The Savings Problem in the Original Position:

Assessing and Revising a Model

Eric Brandstedt

Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, United Kingdom

CPNSS, Lakatos Building, London School of Economics and Political Science, Houghton Street, London WC2A 2AE

Email corresponding author: e.brandstedt@lse.ac.uk

This is an accepted manuscript of an article published by Taylor & Francis in *Canadian Journal of Philosophy* on 25/10/16, available at http://www.tandfonline.com/10.1080/00455091.2016.1250202

The Savings Problem in the Original Position:

Assessing and Revising a Model

The common conception of justice as reciprocity seemingly is inapplicable to relations between non-overlapping generations. This is a challenge also to John Rawls's theory of justice as fairness. This text responds to this by way of reinterpreting and developing Rawls's theory. First, by examining the original position as a model, some revisions of it are shown to be wanting. Second, by drawing on the methodology of constructivism, an alternative solution is proposed: an amendment to the primary goods named 'sustainability of values'. This revised original position lends support to intergenerational justice as fairness.

Keywords: intergenerational justice; fairness; the original position; models; Rawls; constructivism; sustainability of values

I. Introduction

What we do today will often, via more or less complex causal pathways, have effects on, as of yet, non-existing future people. Driving cars to work, taking long-haul flights to holidays, constructing high-speed rail infrastructure, clearing forest for crop production, allowing car manufacturers lax emission controls, giving the go-ahead for the construction of a nuclear plant, and accepting a certain level of government debt are all future-oriented actions. Some of them will clearly have negative effects on future people, for instance, by contributing to climate change and the risks associated with that. Seemingly innocuous actions threaten to bring about a future world in which some people will have to struggle to maintain their life, health and security, and where some essential natural resources are irreversibly gone. Pondering such possibilities raises questions about how to account for the intuitive wrongness of such actions and how to justify obligations to future people. These questions are, however, hard to clearly analyse or even just to conceptualise in the first place.

The intergenerational setting appears to fundamentally challenge traditional conceptions of morality, such as the commonsensical idea of justice as reciprocity, expressed in judgments on the fairness of returning a favour and of doing one's part in a cooperative enterprise, as well as on the unfairness of not getting what one deserve, such as fair pay for a job done.[1] The standard understanding of the intergenerational setting (see Meyer 2015) seemingly makes this view of justice inapplicable. The directedness of time – that is, time's arrow being unidirectional – guarantees that while present people can affect future people, the inverse relation cannot hold. Add to this the, in philosophical discussions commonplace, assumption of non-overlapping generations, and the 'non-reciprocity problem' presents itself. The relation between contemporary and future people seemingly is not one of reciprocity, and thus not one of justice so understood. In fact, this of course depends on the meaning and delineation of different generations. If generations, for instance, are conceptualised as age cohorts, then there can be reciprocal relations between them to the extent that they coexist.[2] I will, however, follow the standard practice and frame the relevant question to be discussed in the following way: what is the present generation morally required to save[3] to future generations it does not and will not coexist with?

The non-reciprocity problem is general, but the focus here is on a specific reciprocity-based theory, one that sees society as 'a fair system of cooperation over time from one generation to the next', that is, John Rawls's 'justice as fairness' (Rawls 2001, 4; cf. Rawls 1971). This is a theory anchored in reciprocal relations because of its starting point in the fair balancing of competing claims over relatively scarce resources by persons roughly equal in power. Despite, or maybe because of, that, Rawls provides the first systematic treatment of intergenerational justice in the contemporary debate (Meyer 2015). But for all the clarity he brings to the analysis of this novel dimension of justice, he clearly struggles to develop this 'extension' of his general theory of justice.

Rawls intends his well-known model of a fair bargaining situation, 'the original position', to be applicable also to thinking about intergenerational obligations: the parties, behind 'the veil of ignorance', are deprived of knowledge of, among other things, what generation they are part of, if it is in an earlier or later stage of development. But whereas that in the standard case guarantees that no party, or group of parties, can tailor principles that unfairly favour them, it does not in the special case of justice between generations. 'The one case where this conclusion fails', Rawls (1971, 140) concedes, 'is that of [intergenerational] savings'. The intergenerational setting, as characterised above, turns the original position into a kind of prisoner's dilemma, where it is rational for the parties to not save for future generations, notwithstanding what other generations do and their place in the chain of generations. For, Rawls (1971, 140) notes, 'previous generations have saved or they have not; there is nothing the parties can do now to affect that'. Given the directedness of time, depriving the parties of knowledge about their generational belonging does not prevent the unwanted result. The lack of reciprocity makes it rational for them to not save either way. For future generations can do nothing for them, and even if previous generations can benefit them by choosing to save, such a choice is already a fact, and although the parties do not know its content, they know that whatever it is, it cannot be affected by their present decision. I will refer to this as the *savings problem in the original position*.[4]

This problem has been widely discussed, not least by Rawls who elaborated on several different solutions, as we will see below.[5] There is, however, something unsatisfactory about the discussion. It seems to not have fully grappled with the idea of the original position as the model Rawls took it to be. The savings problem in the original position is a theoretical one emanating from a model-level, which makes it essentially different from real-world intergenerational problems, such as hyperinflation and climate change. In order to solve the savings problem in the original position, or for that matter to evaluate proposed solutions, we

thus need to know what it is a model of and what it is supposed to do. Otherwise, any freely modelled solution, though bearing the appearance of one, may turn out just *ad hoc*.

I shall proceed in the following way. In section two, I present three different perspectives found in Rawls's theory: that of the parties in the original position, that of citizens in a 'well-ordered society', and that of you and me assessing the virtues of the model. I present a distinction between a 'model-level', corresponding to the original position, and a 'target-level', corresponding to the well-ordered society, on the basis of which I relate to the third point of view and to what would there be a useful model. Drawing on Rawls's methodological approach of constructivism, I argue that the original position is not just a workable device of representation of a choice situation, but also of persons who choose. Furthermore, any revision of the model must be grounded in the public culture of a democratic society. With these desiderata of the model in place, I assess three different solutions that have been proposed to the savings problem in the original position, and argue that they are all found wanting on account of being more or less *ad hoc*. In section four, I sketch an alternative revision of Rawls's theory, which concerns what he calls 'the primary goods', and which can be understood as a representation of a conception of the person found in democratic societies on due reflection. It is a proposal of an amendment to these, which I call *sustainability of values*. Given that many values and projects are conditioned on their future continuation, sustainability of values is an all-purpose means it is rational for the parties to advance. Finally, in section five, I briefly discuss the merits of using the original position for organising thinking of justice in the intergenerational setting.

II. The Original Position and what it Represents

What kind of problem is the savings problem in the original position? Is it a reflection of a general obstacle to our thinking about intergenerational justice, or merely a queer consequence of Rawls's theoretical construct? How might we begin to understand the relevance of the highly abstracted features of this philosophical problem to the concrete

questions policy-makers face in thinking about future-oriented actions? I suggest that we begin by clarifying the role and function of the original position before turning to the problem and the various revisions that have been proposed. This will help us not lose sight of the relevant practical problems – such as if, and to what extent, we are obliged to consider the interests of future people in contemporary decision-making – when treading through some rather abstract theoretical intricacies. It may also help us see why it is so hard to visualise a fair bargaining situation for negotiating principles of justice while keeping in mind the idea of society as a fair system of cooperation over time.

There is no need to reiterate all of its features, but we do need to recall the underlying idea of the original position, as this seems so easily lost in discussions. The first thing to note, then, is that Rawls intends it to be 'an expository device' that helps us, that is, you and me, to organise our intuitions relevant to justice. 'The idea', he writes, 'is simply to make vivid to ourselves the restrictions that it seems reasonable to impose on arguments for principles of justice, and therefore on these principles themselves' (Rawls 1971, 18). The original position models what we consider, on due reflection, reasonable conditions for reasoning about such matters; or, as Rawls also puts it, the reasoning of citizens of a well-ordered society, that is, a society effectively regulated by principles of justice that all citizens publicly accept.[6]

We are, however, not citizens of such a society, nor do we always apply criteria of reasonableness in normative discussions. We do not only fall short of what we would consider good standards of reasoning – for instance, by being biased or making inconsistent judgments – but also, just as often, are blind to our own failures. In order to mitigate such shortcomings, which we must if we are to organise our society by justified principles of justice, we could try to simulate the process of reasonable choice. Thus, Rawls writes: '[a]t any time we can enter the original position, so to speak, simply by following a certain procedure, namely, by arguing for principles of justice in accordance with these restrictions' (Rawls 1971, 19).

This understanding of the original position brings out just how demanding Rawls's approach is, both for the theorist and for, let us say, the practitioner: it requests of the former that it gather and assemble what, on due reflection, can be considered criteria of reasonable or fair argumentation; and of the latter that it be competent enough to simulate reasoning in keeping with those criteria. This should humble us somewhat as theorists. We may be mistaken about what the criteria for a fair choice situation are, especially so when factoring in the novel features of the intergenerational setting, as we must for future-oriented decisions. We should also bear in mind that a desideratum for an interpretation of the original position thus is that it is feasible for people to simulate reasoning in accordance with it.

Describing the original position as a device of representation that people can adopt to simulate a target system is, however, not satisfactory. The two desiderata encapsulated in that – that it be descriptively adequate and facilitate the simulation of the target system – are at any rate insufficiently specified. This can be gathered by reflecting on the methodology, called constructivism, Rawls develops in his later writings (1980; 2005).[7] I will soon explain what I take constructivism to mean, but consider first how it influences the presentation of the original positions. Rawls writes: '[the] role [of the original position] is to establish the connection between the model-conception of a moral person and the principles of justice that characterize the relations of citizens in the model-conception of a well-ordered society' (Rawls 1980, 520). We can see here how the envisaged role for the original position now embodies a further function. It functions as a 'mediating conception' between the conception of the person and the normative principles and so establishes a connection between the two (cf. Rawls 1980, 516). Rawls intends the original position to carry out this function by being a model, not only of an idealised choice situation, but also of an idealised chooser. The moral conception of the person as free and equal, for instance, is part of the model conception of justice as fairness (cf. Rawls 1982, 182).

Although there is no unison and undisputed meaning of constructivism, some of its core ideas can elucidate the revised role of the original position.[8] Constructivism, Rawls (1980, 516) explains, 'specifies a particular conception of the person as an element in a reasonable procedure of construction, the outcome of which determines the content of the first principles of justice'. This is not only a paraphrase of his earlier method (that is, the method of reflective equilibrium), but also a substantial addition to it, namely the emphasis put on the conception of the person for justifying principles of justice. This methodological recommendation thus underlines the role played by the person in the making of fair decisions. In other words, to model a fair decision situation involves modelling what persons would have to be like in order to make such fair decisions. Constructivism accordingly qualifies the first desideratum of the original position: for it to be descriptively adequate of its target system, it needs a richer description that includes also a conception of the person.

There is also a second qualification. To get to that, let us first note some things about what constructivism is not. First, it is not a method for discovering normative ideals; rather, it assumes some such ideal, such as the need for justice. What it does is to pose two preliminary methodologically questions: (1) what conditions must obtain for a normative justification of the considered ideal to be acceptable; and (2) how must persons be constituted for that to happen? Constructivism thus complicates the justification of any normative ideal: it does not accept any conditions as conducive to normative justification, and not any person as able to reason accordingly. The theorist's job is to assess not the acceptability of individual principles, but rather the conditions for any acceptable justification and what ideal person such conditions assume. The second thing to note is that constructivism is not, or at least not directly, a method for normative justification. The ultimate justification of any normative ideal is if it can be accepted in reflective equilibrium, possible only in a well-ordered society (see Rawls 1980, 534; cf. Rawls 2005, 28). Furthermore, not only principles of justice but also criteria of fairness are ultimately justified by practitioners rather than theorists. It is not

the case, Rawls argues, that criteria of reasonable or fair decisions can be taken out of thin air, nor simply be invented by the theorist.

That said, the theorist still has a crucial role to play, which takes us back to the revised function of the original position:

> The aim of political philosophy, when it presents itself in the public culture of a democratic society, is to articulate and to make explicit those shared notions and principles thought to be already latent in common sense; or, as is often the case, if common sense is hesitant and uncertain, and does not know what to think, to propose to it certain conceptions and principles congenial to its most essential convictions and historical traditions (Rawls 1980, p. 518).

Rawls gives the theorist the role of assembling the most essential convictions of what makes for a fair decision and to connect these in ways conducive to the justificatory process (cf. Rawls 2005, 108). Here we get to the second qualification, which regards the second desideratum of the original position. Any revision of it must resonate with conceptions of society and person found in the public culture of a democratic society. This is important both to get any simulation off the ground and for simulators to learn anything from reasoning in accordance with the model conception. Although the original position cannot fully justify principles of justice, it can approximate such a justification, rightly characterised.

Before we move on to assess proposed solutions to the savings problem in the original position, let us add one other clarification. If the original position is a representation of an ideally fair choice situation, the well-ordered society, then the savings problem in the original position too has a counterpart. That the parties in the original position would not adopt a just savings principle represents a possible way in which a society could fall short of being well-ordered. That is, the possibility that a purportedly fair reasoning process leads to the undesirable conclusion that society should be organised such that nothing is saved to future generations. Let us call this the *target problem*. This, of course, takes for granted a judgment about that possibility, namely that the decision to neglect the interests of future people is

unfair and so wrong. Although also this judgment could be challenged, I will proceed under the assumption that it is valid, or as Rawls (2005, 123) would put it, a fact about 'the possibilities of construction'.

However, granted the assumption, against whom would it be unfair? There is, to wit, an ambiguity in the representational relation between the savings problem in the original position and the target problem: either we understand the original position as modelling the conditions for a fair agreement among contemporaries of a well-ordered society, or the conditions for a fair agreement among its contemporary but also future members (or still more inclusively, all affected). This ambiguity comes from the subjects to the allegedly unfair agreement being unspecified: it is unclear whom the theorist should take into account in conjecturing fair conditions. Thus, there are at least two versions of the target problem: (1) unfairness among contemporary citizens, and (2) unfairness among non-contemporary citizens. Deciding which one is relevant is challenging, because although accounting for the problem in the second sense is more intuitive – a no savings decision will, after all, primarily affect future people – there are good countervailing reasons for sticking to the first. It is, for instance, natural to think that the original position represents the activity of collective reasoning under certain conditions, and such reasoning can, of course, only comprise existing persons. Relatedly, if the explanation for why a policy or proposal is unfair is that someone subjected to it have reasons to not accept it, then policies and proposals cannot be unfair against future people, as they can neither accept nor reject things in their current non-existing state.

There is, however, an important, though subtle, difference in conceptualising decision-making of relevance to future generations as either taking into account the likely impact on future people's interests or, as it were, striking a fair deal with future people. In other words, between acknowledging a moral obligation *regarding* future people and a moral obligation *owed to* future people (cf. Mazor 2010, 381; Brandstedt 2015). Thus, even though future

people are not subjects to contemporary agreements, they will in many cases be affected. In other words, they are third parties (cf. Rawls 1971, 128) to future-oriented decision-making of today, and accordingly their interests are made relevant. To the extent that present people represent the interests of future people, their interests are turned into proper claims that need to be balanced against competing ones. As I will argue in section four, there are surprisingly many ways in which future people's interests are represented in projects and practices of today. If that argument is made convincing, it shows that there are reasons for us – here and now – to reject indifference to the interests of future people, even if those reasons ultimately derives from a justificatory process only involving contemporaries.

III. Assessing Solutions on a Model- and Target-Level

Now, the question is not if there is a way in which the original position can be revised so that it does not produce the counterintuitive result. There are indeed several such possible revisions, as we shall soon see. The relevant question is if a revision can be justified in a non-*ad hoc* way.

A first possible revision is mentioned by Rawls (1971, 139-140) already when he presents the savings problem in the original position, and although he quickly dismisses it, it has been picked up by some others as promising (e.g. Barry 1989, 505-510; Gardiner 2009, 114-115). It is to drop the assumption, which we have been working under but not yet spelled out, namely that the parties in the original position all are contemporary: the so-called 'present-time-of-entry interpretation'. To do so would transform the original position into a 'general assembly' composed of representatives of both the actual and merely possible generations, negotiating the terms of a truly intergenerational contract. Rawls (1971, 139) dismisses this idea for the good reason that it would 'stretch fantasy too far; the conception would cease to be a natural guide to intuition'.[9] If the original position is an expository device then it is essential that we *can* adopt it, so that it can facilitate the organisation of our intuitions. But it is difficult if not impossible to imagine oneself in a negotiation, which has

on the table principles threatening the existence and identity of those represented. Furthermore, how could any principle be accepted in such a situation, when each decision results in the non-existence of some represented possible people?[10]

A second revision is to make a 'motivational assumption' of the parties in the original position. In the first configuration, the only requirements placed on their conception of the good are ones of rationality (e.g. that they take effective means to their ends). The idea of a motivational assumption is to introduce a further, and different kind of, constraint on their conception of the good. If we think of the original position as a kind of dynastical model, where the parties as 'heads of families' represent not only their own conception of the good but also that of their closest descendants, then we might avoid the unwanted conclusion (Rawls 1971, 128-129). In this reconfigured model, the parties have reason to save for future generations, because it is instrumental to the furthering of their rational self-interests in this newly specified role.[11]

I will not here repeat all of the criticism that have been raised against this proposal, but just mention one salient point, in a slightly new take. Assuming that the motivational assumption solves the savings problem in the original position, it is far from clear that it solves the target problem. If it does not, maybe that justifies Jane English's (1977) judgment that the solution is *ad hoc.* On the target level, the motivational assumption seems to suggest that the unfairness comes from lack of care. Must we, accordingly, not only be reasonable to our fellow citizens, but also empathic towards our closest relatives when we justify the conditions of a just society? Think of the, seemingly well-ordered, society organised fully negligent of the interests of future people, and of how its citizens relate to their descendants. The citizens certainly seem to exhibit carelessness and lack of empathy, even to the extent that it is hard to understand their behaviour, but is it unfair that they do so, or just puzzling?[12] Conversely, most real people, of course, care greatly about their offspring and are highly motivated to further their wellbeing, yet we have climate change and a whole set of

intergenerational problems. The target problem is unlikely due to lack of care, and so a motivational assumption is unlikely to solve the savings problem in the original position.[13]

Even if the above reasoning is an elaboration thereof, it is in full agreement with Rawls's argumentation. There are indeed good reasons for dismissing the general assembly interpretation and for not opting for the motivational assumption. We now come to a point, however, where there are reasons to deviate from his course of reasoning. Rawls presents his considered response as a simple solution to the savings problem in the original position, whereas it is unclear to me both if it solves the problem and if simplicity should be the arbiter of that. The proposal is of another kind of revision of the original position. If the motivational assumption is an addition to the parties' (thin) conception of the good, this is proposing to revise the choice situation itself, that is, the set of conditions – such as, that principles must have general form (see Rawls 1971, 130-142) – it is composed of.[14] The proposal is the following:

> [T]he parties are to agree to a savings principle subject to the condition that they must want all previous generations to have followed it. They are to ask themselves how much (what fraction of social product) they are prepared to save at each level of wealth as society advances, should all previous generations have followed the same scheme (Rawls 2001, 160; see also Rawls 2005, 274).

I will refer to this as the *additional constraint*. Rawls's description of it as a simple solution presumably comes from thinking of it as preserving both the idea of society as a fair system of cooperation between generations and the present-time-of-entry interpretation of the original position. However, if we fully take into account the idea of the considered model presented above, this solution begins to look less satisfactory.

I should first clarify that additional constraint is not a proposal to shift from, what Rawls calls, 'nonideal theory' to 'ideal theory', that is, to introduce a condition of general (or strict) compliance, as some have thought (e.g. Heyd 2009, 178-181; Attas 2009, 201-203). It is not because that assumption is already in place: this is ideal theory. Rawls has all along

assumed that the principles of justice are worked out under the assumption that all subjected comply – and this has not prevented the savings problem in the original position. This is, of course, premised on the first version of the target problem. According to this, when negotiating a conception of justice, it is irrelevant whether or not past generations have complied and similarly whether or not future generations will comply – except if such things can be suspected to affect the stability of the conception. Apart from that the fairness of the choice is decided by considering only relations between contemporaries. Thus, the reason for why it seems rational, though unfair, for the present generation to not save for future generations is not because of noncompliance, either on part of some members of the present (that has been assumed away) or on part of some other generation (that is not relevant). Strict compliance is not the solution.

Consider now how additional constraint seemingly solves the savings problem in the original position. In the revised model there *are* reasons for the parties to adopt a just savings principle. Any decision to not save would defy additional constraint, as it would not be compatible with wanting previous generations to have followed it. In other words, it is inconsistent for the parties to uphold simultaneously the following two desires: to want previous generations to have saved *and* to want to not save for future generations. Consider now how this could be reconstructed on the target level. Basically, the idea is that what is missing, and what accounts for the unfairness of the considered problem, is the application of something like an universalizability test to proposed savings policies. What accounts for the shortfall of fairness in the less-than well-ordered society accordingly is that the savings decision (i.e. of zero savings or even dissaving) contradicts what its citizens implicitly want previous generations to have done for them. I write 'implicitly' because the citizens presumably need not make the desire about past savings explicit: mere appreciation and consent implicit in accepting the institutional order of society could be enough to contradict a desire for no future savings. On yet another formulation of this proposal, fairness demands

that all desires relevant to intergenerational savings have the form of a law, such that it could be held by any, and so all, generations

There are, as I see it, at least three reasons for being somewhat sceptical of additional constraint as a solution.[15] The first is that it effectively requires a particular desire, namely to want previous generations to have saved at a certain rate. The proposal thus is not just that fairness demands of citizens that they iron out any rational inconsistencies in their sets of desires before allowing them to influence the fabric of society. That is another requirement already assumed, namely ordering or completeness (Rawls 1971, 133-134). As noted, the proposal does amount to rationally require a particular desire. That is, however, a controversial point to make, for instance, in standard decision theory. The second reason for scepticism comes from additional constraint being more specific, that is restricted, than any of the other constraints of the original position. Whereas the others are formulated in general terms and all are plausible candidate conditions for an explication of the concept of right, additional constraint seems to have only one function, that is, to solve the savings problem. The worry thus is that it is *ad hoc*, which would be mitigated only by independently justifying it as being in concordance with what we would, on due reflection, consider fair. Until such justification is supplied, there is something wanting about additional constraint.

Finally, the third reason for scepticism relates to the previously discussed desiderata of the original position as an expository device: how does additional constraint stand up to an evaluation of the model in terms of its function? To repeat, Rawls thinks additional constraint is justified because it is a reconfiguration of the original position that produces the intuitively desirable result while preserving the present-time-of-entry interpretation. Taking for granted that it should also be descriptively adequate, the seemingly only other desideratum of a solution, we can infer, is that it maintains the present-time-of-entry interpretation so that it *can* be adopted. But as was argued in section two, this is to underappreciate the role the original position could play. It is not just a workable device of representation, but more

specifically of a particular choice situation with citizens suitably related to a certain self-understanding. So, although we can enter the original position with additional constraint and reconstruct the reasoning to the intuitive result of a just savings principle, to be content at that would be to short-change the potential of the model.

IV. Sustainability of Values – Another Primary Good

On the basis of the discussion in section two, one aspect of Rawls's theory stands out as especially interesting in responding to the savings problem in the original position and its target problem. It has not yet been discussed, although it plays a crucial role in motivating the choice of the parties in the original position. It is Rawls's account of 'primary goods'.

Rawls first defines primary goods as 'things that every rational man is presumed to want' (Rawls 1971, 62), but later revises it to mean 'persons' needs as citizens' (Rawls 2005, 179; cf. Rawls 1982). These are either social or natural, but since it is only goods of the former kind that are under direct control, it is those he focuses on. Rawls (2005, 181) writes: 'The basic list of primary goods (to which we may add should it prove necessary) has five headings as follows:

a. basic rights and liberties […];
b. freedom of movement and free choice of occupation against a background of diverse opportunities;
c. powers and prerogatives of offices and positions of responsibility in the political and economic institutions of the basic structure;
d. income and wealth; and finally,
e. the social bases of self-respect.'

These are specified by 'identifying a partial similarity in the structure of citizens' permissible conceptions of the good' (Rawls 2005, 180).[16] No matter what they desire or value, these are things citizens need to advance their conception of the good. It is thus 'a shared idea of rational advantage', or a standard of legitimacy that allow for interpersonal comparisons in a

situation of opposed and incommensurable interests, such as that of a contemporary democratic society (Rawls 2005, 180; cf. Rawls 1982, 161). The underlying problem of finding a public understanding of what are appropriate claims to make in a pluralist society is solved by the account of primary goods. There is also a counterpart problem on the level of the original position. The veil of ignorance deprives the parties of the possibility of unfair bargaining, but also of the possibility of a rational agreement because of the scant information allowed. The idea of primary goods solves this too. If we assume that the parties prefer as large shares as possible of primary goods, then, it is rational for them to agree on principles of justice, even without having to assume anything about the determinate conceptions of the good of those represented.

I now want to defend the claim that the savings problem in the original position, that is, the legitimising of no savings to future generations, can be dealt with by an addition to the primary goods. As a sixth category to the existing five, I propose: *sustainability of values*. Sustainability of values is a primary good in the sense that it gives values broadly conceived – underlying preferences, commitments, projects and traditions – point and meaning by guaranteeing their continuation over time. I will soon motivate this proposal further, but consider first how it solves the problem, on both model- and target-level. With sustainability of values as a primary good, it is rational for the parties to agree on a just savings principle, because that advances this primary good irrespective of what projects, traditions and preferences the citizens represented have once the veil of ignorance is lifted. As to the target problem, it can be solved in the following way: the group of people treat one another unfairly by choosing not to save anything for future generations, as doing so jeopardises the preferences, commitments, projects, and traditions whose value and meaningfulness are conditioned on a future continuation and maintenance of society. Put in another way, to not save is unfair because it treats persons as severely limited temporal beings, like mayflies if you will.

The primary good sustainability of values can be motivated by drawing attention to what Samuel Scheffler (2013) recently has put forward as 'the afterlife conjecture'.[17] Scheffler uses 'afterlife' in an unconventional sense, as the continuation of the material and social world after our personal death. His idea is that by reflecting on scenarios where it is common knowledge that there will be no such afterlife, we learn something about the concept of valuing and what it presupposes. The main point is presented as a conjecture: 'people would lose confidence in the value of many sorts of activities, would cease to see reason to engage in many familiar sorts of pursuits, and would become emotionally detached from many of those activities and pursuits' (Scheffler 2013, 44). The reason for this, Scheffler argues, is that the afterlife is a condition for many, if not all, of the things that matter to us to continue to matter, be important and valuable.

The afterlife conjecture thus reveals hidden dimensions of valuing, for example, that it is to some extent conservative: 'to value X', Scheffler (2013, 60) explains, 'is normally to see reasons for trying to preserve or extend X over time'. There are some examples in which this is especially evident. On a general level, these examples are of institutions, practices, activities, ways of life, traditions, and group-dependent projects that have temporally distant goals and diffused benefits. More concretely, think of, for instance, cancer research, social activism, and large infrastructure projects. These are projects and practices whose point and purpose directly depend on the afterlife. Less obviously participation in cultures, religions and traditions, such as taking on a national identity or carrying on a particular cultural celebration, also have such dependence. And most far-reaching, one may follow Scheffler's proposal to the conclusion that the afterlife is a condition for any thing at all to matter.

Setting aside that strongest version of the thesis[18], the gist of the afterlife conjecture can enrich our understanding of agency in the intergenerational setting. The dependence on the continuation and preservation of things of a transboundary character, expressed by the afterlife conjecture, suggests that present-day individual interests are temporally interlaced

with the interests of future people. That there will be a future, in which people continue on and maintain contemporary projects and practices, is a condition for many things mattering today. Thus, measures taken to increase the chances of such maintenance are purposeful irrespective of what specific transgenerational projects and practices we invest in.

The introduction of sustainability of values in the original position is a revision or, if you will, an update of the general knowledge allowed to the parties, but leaves intact their motivation. It is, in other words, compatible with the (thin) conception of the good, Rawls calls 'goodness as rationality'. It does not assume that the parties, or those represented, have any particular determinate conception of the good, other than that it is a reasonable one, subordinated to the fairness norms of the choice situation. Thus, although the proposal seemingly is similar to the motivational assumption dismissed above, it clearly differs in this respect. The motivational assumption ascribes a determinate, if only partial, conception of the good, which includes care for ones descendants. Even though such an ascription would make the model descriptively more adequate of actually existing people – most of us certainly care about our descendants – it would distort the underlying target problem. The original position should be descriptively adequate of how fair choices are made in the well-ordered society, not of how actual choices are made in our actual society. It is, in this sense, a normative model. To repeat, the theorist's job is to make explicit common sense normative judgments and connect them in ways which facilitate the justificatory process. That cannot always, or even very often, be carried out by surface level descriptions of what people actually care about and prefer. A better approach is the constructivist one that informs the development of the account of primary goods. That is, to think about what is in a type situation a salient conception of the person and about what that involves.

Given the intimate connection between normative ideals and conceptions of the person, as stressed by constructivism, we can work from two directions. The savings problem in the original position could be understood as the result of assuming a too sparse conception

of the person; and the other way around, reconsidering the conception of the person could facilitate the elaboration of novel normative requirements. With this we get a kind of openness lacking in additional constraint. In addition, we avoid the *ad hoc*-charge, as sustainability of values is independently justified rather than just assumed to solve the modelling problem. The conception of the person that goes with it is not created out of nothing.

The proposal can be elaborated on somewhat by briefly discussing three worries one might have in relation to it. First, the way in which sustainability of values was presented can suggest that it is a claim about all conceptions of the good. That is, a proposal that whatever is valuable to people, they have reason to maintain it over time. But is this really true of all values and of all conceptions of the good? Can we not imagine, say, a hedonist, who only values net pleasure satisfaction, and that such a person could rejoice without being the least worried about or dependent on a future continuation of the world? Although, as noted above, Scheffler also gives us reasons to doubt that hedonistic values are correctly described in this way, this is not my response (see fn. 20). We might well imagine a rational 'presentist hedonist' – like the Ancient philosopher Aristippus (Wolfsdorf 2013, 22) – as one conception of the good in a well-ordered society. But in order for rational hedonists to be reasonable in the Rawlsian sense, they must blind themselves to that particular conception of the good in negotiating principles of justice. Although their own conception of the good may not mandate future savings, they do not know this. But they do know that sustainability of values is a primary good. Thus the presentist hedonist is not a challenge to the discussed solution to the savings problem in the original position.

The second worry, or rather set of worries, concerns how sustainability of values seems to bring Rawls's theory closer to communitarian theories (e.g. de-Shalit 1995). One might, for instance, be concerned that it amounts to a thick and controversial conception of the good, which would make the theory parochial, or at least less than universally applicable.

Or, alternatively, that it makes the theory conservative. The response to such worries can only be sketchy here as there is insufficient space to develop the more concrete discussion of the just savings principle a full answer would require. But note, as a tentative response, that the relevant sense of value neutrality is not jettisoned by the proposal, as was hopefully made clear in the reasoning above. Note also that the Rawlsian theory developed here is on a higher level of abstraction than communitarian theories.[19] Sustainability of values summarises reasons to preserve values underlying various practices and projects, but these are not necessarily reasons to preserve any particular culture or tradition that may instantiate them. Because of that, the resultant theory cannot be said to be improperly conservative.[20]

The third worry relates to the previous two. Perhaps the arduousness of finding a solution to the savings problem in the original position has been in vain. Joseph Heath (2013) argues that the non-reciprocity problem is not a genuine problem for contractualist theories. We see this once we drop the abstraction of non-overlapping generations and work out the implications of the relevant sense of (indirect) reciprocity, he argues. The easiest way to demonstrate his point – without the game theoretical language in which it is expressed – is by using the example he uses consistently, namely that of a pay-as-you-go pension scheme. Such an arrangement is unfunded in the sense that those who currently benefit, the old generations, are paid not by their previous savings but by direct payments from the young generations. The question thus arises for the young generations: why should they dispense a fraction of their salary to the old generations' pensions when they cannot expect anything in return from the elderly? The answer comes from understanding the structure of the scheme, how benefits flow 'upstream' from younger to older generations. Granted the continuing existence of the system, the young can expect that just as they benefit the old generation, one day they will be the beneficiaries of their future successors. On the assumption that the benefits of the system outweighs the costs, and that it is expected to continue indefinitely, it is rational for each participant to contribute to its maintenance, in spite of the lack of direct reciprocity.

A conclusion from Heath's argument, one may think, is 'if it ain't broke, don't fix it'. But that would be a jump. Heath's argument is convincing with regards to existing intergenerational projects, such as pay-as-you-go pension schemes, but not more generally. Once such systems are set up, each participant stands to benefit from the preservation of the system and the lack of direct reciprocity between payment and benefit is no obstacle to that. This is so even on the standard assumption of a narrow conception of self-interests, such as revealed preference, something for which the existence of such systems is proof of. However, it is much less clear how it can account for a more general sense of intergenerational unfairness, which covers also the unfairness of, say, unmitigated greenhouse gas emissions. As there is yet no effective mechanism for abating climate change, it is not in anyone's narrow self-interest to pay for climate change mitigation. The challenge with respect to many of the most pressing intergenerational problems is to justify why the present generation is morally required to reserve resources, save or invest in the absence of anything like a mechanism for indirect reciprocity. This can be met with what has been proposed above. Sustainability of values is an explication of our self-interest, from which we could learn something that cannot be read straight off of revealed preference, namely that the future matters greatly to us.

V. Conclusion

I have argued for a previously neglected way of adapting Rawls's justice as fairness to questions about justice between generations. It begins by recognising the original position as a model-conception of a specific target-problem in the well-ordered society. Accordingly, there is not only the savings problem in the original position, but also a way in which *that* society falls short off fairness. The argument continued by following Rawls's constructivist methodology. I argued, first, that any revision of the original position must meet two desiderata: it must preserve an accurate representation of the target and allow for feasible simulations. Second, that it is not only a model of a choice situation, but also of relevantly

situated choosers, and that the theorist developing the model must only make conjectures grounded in the public culture of a democratic society. On the basis of these methodological remarks, I argued that proposed revisions of the original position in the intergenerational setting are all wanting. A more promising aspect of the theory to revise is the account of primary goods, that is, the general knowledge allowed to the parties in the original position. I proposed sustainability of values, which is an all-purpose means, rational for the parties to advance no matter the determinate conception of the good of those represented.

By way of conclusion, consider what the proposal means on a target level and what implications it has from the point of view of you and me. The target problem is explained by the narrow conception of the person assumed. If persons are assumed to be atomistic and short-sighted, then one should not be surprised that the theory makes rational no intergenerational savings. But should we, appreciative of the crucial role played by the conception of the person, instead provide a richer description of the person and its interests, including projects, cultural traditions and values entertained, then a rational and fair choice involves intergenerational savings. In other words, once we as theorists think about the relation between present and future people, we uncover a ground from which intergenerational savings can be justified. This methodological observance also has another positive effect. It sheds light on what is at stake when we go about organising our intuitions around the novel features of the intergenerational setting, namely the relation between our interests and those of future people.

Acknowledgements: I am grateful to Clare Heyward, Maxime Desmarais-Tremblay, Nicolas Wüthrich, Joe Mazor, and the anonymous reviewers of this article for their helpful comments on earlier drafts. I also wish to thank the audience of the PhD seminar at the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science for helpful discussion. I am also indebted to the Swedish Research Council for their generous funding of my postdoctoral project in which this article really is a sidetrack.

Endnotes

[1] For a general introduction to 'justice as reciprocity' (see Rawls 2009, 190-224; cf. Barry 1989, 463-493); for a presentation of some of the problems it faces in the intergenerational setting (see Gosseries 2009).

[2] There are models that do not assume non-overlapping generations, not least ones developed by economists. The best known probably is Paul Samuelson's (1958) overlapping generations model, which shows that 'cold and selfish competitive markets' (Samuelson 1958, 473) do not lead to intergenerational savings without further complements. Also some philosophers have assumed overlapping generations (e.g. English 1977; Mazor 2010; Heath 2013; cf. Gosseries 2009). Among them, it is common to argue that the more 'realistic' starting point saves the contract theory from the above-mentioned problem. There is merit to this claim, which is something I will come back to below in the main text. But I will first consider the possibilities of working out intergenerational obligations under the more challenging assumption of non-overlapping generations.

[3] 'Save' is a technical term in the text, which covers all kinds of positive intergenerational transfers. Apart from direct savings, it includes various investments (e.g. in technology and in infrastructure projects), but also measures aimed at maintaining the existing stock of goods and preventing future threats (e.g. climate change mitigation and measures to control inflation).

[4] This problem should not be confused with a more general one, sometimes called 'the problem of the first generation'. This is a challenge not primarily to reciprocity-based views, but to any inequality aversive view of intergenerational ethics, such as egalitarianism, prioritarianism and utilitarianism. It arises out the same features as noted above, plus the assumption of economic growth. Thus, people can only improve the situation of their descendants (not that of their predecessors), and if they do, then their descendants will, needless to say, end up better off than them. According to the just mentioned ethical views, there are moral reasons to redistribute resources to improve the situation of those in the worst off position. Now, in this setting this seems to imply that resources *ought not* to be invested in intergenerational savings at all. That is because to not save seems to prevent the worst off, first generation, from being exploited by having to give to their better off descendants. It thus seems as if not only is it the case that the first generation have no rational reasons to save because of the lack of

reciprocity, but also that they neither have any moral reasons to save, or if they do, it is one to not save. The problem can be generalised in the following way: for any subsequent generation, if its predecessors have not saved for it or (even worse) debited it, then it is effectively the worst off generation and so morally required to not save in order to not further worsen its situation.

If Rawls were to apply his moderately egalitarian 'difference principle' intergenerationally, this would clearly be a problem for him too. However, for just this reason, he makes no such attempt (Rawls 1971, 291). It is potentially a problem even for the sufficientarian just savings principle, that is, on the so-called 'general assembly' interpretation of the original position considered below. But as there are other reasons for discarding that interpretation, there is no need for a separate treatment of the problem of the first generation.

[5] From when it was first presented until just recently, Rawls's account of intergenerational justice has been discussed and criticised, see e.g. (Hubin 1976; English 1977; Barry 1978; Gardiner 2009; Heyd 2009; Gardiner 2011). But, I will argue that something is missing in the discussion. That is, to develop the extension on the basis of the general approach to normative theorising Rawls calls constructivism. This has not been pursued in the debate; even Rawls himself seems to have missed the potential of his constructivism to facilitate a more natural extension of his theory to the intergenerational setting.

[6] Rawls's presentation of the original position changes somewhat over time (for an overview see Freeman 2014). For instance, he answers the question of what is modelled differently: at one point, it is 'the way in which the citizens in a well-ordered society, viewed as moral persons, would ideally select first principles of justice for their society' (Rawls 1980, 520); at another it is 'what we [which just before has been qualified with 'here and now'] think on due reflection are the reasonable considerations to ground the principles of a political conception of justice' (Rawls 2001, 17). I do not think that these are substantial changes though, just shifts of emphasis.

[7] In the introduction, I noted that no one has made use of the constructivist methodology to develop the revisions required of Rawls's theory of justice in the intergenerational setting. I should, however, add that Attas (2009) at least mentions it. He writes: 'This kind of tinkering with the details of the original position is in line with Rawls's general methodological approach of constructivism […] Much of Rawls's discussion on justice between generations in *Theory* and the critical literature that

addressed it can be seen as an exercise in constructivism: a way of modifying the initial situation so that principles generated will take adequate account of future generations' interests too' (Attas 2009, 192). This is not very helpful though. Attas assumes an understanding of constructivism that is much too truncated to provide a useful approach to these questions. If anything, it comes out as question-begging by giving the impression that the 'tinkering' is done to produce an already decided outcome.

[8] For an early, but still very illuminating discussion of constructivism see (Darwall, Gibbard, and Railton 1992, 137-144).

[9] But see Brian Barry (1989, 506) who claims, in a kind of *tu quoque*-argument, that the difficulties are equally featured in both the present-time-of-entry and general-assembly interpretations, and that it is 'a point in favour of' the latter 'that it brings out the difficulties graphically'.

[10] See also (Heyd 2009, 173). For two different attempts at responding to similar problems on behalf of the general-assembly interpretation (see Barry 1989; Gardiner 2009).

[11] It should be noted that the motivational assumption leaves intact another assumption Rawls makes of the parties, namely that they are mutual disinterested, that is, 'conceived as not taking an interest in one another's interests' (Rawls 1971, 13). That the parties are seen as heads of families, caring about the wellbeing of their closest descendants, does not mean that they are assumed to take an interest in any other party's interests. In other words, they are only assumed to care about their own descendants. Some commentators (e.g. Attas 2009, 197-198) have mistakenly pinpointed mutual disinterestedness as the object for the revision; interestingly, also Rawls (e.g. Rawls 2001, 160).

[12] One can also think about people who choose not to have children: are they still required to care about the future? One possible remedy to this is Clayton Hubin's proposal of a 'psychological' rather than motivational assumption (Hubin 1976). That is, to assume that it is common knowledge that people in general care greatly about their offspring. If this psychological fact were known among the parties in the original position, then it would be rational for them to choose a savings principle to prevent the event that they turn out to be a caring parent ones the veil of ignorance is lifted. This solution is an improvement, but still problematic in that it tethers justice too closely to the limits of care (cf. Barry 1978, 227-228).

[13] One possibility is that, even if the motivational assumption does not fully solve the problems considered, it is the best way can hope for given the circumstances at hand. We should perhaps be content with a statement about the value of maintaining background justice, as David Heyd (2009) argues. This might be right if we consider the prospects of a truly intergenerational contract, grounded on intergenerational cooperation; but it is wrong for the relevant sense of the target problem, as I will argue in the subsequent section.

[14] Another possibility is pursued by Daniel Attas (2009). He argues that the solution can be found in the already existing constraint of universality, which, in Rawls's (1971, 132) words rules out a principle that is 'self-contradictory, or self-defeating, for everyone to act upon it', as well as a principle that is 'reasonable to follow only when others conform to a different one'. The problem with this possibility is that a no-savings choice is not self-contradictory or self-defeating: there is no contradiction in all generations, or all individuals within one generation choosing to not save. Nor is such a choice only reasonable on the condition that previous generations have saved: the prisoner's dilemma-structure of the problem makes it rational for the parties to not save irrespective of what previous generations have done, and that is true for each generation. In other words, each generation can adopt the universal principle of no savings without having to assume anything about what other generations have or have not done.

[15] A fourth reason, or set of reasons, concerns what looms in the background of additional constraint, namely Kantian ethics. Barry (1978) argues that Rawls exhibits his Kantian persona here and that it is in tension with his Humean persona, as manifested in the doctrine of the circumstances of justice. If this is right, there are perhaps other reasons to worry about additional constraint. Both because it gives up the reciprocity-ground and because it takes on general problems related to Kantian ethics. Think, for instance, of the common critique of empty formalism: can we derive substantial norms, such as principles of intergenerational justice, just by considering formal features of agency?

[16] This is, of course, premised on the specific conception of the person as free and equal, with the two requisite moral powers, Rawls assumes in justice as fairness. There may be reasons to doubt some of the details of that conception, but it would take as too far astray to that now.

[17] See also Raz (2003). It should be noted that Scheffler does not relate his proposal to Rawls's theory. The idea of sustainability of values as a primary good is, though inspired by Scheffler's book, my own proposal.

[18] Which I believe is too strong. There are plenty of values (e.g. the goodness of wine or the beauty of sunsets) that are intrinsically satisfactory (cf. Raz 2003, 128-131) and would continue to be so even in the afterlife scenario. See also the discussion of the first worry about sustainability of values in the main text.

[19] I thank an anonymous referee for pushing me to clarify this point.

[20] For a more elaborated discussion of a similar point see Raz (2003, 119-121).

References

Attas, Daniel. 2009. “A Transgenerational Difference Principle.” In *Intergenerational Justice*, edited by Axel Gosseries and Lukas H Meyer. Oxford: Oxford University Press.

Barry, Brian. 1978. “Circumstances of Justice and Future Generations.” In *Obligations to Future Generations*, edited by R I Sikora and Brian Barry. Philadelphia.

Barry, Brian. 1989. “Justice as Reciprocity.” In *Democracy, Power, and Justice*. Oxford: Clarendon Press.

Brandstedt, Eric. 2015. “The Circumstances of Intergenerational Justice.” *Moral Philosophy and Politics*, no. 2: 33–55. doi:10.1515/mopp-2014-0018.

Darwall, Stephen, Allan Gibbard, and Peter Railton. 1992. “Toward Fin De Siecle Ethics: Some Trends.” *The Philosophical Review* 101 (1): 115–89. doi:10.2307/2185045.

de-Shalit, Avner. 1995. *Why Posterity Matters*. London: Routledge.

English, Jane. 1977. “Justice Between Generations.” *Philosophical Studies* 31 (2). Kluwer Academic Publishers: 91–104. doi:10.1007/BF01857179.

Freeman, Samuel. 2014. “Original Position.” Edited by Edward N. Zalta. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2014/entries/original-position/.

Gardiner, Stephen. 2011. “Rawls and Climate Change: Does Rawlsian Political Philosophy Pass the Global Test?.” *Critical Review of International Social and Political Philosophy* 14 (2): 125–51. doi:10.1080/13698230.2011.529705.

Gardiner, Stephen M. 2009. “A Contract on Future Generations?.” In *Intergenerational Justice*, 77–118. Oxford: Oxford University Press.

Gosseries, Axel. 2009. “Three Models of Intergenerational Reciprocity.” In *Intergenerational Justice*, edited by Axel Gosseries and Lukas Meyer, 119–46. Oxford: Oxford University Press.

Heath, Joseph. 2013. “The Structure of Intergenerational Cooperation.” *Philosophy & Public*

Affairs 41: 31–66.

Heyd, David. 2009. “A Value or an Obligation? Rawls on Justice to Future Generations.” In *Intergenerational Justice*, edited by Axel Gosseries and Lukas H Meyer, 167–88. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199282951.003.0007.

Hubin, D Clayton. 1976. “Justice and Future Generations.” *Philosophy & Public Affairs* 6 (1). Wiley: 70–83. doi:10.2307/2265063?ref=no-x-route:4d52d424cb95ae150d64058bda739e23.

Mazor, Joe. 2010. “Liberal Justice, Future People, and Natural Resource Conservation.” *Philosophy & Public Affairs* 38: 380–408.

Meyer, Lukas. 2015. “Intergenerational Justice.” Edited by Edward N. Zalta. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/justice-intergenerational/.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA.: Harvard University Press.

Rawls, John. 1980. “Kantian Constructivism in Moral Theory.” *The Journal of Philosophy* 77 (9). JSTOR: 515–72.

Rawls, John. 1982. “Social Unity and Primary Goods.” In *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams, 159–86. Cambridge. doi:10.1017/CBO9780511611964.

Rawls, John. 2001. *Justice as Fairness*. Edited by Erin Kelly. Cambridge, MA.: Harvard University Press.

Rawls, John. 2005. *Political Liberalism*. Expanded edition. New York: Columbia University Press.

Rawls, John. 2009. “Justice as Reciprocity.” In *John Rawls. Collected Papers*, edited by Samuel Freeman, 190–224. Harvard University Press.

Raz, Joseph. 2003. *The Practice of Value*. Edited by R Jay Wallace. Oxford: Clarendon Press.

Samuelson, Paul A. 1958. “An Exact Consumption-Loan Model of Interest with or Without the Social Contrivance of Money.” *The Journal of Political Economy* 66: 467–82.

http://www.jstor.org/stable/1826989.

Scheffler, Samuel. 2013. *Death and the Afterlife*. Edited by Niko Kolodny. Oxford: Oxford University Press.

Wolfsdorf, David. 2013. *Pleasure in Ancient Greek Philosophy*. Cambridge: Cambridge University Press.