How should we reconcile self-regarding and pro-social motivations? A renaissance of "Das Adam Smith Problem"

Natalie Gold, Faculty of Philosophy, University of Oxford

Abstract. Das Adam Smith Problem is name given by eighteenth century German scholars to the question of how to reconcile the role of self-interest in the *Wealth of Nations* with Smith's advocacy of sympathy in *Theory of Moral Sentiments.* As the discipline of economics developed, it focused on the interaction of selfish agents, pursuing their private interests. However, behavioural economists have rediscovered the existence and importance of multiple motivations and a new Das Adam Smith Problem has arisen, of how to accommodate self-regarding and pro-social motivations in a single system. This question is particularly important because of evidence of motivation crowding, where paying people can backfire, with payments achieving the opposite effects from those intended. Psychologists have proposed a mechanism for the crowding out of "intrinsic motivations" for doing a task, when payment is used to incentivise effort. However, they argue that pro-social motivations are different from these intrinsic motivations, implying that crowding out of pro-social motivations requires a different mechanism. In this paper I present an answer to the new Das Adam Smith problem, proposing a mechanism that can underpin the crowding out of both pro-social and intrinsic motivations, whereby motivations are prompted by frames and motivation crowding is underpinned by the crowding out of frames. I explore some of the implications of this mechanism for research and policy.

Keywords: Altruism, Das Adam Smith Problem, Framing, Institutions, Markets, Moral Sentiments, Motivation Crowding, Pro-sociality, Self-Interest, Self-Regard, Trust

1. The new Das Adam Smith Problem: There and back again

Das Adam Smith Problem is the name given by eighteenth century German scholars to the question of how to reconcile the role of self-interest in *The Wealth of Nations* with Smith's advocacy of sympathy in *The Theory of Moral Sentiments.* It seemed to them that Adam Smith had written two very different books. Their (now disputed) reading was that *The Wealth of Nations* is founded on an egoistic theory of behaviour, showing how the interaction of self-interested individuals could lead to benefits for all. In contrast, *The Theory of Moral Sentiments* not only espouses a theory of human nature in which we have multiple motivations, especially "sympathy", which can underpin moral judgments and virtuous actions, but Smith argues that we ought not to be purely self-interested: "And hence it is, that to feel much for others and little for ourselves, that to restrain our selfish, and to indulge our benevolent affections, constitutes the perfection of human nature; and can alone produce among mankind that harmony of sentiments and passions in which consists their whole grace and propriety" (Smith, 1759, Part I, Ch. 1). In the twenty-first century, few scholars believe there is a contradiction between the two books; however, there is no consensus about the right way to solve Das Adam Smith Problem (see Montes, 2003 for a survey of the current debate).

Regardless of the status of Das Adam Smith Problem, the point remains that in eighteenth-century it was standard to acknowledge that multiple motivations are relevant for the study of political economy. But this picture was on the wane. The lure of Smith's idea that an agent who intends only his own gain is led by an "invisible hand" to pursue the good of society, "more effectually than when he really intends to promote it" (Smith, 1776, Book IV, Ch.2) proved compelling for many. By the nineteenth century, James Stewart Mill wrote of political economy that, "It is concerned with [man] solely as a being who desires to possess wealth, and who is capable of judging the comparative efficacy of means for obtaining that end." (Mill, 1836/ 1874, essay 5, paragraphs 38 and 48). Nevertheless, political economists did not offer the pursuit of wealth as a complete theory of human nature. Mill wrote that the desire for wealth was not the whole of Man's nature; that there are other human motives, such as "the affections, the conscience, or feeling of duty, and the love of approbation". However, he considered these to be the subject matter of philosophy. This prefigures the turn of economics towards treating people as solely pursuing their own private and selfish material interests, which I will call *the principle of self-regard*.

The principle of self-regard became increasingly important in the late 19th century, as economics replaced political economy as the subject that studies production,

exchange, and the distribution of resources. Economists such as Alfred Marshall and Francis Edgeworth emphasized the way in which the interaction of individual agents causes economic outcomes. They pioneered a theory of behaviour in which individuals maximise utility and firms maximise profits, subject to constraints on their budgets and resources. This is the core of neo-classical economics, the current mainstream of the subject. Strictly speaking, "utility" is an empty placeholder which includes anything that might make an agent choose one option over another. However, in practice it is usually taken to be a function of the agent's own consumption of goods and services. A standard graduate textbook in microeconomic theory states that, "A defining feature of microeconomic theory is that it aims to model economic activity as an interaction of individual economic agents pursuing their private interests" (Mas-Colell, Whinston and Green, 1995, p.3). This approach arguably has its roots in *The Wealth of Nations;* it discards Smith's insights about other sources of motivation in *The Theory of Moral Sentiments.*

However, in the twenty-first century, economics is seeing a renaissance of some of the traditional themes of political economy. It has rediscovered the existence and importance of pro-social motivations, both for the design of institutions and in market settings. Economists have studied altruism, fairness, equity, kindness, reciprocity, and trustworthiness, to name a few. These are studied alongside the principle of self-regard, which is still acknowledged as an important driver of behaviour in many circumstances. Therefore we have a renewed Das Adam Smith Problem for the twenty-first century: how do we integrate the fact that much economic analysis is based on self-regard (via the price mechanism) with renewed interest in and evidence of the importance of pro-social motivations? This acuteness of this problem is demonstrated by evidence that paying people can backfire if they are driven by pro-social motivations. A synthesis would provide directions and instructions for the designers of institutions. Which motivations people use—and should use—in a given context has implications for how to structure institutions and incentives.

In order to set up the problem (and to introduce some of the distinctions that will play a part in later discussion), in Section 2 I present a taxonomy of motivations from psychology and relate it to evidence from behavioural economics. In Section 3, I explain why the problem is of more than theoretical interest. There is a large literature, which originated in psychology, that shows that paying people can have perverse effects on their behaviour, the so-called motivation crowding effect. The original demonstrations of motivation crowding involved payments for effort, but economists have tended to

assume that motivation crowding also applies to pro-social behaviour. However, psychologists have argued that pro-social behaviour is relevantly different from payment for effort, in a way that means their standard theoretical explanation does not apply, leaving a question about the mechanism behind the crowding out of pro-social motivations. In Section 4, I propose a mechanism, drawing on framing, that can explain why payments affect both effort and pro-social motivations. In Section 5, I explore its implications for research and institutional design.

2. Evidence for pro-social behaviour and pro-social motivations

The principle of self-regard makes mistakes about the ends that people pursue and the reasons for which they pursue them: they may be concerned with ends other than their own outcomes, and their reasons for pursuing them need not be completely self-interested. In contrast to the assumption of the principle of self-regard, people's behaviour may be *pro-social*, promoting the well-being of others. (Note that this can include promoting the well-being of specific others, which may not promote the well-being or interests of society as a whole. For example, a mafioso can act pro-socially towards other members of the cosa nostra, but that can lead to bad outcomes for society.) There is a vast amount of evidence, from behavioural economics as well as psychology, that people are not only concerned with their own outcomes. Participants in experiments give money in dictator games and return money in one-shot trust games. Psychologists have also studied helping behaviour in a more contextualized manner, putting subjects in actual helping situations, which they do not realise are an experimental set-up.

Pro-social behaviour promotes the wellbeing of others. Pro-social motivation is a motivation to promote the wellbeing of others. Motivation is a slippery concept, it means different things to different writers. For instance, in psychology a motivation could be a goal-directed force (Batson, 1994), while in philosophy a motivation might be shorthand for a motivating reason (Parfit, 1984). For the purposes of this paper, either of these two ways of casting motivation would do and they could be used interchangeably. Indeed, one psychologist describes motivations in a manner that combines these two ideas, as "the reasons that drive actions" (Grant, 2008, p.48).

There can be chains of motivations. If we ask of any individual's behaviour "Why did she do that?" we can often take the answer and run at least one more iteration of the question. For instance, if someone gives to a food bank then we can ask

of our donor "Why did she give food?" Our answer might be "Because she was concerned with the welfare of people who cannot afford to feed themselves." But then we can ask the further question "Why was she concerned with the welfare of people who cannot afford to feed themselves?". One possible answer is "Because she takes pleasure in others' welfare gains"; another possibility is that there is no further answer, improving people's welfare is her ultimate motivation or her ultimate goal. Some motivations or goals may be seen as instrumental, pursued for the sake of a higher motivation or goal. The ultimate motivation or goal is the place where the buck stops.

As well as debates about the possibility of pro-social behaviour and proximate pro-social motivations and goals, there is also debate about the nature of ultimate motivations and goals. Some researchers argue that all behaviour is ultimately self-interested, that pro-social behaviour is really enlightened self-interest, a position that is known as psychological egoism (Feinberg, 1978). This position has seemed attractive to some because the reasons I act for are my reasons and the goals I pursue are my goals. However, nothing follows from this: There can be a separation between my goals and my welfare; it is not true that pursuing my goals and my reasons will always make me better off. (Butler argued that it is not in one's self-interest to be self-regarding long ago in his *Sermons*; for a more recent argument against psychological egoism see Sober and Wilson, 1998.)

We can understand this distinction—between goals that are motivated by enlightened self-interest, where helping others positively impacts the agent's own welfare, and goals that do not promote the agent's welfare—in the context of Sen's (1977) distinction between sympathy and commitment. Sympathy is when the concern for others directly affects the agent's own welfare, an idea that Sen takes from Smith and Edgeworth, although arguably it is closer to our modern notion of empathy: the agent takes pleasure in others' gains and pain in others' losses. Commitment is when the outcome that the agent is concerned about does not directly impact on his or her welfare, but the agent is never-the-less motivated to achieve it. Commitment covers a class of reasons for acting that result from normative imperatives including, but not limited to, moral imperatives. People's actions can be over-determined. An agent who is committed to making a charitable donation might also take pleasure in it, even though that wasn't her reason for contributing. Therefore, deciding whether or not an agent acts from commitment may require making judgments about counter-factual cases. An agent who acts with commitment is one who would have made the donation even if it had not made her better off by giving her pleasure.

In the same way that some researchers argue that all behaviour is ultimately self-interested, some philosophers might argue that all behaviour ought ultimately to be underpinned by morality. For instance, for a Utilitarian, the ultimate goal is the maximization of utility. For a Kantian, one should always ask whether one is acting on a principle that could be willed as a universal law.

But there are also other possible ultimate motivations. Batson (1994) identifies four ultimate motivations:

(1) egoism—increasing the actor's own welfare; the benefits can be material, social or self-rewards (e.g., monetary rewards, praise, self-esteem) or the avoidance of material, social or self-punishment (e.g. fines, social censure, guilt, shame)

(2) collectivism— increasing the welfare of a group or collective

(3) altruism—increasing the welfare of one or more individuals other than oneself

(4) principlism—upholding some standard or principle; Batson specifies moral principles, but it is possible to act to uphold standards and principles that are not moral, e.g., professionalism involves working to a professional standard, or one might act to uphold the law and legal principles.

For Batson, these are all at least potentially ultimate motivations. He has spent his career studying pro-social behaviour and showing that altruism can be an ultimate motivation. His hypothesis is that we are motivated by empathy-induced altruism, that "feeling empathy for [a] person in need evokes motivation to help [that person] in which these benefits to self are not the ultimate goal of helping; they are unintended consequences" (Batson & Shaw, 1991, p.14). Thus, empathy-induced altruism is a form of commitment. Batson's strategy is to take instances of helping behaviour and to show that they are not caused by plausible egoistic motives; that high empathisers continue to help even when the egoistic motivation is neutralized (Batson, 2011; Batson, 1992; Batson & Shaw, 1991). It is not a direct test of the hypothesis that altruism is caused by commitment but, by excluding a variety of egoistic explanations and showing that there is helping behaviour that they cannot account for, Batson increases the probability that altruistic behaviour is caused by commitment rather than being "a subtle and sophisticated form of egoism" (Batson, 2011, p.224).

Examples of all four types of motivation can be found in experimental and behavioural economics. Egoism is the standard currency of economists and behaviour in the lab varies a lot by individual, so any experiment that shows that at least some

subjects are pro-socially motivated also has some subjects who are egoists. Therefore I do not address it specifically.

Altruism: The classic example of altruism in experimental economics is giving in dictator games. In a paradigm set-up, a subject is given $10 and can choose how much of it to give to another anonymous subject. Usually more than 60% of subjects give some money, with the mean transfer being approximately 20% of the total (Camerer, 2003).

Collectivism: When group identity is manipulated, people are more favourable to ingroup members (Chen, Yan & Li, 2009). Some economists have argued that groups can be agents and individuals in groups use 'team reasoning', asking themselves the question "what should *we* do?", and that this is the best way of explaining the vast empirical literature showing that people cooperate and coordinate in ways that standard individualist economic theory cannot explain (Sugden, 1993; Bacharach, 2006).

Principlism: An example of principlism can be found in the literature on tax compliance. According to the principal of self-regard, tax evasion—like all other criminal behaviour—should be viewed simply as a choice whether to take a gamble that has a positive payoff if successful but a penalty if caught (Becker, 1968). However, subjects in the lab do not act according to this model: subjects are less likely to take gambles if they are presented as a tax evasion decision (Baldry 1986; Baldry, 1987) and their behaviour is affected by moral constraints (Bosco, & Mittone, 1997). Of course, the lab is an artificial environment (and, one might argue, subjects could be influenced by "experimenter demand effects") but self-regard cannot explain actual tax evasion behaviour either, whilst the hypothesis that at least some tax payers are motivated by moral principles can (Gordon, 1989).

Although Batson (1994) does not mention social norms in his taxonomy, they have been prominent topics of research in behavioural economics (Fehr & Fischbacher, 2004; Bicchieri 2005). However, this is not an important omission for Batson, given that he is concerned with ultimate motivation. Many researchers think that social norms are enforced by social approval and disapproval, or similar, in which case they are ultimately an egoist motivation, according to Batson's typology. Alternatively, we can imagine someone who had completely internalized social norms (someone who, if she found herself alone on a desert island, would still follow conventions such as "walk on the right, stand on the left" or continue to keep up her manners, things that have conventionally been instilled in her as "the right thing to do"). This would seem to be a variety of principlism, albeit a slightly strange one. So while social norms are an

important form of proximate motivation, they can be subsumed within Batson's categories of ultimate motivations.

A similar thing could be said about other types of non-self-regarding behaviour that experimental and behavioural economists have been interested in, such as equity, reciprocity, and trust and trustworthiness. Batson (1994) has a concise list because it is a list of ultimate motivations. In contrast, economists are better thought of as investigating proximate motivations and their models need not imply anything about ultimate motivations. The standard way of representing motivations in economics is as arguments in a utility function. Despite the terminology, the utility function only represents an agent's goals; it is a functionalist method of predicting action. The same function could represent either a "warm glow" from sympathy or a non-sentimental commitment. Further, there is no presumption that agents know their own utility functions: utility theory describes how people act but does not presume that people are aware of their own motivations.

Economists now agree there is a multiplicity of types of proximate motivation, including many that are not self-interested, and arguably there are multiple types of ultimate motivations. Economists tend to study each motivation in a particular setting or laboratory game. In order to rationalise he number of explanations of behaviour, they have developed hybrid models, which include multiple motivations, that aim to explain behaviour in multiple types of experiments. But even these models cannot explain all the empirical evidence (Fehr & Schmidt, 2006). The new Das Adam Smith problem, as I investigate it here, is a question about how these different motivations are related; it arises for proximate as well as ultimate motivations, so the question requires answering regardless of one's view on ultimate motivations.

3. The importance of the problem: The motivation crowding effect

It's important to have a theory of motivation because different motivations respond to different incentives, and using monetary incentives when people are acting on non-self-regarding motivations can be counter-productive. Well known examples of financial incentives backfiring include: payment for blood leading to less blood being collected (Mellström & Johannesson, 2008); fines for parents who failed to pick up their children on time from daycare leading to increased lateness, which persisted even after the fine was removed (Gneezy & Rustichini, 2000); the offer of financial compensation increasing NIMBY-ism, when people were asked if they would permit a nuclear waste repository to

be sited in their community (Frey & Oberholzer-Gee, 1997); the use of financial penalties for untrustworthy behaviour increasing the amount of untrustworthy behaviour (Fehr & List, 2004); and the use of financial penalties to enforce contracts leading to more contracts being breached (Fehr & Gächter, 2002). Why does this happen and what are the implications for institutional design?

The examples I just gave are all instances where payment affects pro-social behaviour. They are also often given by economists as examples of the *motivation crowding effect*, where payment for a task crowds out intrinsic motivation (e.g. Frey, 1997a; Bowles, 2008; Bowles 2016). The concept of intrinsic motivation is slippery. An early definition in the literature is that "[o]ne is said to be intrinsically motivated to perform an activity when one receives no apparent reward except the activity itself." (Deci, 1975, p.175). Conversely, one is extrinsically motivated when one does something to receive a reward or avoid a punishment. Let us call this Definition 1. (This is already slippery: what is an "apparent reward"? My reading is that it is a tangible, physical reward, i.e. it does not include intangible rewards like esteem.) Another way of thinking about the difference between intrinsic and extrinsic motivation is that one is intrinsically motivated when one does something for its own sake. Let us call this Definition 2. So while Definition 1 characterizes motivation crowding according to the environment in which the behaviour occurs, Definition 2 characterizes it in terms of ultimate motivations, which has some different implications, as we will see below

The original and paradigm example of motivation crowding from psychology involved payment for effort. Subjects who had been paid to solve puzzles were less likely to return to them later, after payment had been withdrawn, than a control group who had received no reward for their activity in the first period; and the paid subjects also reported a lesser interest in the task than the unpaid (Deci, 1975, 1971). Amongst psychologists, the predominant explanation for motivation crowding is the over-justification of the agent, where the payment is seen as controlling, and the external intervention therefore undermines feelings of self-determination and autonomy, which causes the agent to relinquish the intrinsic motivation (Deci and Ryan, 2000; Ryan and Deci, 2000).

In this early work, intrinsic motivation was defined in contrast to extrinsic motivation, as anything that is not done for a tangible reward (Definition 1). So it was natural to interpret intrinsic motivation as encompassing many different sorts of motivations for undertaking an activity, including both enjoyment of a task and pro-social motivations. As we saw in the examples at the beginning of this section, payments can crowd out pro-social motivations as well as effort. However, in later work, Ryan and Deci

(2000) provide a rather more refined definition of intrinsic motivation. They say that it is "the doing of an activity for its inherent satisfactions rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external prods, pressures, or rewards." (Ryan & Deci, 2000, p.56). They take Definition 2 and extend it, by specifying the exact motivation: for fun or challenge. Therefore, according to Ryan & Deci's definition, pro-social motivation is not an intrinsic motivation, since it is based on benefitting others rather than on interest in and enjoyment of a task (see also Grant, 2008).

Even if we discard the stipulation that intrinsic motivation involves acting for fun or challenge, retaining only the idea that it involves doing something for its own sake (i.e. adopt Definition 2), psychologists have noted differences between pro-social motivations and the motivation to make an effort, which imply that the over-justification theory does not apply to pro-social motivations. Grant (2008) starts from the position that intrinsic motivation is associated with pleasure and enjoyment, and pro-social motivation with meaning and purpose.[1] He argues that: intrinsic motivation phenomenologically pulls people to do things, whereas pro-social motivation may require people to push themselves, necessitating self-regulation to achieve a goal; intrinsic motivation focusses on the process, whereas pro-social motivation focuses on the outcome or goal;[2] and that—relatedly—intrinsic motivation involves a focus on the present experience, whereas pro-social motivation involves a focus on the future, on the meaningful outcome that will result from the behaviour.[3]

However, this implies that there is problem with using the over-justification theory to explain the effect of incentives on prosocial motivations. For Grant (2008, p.49), it follows from the differences between them that intrinsic and pro-social motivations involve different levels of autonomy. He says that intrinsic motivation is "fully volitional,

[1] In his paper, Grant refers to motivations as desires, so that intrinsic motivation is "the desire to expend effort based on interest in and enjoyment of the work itself" and pro-social motivation is "the desire to expend effort to benefit other people" (p.49). I have not repeated this full definition because it is pretty clear to me that it is incorrect to define a motivation as a desire; at the very least this needs to be amended to the desire that is acted on, since we may have plenty of desires that are latent or not acted on.

[2] It is not clear that motivations like fairness fit so neatly into this dichotomy, since fairness can be about following correct processes (procedural) as well as about fair outcomes.

[3] We might note that it is not so clear whether this contention is true. Grant's (2008) position is consistent with that of other researchers who have hypothesized that self-control and cooperation (especially in prisoner's dilemmas) both require the subjugation of short-term goals for long-term ones (Dewitte & Cremer, 2001). However, there is evidence that cooperation in prisoner's dilemmas is the spontaneous, intuitive response, which is reigned in by reflective decision-making (Rand & Nowak, 2013), which suggests that self-control doesn't require self-regulation so much as not thinking.

self-determined and autonomous" whereas pro-social motivation "is less autonomous, as it is based more heavily on conscious self-regulation and self-control to achieve a goal". If pro-social motivations are not associated with autonomy, then the explanation for the crowding out of pro-social motivations cannot be that autonomy is impaired.

There are two possible counters, neither of which entirely solves the problem. First, one could get into a philosophical debate about what constitutes autonomy, arguing that self-regulation is a form of Kantian autonomy, where one follows a rule that one makes for oneself. (Grant, 2011, discussing a slightly different question around the ethics of incentives, takes this sort of line.) However, this response misses that Grant's (2008) point is really about the phenomenology of behaviour: if the mechanism of motivation crowding is that applying incentives makes people lose their feeling of being autonomous, and if pro-social behaviour often does not feel autonomous in the first place, then there is no reason to expect pro-social behaviour to respond to the mechanism—even if it belongs to the pholosophical category of Kantian autonomy. Second, psychologists allow that, to the extent that we value and identify with pro-social behaviours, we may experience greater autonomy in their performance (Ryan and Deci, 2000). But they also make it quite clear that they consider pro-social motivation a type of extrinsic motivation because acting for the benefit of others, even if that fulfils core values and identities, is a type of external goal. (Though this would seem to conflate having an external goal and wanting to achieve that goal for its own sake, as an ultimate goal).

One response would simply be to follow Bowles (2008) in endorsing a variety of mechanisms of motivation crowding, so different instances of crowding are explained by different mechanisms. But that leaves us in a place where we cannot conclude much. Bowles' recommendations to policy makers are: to use more realistic psychological assumptions when doing mechanism design, and that good policies and constitutions will support socially valued ends by evoking, cultivating, and empowering public-spirited motives. These are all very sensible, but not very specific. It would be nice to be able to say something more specific about institutional design or the direction of research needed to do good design.

Instead, I will propose a different mechanism, which can explain both the crowding out of intrinsic motivations and the crowding out of pro-social motivations, and explore the implications for policy and research.

4. Framing and motivation crowding

One solution to the original Das Adam Smith Problem is that different motivations are used in different spheres, and with different people (Nieli, 1986; Roberts 2014). That has an intuitive plausibility about it. I want to think about this in the context of research on the prisoner's dilemma, which is extensive, and suggests a more specific mechanism.

It should not come as a surprise to anyone that there is a higher rate of cooperation in the prisoner's dilemma when it is called the "Community Game" rather than the "Wall Street Game" (Ross & Ward,1997). Changing the labels on a decision-problem and observing that this causes people to choose differently is an example of a *framing effect*. Framing is often implicitly and sometimes explicitly offered as an explanation of the effect of payments on effort (e.g., Gneezy & Rustichini, 2000; Heyman & Ariely, 2004). Lindenberg and Frey (1993) claim that when motivation crowding occurs a "gain frame" crowds out a "normative frame", but this is not explained in any further detail.

We can think of the agent's frame as the set of concepts that she uses to think about her situation (Bacharach, 2003). Framing is notorious because of Tversky and Kahneman's (1981) work on framing effects, where two groups of subjects were put in the position of policy makers facing an epidemic and asked to choose between two vaccination programs. Subjects who were given the decision problem in terms of "lives saved" by each program tended to choose a different program to those who were given the problem in terms of "lives lost" by each program. Similarly, Ross & Ward (1975) took a laboratory prisoner's dilemma but for one set of subjects they referred to it as the "Community Game" and for another they referred to it as the "Wall Street Game". Two-thirds of subjects cooperated in the Community Game, compared to one-third in the Wall Street Game.

There is another framing effect involving prisoner's dilemmas that researchers have hypothesized is caused by a change in motivations. The standard way of presenting prisoner's dilemma is as a 2x2 payoff matrix. However, it is possible to "decompose" the payoffs and present them as a choice between two different allocations of payoffs between Player 1 and Player 2 (Messick & McClintock, 1968; Pruitt, 1967). Figure 1 gives an example of a prisoner's dilemma matrix and an associated decomposed game. Both players choose a payoff allocation and then each gets the total of the payoff they awarded to themself plus the payoff s/he was awarded by the other player. For instance, if Player 1 chooses allocation C and Player 2 chooses allocation D then Player 1 has assigned 0 to herself and 12 to Player 2 whilst Player 2 has assigned 6 to herself and 0 to Player 1. So Player 1 gets 0 from herself and 0 from Player 2, a total of 0. Player 2 gets 12 from his choice and 6 from player one, a total of 18. The outcome is 0 for Player 1 and 18 for Player 2, which is the same as the payoffs for (C, D) in the game matrix. The totals from each

combination of allocations is the same as the payoff from the equivalent strategy combination in the prisoner's dilemma. Therefore, in any decomposed game, it is possible to work out the payoff matrix from the choices in the allocation decision.[4] The decomposition and the parent game are two different ways of presenting the four possible payoff outcomes. However, experimenters have found higher rates of cooperation with the decomposed game compared to the matrix presentation (Pruitt, 1967; Komorita, 1987; Cookson, 2000).

In an investigation designed to discover why behavior was different in the decomposed games, Pruitt asked subjects to record the thinking behind their decisions (Pruitt, 1970). He discovered that, in accordance with expectations derived from game theoretic reasoning, those who played D in the above games were motivated by the payoff they could get by doing so. In the decomposed game, responses to open-ended questions showed that many subjects viewed alternative C as a way of being "helpful" or "generous". Pruitt (1970, p.235) postulated that "the games produce differing motives, which in turn produce differing behavior", a suggestion that has also been echoed by Colman (1995).

Kahneman and Tversky (1979) explained their framing effect using Prospect Theory, drawing on the idea that people display different risk preferences depending on whether options are framed as losses or gains. However, it is hard to see how Prospect Theory could explain the difference in play between these differently framed prisoner's dilemmas—or, for that matter, the examples of collectivism and principlism in Section 2, which were also demonstrated by taking a laboratory game and changing the framing: manipulating the group identity of the players (collectivism) or calling a gamble a tax evasion decision (principlism). To explain these examples, we need a more general theory of framing effects.

Decision theorists have given explanations of framing effects that relate them to reasons (Dietrich & List, 2016; Weirich, 2010; Gold & List, 2004; Schick, 2003). What the different models have in common is that the reasons that underpin an agent's choices depend on how they frame or, in the case of Schick (2003), "understand" the decision. Note that acting and choosing for a reason does not have to be understood as involving a conscious reasoning process. A minimal requirement is that the agent is disposed to be responsive to reasons, where these are based on facts that count in favor of a particular

[4] However, we cannot assume that a player would see any given decomposition from the parent game because, if a prisoner's dilemma is decomposable, then there are an infinite number of possible decompositions (Messick & McClintock, 1968).

decision or action. In this paradigm, framing effects occur when there are reasons in favor of both options and the reason that the agent responds to depends on the way in which the decision is presented or described. In effect, these agents are not weighing all their reasons, but act on the basis of a single reason. If they have an acceptable reason to hand, then they do not search for others. This has psychological plausibility. It is consistent with evidence of "concrete thinking", whereby decision-makers appear to use only surface information and information that has to be inferred from the display or created by some mental transformation tends to be ignored (Slovic, Fischoff, & Lichtenstein, 1988; Fischhoff, Slovic, & Lichtenstein, 1978). Concrete thinking may be connected to people's desire to justify decisions by saying that they chose for a (single) reason, even to the extent of constructing and selecting choice situations such that there is always a dominant reason for choice (Montgomery, 1983). Once a reason for choice has presented itself, people are not motivated to seek out further reasons. Call this "one-reason decision-making".

In Tversky and Kahneman's problem, the fact some people will die for sure is a reason not to choose the policy with certain outcomes, while the fact that there is a possible outcome where no-one is saved is a reason not to choose the risky policy. The two different ways of framing the decision make these different reasons salient, which affects people's choices (Gold & List, 2004). This explanation is in accordance with the psychological literature on "reason-based choice". A classic example there is the custody decision, where the question of which parent *should* get custody elicits the same answer as the question of which parent *should not* get custody: the questions elicit a search for positive and negative attributes respectively, which would be reasons for giving or not giving custody, and one parent has both more positive and more negative attributes (Shafir, Simonson, & Tversky, 1993).[5]

The idea that the presentation of the decision affects the reason that people act on can explain a wide class of framing effects, including ones that involve motivations. Reasons are connected to motivations. We can think of the reason for which an agent acts as her motivating reason, so framing can affect an agent's motivating reason.

In the decomposed prisoner's dilemma, game-theoretic reasoning about monetary payoffs conflicts with being helpful or generous. This is sometimes referred to as "might versus morality" (Liebrand et al, 1986). According to Pruitt (1970) and Colman (1995), the decomposition makes helpfulness, or the moral side of the coin, more salient. Their

[5] It is possible to translate between the "value-based" model given by Tversky and Kahneman and the "reason-based" tradition. Roughly, what Kahneman and Tversky describe as a change in curvature of the utility function becomes a difference in how people value the options. See also Gold & List (2004).

suggestion is supported by evidence that the way that subjects frame the prisoner's dilemma correlates with the move they make. Subjects who perceive playing C as co-operative and playing D as non-co-operative are more likely to play C (Baranowski, & Summers, 1972). Similarly, co-operative types (defined as such because they behave co-operatively) tend to frame the dilemma in terms of morality (Liebrand et al, 1986). If moral reasons support a different choice from game-theoretic dominance reasoning and the salience of these reasons can be affected by the presentation of the decision, then people will make different choices in different frames. Further, Bruner (1957) postulates that once an agent has categorized a situation, incongruent cues may be "gated out". Bruner does not say how or why gating out occurs but, in cognitive psychology, there is a well-known effect called *assimilation*, where an agent perceives an object's attributes as more typical of the category that is being used than it actually is (Herr, Sherman,& Fazio, 1983).

The reason-based explanation of framing effects is consistent with the evidence of a connection between framing and behavior in prisoner's dilemmas, but refines it by offering a direction of causality, namely that framing the game in moral terms may lead to co-operative behavior by increasing the perception of, and hence the chance that people act on, moral or other-regarding reasons.

A framing theory of motivation crowding can also explain the paradigm examples of motivation crowding, where payment crowds out intrinsic motivations. These do not involve changes in explicit descriptions. However, the monetary payment may still affect the way subjects frame the situation. Take Deci's (1975, 1971) experiments, where subjects were given puzzles to solve. There were two periods in this experiment. In the second period, subjects were left alone in the room with the puzzles. This was the same for all subjects. In the first period, half of the subjects were paid to solve puzzles and half of the subjects played with them without payment. Deci found that first period activity affected second period behavior, even though naïve theory suggested that the second period was the same for all subjects. The first period activity may have served as an implicit framing task. The puzzles were supposed to be interesting to solve for their own sake. Subjects who were paid were given another way to think about the puzzles: as an activity engaged in to make money. In the second period, the monetary payment was withdrawn. If the subjects who had been paid framed the task of solving them in terms of money, and acted on their monetary motivations, then their reason for solving the puzzles would have gone.

Concrete thinkers, who do not generally search for information, would not investigate whether there were other reasons to carry on solving the puzzles.[6]

When an agent is performing a task that she has intrinsic reasons or other-regarding reasons to do and she is also being paid, then her action is over-determined. There is a sense in which she is over-justified—because she has multiple reasons in favor of her action, not because the price is seen as an instrument of control. If people are one-reason decision-makers, then one of the motivations will become the primary motivating reason, at the expense of any others. Why should the monetary rather than the non-monetary reason become the motivating reason?

We can answer this question by drawing on *attribution theory*, according to which actors are more likely to attribute their behavior to external factors than internal ones (Jones et al, 1972; Heider, 1958). So attribution theory would predict that, if agents are offered payment, then they will attribute their motivation to the payment, rather than any intrinsic or pro-social motivation they may also have had. This is also supported by evidence from the Fundamental Attribution Error (Ross, 1977). The *Fundamental Attribution Error* is an asymmetry in the way people explain behaviour, with people explaining their own behaviour differently from the way they explain the behaviour of others. The important thing for us is that people tend to attribute the causes of their own behaviour to their external situation (whereas they tend to attribute other people's behaviour to internal traits). So if I send in my paper late, then I explain it by saying things such as "I had some important emergencies that prevented me from finishing on time" (whereas if your paper is late I am more likely to say that you are bad at time management or cannot stick to deadlines). So the Fundamental Attribution Error supports the idea that if we offer someone a reward for performing a behaviour, then she is likely to attribute her behaviour to the presence of the reward. In that case, it is not surprising that she would stop the behaviour when the reward is withdrawn.[7]

The cases I have discussed so far have all been examples of both framing effects and motivation crowding. However, the framing mechanism can also explain examples where there is an actual change in the situation, as well as a change in framing, i.e. which are not

[6] In further support of the framing theory of motivation crowding for the paradigm effects, we know that the salience of the reward affects motivation crowding. Ross (1975) has shown that a highly salient reward is more detrimental to intrinsic interest than the same reward when it is relatively non-salient. He also showed that reward is less detrimental when the subject's attention is distracted from it.

[7] Interestingly, the Fundamental Attribution Error might also explain why people do not predict motivation crowding effects in others, tending to choose to use incentives even when their effect is counterproductive (Fehr & List, 2004).

framing effects. So I am not claiming that all motivation crowding effects are framing effects and I do not mean to make any claim about the rationality of motivation crowding. But the process of framing, which operates in framing effects, also operates in cases of motivation crowding. Introducing a reward also introduces a new way to think about the behaviour, as being done for a reward. There is a change in the agent's frame. Changing the way that people frame a problem may change their motivating reason. Once someone has the concept of doing something for a reward in their frame (or, as Lindenberg and Frey put it, uses a "gain frame"), then it becomes likely that withdrawal of payment leads to cessation of the activity. This mechanism explains the contention of Lindenberg and Frey (1993), that a "gain frame" will "crowd out" other ways of framing the task.

5. Implications for research

There has been a tendency for economists to resist adding frames as a primitive to their theories and a tendency to think of framing as irrational. Both of these tendencies are mistakes.

Frames are usually thought of as a purely cognitive feature. In the classic accounts of framing effects, frames may affect the attractiveness of options, quite literally. For instance, describing beef as 25% fat instead of 75% lean makes people rate it as less likely to be tasty (Levin & Gaeth, 1988). And the standard question that follows from framing effects is how can people be so irrational as to change what they want when all that has changed is the description? If the decision is whether to have a surgical procedure with a 90% survival rate and a 10% mortality rate (McNeill et al, 1992), then there are serious consequences that follow from the choice. In my account of motivation crowding as involving framing, frames also have normative features. A change in frame is not just about changing the attractiveness of an option, but it also changes what motivations and behaviours are seen as appropriate.

Framing is a part of the decision-making process, prior to assessing options and making choices. Most motivation crowding effects are not framing effects, even if they do involve framing, so the rationality or otherwise of framing effects is orthogonal to this discussion. But we might note that these general effects on motivation cast doubt on at least some of the reasons for declaring framing effects irrational. The core assumption is that it is irrational for one's choice to depend merely on the description. One reason that has been given for this is that the sort of selective seeing of a situation that is involved in framing is irrational; that rationality requires us to see all possible ways of framing a situation and that this requirement is imposed by orthodox decision theory (Skyrms,

1998). The classic examples of framing effects have two obvious frames, the opposition between positive and negative. However, once we move to a theory where frames can activate motivations, then there are an infinite number of ways of framing the situation. For instance, in the prisoner's dilemma, if there is one way of decomposing a matrix then there are an infinite number of possible decompositions. We have finite minds so we cannot see them all. This casts some doubt on whether it really is irrational not to see all the decompositions, unless rationality is merely a standard to which we aspire rather than a state we have any hope of achieving. Separating discussion of framing from the presumption of irrationality is a good thing because the presumption of irrationality may be a barrier to economists incorporating framing in their models.

Another unsuccessful argument against adding frames to the primitives of rational choice theory is that we can do all of the work using expectations. Some of the examples I discussed above might involve a change in expectations. For instance, changing the framing in the decomposed prisoner's dilemma by decomposing the game or by calling it the "Community Game" may change a player's expectations about what the other player will do. In many theories this change in expectations will cause a change in behaviour. For example, in the Rabin (1993) model of reciprocal fairness, agents want to be kind to agents who they expect will be kind to them. If a player vcares about Rabin-kindness and the decomposed dilemma leads her to expect that her co-player will cooperate, then the change in expectations could lead her to cooperate. But why would decomposing the dilemma increase a player's expectation that her co-player will be kind, without including an increased perception of the possibility of kindness in the explanation?

To see more clearly why this must be the case, consider an alternative theory that gives a prominent role to expectations, the idea that people are acting on social norms, whereby they have a conditional preference that they conform given that others will too (Bicchieri, 2005). In order for a social norm to lead to cooperation in a prisoner's dilemma, a player needs to know that a social norm exists and have an expectation that other players will cooperate. So there are two routes by which a change in presentation could lead to a change in behaviour. Either it could directly cause a player to perceive that they are in a situation that is governed by a social norm, when s/he did not see that before, or it could change expectations about the other player's behaviour. If a player's own frame has not changed but the change in presentation has changed her expectations about others, then the player's beliefs about whether the other player has perceived the norm must have changed. So either her frame or her beliefs about the other player's frame have changed; either way, we cannot dispense with the notion of a frame. The same applies to the case of

Rabin-kindness; just replace "norm" with "Rabin-kindness" in the argument. In both cases, the change in expectations of behaviour occurs because there is a change in expectation about how the other player frames the decision.

There is also evidence that framing can affect honest behavior without changing expectations. Cohn, Fehr, & Maréchal (2014) ran an honesty experiment, where subjects' payments depended on the outcome of a coin toss, which they self-reported, giving them the incentive to report dishonestly. (In this type of experiment, the subjects' actions are anonymous—with a large number of participants we would expect the distribution of heads and tails to follow a binomial distribution, for instance with a single coin toss we would expect heads and tails each to come up 50% of the time, so if the distribution of the subjects' reports is skewed away from that, it is a sign of dishonest reporting, and one can compare dishonesty between experimental conditions by comparing the distribution of the number of heads reported.) The subjects were bank employees and the researchers found that making subjects' professional identities as bank employees salient increased dishonest reporting. However, they also measured subjects' beliefs about other bank employees' reporting behaviour and this was not affected by the framing. The change in behaviour seems to have been caused by the framing, not by the expectations of what others would do.

When discussing situations with normative features, there has been a tendency for behavioural economists to focus on expectations, even when those expectations are connected to contexts. For instance, in Bicchieri's (2005) theory of social norms, a norm is triggered by the context, but the existence of a norm is defined as a network of expectations and the tests of the theory involve fixing a context and testing the effect of changing expectations. Similarly, List (2007, p.84), when discussing his finding that the amount sent in dictator games is sensitive to whether the experiment also includes the option to take money, concludes that the traditional set-up 'evokes expectations of the "givers" and "receivers" that seemingly demand a positive gift.' He concludes that the different choice sets invoke different social norms. One research implication that follows from the importance of framing is that we should be investigating what frames people bring to the situations we are studying, how that connects to their motivations, and using that knowledge to formulate testable hypotheses about what motivates their behaviour and what will induce behaviour change.[8] We need to focus on frames, not just on expectations.

[8] Lakoff (2014) has already been doing this sort of thing, not in the context of incentives, but in the context of political persuasion.

Some of what we find might surprise us. For instance, the market frame is associated with the efficacy of financial incentives and the pursuit of self-interest, but it is also associated with fairness and trust. Societies with market structures are more likely to be cooperative (Henrich et al., 2001); priming markets leads to senders sending more money in a trust game, but not in a dictator game (Al-Ubaydli, 2013). Markets are not only about the pursuit of narrowly defined self-regard, they are constrained by rules. Exchanges are not simultaneous, someone usually has to be the first mover, but when you hand your money to the butcher, the brewer, or the baker, you are confident that they will hand over the goods (Gold, 2014). Many people do not count their change. They can feel safe doing that because, in markets, bargaining is permitted but cheating is not.

We need framing even if we think both market and non-market behaviour are encompassed in a single overarching theory of behaviour. One solution that has been offered to Das Adam Smith Problem is that, although there seem to be two spheres, the principles of action are actually the same in each (Otteson, 2002; Smith, 1998). For instance, Smith (1998) argues that we maximise the gains from exchange in both markets and personal exchange, but in markets this is done through non-cooperative self-interest and in personal exchange through reciprocity. However, even in this theory, people need some way of identifying when market exchange is appropriate and when they should be engaging in personal exchange. People create and maintain strong distinctions among different kinds of social relations and meaning systems, to convey whether exchanges are gifts, entitlements, or payments (Zelizer, 1997). If there is a mismatch between the two sides of an exchange, between people who are pro-socially motivated and people who are not, then there is the opportunity for exploitation (Andersen, 1990).

The picture of decision-making proposed here recommends a different role for framing in institutional design than that suggested by the way framing if perceived in the "heuristics and biases" programme. Instead of setting up framing in order to enable people to be rational, we should design institutions that support framings that produce good outcomes. Sometimes that will involve a market frame and financial incentives, but other times it will involve supporting non-market ways of seeing the interaction and pro-social motivations. Institutional designers need to ensure that incentives offered are congruent with motivations, if they are to achieve their required results. And when making institutional changes, the impact on frames should be considered. It may be the case that treating people as though they are extrinsically motivated will actually cause them to be so motivated, creating the need for incentives and rewards where none existed before. (Further discussion of the idea that designing institutions as if people are knaves

causes them to behave as such can be found in Frey, 1997b). Or, when we desire to change the culture of institutions, designers could consider what frames are in play and how to change them.

6. Conclusion

There was something lost at the origins of political economy that we are rediscovering: the importance of pro-social motivations, and how they interact with and can be a corrective to self-regard. I have argued that we should understand motivations as being prompted by different normative frames. This leads to a new research direction, investigating a broader range of frames, and a policy recommendation, to design institutions to support frames that we consider desirable and efficacious. The currency of reward needs to be appropriate to the motivation, and feedback from rewards to frames should be considered. But we need to do this in ways that do not promote the exploitation of those who are pro-social, either because they are not adequately financially rewarded or because they are exploited by a self-regarding partner in the interaction.

One question I have not addressed is why we should design institutions to support pro-sociality, rather than letting market incentives take their course. One answer is that sometimes pro-social outcomes are more effective. Another might speak of the type of society we want to live in. A third brings me back to Adam Smith. In this paper, I have spoken quite narrowly of self-regard. But people's enlightened self-interest includes behaving pro-socially. Helping others is welfare increasing (Grant, 2013). The idea that social well-being is a part of our self-interest would have been familiar to Smith and is another return to the origins of political economy (Schmidtz, 2016; Paganelli, 2016).

References

Al-Ubaydli, O., Houser, D., Nye, J., Paganelli, M. P., & Pan, X. S. (2013). The causal effect of market priming on trust: An experimental investigation using randomized control. *PloS one*, 8(3), e55968.

Anderson, E. S. (1990). Is women's labor a commodity?. *Philosophy & Public Affairs*, 71-92.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Bacharach, M. (2003) Framing and cognition: the bad news and the good. In: N. Dimitri, M. Basili & I. Gilboa (eds.), *Proceedings of ISER Workshop XIV: Cognitive Processes in Economics*, London: Routledge, pp. 63-74.

Baldry, J. C. (1986). Tax evasion is not a gamble: A report on two experiments. *Economics Letters*, *22*(4), 333-335.

Baldry, J. C. (1987). Income tax evasion and the tax schedule: Some experimental results. *Public Finance= Finances publiques*, *42*(3), 357-383.

Baranowski, T.A. and Summers, D.A. (1972) Perception of response alternatives in a Prisoner's Dilemma game. *Journal of Personality and Social Psychology*, 21(1): 35.

Batson, C.D. (2011). Altruism in Humans. New York: Oxford University Press.

Batson, C. D. (1994). Why act for the public good? Four answers. *Personality and Social Psychology Bulletin*, *20*(5), 603-610.

Batson, C. D. (1992) Experimental tests for the existence of altruism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.

Batson, C.D.; Shaw, L.L. (1991). "Evidence for Altruism: Toward a Pluralism of Prosocial Motives". Psychological Inquiry 2 (2): 107–122.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime* (pp. 13-68). Palgrave Macmillan, London.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bosco, L., & Mittone, L. (1997). Tax evasion and moral constraints: some experimental evidence. *Kyklos*, *50*(3), 297-324.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *Science*, *320*(5883), 1605-1609.

Bowles, S. (2016). The moral economy: why good incentives are no substitute for good citizens. Yale University Press.

Bruner, J. (1957) On Perceptual Readiness. *Psychological Review,* 64(2):123-152.

Camerer, C. F. (2011). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.

Chen, Yan, and Sherry Xin Li. "Group identity and social preferences." *American Economic Review* 99, no. 1 (2009): 431-57.

Cohn, A., Fehr, E. and Maréchal, M.A., (2014). Business culture and dishonesty in the banking industry. *Nature* 516(7529) .86-89

Colman, A. M. (1995). *Game Theory and its applications in the social and biological sciences* (2nd ed.). Oxford: Butterworth-Heinemann & London: Routledge.

Cookson, R. (2000) Framing effects in public goods experiments. *Experimental Economics* 3(1):55-79.

Deci, E. L. (1975) *Intrinsic motivation*. New York: Plenum Publishing Co.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, *11*(4), 227-268.

Dewitte, S., & Cremer, D. D. (2001). Self-control and cooperation: Different concepts, similar decisions? A question of the right perspective. *The Journal of Psychology*, *135*(2), 133-153.

Dietrich, F., & List, C. (2016). Reason-based choice and context-dependence: An explanatory framework. Economics and Philosophy, 32(02), 175-229.

Fischhoff, B., Slovic, P. and Lichtenstein, S., 1978. Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance* 4(2): 330.

Frey, B. S. (1997a) *Not Just for the Money* London: Edward Elgar.

Frey, B. S. (1997b) A Constitution for Knaves Crowds Out Civic Virtues. *The Economics Journal* 107(443): 1043-1053.

Fehr, E., & List, J. A. (2004). The hidden costs and returns of incentives—trust and trustworthiness among CEOs. *Journal of the European Economic Association*, *2*(5), 743-771.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87.

Fehr, E., & Gächter, S. (2002). Do incentive contracts undermine voluntary cooperation?.

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity*, *1*, 615-691.

Feinberg, J. (1978). Psychological egoism. In Russ Shafer-Landau & Joel Feinberg (eds.), *Reason and Responsibility*. Wadsworth.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American economic review*, *87*(4), 746-755.

Gold, N. and List, C. (2004) Framing as path dependence. *Economics and Philosophy* 20(02):253-277.

Gold, N. (2014). Trustworthiness and motivations. In Vines, D. & Morris, N. (eds.) *Capital failure: Rebuilding trust in financial services*, 129-53. Oxford: OUP.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, 29(1), 1-17.

Gordon, J. P. (1989). Individual morality and reputation costs as deterrents to tax evasion. *European economic review*, 33(4), 797-805.

Grant, A. M. (2013). Give and take: A revolutionary approach to success. Penguin.

Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. *Journal of applied psychology*, *93*(1), 48-58.

Grant, R. W. (2011). Strings attached: Untangling the ethics of incentives. Princeton University Press.

Heider, F. (1958) *The Psychology of Interpersonal Relations*. New York: Wiley.

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. (2001) In search of homo-economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review* 91: 73–78.

Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, et al. (2010) Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327: 1480–1484.

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of experimental social psychology*, *19*(4), 323-340.

Heyman, J., & Ariely, D. (2004). Effort for payment a tale of two markets. *Psychological science*, *15*(11), 787-793.

Jones, E.E., Kannouse, D., Kelley, R., Nisbett, R., Valins, S. and Weiner, B (eds.) (1972) *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Kahneman, D. and Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2): 263-291.

Komorita, S.S. (1987) Cooperative choice in decomposed social dilemmas. *Personality and Social Psychology Bulletin* 13(1): 53-63.

Lakoff, G. (2014). *The all new don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of consumer research*, *15*(3), 374-378.

Liebrand, W.B., Jansen, R.W., Rijken, V.M. and Suhre, C.J., (1986) Might over morality: Social values and the perception of other players in experimental games. *Journal of Experimental Social Psychology* 22(3): 203-215.

Lindenberg, S. and Frey, B.S. (1993) Alternatives, frames, and relative prices: A broader view of rational choice theory. *Acta sociologica* 36(3): 191-205.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, *115*(3), 482-493.

McNeill, B. J., Pauker, S. G., Sox, H., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England journal of medicine*, *306*(2), 1259-1262.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995) *Microeconomic Theory* Oxford University Press

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation: was Titmuss right?. *Journal of the European Economic Association*, *6*(4), 845-863.

Messick, D.M. and McClintock, C.G. (1968) Motivational bases of choice in experimental games. *Journal of experimental social psychology* 4(1): 1-25.

Mill, John Stuart. (1836) "On the Definition of Political Economy, and on the Method of Investigation Proper to It," London and Westminster Review, October 1836. Reprinted in *Essays on Some Unsettled Questions of Political Economy*, 2nd ed. London: Longmans, Green, Reader & Dyer, 1874.

Montes, L. (2003). Das Adam Smith Problem: its origins, the stages of the current debate, and one implication for our understanding of sympathy. *Journal of the History of Economic Thought*, *25*(1), 63-90.

Montgomery, H. (1983) Decision rules and the search for a dominance structure: Towards a process model of decision making. *Advances in psychology* 14: 343-369.

Noe, T. & Young, H. P. (2014). "The Limits to Compensation in the Financial Sector". In: N. Morris and D. Vines, ed., *Capital Failure: Rebuilding trust in financial services*. Oxford: OUP.

Nieli, R. (1986). Spheres of intimacy and the Adam Smith problem. *Journal of the History of Ideas*, *47*(4), 611-624.

Otteson, J. (2002). Adam Smith's marketplace of morals. *Archiv fur Geschichte der Philosophie*, *84*(2), 190-211.

Paganelli, M. P. (2008) 'The Adam Smith Problem in Reverse' *History of Political Economy 40(2),* 365-82.

Parfit, D. (1984). *Reasons and persons*. OUP Oxford.

Pruitt, D. (1967) Reward structure and cooperation: The decomposed prisoner's dilemma game. *Journal of Personality and Social Psychology* 7(1): 21-7.

Pruitt, D.G. (1970) Motivational processes in the decomposed Prisoner's Dilemma game. *Journal of Personality and Social Psychology* 14(3): 227-38.

Rabin, M. (1993) Incorporating fairness into game theory and economics. *The American economic review* 83(5): 1281-1302.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, *17*(8), 413-425.

Roberts, R. (2015). How Adam Smith can change your life: An unexpected guide to human nature and happiness. Portfolio Trade

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process1. In *Advances in experimental social psychology* (Vol. 10, pp. 173-220). Academic Press.

Ross, L., and Ward, A. (1997). Naive realism in everyday life: Implications for social conflict and misunderstanding. In: Reed, E., Turiel, E., and Brown, T. (eds) *Values and knowledge*. Psychology Press, pp. 103-135.

Ross, M., (1975). Salience of reward and intrinsic motivation. *Journal of Personality and Social Psychology*, 32(2): 245-254.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Schick, F. (2003) *Ambiguity and logic*. Cambridge: Cambridge University Press.

Schmidtz, D. (2016). Adam Smith on Freedom. *Adam Smith: His Life, Thought, and Legacy*, 208-227.

Sen, Amartya (1977). 'Rational fools: a critique of the behavioural foundations of economic theory', Philosophy and Public Affairs 6: 317-344.

Shafir, E., Simonson, I. and Tversky, A. (1993) Reason-based choice. *Cognition* 49(1-2): 11-36.

Skyrms, B. (1998) "Review of Frederick Schick's Making Choices" *The Times Literary Supplement*. 949: 30

Slovic, P., Fischoff, B., and Lichtenstein, S. (1988) Response Mode, Framing, and Information-Processing Effects in Risk Assessment. In: D. Bell, H. Raiffa & A. Tversky (eds.), *Decision making: Descriptive, normative and prescriptive interactions* Cambridge: Cambridge University Press, pp.152-66.

Smith, A. (1759). *The Theory of Moral Sentiments.* London: A. Millar and Edinburgh: J. Bell. Reprinted at Metalibri v1.0s., 2005, <http://metalibri.wikidot.com/ title:theory-of-moral-sentiments:smith-a>, Sálvio M. Soares (ed.), accessed 11 February 2019.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations.* London: W. Strahan and T. Cadell, Reprinted by Project Gutenberg at <http://www.gutenberg.org/ebooks/3300>, accessed 11 February 2019.

Smith, V. L. (1998). The two faces of Adam Smith. *Southern economic journal*, 2-19.

Sober, E., & Wilson, D. S. (1998). *Unto other*s. Cambridge Ma: Harvard University Press.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, *10*(1), 69-89.

Tversky, A. and Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453-458.

Weirich, P. (2010) Utility and framing. *Synthese* 176(1): 83-103.

		Player 2 C	Player 2 D
Player 1	C	12,12	0,18
	D	18,0	6,6

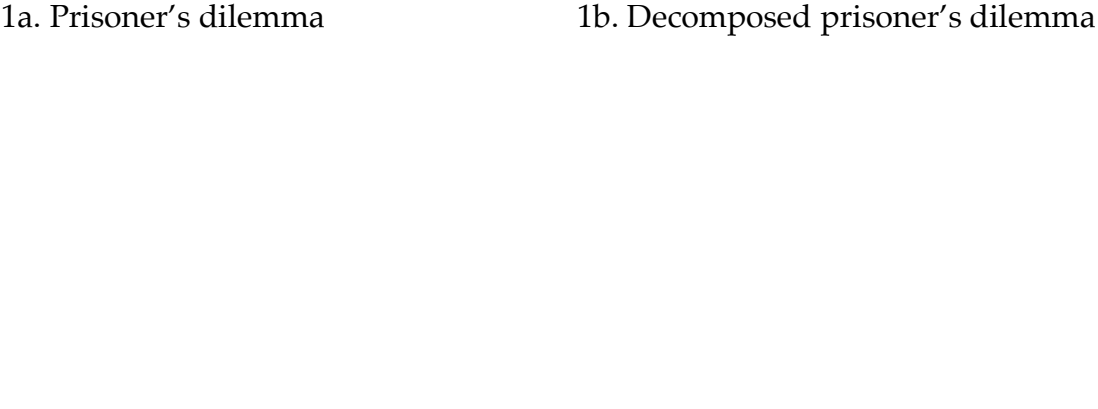
		Your Gains	Other's Gains
Your Choice	C	0	12
	D	6	0

Williamson, O. E. (1993). Calculativeness, trust, and economic organization. *The journal of law and economics*, *36*(1, Part 2), 453-486.

Zelizer, V. A. R. (1997). *The social meaning of money*. Princeton: Princeton University Press.

Figure 1. Alternative presentations of the prisoner's dilemma.

1a. Prisoner's dilemma



1b. Decomposed prisoner's dilemma