Strong reciprocity and welfare (workshop): list of abstracts

Antoinette Baujard (with Muriel Gilardone): Thinking differently about public action. The promising potential of Sen's idea of justice.

In a welfarist framework, individual welfare – or individual preferences – is the only relevant information to measure social welfare. Moreover, a theory of justice determines which and whose preferences are relevant to take into account; and it also characterizes how they should be aggregated. While this double framework frames a standard way to help public action, clarifying Sen's idea of justice allows to think out of this box. We claim that Sen's commentators misconceive his contributions precisely because they read him from the prisms of welfarism and of the theories of justice. Although our aim is not primarily to provide an exegetic review of Sen's works, it is a necessary step to renew the perspective as to how a normative philosophy is able to tackle public action with respect to agency and democratic principles. The structure of the paper is as follows. It first recalls how the two frameworks, welfarism and the theories of justices, works as two prisms that we need to be aware of to understand the contributions of Sen. Second, the value of agency, which justifies the focus on positional views rather than on individual preferences according to our reading of Sen's works, imposes to go beyond the dichotomy between subjectivity and objectivity. Third, we show that Sen reconciles agency and impartiality by acknowledging positional bias and by building upon a relational ethics.

Maria Bigoni (with Marco Casari, Andrea Salvanti, Andrzej Skrzypacz, and Giancarlo Spagnolo): The Importance of Being Even: Restitution in the Repeated Prisoner's Dilemma

We explore the role of restitution as a means to restore cooperation in repeated social dilemmas. In contrast to the memory-one strategies on which the recent experimental literature has focused, restitution strategies "propose" returning to cooperation by restoring payoffs lost for a past breach. They can also be seen as "fair", as they close the payoff gap created by deviations, making subjects even. We study metadata from experiments on finitely and infinitely repeated Prisoner's Dilemma, with perfect and imperfect monitoring, and find evidence in support of these strategies in all these classes of games. We then study the theoretical properties and empirical validity of the simplest restitution strategy we could identify - termed Payback - in the imperfect monitoring environment of Fudenberg, Rand and Dreber (2012). Two "puzzling" findings - that Tit-for-Tat is common even though it is not an equilibrium strategy, and that risk dominance loses its predictive power compared to environments with perfect monitoring – disappear once Payback is accounted for.

Urs Fischbacher: Rights, Duties, and Taboos: The Social Codex of Peer Punishment

Cooperation is a central part of human life but it is difficult to establish and maintain on its own. Peer punishment can promote cooperation when defectors are punished and cooperators are spared. However, peer punishment can also harm cooperation if punishment is used in a dysfunctional way. In an experiment, we investigate how people assess the appropriateness of different forms of punishment, including (several types of) second-order

punishment and counter-punishment. By assessing the appropriateness of both punishment and non-punishment, we can distinguish between when punishment is a right, when it is a duty, and when it is a taboo. We find that people generally see punishment as an obligation to do or not to do, rather than a right. This is particularly pronounced in the case of third-party punishment. We can summarise the assessments of appropriateness across all the different forms of punishment in three "commandments": 1. do not punish cooperators. 2. if you are a defector, do not punish. 3. punish people who break rules 1 or 2. People do not see punishing free riders as a form of second-order public good that should be enforced by punishment: Punishing people who do not punish free riders is considered socially inappropriate. Finally, retaliation can be a duty, in particular when the potential retaliator has cooperated.

Shaun P. Hargreaves Heap: The case for constitutional welfare economics and a Millian illustration

The case for a constitutional (or procedural) approach to welfare economics turns on two arguments/observations. First, there is considerable behavioural evidence that we do not always act to satisfy preferences and, when we do, these preferences cannot always be taken to be exogenous. Second, the prediction of the full consequences of any policy becomes less accurate, the more complex is the economy. The alternative constitutional approach judges policy by the character of the proposed rules or procedures that will constrain actions in themselves and not, as in the more usual approach, by their consequences. As an illustration, the paper considers JS Mill's proposal for rules of liberty. One cannot in general predict what will occur as the result of individuals' exercise of liberty. Instead, the rules of liberty are to be valued because they are intrinsically desirable. Mill's rules have a further desirable property: if individuals actually had exogenously given preferences and their actions to satisfy these preferences were fully predictable, the outcome would be Pareto efficient. In the more usual case, where these two conditions are not satisfied, the central issue for policy (and politics) becomes the determination of what constitutes a 'harm'.

Tobias Henschen: Strong reciprocity and welfare non-consequentialism

Traditionally, economists subscribe to a position of welfare consequentialism when ranking welfare policies: they use social welfare functions (SWFs) to derive ethical rankings of policies from rankings of the consequences that these policies have for the welfare of individuals. I will argue that the satisfaction of other-regarding preferences can amount to welfare, and that welfare policies cannot be ranked in the fashion of welfare consequentialism if individuals have other-regarding preferences in the shape of strong reciprocity. While other-regarding preferences in the shape of inequity aversion can be aggregated to form a specific (rank-weighted) SWF, other-regarding preferences in the shape of strong reciprocity cannot be aggregated in this way. There is no direct path from "reciprocity equilibria" to Pareto optima, and it is unclear whether individuals exhibiting strong reciprocity would be willing to accept compensating variations that lead to Pareto improvements. I will also argue that individuals exhibiting strong reciprocity have internalized (something like) a deontological constraint, and that rankings of welfare policies derive from this constraint if individuals exhibit strong reciprocity.

Michiru Nagatsu: (How) should behavioral policy and economic engineering exploit strong reciprocity? A virtue ethics approach

Economic experiments have established a class of phenomena called strong reci-procity in non-cooperative games such as the public goods game and the Ultimatum game. Unlike weak reciprocity, which refers to reciprocal strategies (e.g. tit-for-tat in prisoner's dilemma) that constitute a Nash equilibrium among self-regarding players, strong reciprocity implies sub-optimal strategies that do not maximize individual payoffs. Such strategies, however, have been modelled as maximizing a type of social preferences, such that strongly reciprocal strategies—reciprocating (perceived) good and bad intentions of other players with generous and mean responses, respectively— also maximize something. This technical innovation gives rise to problems, such as the incompatibility with the basic framework of expected utility theory (Guala 2006) and the challenge to square it with the standard welfare analysis (as pointed out in the workshop outline).

In this paper, I address welfare implications of exploiting strong reciprocity from a broader, systemic perspective, paying close attention to its psychological mechanisms and long-term aggregate implications, rather than formal properties of models of strong reciprocity. To do so, I highlight two types of asymmetries between strong positive and negative reciprocity. First, at the psychological level, the former seem to be caused by positive affect and emotion, such as gratitude, warm glow, trust and so on, whereas the latter seem to be caused by negative affect and emotion, such as anger, resentment, and so on. One could even argue that positive strong reciprocity is caused by basic social motivations for collaborative joint action, although the action is not temporarily synchronized (Godman et al. 2014). From the psychological perspective, then, social design should promote strong positive reciprocity while discouraging strong negative reciprocity, other things being equal. Second, at the aggregate level in a naturally occurring set-up of cooperation with elements of conflict, negative reciprocity such as uncoordinated voluntary punishment seem to have strong undesirable side effects such as revenge and feuds (Guala 2012). In contrast, positive reciprocity does not seem to have similar side effects. Therefore, the same conclusion seems to hold from the aggregate perspective. However, I argue that such conclusions cannot be derived from a utilitarian framework, either in a formal preference satisfaction variant or in a hedonist variant. Rather, sufficient normative evaluation of the strong-reciprocity-based social design requires a more systemic framework that takes into account endogenous and inter-dependent relationships between institutions and individual dispositions. I propose that virtue ethics as interpreted by Bruni and Sugden (2013) might offer such a normative framework, and assess this conjecture against cases in which the asymmetry between positive and negative strong reciprocity do not play out as suggested above (Bolt and Ockenfels 2012; Nagatsu et al. 2018).

Adam Oliver: Desert, Reciprocity and 'Das Adam Smith Problem'

In this talk I will argue that justice and desert are connected, and that desert is linked to the concept of reciprocity. That is, in our social interactions, we reciprocate (both positively and negatively) with those who we think 'deserve'. Although desert is underpinned by many possible considerations, I will contend that in our social interactions the intentions behind our

actions and the outcomes resulting from those actions are crucial components when apportioning blame and assigning credit. In basic economic, as opposed to social, interactions, however, the *expectation* that another party intends and thus undertakes an action is absent. There is thus no moral blame attached to people who do not intend to engage in any particular market transaction. This, for the most part, reduces economic interactions to two parties who intend to, and do, benefit themselves (and by extension, benefit each other), with no moral opprobrium directed towards those who do not intend to take part in the exchange. The considerations of desert that underpin many social interactions are thus absent from basic economic exchanges. I will offer the distinction between social and economic transactions in this regard as a resolution to *Das Adam Smith Problem*.

Julian Reiss: Three arguments against paternalism

Once again, paternalism is rearing its ugly head. Whether in academia — by behavioural economists of 'soft paternalism' conviction and more hardcore paternalist philosophers such as Sarah Conly — or in politics — by all those who advocate bans on meat, sugar, combustion engines, flying or, indeed, at times, leaving the house — paternalist proposals have gained much popularity in recent years.

In this paper I argue that an agent A's interference with another agent P's liberty is justified at best when three necessary (but not sufficient) conditions are met:

1. A knows what constitutes P's well-being or what his goals or interests are.
2. A knows that Z promotes P's well-being, goals, or interests.
3. Of all available actions that might promote P's well-being, goals, or interests, Z is the action that least interferes with P's liberty or autonomy.

I argue that for most agents (and governments in particular) deciding about the implementation of possibly welfare-enhancing interventions, typically, conditions 1)-3) are not met.

Maj-Britt Sterba (with Sören Harrs): Fairness Preferences and Support for Welfare Policies

Abstract: People disagree about what is fair. But how important are fairness preferences for understanding people's political disagreements about the design of welfare policies? And do exogenous shocks change what people perceive as fair? In this paper we study these questions in a large and representative sample of US Americans in the context of the coronavirus pandemic. First, we show descriptively that fairness preferences - identified using a novel experimental design - are a stronger predictor of people's support for welfare policies than their income. Fairness preferences also prove to be fundamental because they shape how much weight an individual attaches to beliefs about the causes of inequality. Second, employing individual-level panel data and an experimental manipulation, we find evidence that changes in support for welfare policies during the pandemic are rather caused by changes in beliefs about the causes of inequality or by self-interest than by changes in fundamental fairness preferences. Our results have implications for major models in political economy and for understanding the mechanisms shaping people's demand for fair policies.

Robert Sugden: Why psychological game theory is not a theory of welfare

My paper addresses the scientific goal of the workshop, i.e. 'to investigate the sense in which the satisfaction of other-regarding preferences in the sense of strong reciprocity qualifies as welfare'. I interpret this as the question: 'What, if anything, is the conceptually most coherent way of combining the empirical model of strong reciprocity with the normative model of welfarism?' My answer is there is no coherent way of doing this. Either we need a different theory of reciprocity (e.g. one based on team reasoning) or we need a different normative theory (e.g. Benthamite utilitarianism). Since the seminal paper by Rabin (1993), most social preference theories of reciprocity have treated reciprocity as being kind (unkind) to people who act on kind (unkind) intentions towards you, and use the framework of psychological game theory. This framework defines preferences over 'outcomes' which refer to a player's own beliefs, his beliefs about his coplayers' beliefs, and so on. But the interpretations of 'preference' and 'welfare' in welfarism require that a person's preferences are defined over *potential objects of choice*. The decision theory axioms that underlie the game-theoretic concept of 'payoff' (e.g. Savage's axioms) require that decision problems for individuals can be constructed by arbitrarily assigning outcomes to states of the world (to create acts) and then arbitrarily assigning acts to opportunity sets. Despite what many game theorists claim, the terminal node of a path of play in a game is *not* a legitimate carrier of utility. (This is a reprise of an argument in Sugden, *Economic Journal*,1991.)