Falk find that many workers perform more effort if they are paid more, even though this is contrary to their self-interest. Wages are higher than they would be in a market of self-interested agents but, given the actual behaviour of workers, employers act according to *their* self-interest. This suggests an asymmetry between trust and trustworthiness, as in the Paradox of Trust described in Section 3.1: the player who trusts (i.e., the employer) acts in his self-interest, but the player who returns that trust (i.e., the worker) acts contrary to hers. If workers view high effort as a practice, this pattern is similar to that described by Implications 2 and 5 of our model.

Discussing a similar asymmetry, Bicchieri et al. (2011) report a norm-elicitation experiment based on Berg et al.'s (1995) original Trust Game design. In Bicchieri et al.'s experiment, each participant was asked to consider various hypothetical combinations of decisions in the Trust Game and asked whether, if given the opportunity, they would choose to impose a 'payoff cut' on a specific player. Each participant reported guesses about the distribution of responses to this question and was paid for correct guesses. Interpreting these guesses as evidence about social norms, Bicchieri et al. report that punishment was typically judged appropriate for untrustworthy Player 2s, but not for untrusting Player 1s; they conclude that 'trustworthiness is a social norm, but trust is not'. Smith and Wilson (2019: 149–156) find similar patterns in actual behaviour in experimental Trust Games with costly punishment options.[22]

### *6.4 Crowding in*

List (2006) reports a field experiment designed to investigate the trustworthiness of dealers at sports card shows. In the 'Tickets' treatment, dealers were approached by experimental assistants who acted as buyers, specifying a particular type and quality of ticket and offering to pay a predetermined amount of money for this; the offer was a reasonable price for the requested quality. Whatever tickets the dealers supplied were bought without question and then graded by experts. This treatment compared behaviour before and after the public announcement of the introduction of a professional service of third-party grading of tickets. List separates dealers into two groups – those who had some scope for building reputations ('locals') and those who did not ('nonlocals'). Prior to the announcement of the introduction of third-party grading, both groups supplied low quality tickets irrespective of the quality requested and paid for. After the announcement, nonlocals behaved much as before, but locals behaved more honestly. List interprets these results as evidence in favour of the hypothesis that honesty in markets results from traders' self-interested concerns for reputation and not from social preferences for gift exchange.

The structure of this experiment is similar to our model of a Mutual Benefit Game, with Player 1 as the buyer, Player 2 as the dealer, and $X$ and $Y$ respectively as the delivery of the specified and inferior qualities. The introduction of third-party grading is equivalent to a

[22] In a related experiment, Smith and Wilson (2018) study a one-shot Ultimatum Game that is preceded by a node at which the Responder can take an outside option that is slightly better for both players than the outcome of a rejected offer. In the most frequent path of play, the Responder enters, the Proposer offers an unequal distribution that favours them, and this is accepted. Responders may have been willing to accept this inequality because they entered voluntarily, expecting and implicitly consenting to it.

downward shift in local dealers' payoff parameter distributions. Thus, to the extent that ⟨*X*⟩ was understood to be the practice in this market, our Implication 7 is applicable: the observed increase in honesty might be the result of two complementary effects – an increase in material incentives to honesty and crowding-in.

## 7. New experimental evidence

In this section, we report a new experiment which investigates behaviour in a setting that is similar to the model analysed in Sections 4 and 5, but is explicitly framed in terms of a potential market transaction.

### *7.1. Experimental design*

In our theoretical analysis of the Mutual Benefit Game, actions were given neutral labels (such as '*in*', '*out*', '*X*' and '*Y*') and the distributions of payoff parameters were described abstractly. However, if the implications of a theoretical model are to be interpreted as hypotheses about general tendencies in the real world, a useful experimental test of those hypotheses may require some translation from abstract theoretical concepts to concrete realities that experimental participants may easily relate to. Since we are presenting our work as a contribution to the 'markets and morals' literature, we chose a design with an explicit market frame.

Building on an intuitive idea introduced in Section 4, the experimental interaction was framed as an exchange transaction which takes place only if both players choose to enter, and in which one player (the 'second mover') may be able to gain by cheating. To give a concrete representation of the opportunity to cheat, we presented it as a specific action (not completing the exchange) available to the second mover with a specified probability. By letting the second mover, but not the first, know whether this opportunity was available before making their entry decision, we reproduced the trust-and-trustworthiness logic of the Mutual Benefit Game. In every interaction, each player's decisions involved either trust (as in the case of Player 1 in the Mutual Benefit Game) or trustworthiness (as in the case of Player 2).

The core component of our design is the following *Market Trust Game*, which we initially describe in one-shot form. The background story is that Owner owns a good that has zero value to them, but has a positive *exchange value*, *e*, to Buyer. The value of *e* is known to Buyer but not to Owner. There is an exogenous *trade price t*, known to both players. For an exchange to take place, Buyer must send the trade price to Owner, and Owner must pay a *delivery charge d* (a deadweight loss) to send the good to Buyer. The value of $d$ is known to Owner but not to Buyer. The values of $d$, $t$ and $e$ are independent random draws from (respectively) the following sets of money amounts (in UK pounds), which are common knowledge:

$D$ = {£0.50, £1.50, £2.50, £3.50, £4.50};

$T$ = {£3.00, £4.00, £7.00, £10.00, £11.00};

$E$ = {£9.50, £10.50, £11.50, £12.50, £13.50}.

For each player, the *profit* from exchange ($t – d$ or $e – t$ for Owner and Buyer respectively) may be positive or negative (but not zero); the *surplus* ($e – d$), i.e., the sum of the profits, is always strictly positive. With strictly positive probability, exchange is mutually beneficial.

At the start of the game, each player makes an *entry* decision, choosing either *in*, indicating willingness to enter a *transaction* with the other, or *out*. These decisions are made simultaneously. If either chooses *out*, the game ends and each player gets their *outside option*, with a payoff of zero. Otherwise, one of the following *transaction procedures* is implemented:

*Move together*: Owner is required to *send* (i.e., pay $d$ to send the good to Buyer). Simultaneously, Buyer is required to *send* (i.e., pay $t$ to Owner). The payoff profile (with Owner's payoff shown first) is ($t – d$, $e – t$).

*Owner first*: Owner is required to *send*. Buyer then chooses either *send*, with payoff profile ($t – d$, $e – t$), or *keep*,[23] with payoff profile ($–d$, $e$).

*Buyer first*: Buyer is required to *send*. Owner then chooses either *send*, with payoff profile ($t – d$, $e – t$), or *keep*, with payoff profile ($t$, $–t$).

The *move together* procedure models a market in which contracts are fully enforceable. The *Owner first* and *Buyer first* procedures model trading situations in which one party (Buyer or Owner respectively) has an opportunity to gain by cheating the other. Given the distributions of $d$, $t$ and $e$, being cheated always results in a negative payoff. We will say that a player is a *first mover* (he) in any transaction in which he is required to *send* (i.e., Owner in *Owner first*, Buyer in *Buyer first*, or either player in *Move together*). Otherwise (i.e., Owner in *Buyer*

[23] In the experiment, this action was called 'Do not send'.

*first*, or Buyer in *Owner first*) the player (she) is a *second mover* and makes a *completion* decision between *send* and *keep*. We will sometimes use the term 'cheating' as a shorthand for 'choosing *keep* as a second mover', and 'trading honestly' to indicate 'choosing *send* as second mover' (neither term was used in the experiment).

The transaction procedure is determined randomly. *Move together* occurs with probability $v$, where $0 < v < 1$; each of the other procedures occurs with probability $(1 - v)/2$. The value of $v$ is common knowledge. However, when players make their entry decisions, each player knows only whether they are a first mover (i.e., with no opportunity to cheat, but matched with a second mover who can cheat with probability $(1 - v)/(1 + v)$) or a second mover (i.e., sure to have the opportunity to cheat with no risk of being cheated). After the entry decisions, the actual transaction procedure is revealed to both players.

If there was common knowledge that the *Move together* procedure was operative, the roles of Owner and Buyer would be isomorphic.[24] However, the possibility of cheating introduces an asymmetry between those roles. In *Owner first*, the initiation of the interaction creates surplus; cheating by Buyer affects how that surplus is distributed. In *Buyer first*, surplus is created only if Owner does not cheat.

The game was played recurrently for 60 periods within anonymous groups of four experimental participants, so that norms could emerge spontaneously through experiential learning. Each participant faced a series of *episodes*, in each of which they played with a randomly selected other member of their group, in one of two *markets*. In each episode, the two players were randomly assigned to the roles of Owner and Buyer. The values of the parameters $e$, $d$ and $t$, and the transaction procedure were determined independently for each episode. The two markets differed in only one respect: $v$ took the value 0.7 in the *high-reliability* market and 0.3 in the *low-reliability* market.[25] Thus, opportunities for cheating were more frequent in the low-reliability market. For each participant, the sequence of

[24] Both players have the same outside option (zero). If $t$ = £7.00, both players have the same expected profit (£4.50). If $t$ = £4.00, Owner's expected profit is £2.50 and Buyer's is £5.50. Symmetrically, if $t$ = £10.00, Owner's expected profit is £5.50 and Buyer's is £2.50. There is a corresponding symmetry between $t$ = £3.00 and $t$ = £11.00.

[25] In the experiment, the distinction between the two markets was achieved using two different picture styles, associated with two fictitious artists, Circlinsky and Trianglich. Circlisnky pictures were randomly generated using coloured circular patterns; Trianglich pictures were randomly generated using coloured triangular patterns. Which picture type was associated with the high-reliability market was counterbalanced across sessions.

episodes included games in both markets, interspersed with one another and presented in random order.[26] In any given episode, the value of $v$ was common knowledge.

By the end of each episode, each player knew which moves their coplayer had made in that episode, but did not know the coplayer's payoff. Participants received no information about moves or payoffs in episodes in which they were not players. Over time, however, each participant accumulated experience of how their coplayers (considered in general) had behaved, and could potentially break down that experience according to whether coplayers were Owners or Buyers, whether they were first or second movers, and whether the value of $v$ was high or low. Thus, there was scope for individuals to form experience-based beliefs about the behaviour of subsequent coplayers, and thereby for practices of cheating or non-cheating to emerge.

Full details of the experimental procedures can be found in the Supplemental Appendix.

### *7.2. Testable hypotheses*

In this design, a first mover chooses whether to enter an exchange relationship with a coplayer, not knowing whether the coplayer will have an opportunity to gain by cheating him. Thus, a first mover's decision problem is analogous with that of Player 1 in the Mutual Benefit Game. A second mover chooses whether to enter an exchange relationship with a coplayer, already knowing that she will have an opportunity to gain by cheating him. Her decision problems are analogous with those of Player 2 in case C of the Mutual Benefit Game (see Section 4) if her profit from exchange is positive (i.e., *in/send* is second-best) or in case D if it is negative (i.e., *out* is second-best).

Using the term 'practice' in an intuitive sense, there is a potential practice of 'honest trade', namely that, conditional on the initiation of the interaction, Owner and Buyer both take their *send* actions. Irrespective of which transaction procedure applies, the outcome of these actions is that Owner receives the trade price minus the delivery charge and Buyer receives the exchange value minus the price. We will say that these are the outcomes of

[26] In each episode, there was a 30 percent chance of $v = 0.7$ and a 70 percent chance of $v = 0.3$, equalising across the two markets the expected number of episodes in which one of the players was a second mover. This within-subject design controls for income effects. In a between-subjects design, the high-reliability market would tend to induce higher expected earnings, which might affect first movers' willingness to risk being cheated and second movers' willingness to forgo opportunities to cheat.

*completed exchange*. The transaction is voluntary in the sense that it takes place if and only if both players choose *in*. If honest trade is in fact the practice, the transaction is voluntary in a further sense. When making their entry decision, each player knows that *out* will certainly lead to their outside option and that, if both players conform to the practice, *in* will lead either to the outside option or to completed exchange. Thus, *in* can be interpreted as an implicit promise to trade honestly, conditional on the other player doing the same. In the revealed-preference sense explained in Section 2, this practice is mutually beneficial.

Formally, the experimental game does not fully satisfy the properties of an IVI game as defined in Section 1, because it contains chance nodes. However, as we show in the Supplemental Appendix, it is possible to generalise the definition of an IVI game to include chance nodes, and to generalise the definition of a practice in a corresponding way. In terms of these definitions, the Market Trust Game is a 'generalised' IVI game and 'Owner *sends*, Buyer *sends*' is a potential practice. For the purposes of the experiment, we treat the high-reliability and low-reliability markets as distinct games, each with its own potential practice of honest trade. In each market, we define the rate of conformity to this practice (the *honest trading rate*) as the ex ante probability that, in a randomly selected episode in which both players choose *in*, both players take the action *send* (either by necessity as a first mover or by choice as a second mover).

Although the implications of our model were formulated in terms of equilibrium states of processes of experiential learning in large populations, for experimental purposes it is more useful to focus on how a model represents *individual* behaviour, irrespective of whether an equilibrium is reached. Thus, hypotheses about players' behaviour are best formulated in relation to *empirical optimality* (Weizsäcker, 2010). In an environment of recurrent play, a player's strategy is empirically optimal if it is a best response to the (relative) frequencies of the strategy choices of her potential coplayers. As we will show, weak assumptions about what players can infer from their experience of the game are sufficient to generate sharp implications about the behaviour of players who adhere to PMB.

Clearly, our design can test hypotheses about the prevalence of cheating only if first movers sometimes choose *in/send*, which is not guaranteed as the game has a (subgame-perfect) equilibrium in which first movers always play *out* and second movers always play *in/keep*.[27] We chose the two values of *v* in the expectation that both would induce sufficient *in/send* choices by first movers to allow us to investigate second-mover behaviour – an expectation that proved correct.

Varying *v* is important for a test for a crowding-in effect. In principle, there is crowding in if, other things being equal, second movers are more likely to conform to the honest trading practice when they believe the honest trading rate to be higher. However, given the small size of the experimental groups, there is a danger of circularity in that formulation, because each player's decisions can affect the honest trading rate of their group, both directly (as constituents of that rate) and indirectly (by affecting other group members' subsequent behaviour). To avoid this problem, we treat 'trading in the high-reliability market' as a dummy variable that serves as a proxy for factors that are outside the control of individual players and that might influence their beliefs about how the honest trading rate differs between the two markets. Our design was premised on the expectation that the overall honest trading rate would be higher in the high-reliability market. This is *Precondition 1*.

A similar method can be used in testing whether first movers act on self-interest. Other things being equal, a self-interested first mover is more likely to choose *in/send* the higher the *completion probability* he assigns to his coplayer playing *send*, conditional on the latter playing *in*. 'Trading in the high-reliability market' can serve as a proxy for factors that might influence this belief. Our prior expectation was that, overall, observed *completion rates* (i.e., the empirical counterparts of completion probabilities) would be higher in the high-reliability market. This is *Precondition* 2. Preconditions 1 and 2 are expectations that do not depend on players' motivations; they are based on the higher proportion of first movers in the high reliability market and the fact that first movers cannot cheat.

Under the assumption that Preconditions 1 and 2 are in fact satisfied, and that players adhere to PMB and make empirically optimal choices, we derive seven hypotheses that closely parallel Implications 1 to 7 for the Mutual Benefit Game. We distinguish player roles by three independent binary criteria: whether a player is a first (F) or a second mover (S), whether they are Owner (O) or Buyer (B), and whether their profit is positive (+) or negative

[27] We did not use the 'strategy method' of eliciting conditional strategy choices, since that could have produced confounding effects. If the consequences of an individual's decision are conditional on an uncertain event, the moral or emotional salience of the choice problem may be attenuated, for example, by creating 'moral wiggle room' (Dana et al., 2007), or by weakening desires to punish bad behaviour (Brandts and Charness, 2011). Eliciting only direct responses also made the experimental tasks more transparent to participants.

(–). (For example, ‘FO$^+$’ denotes ‘first-mover Owner with positive profit’.) For each role, we present hypotheses about the proportion of the occupants of that role who play specific strategies. We derive these hypotheses by assuming that players act on self-interest whenever this does not involve cheating. In defining ‘self-interest’, we do not impose any restriction on attitudes to risk, but take account of unambiguous effects of variation in *d*, *t*, *e* and *v* (interpreted as a proxy for expectations about completion) on the players’ material incentives.[28] We assume that, for each participant *i*, the moral cost of cheating (expressed in units of material payoff) is non-negative; it is given by a weakly increasing function $m_i(\rho)$, where ρ is the honest trading rate in the relevant market.

Our hypotheses are as follows:

**Hypothesis 1:** (i) The proportion of FO$^+$ **(**FB$^+$**)** players choosing *in* is higher at lower values of *d* (higher values of *e*); and (ii) these proportions are higher at the higher value of *v*.[29]

**Hypothesis 2:** FO$^-$ and FB$^-$ players do not choose *in*.

**Hypothesis 3:** SO$^+$ and SB$^+$ players choose *in*.

**Hypothesis 4:** The proportion of SO$^-$ **(**SB$^-$) players choosing *in* is higher at higher values of *t* (higher values of *e*).

**Hypothesis 5:** The proportion of SO$^+$ (SB$^+$) players choosing *send* after entry is higher at lower values of *d* (lower values of *t*).

**Hypothesis 6:** SO$^-$ and SB$^-$ players do not choose *send* after entry.

**Hypothesis 7:** (i) The proportions of SO$^-$ and SB$^-$ players choosing *in* are lower at the higher value of *v*; and (ii) the proportions of SO$^+$ and SB$^+$ players choosing *send* after entry are higher at the higher value of *v*.

[28] In the theoretical analysis presented in Section 5, it was convenient to use risk neutrality as a modelling simplification. But if risk-averse or risk-loving behaviour is found in an empirical test, that in itself should not count as evidence against PMB.

[29] Hypothesis 1 does not refer to *t* because its effect on first movers is ambiguous. For an FO$^+$ player, for example, higher values of *t* imply higher profit, but they also imply lower profit for the coplayer, which may affect the completion probability. (This problem will not arise for Hypotheses 4 and 5 because, in our model, second movers’ strategy choices depend only on their own material payoffs and the conformity rate in the relevant market.)

These hypotheses allow two forms of non-self-interested behaviour by second movers (for whom self-interest always dictates *in/keep*): choosing *out* when profit is negative (Hypothesis 4), and choosing *in/send* when profit is positive (Hypothesis 5). But they rule out all cases in which a player takes the action *in* when self-interest dictates *out* (Hypothesis 2), or in which a second mover chooses *in/send* when *out* would give a higher material payoff (Hypothesis 6). Hypothesis 7 states that adherence to PMB is crowded in by an increase in *v*.

### *7.3 Results*

We recruited 204 participants from the general population of the University of East Anglia. The experiment took approximately 80–90 minutes to complete, with an average payment of £12.70.

Table 2 shows an overall summary of the Market Trust Game episodes in the two markets, with a breakdown based on the type of transaction procedure.

*[Insert Table 2 here]*

There were a total of 6,120 episodes, 4,261 (69.6 percent) in the high-reliability market and 1,859 (30.4 percent) in the-low reliability market. Overall, 1,341 episodes (31.5 percent) resulted in an interaction being initiated in the high-reliability market, and 368 (19.8 percent) in the low-reliability market. Since in both markets interactions were initiated more often for the *Buyer first* or *Owner first* than for the *Move together* transaction procedure, the data provide extensive opportunities to study the behaviour of second movers.

In line with Precondition 1, the honest trading rate was much higher in the high-reliability market (70.2 percent) than in the low-reliability market (42.7 percent), $p < 0.001$ in a two-tail test of proportions.

In line with Precondition 2, the overall completion rate was much higher in the high-reliability market (0.808) than in the low-reliability market (0.498), $p < 0.001$ in a two-tail test of proportions.[30]

Our seven hypotheses concern the entry decisions by all players and second movers' completion decisions in the interactions that are initiated.

[30] The rates reported in Table 2 aggregate over Owners and Buyers, and over different values of the trade price. In principle, players could condition their decisions on these aspects. Completion rates conditional on player role and trade price are reported in the Supplemental Appendix.

Table 3 reports a summary of the entry decisions in the two markets – i.e., the number and proportion of first and second movers choosing *in* – conditional on whether they were Owners or Buyers, and on whether the potential profits from honest trade were positive or negative.

*[Insert Table 3 here]*

Table 4 reports the number and proportion of positive- and negative-profit second movers who chose *send* when the interaction was initiated.

*[Insert Table 4 here]*

Table 5 reports the estimated marginal effects from a set of repeated-measure random-effect logistic regressions that investigate the relationship between entry and completion decisions and the parameters that were exogenously manipulated in the experiment ($d$, $t$, $e$ and $v$).[31] To match the statements of our hypotheses, we conduct separate regressions conditioning on whether players were first or second movers, whether they were Owners or Buyers, and whether their profits from honest trade were positive or negative. Each player could condition their decision on the *revenue* from trading ($t$ for Owners, $e$ for Buyers), on the corresponding *cost* ($d$ for Owners, $t$ for Buyers), and on whether they were trading on the high- or low-reliability market (with low-reliability as the base case). These variables are included as regressors. In order to allow for time effects, we also control for the experimental Period (1 to 60).

*[Insert Table 5 here]*

**Entry decisions by first movers.** Hypothesis 1 concerns the entry decisions by first movers with positive profits.

According to part (i) of that hypothesis, Owners should be more likely to enter the interaction when their cost of trading ($d$) is lower, and Buyers when their revenue ($e$) is higher.[32] The significantly negative effect for the cost variable in model (1) and the

[31] Our main regressions use the `xtlogit` command in Stata. As a robustness check, in the Supplemental Appendix we report the corresponding results obtained using simple logit regressions with standard errors clustered at the group level. The two sets of models always produce effects in concordant directions, although the strength and statistical significance of those effects sometimes vary.

[32] Because of the correlation introduced by the common value of $t$, the predictions for the effects of the Owner's revenue and the Buyer's cost are ambiguous (see footnote 29).

significantly positive effect for the revenue variable in model (2) in Table 5 confirm that both trends were observed ($p < 0.01$ in both cases), strongly supporting the hypothesis. Revenue had also a significantly positive effect on Owners' decision to enter the interaction, and cost a significantly negative effect on Buyers'.

According to part (ii) of Hypothesis 1, first movers should be more likely to enter the interaction at higher values of $v$, i.e., in the high-reliability market. Table 3 clearly shows that first movers were indeed more likely to enter in that market (between 50 and 70 percent of the cases compared to 30–45 percent in the low-reliability market). The strongly significant positive effects for the high-reliability market dummy in models (1) and (2) confirm that this hypothesis is strongly supported for both Owners and Buyers ($p < 0.01$ in both cases).

In line with Hypothesis 2, Table 3 shows that first movers virtually never chose *in* when profits were negative. Across the two markets and the two roles, this never happened more than three percent of the time.[33]

**Entry decisions by second movers.** According to Hypothesis 3, positive-profit second movers should always choose *in*. Table 3 shows that this happened in more than 90 percent of the cases in both markets and for both player roles, providing strong support for that hypothesis.[34]

Hypothesis 4 describes a distinctive form of non-self-interested behaviour: second movers with negative profit, who would benefit from entering and cheating, may choose *out* if their moral cost is sufficiently large. In Table 3, we find ample evidence of this behaviour: overall, between 50 and 59 percent of second movers chose *out* when their profits were negative. Models (3) and (4) in Table 5 test the hypothesis that these players should choose *in* more often when their revenue ($t$ for Owners and $e$ for Buyers) is higher. Model (4) shows that this was the case for Buyers (p < 0.01), for whom there was also a tendency to enter the interaction less often when their cost was higher ($p < 0.01$). However, the hypothesis is not

[33] Table S3a in the Supplemental Appendix shows that, for both Owners and Buyers, these instances were concentrated in the first 20 periods of the experiment, and were extremely rare in the last 20 periods. The most plausible interpretation is that these were occasional mistakes participants learnt to avoid as the experiment progressed.

[34] There was a tendency, for both $SO^+$ and $SB^+$, to choose *in* more often in the last third of the experiment (94.1 and 92.8 percent of the episodes respectively) than in the initial third (89.9 and 89.5 percent) – see Table S3a in the Supplemental Appendix.

satisfied for Owners. Model (3) shows a marginally significant effect in the opposite direction ($p < 0.1$): second-mover owners with negative profit were less likely to enter the interaction when their revenue was higher.

**Completion decisions by second movers.** According to Hypothesis 5, positive-profit second movers may act contrary to self-interest by choosing *send* in an interaction if the moral cost is sufficiently high. In Table 4, we see that this form of non-self-interested behaviour was observed between 32 and 42 percent of the time. Because of the trade-offs between self-interest and morality built into our model, there should be a negative relationship between the cost of honest trade ($d$ for Owners, $t$ for Buyers, i.e., their temptation to cheat) and the likelihood of choosing *send*. This prediction can be tested with the aid of models (5) and (6), which find support for it for both player roles ($p < 0.01$).

Hypothesis 6 concerns the behaviour of negative-profit second movers in interactions that are initiated. Because choosing *out* has no moral cost, the prediction is that those who enter would do so to cheat by choosing *keep*. Table 4 shows that this prediction is strongly supported. Out of the 41 interactions that were initiated when second movers' profits were negative, there were just 3 *send* decisions.

**Crowding in.** In our design, the crowding in prediction encapsulated in Hypothesis 7 can be tested by comparing the behaviour of second movers in the two markets. According to part (i) of that hypothesis, negative-profit second movers should choose *in* less often the high-reliability market. Table 3 shows that they were equally likely to enter in the two markets, a finding corroborated by the non-significant marginal effect of the high-reliability market dummy in models (3) and (4). According to part (ii), positive-profit second movers should choose *send* more often in the high-reliability market. Again, Table 4 shows that, for both Owners and Buyers, *send* decisions were equally likely in the two markets, as confirmed by the non-significant marginal effect of the high-reliability market dummy in models (5) and (6). This prediction is not supported.

**Learning**. We find evidence of learning over the course of the 60 rounds. The period variable in Table 5 shows that, during the experiment, positive-profit first movers became progressively less likely to choose *in* ($p < 0.01$ in models (1) and (2)),[35] negative-profit second movers more likely to enter to cheat ($p < 0.01$ in models (3) and (4)),[36] and positive-profit second movers less likely to choose *send*.[37] Although both forms of non-self-interested behaviour by second movers showed some tendency to decay over the course of the experiment, they remained very common after dozens of rounds.

**Heterogeneity.** By looking at the relationship between individuals' tendencies to behave contrary to self-interest as second movers, we see potential signs of heterogeneity. Choosing *out* with negative profits correlated significantly with choosing *send* with positive profits (Spearman rho 0.24, $p < 0.01$). Although this may be an indication that participants differed in their non-self-interested inclinations, it may also be partly due to heterogeneity in their experiences. Both forms of behaviour were correlated with choosing *in* as positive-profit first movers (rho 0.23 and 0.35 respectively, $p < 0.01$ in both cases); and all three behaviours correlated with players' experiences of being cheated as first movers (rho 0.41, $p < 0.01$ for *in* choices by positive-profit first movers; rho 0.18 and 0.20 for *in* decisions by negative-profit second movers and *send* decisions by positive-profit second movers respectively, both $p < 0.05$). These correlations are consistent with the possibility, compatible with our model, that a common belief about the prevalence of honest trading guided players' decisions as both first and second movers.

### *7.4 Unexpected findings*

The most significant unexpected finding is the lack of evidence of crowding in. The behaviour of positive-profit first movers demonstrates that participants recognised the difference between the two markets. Why did we see no evidence of crowding in? One possibility is that, given the distinct market framing of the task, participants may have applied a single, home-grown norm in both markets. But a pre-existing norm seems incompatible with the evidence of learning and heterogeneity discussed in Section 7.3. A more convincing alternative is that a single norm of honest trade emerged in the experiment as a whole. In

[35] Table S3a in the Supplemental Appendix shows that Owners entered in 67.7 percent of the episodes in the first 20 periods, and just in 60.7 percent in the last 20. Buyers entered 56.8 percent of the first 20 episodes and 40.0 percent of the last 20.

[36] In the first 20 periods, $SO^{-}$ (respectively, $SB^{-}$) did so in 34.7 (45.8) percent of the cases, which went up to 46.0 (52.7) percent in the last 20 periods. See Table S3a.

[37] $SO^{+}$ chose *send* in 55.1 percent ($SB^{+}$ in 40.1 percent) of the cases in the first 20 periods, dropping to 31.1 percent (26.2) in the last 20 periods. See Table S3b.

other words, participants interpreted 'Both players choose *send*' as a single practice, to be followed in both markets, conditional on the initiation of the voluntary interaction. If the practice is defined in this way, the rate of conformity was 0.64.

While the within-subject nature of our design may have limited the scope for observing crowding in, it is not obvious that a between-subject design would have been preferable. As explained in footnote 27, this design feature was chosen to control for income effects. Identifying the right setup, experimental or otherwise, to test this distinctive prediction of our model is an important challenge that remains open for further research.

Our model implies that entry decisions by negative-profit second movers are independent of cost (because, whether the player cheats or not, the cost is not incurred) and that completion decisions by positive-profit second movers are independent of revenue (because, whether the player cheats or not, the revenue is received). However, Table 5 shows that there were strong tendencies for SB$^{-}$ players to be more honest when cost (*t*) was higher, and for SO$^{+}$ players to be more honest when revenue (*t*) was higher. There was no corresponding effect for SO$^{-}$ players (for whom cost was *d*), or for SB$^{+}$ players (for whom revenue was *e*). This pattern may be the result of a motivating factor that the current form of our model excludes. In our model, the moral cost that a second mover incurs by cheating is independent of her beliefs about the consequences *for the other player* of being cheated. For an SB$^{-}$ player who cheats, *t* is a payment that Owner has trusted her to make, but which she fails to make. For an SO$^{+}$ player who cheats, *t* is a payment that Buyer has made to her in the expectation that she will send the good – which she fails to do. In each case, the amount of this payment is common knowledge between the players. In contrast, the second mover's value of *d* or *e* provides no information about the first mover's payoffs. A plausible conjecture is that the perceived moral cost of cheating was higher at higher values of *t*.[38] A possible generalisation of our model would be to assume that the moral cost of violating PMB is weakly increasing in the losses that the violation imposes on co-players.

Finally, a minor unanticipated regularity in our data. The similar size of the marginal effects for the revenue and cost variables in models (1) and (2) implies that, at given values of revenue and cost, first-mover Owners and Buyers behaved in a similar way. It seems that first movers did not recognise that, because of the correlation induced by the common value

[38] This conjecture might also help to explain the negative marginal effect for revenue in model (3).

of *t*, when entering was more profitable for them second movers were more likely to enter to cheat.

## 8. Conclusion

Recent theoretical and experimental research has identified a range of non-self-interested motivations – for example, reciprocal kindness and unkindness, guilt aversion and inequality aversion – that are revealed in many forms of human behaviour. This research programme has also raised questions about whether market environments tend to inhibit such motivations. We have argued that this literature has failed to recognise the distinction between two kinds of interaction – those that are voluntary, taking place only with the prior consent of all participants, and those that are not.

The distinction is important because there may be norms of behaviour that are specific to voluntary interactions. We have characterised one such norm – the Principle of Mutual Benefit – which prescribes that people follow mutually beneficial practices in interactions that they enter voluntarily, as long as sufficiently many others do likewise. We have analysed, theoretically and experimentally, the implications of this norm in the context of games in which two players may or may not benefit relative to their outside options from entering a voluntary interaction and following an established practice within it.

Our theoretical analysis has shown that people who adhere to the principle display a distinctive combination of non-self-interested and self-interested behaviours: they act according to self-interest when deciding whether to take part in a voluntary interaction, but follow a practice within it even if this is contrary to their self-interest. Because one's obligation to follow a practice derives from the observation of the behaviour of others without scrutiny of their intentions, non-self-interested behaviour may be crowded in by material incentives to follow a practice. Our experimental work suggests that observing crowding in may not be straightforward in the laboratory, but it does provide additional evidence of the systematic effects of voluntariness which align well with our theoretical analysis.

Since market transactions are paradigm cases of voluntary interaction, our approach may be better able to explain trust and trustworthiness in markets than approaches that treat self-sacrifice as the fundamental form of non-selfish motivation, or that treat apparently non-self-interested behaviour as reluctant conformity to social norms. More fundamentally, our approach calls into question the deep-rooted idea that markets are hostile environments for any motivations that are not based on self-interest.

## References

Akerlof, George (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97: 543–569.

Anderson, Elizabeth (1993). *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Andreoni, James, Justin Rao and Hannah Trachtman (2017). Avoiding the ask: a field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy* 125: 625–653.

Bacharach, Michael (2006). *Beyond Individual Choice.* Princeton: Princeton University Press.

Bacharach, Michael, Gerardo Guerra and Daniel Zizzo (2007). The self-fulfilling property of trust: an experimental study. *Theory and Decision* 63: 349–388.

Bartling, Björn, Roberto Weber and Lan Yao (2015). Do markets erode social responsibility? *Quarterly Journal of Economics* 130: 219–266.

Battigalli, Pierpaolo and Martin Dufwenberg (2007). Guilt in games. *American Economic Review: Papers and Proceedings* 97: 171–176.

Becker, Gary (1974). A theory of social interactions. *Journal of Political Economy* 82: 1063–1093.

Bem, Daryl (1967). Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74: 183–200.

Bénabou, Roland and Jean Tirole (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies* 70: 489–520.

Berg, Joyce E., John W. Dickhaut and Kevin McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122–142.

Bicchieri, Cristina (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Bicchieri, Cristina, Erte Xiao and Ryam Muldoon (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy and Economics* 10: 170–187.

Bolton, Gary and Axel Ockenfels (2000). ERC: A theory of equity, reciprocity and competition. *American Economic Review* 90: 166–193.

Brandts, Jordi and Gary Charness (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics* 14: 375-398.

Brekka, Kjell Arne, Snorre Kverndok and Karine Nyborg (2003). An economic model of moral motivation. *Journal of Public Economics* 87: 1967–1983.

Charness, Gary and Martin Dufwenberg (2006). Promises and partnership. *Econometrica* 74: 1579–1601.

Charness, Gary and Matthew Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117: 817–869.

Cialdini, Robert, Raymond Reno and Carl Kallgren (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58: 1015–1026.

Cubitt, Robin and Robert Sugden (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. *Economics and Philosophy* 19: 175–210.

Dana, Jason, Daylian Cain and Robyn Dawes (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100: 193–201

Dana, Jason, Roberto Weber and Jason Xi Kuang (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory* 33: 67–80.

DellaVigna, Stefano, John List and Ulrike Malemendier (2012). Testing for Altruism and Social Pressure in Charitable Giving. *Quarterly Journal of Economics* 127: 1–56.

Dufwenberg, Martin and Georg Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior* 47: 268–298.

Ederer, Florian and Alexander Stremitzer (2017). Promises and expectations. *Games and Economic Behavior* 106: 161–178.

Ellingsen, Tore and Magnus Johannesson (2004). Promises, threats and fairness. *Economic Journal* 114: 397–420.

Erev, Ido and Alvin Roth (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88: 848–881.

Falk, Armin and Urs Fischbacher (2006). A theory of reciprocity. *Games and Economic Behavior* 54: 293–315.

Falk, Armin and Nora Szech (2013). Morals and markets. *Science* 340 (10 May): 707–711.

Fehr, Ernst and Klaus Schmidt (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114: 817–868.

Fehr, Ernst and Armin Falk (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy* 107: 106–34.

Frey, Bruno (1994). How intrinsic motivation is crowded in and out. *Rationality and Society* 6: 334–352.

Gibbard, Allan (1973). Manipulation of voting schemes: a general result. *Econometrica* 41: 587–601.

Gilboa, Itzhak and David Schmeidler (1995). Case-based decision theory. *Quarterly Journal of Economics* 110: 605–639.

Gold, Natalie and Robert Sugden (2007). Collective intentions and team agency. *Journal of Philosophy* 104: 109-137.

Gneezy, Uri (2005). Deception: the role of consequences. *American Economic Review* 95: 384–394.

Gneezy, Uri and Aldo Rustichini (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics* 115: 791–810.

Heyes, Anthony (2005). The economics of vocation, or 'Why is a badly-paid nurse a good nurse?' *Journal of Health Economics* 24: 561–569.

Hart, Sergiu (1992). Games in extensive and strategic forms. In Sergiu Hart and Robert Aumann (eds), *Handbook of Game Theory, Volume 1*: 20–40. Amsterdam: Elsevier.

Isoni, Andrea and Robert Sugden (2019). Reciprocity and the Paradox of Trust in psychological game theory. *Journal of Economic Behavior and Organization* 167: 219–227.

Karpus, Jurgis and Mantas Raszvilas (2018). Team reasoning and a measure of mutual advantage in games. *Economics and Philosophy* 34: 1–30.

Katz, Eliakim and Femida Handy (1998). The wage differential between non-profit institutions and corporations: getting more by paying less? *Journal of Comparative Economics* 26: 246–261.

Lazear, Edward, Ulrike Malmendier, and Roberto Weber (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics* 4: 136–163

Levine, David (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1: 593–622.

Lewis, David (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

List, John (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy* 114: 1-37.

Lundquist, Tobias, Tore Ellingsen, Erik Gribbe and Magnus Johannesson (2009). The aversion to lying. *Journal of Economic Behavior and Organization* 70: 81–92.

Maskin, Eric and Tomas Sjöström (2002). Implementation theory. In Kenneth Arrow, Amartya Sen and Kotaro Suzumura (eds), *Handbook of Social Choice and Welfare, Volume 1*: 238–288. Amsterdam: Elsevier.

Moore, John and Rafael Repullo (1988). Subgame perfect implementation. *Econometrica* 56: 1191–1220.

Pelligra, Vittorio (2005). Under trusting eyes: the responsive nature of trust. In Benedetto Gui and Robert Sugden (eds), *Economics and Social Interaction*. Cambridge: Cambridge University Press: 195–124.

Rabin, Matthew (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.

Sandel, Michael (2012). *What Money Can't Buy: The Moral Limits of Markets*. New York: Farrar, Straus and Giroux.

Smith, Adam (1759/ 1976). *The Theory of Moral Sentiments*. Oxford: Oxford University Press.

Smith, Adam (1763/1978). *Lectures on Jurisprudence*. Oxford: Oxford University Press.

Smith, Adam (1776/ 1976). *The Wealth of Nations*. Oxford: Oxford University Press.

Smith, Vernon and Bart Wilson (2018). Equilibrium play in voluntary ultimatum games: beneficence cannot be extorted. *Games and Economic Behavior* 109: 452–464.

Smith, Vernon and Bart Wilson (2019). *Humanomics*. Cambridge: Cambridge University Press.

Sugden, Robert (1993). Thinking as a team: toward an explanation of nonselfish behavior. *Social Philosophy and Policy* 10: 69–89.

Sugden, Robert (2002). Beyond sympathy and empathy: Adam Smith's concept of fellow-feeling. *Economics and Philosophy* 18: 63-87.

Sugden, Robert (2018). *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

Sutter, Matthias, Jürgen Huber, Michael Kirchler, Matthias Stefan and Markus Walzl (2020). Where to look for the morals in markets? *Experimental Economics* 23: 30–52.

Titmuss, Richard (1970). *The Gift Relationship*. London: Allen and Unwin.

Vanberg, Christoph (2008): Why do people keep their promises? An experimental test of two explanations. *Econometrica* 76: 467–480.

Weizsäcker, Georg (2010). Do we follow others when we should? A simple test of rational expectations. *American Economic Review*, 100: 2340–2360.

## Figure 1 – The Taxi Ride Game

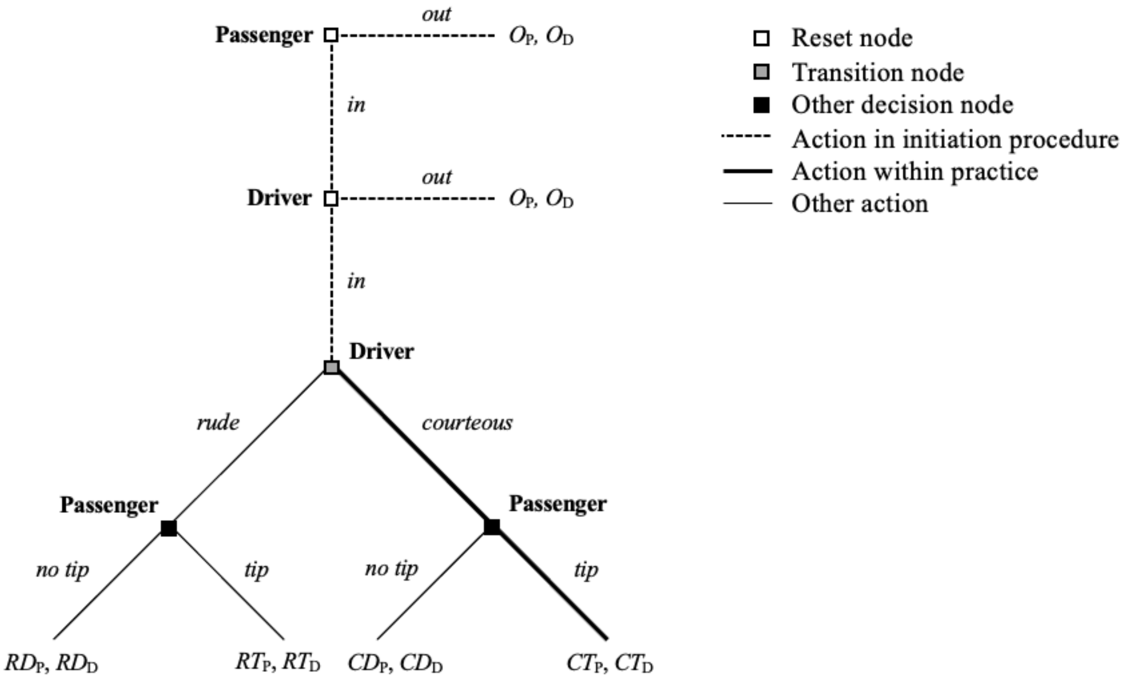
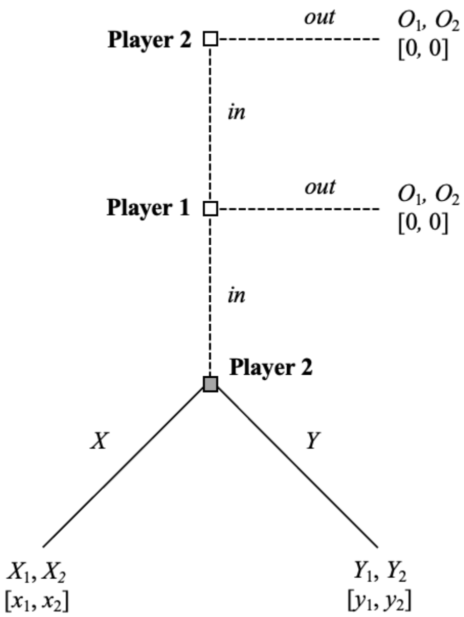
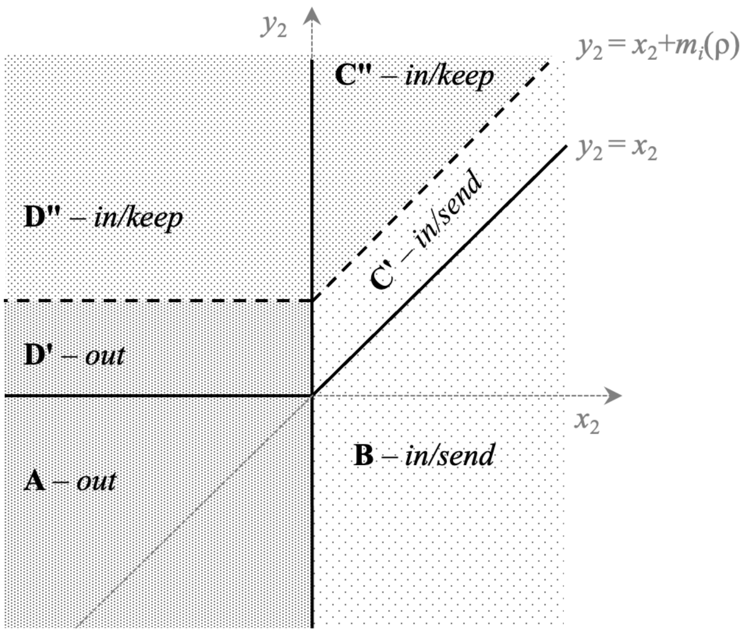


**Figure 2 – The Mutual Benefit Game**



## Figure 3 – Best responses by Player 2, conditional on moral cost



## Figure 4 – Best response functions and mutual benefit equilibria

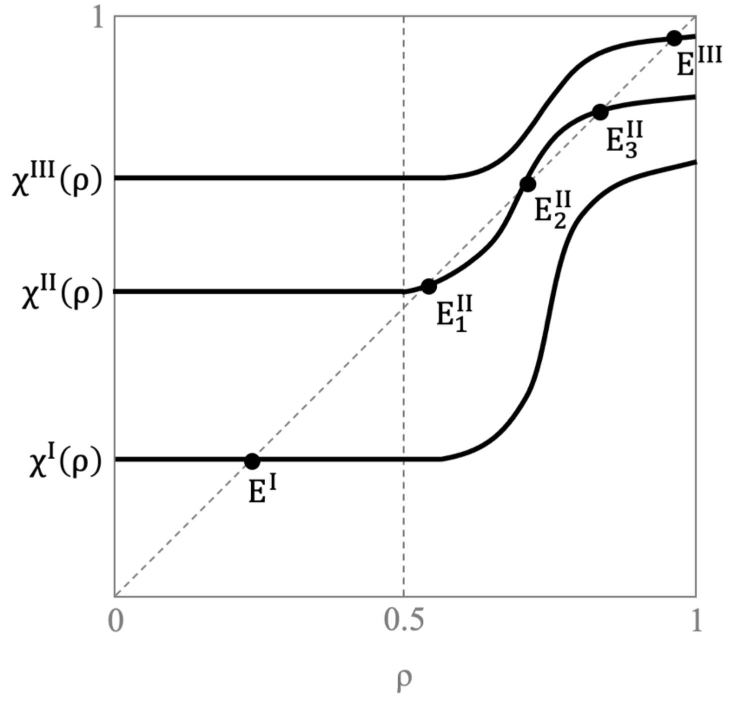


**Table 1 – Best responses by Player 2**

| Case | Probability | Restriction on payoff parameters | Best response by Player 2 if: | |
|---|---|---|---|---|
| | | | unconditionally self-interested | self-interested conditional on not choosing *in/Y* |
| A | $\pi_A$ | $0 > y_2 > x_2$ or $0 > x_2 > y_2$ | *out* | *out* |
| B | $\pi_B$ | $x_2 > y_2 > 0$ or $x_2 > 0 > y_2$ | *in/X* | *in/X* |
| C | $\pi_X$ | $y_2 > x_2 > 0$ | *in/Y* | *in/X* |
| D | $\pi_\Delta$ | $y_2 > 0 > x_2$ | *in/Y* | *out* |

**Table 2 – Summary of the Market Trust Game episodes**

| | High-reliability market ($v$ = 0.7) | | | | Low-reliability market ($v$ = 0.3) | | | |
|---|---|---|---|---|---|---|---|---|
| | All | Move together | Buyer first | Owner first | All | Move together | Buyer first | Owner first |
| No. episodes | 4261 | 3012 | 625 | 624 | 1859 | 593 | 654 | 612 |
| Prop. interactions initiated | 0.315 | 0.246 | 0.440 | 0.519 | 0.198 | 0.088 | 0.219 | 0.283 |
| Honest trading rate | 0.702 | 1.000 | 0.385 | 0.290 | 0.427 | 1.000 | 0.385 | 0.289 |
| Completion rate | 0.808 | | | | 0.498 | | | |

*Note*: Honest trading rate = proportion of initiated interactions in which both players choose *send*. Completion rate = proportion of first mover *in* decisions met by a *send* decision by second mover.

**Table 3 – Summary of entry decisions**

| Case (Mover, Role, Profit) | High-reliability market ($v = 0.7$) | | | Low-reliability market ($v = 0.3$) | | |
|---|---|---|---|---|---|---|
| | Obs. | No. *in* | Prop. *in* | Obs. | No. *in* | Prop. *in* |
| $FO^+$ (First, Owner, Positive) | 3216 | 2254 | 0.70 | 1036 | 467 | 0.45 |
| $FB^+$ (First, Buyer, Positive) | 3182 | 1762 | 0.55 | 1111 | 333 | 0.30 |
| $FO^-$ (First, Owner, Negative) | 420 | 11 | 0.03 | 169 | 3 | 0.02 |
| $FB^-$ (First, Buyer, Negative) | 455 | 15 | 0.03 | 136 | 4 | 0.03 |
| $SO^+$ (Second, Owner, Positive) | 549 | 518 | 0.94 | 580 | 530 | 0.91 |
| $SB^+$ (Second, Buyer, Positive) | 552 | 515 | 0.93 | 532 | 484 | 0.91 |
| $SO^-$ (Second, Owner, Negative) | 76 | 31 | 0.41 | 74 | 34 | 0.46 |
| $SB^-$ (Second, Buyer, Negative) | 72 | 34 | 0.47 | 80 | 40 | 0.50 |

**Table 4 – Summary of completion decisions**

| Case (Mover, Role, Profit) | High-reliability market ($v = 0.7$) | | | Low-reliability market ($v = 0.3$) | | |
|---|---|---|---|---|---|---|
| | Obs. | No. *send* | Prop. *send* | Obs. | No. *send* | Prop. *send* |
| SO$^+$ (First, Owner, Positive) | 248 | 105 | 0.42 | 129 | 53 | 0.41 |
| SB$^+$ (First, Buyer, Positive) | 292 | 94 | 0.32 | 151 | 50 | 0.33 |
| SO$^-$ (Second, Owner, Negative) | 27 | 1 | 0.04 | 14 | 2 | 0.14 |
| SB$^-$ (Second, Buyer, Negative) | 32 | 0 | 0.00 | 22 | 0 | 0.00 |

### Table 5 – Effect of exogenous parameters on players' decisions

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | FO$^+$ *in* | FB$^+$ *in* | SO$^-$ *in* | SB$^-$ *in* | SO$^+$ *send* | SB$^+$ *send* |
| Revenue (*t* for Owner, *e* for Buyer) | 0.0533***<br>(0.0019) | 0.0417***<br>(0.0039) | -0.1332*<br>(0.0780) | 0.4025***<br>(0.0741) | 0.0167**<br>(0.0084) | 0.0045<br>(0.0125) |
| Cost (*d* for Owner, *t* for Buyer) | -0.1199***<br>(0.0044) | -0.0721***<br>(0.0016) | 0.1239*<br>(0.0692) | -0.2304***<br>(0.0810) | -0.0722***<br>(0.0156) | -0.0264***<br>(0.0055) |
| High-reliability market ($v$ = 0.7) | 0.2462***<br>(0.0134) | 0.2371***<br>(0.0125) | -0.0449<br>(0.0542) | -0.0899<br>(0.0591) | 0.0382<br>(0.0457) | 0.0196<br>(0.0307) |
| Period | -0.0018***<br>(0.0003) | -0.0039***<br>(0.0003) | 0.0063***<br>(0.0017) | 0.0043***<br>(0.0016) | -0.0044***<br>(0.0013) | -0.0042***<br>(0.0009) |
| Observations | 4,252 | 4,293 | 150 | 152 | 377 | 443 |

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1