Erasmus Journal for Philosophy and Economics,
Volume 14, Issue 2,
Winter 2021, pp. aa–aa.
https://doi.org/10.23941/ejpe.v14i2.495

Social Contract, Extended Goodness, and Moral Disagreement

CYRIL HÉDOIN
University of Reims Champagne-Ardenne

Abstract: This article discusses the role played by interpersonal comparisons (of utility or goodness) in matters of justice and equity. The role of such interpersonal comparisons has initially been made explicit in the context of social choice theory through the concept of extended preferences. Social choice theorists have generally claimed that extended preferences should be taken as being uniform across a population. Three related claims are made within this perspective. First, though it is sometimes opposed to social choice theory, the social contract approach may also consider the possibility of interpersonal comparisons. This is due to the fact that justice principles may be partially justified on a teleological basis. Second, searching for the uniformity of interpersonal comparisons is both hopeless and useless. In particular, moral disagreement does not originate in the absence of such uniformity. Third, interpersonal comparisons should be accounted for both in social choice and social contract theories in terms of sympathetic identification based on reciprocal respect and tolerance, where each person's conception of the good partially takes care of others' good. From the moral point of view, any person's conception of the good should thus be 'extended' to others' personal conceptions. This extension is, however, limited due to the inherent limitations in sympathetic identification and is a long way from guaranteeing the uniformity assumed by social choice theorists.

Keywords: social contract theory, social choice theory, extended preferences, interpersonal comparisons, teleological justification

JEL Classification: D63

I. INTRODUCTION

The social choice approach and social contract theory are two broad traditions that aim at reflecting on possible ways for overcoming the

AUTHOR'S NOTE: This article has been presented at the "Social Justice in a Complex World" international workshop, held in Reims in November 2019. It has benefited from the comments of the participants whom I thank. I also thank two anonymous referees who have provided detailed and sharp comments on a previous version. All remaining mistakes are my own.

plurality of judgments and viewpoints to establish a social or collective agreement over rules and choices. These traditions have been traditionally opposed with respect to moral issues, especially those concerned with justice and equity (Gaus 2011; Sen 2017). This paper addresses an aspect on which they may be thought to be in opposition, i.e., the respective role played and the form taken by interpersonal comparisons (of utility, of goodness) in these two traditions.

Interpersonal comparisons of utility have a long and controversial history in normative economics. Their rejection in the 1930s by influential economists on the ground that they rely on unscientific value judgments is directly responsible for Arrow's (1963) impossibility result in social choice theory. They have been subsequently rehabilitated by Sen (1970) and others. One motivation for this rehabilitation was related to the consideration of issues related to equity and justice: once social choice theory is used to account for moral principles and doctrines such as utilitarianism or Rawls' difference principle, escaping interpersonal comparisons is no longer possible; there must be some ways through which utilities (thought to correspond to individual welfare or any other morally relevant metric) can be compared. The concept of *extended preferences* is the main device by which interpersonal comparisons have been made meaningful within a social choice framework. They correspond to a binary preference relation between pairs of variables (x, i) where x refers to a social alternative or position and i to a personal identity. Then, an extended preference indicates whether one prefers to be individual i in social alternative x or individual j in social alternative y. Interestingly, there are several instances in the literature of 'hybrid' accounts mixing an explicit social choice framework with a broad contractualist commitment over moral matters. Because they use social choice theory as their formal roots, these accounts have tended to rely on the concept of extended preferences. This creates a significant constraint on deriving moral conclusions, i.e., that individuals across a society share the same set of extended preferences. However, arguments justifying such a uniformity assumption are left wanting.

From this perspective, this paper makes three related claims about the comparative status of interpersonal comparisons in social choice and social contract theories. First, I argue that while the concept of extended preferences should be dispensed with altogether, social contract theorists might also consider the need for an appropriate account of the way members of a society settle over a common conception of goodness. This is

due to the fact that some contractualist accounts may partially rely on a teleological form of justification for a moral code. Second, even though establishing such a common conception entails making interpersonal comparisons of goodness possible, searching for the uniformity of interpersonal comparisons is both hopeless and useless. In particular, moral disagreement does not originate primarily in the absence of such uniformity. Third, interpersonal comparisons should be accounted for both in social choice and social contract theories in terms of sympathetic identification based on reciprocal respect and tolerance, where each person's conception of the good partially takes care of others' good. From the moral point of view, any person's conception of the good should thus be 'extended' to others' personal conceptions. This extension is, however, limited due to the inherent limitations in sympathetic identification and is a long way from guaranteeing the uniformity assumed by social choice theorists.

As a result, this article develops a rationale for what can be called a (partially) *teleological contractualism*. This rationale helps to show that the opposition between social choice and social contract approaches to justice and equity issues is less strong than it is generally thought. I proceed through a comparison with other contractual*isms*. Interestingly, although John Rawls (1971) has developed a thoughtful criticism of teleological accounts of justice, it appears that Rawlsian contractualism also has a teleological dimension that materializes in Rawls' thin theory of the good and concept of primary goods. Gerald Gaus' version of contractualism initially also relied on a teleological form of justification (Gaus 1990), but has more recently given up any reference to the good (Gaus 2012). I shall suggest, however, that teleological contractualism is better equipped to deal with the problem that is at the core of both Rawlsian and Gausian contractualism, namely the problem of moral disagreement.

The rest of the paper is organized as follows. Section II briefly presents the concept of extended preferences as developed within social choice theory and the problem of their non-uniformity across a population. Section III presents an argument to the effect that social contract theory, even if taken to constitute an alternative to the social choice approach, may take advantage of considering the role of interpersonal comparisons. This '(partially) teleological contractualism' goes further than Rawls' use of a thin theory of the good. Section IV develops an account of 'extended goodness' and argues that interpersonal comparisons should be accounted for in terms of sympathetic identification. Section V

concludes, reflecting on the fact that extended goodness judgments are unlikely to be uniform and complete. While this a source of moral disagreement, it is, however, not the only one.

II. EXTENDED PREFERENCES IN SOCIAL CHOICE THEORY

The status of interpersonal comparisons of utility in welfare economics has been controversial at least since Lionel Robbins' (1938) claim that such comparisons necessarily involve unscientific value judgments. The ensuing rejection of interpersonal comparisons has considerably restricted the range of welfare criteria available to assess states of affairs. Arrow's (1963) concept of 'social welfare function' defined as a function from a vector of individual ordinal rankings to a social preference ordering effectively implies the impossibility of making interpersonal comparisons. The exclusion of interpersonal comparisons within the Arrowian framework directly leads to the infamous impossibility result that marked the birth of social choice theory: there is no social welfare function satisfying both a Paretian and a non-dictatorship condition that is defined on any vector of individual rankings and that orders two social states only as a function of the individual orderings of these two states.

Social choice theorists have started to reconsider the role and the legitimacy of interpersonal comparisons for two related reasons. On the one hand, allowing for ordinal or even cardinal comparability has proved sufficient at the formal level to avoid Arrow's impossibility result. This research program, opened by Sen (1970), has established that using a broader framework than Arrow's social welfare function allows for a richer informational basis of social choice. On the other hand, as explicitly stated by Sen (1970) and Arrow (1978), while a general theory about collective decision, social choice theory partially overlaps with theories of justice. In particular, social choice may have to select alternative distributions (of welfare, of satisfaction), and hence, "serves the same function as the principle of distributive justice and might be identified with it" (Arrow 1978, 223). It is unclear, however, how social choice could produce an evaluation in terms of distributive justice while prohibiting any kind of interpersonal comparison.

From a purely technical perspective, allowing for ordinal or cardinal interpersonal comparability in a social choice framework is unproblematic. This is easily achieved within the 'social welfare functional' and 'welfarist' approaches of social evaluation that has been dominant since the

1970s.[1] Indeed, within this framework, measurability and comparability assumptions are fully accounted for by the uniqueness properties of the utility functions serving as inputs for the social evaluation (Weymark 2016).[2] However, this approach is silent with respect to the source of the information allowing for the different kinds of interpersonal comparisons: the possibility of interpersonal comparisons is stipulated rather than demonstrated. The concept of extended preferences is constitutive of a whole methodological and theoretical account for making interpersonal comparisons. If successful, it would provide social choice theorists with a way to justify the comparability assumptions made within a social choice framework. Unsurprisingly, this account has been advocated by welfare economists and social choice theorists such as Suppes (1966), Sen (1970), Harsanyi (1977), and Arrow (1978) interested in issues related to social justice. It is still regarded as the main way to give meaning to interpersonal comparisons and to make them eventually operational.[3]

As the name indicates, extended preferences are based on the preference concept that is at the core of modern economics, both positive and normative. I will ignore here the debates surrounding both the definition and the measure of preferences. What matters here is that generally economists regard preference satisfaction either as a proxy for or as constitutive of an agent's welfare. Formally, preferences correspond to a set of binary relations R_i where xR_iy means that individual i weakly prefers social alternative x to social alternative y.[4] These binary relations are generally assumed to be reflexive, complete, and transitive, and thus to define

[1] 'Welfarist' is an ambiguous term. It is understood here in the sense of 'formal welfarism' as characterized by Fleurbaey (2003), i.e., a formal approach that makes the social ordering fully dependent on the individual utility functions of the agents constituting the relevant population. This approach is, however, silent regarding the substantive interpretation of the utility functions, e.g., whether they represent happiness or preference satisfaction. Formal welfarism should not be conflated with 'real welfarism', a substantive moral doctrine which is the target of Sen's criticism in several articles (e.g., Sen 1979).

[2] Take the two following examples. The Rawlsian maximin criterion requires ordinal measurability and full comparability. It is obtained if each individual preference ordering is represented by a utility function unique up to any *common* monotonic positive transformation. On the other hand, cardinal measurability and full comparability is required to define a prioritarian social welfare function. It implies that individual utility functions are unique up to any *common* affine positive transformation.

[3] The main alternative is constituted by money-metric approaches which only allow for indirect and essentially ordinal interpersonal comparisons. In a nutshell, they consist in determining individuals' willingness to pay for achieving a state of affairs or in identifying income equivalents and then using these measures as proxies for individuals' preferences or welfare.

[4] The corresponding relations of strict preference P_i and indifference I_i are defined in terms of R_i: xP_iy if and only if xR_iy and not yR_ix; xI_iy if and only if both xR_iy and yR_ix.

(pre-)orderings of social alternatives. In some cases, additional assumptions can be made. A continuity condition allows each ordering R_i to be represented by a set of utility functions u_i, all positive monotonic transformations of each other. Moreover, if the relations R_i are also defined over probabilistic distributions (i.e., lotteries or prospects) of social states and satisfy a sure-thing or independence axiom, then the functions u_i are cardinal, i.e., they represent the same ordering up to all positive affine transformations of each other.[5] Mathematically speaking, extended preferences will be defined by the same set of properties. They correspond to an ordering and may be represented by a set of ordinal or cardinal utility functions. The difference is the domain over which these relations are defined. Denote X the set of social alternatives to be evaluated and ranked. Any social alternative $x \in X$ is an exhaustive description of everything that is relevant from a normative point of view, including possibly wealth distribution, health states, happiness levels, and so on. Denote N the set of individuals figuring in the relevant population—I will assume here that N is fixed, i.e., we ignore population issues in collective choices. Each individual $i \in N$ is endowed with a preference ordering R_i over X. I will assume that each ordering can be represented by a set of utility functions u_i but put aside for the moment the question of whether additional assumptions are relevant, especially about the cardinality of the functions u_i. A classical social choice exercise is to determine restrictions on the set of possible social orderings R^* given any profile $\{u_i\}_{i \in N}$ of utility functions. As an illustration, we may think it relevant to impose a weak Pareto condition such that if $u_i(x) > u_i(y)$ for all $i \in N$, then xR^*y. As I have stated the problem, the social choice is also restricted by the relative 'thinness' of the informational basis. Indeed, because each function u_i in any profile $\{u_i\}_{i \in N}$ is unique up to any positive monotonic transformation, we are free to use instead, for any person i, any function $v_i = f_i(u_i)$ such that $v_i(x) \geq v_i(y)$ if and only if $u_i(x) \geq u_i(y)$. Because there is no need to apply the same transformation to all individuals, utilities are obviously non-comparable. As I indicate above, this restricts the range of possible social welfare functions.

[5] Roughly, the independence condition states that an agent weakly prefers a lottery L over a lottery L', if and only if she prefers the 'compound' lottery M over the 'compound' lottery M', where M and M' are formed by the same probabilistic distribution of L and L'' on the one hand and L' and L'' on the other hand, with L'' any lottery. Independence then guarantees that the preference relation over any pair of lotteries depends only on the 'non-constant' part of these lotteries.

An extended preference relation R_i^E is (minimally) defined over the Cartesian product $X \times N$, i.e., over all pairs of social alternatives and individuals. Hence, the statement $(x,i)R_k^E(y,j)$ may be read as 'individual k weakly prefers to be individual i in social state x than individual j in social state y'. An alternative and—as it will appear—more satisfactory reading is 'individual k judges as good or better to be individual i in social state x than individual j in social state y'. I shall, however, leave this interpretative issue for the next sections. In any case, the point of defining extended preference relations is that they offer a basis to make interpersonal comparisons of utilities. If, above of the fact of defining orderings, the relations R_i^E are continuous, then they can be represented by sets of utility functions u_i^E such that $u_k^E(x,i) \geq u_k^E(y,j)$ if and only if $(x,i)R_k^E(y,j)$. Hence, from individual k's point of view, individuals i's and j's utilities can be compared, ordinally at least. We may go farther and assume that the extended preference relations are not only defined over $X \times N$ but also over $\Delta(X \times N)$, i.e., the set of all probabilistic distributions of pairs of social alternatives and individuals. Call any such probabilistic distribution $L \in \Delta(X \times N)$ an *extended lottery*. If, in addition to the preceding conditions, each R_i^E also satisfies an independence requirement, then they can be represented by utility functions from which *cardinal* interpersonal comparisons can be derived. As an illustration, suppose individual k has to compare two lotteries L and L'. The former corresponds to an equiprobable distribution of (x, i) and (y, j) and the latter to an equiprobable distribution of (w, i) and (z, j). Now, if LR_k^EL', then that implies $u_k^E(x,i) + u_k^E(y,j) \geq u_k^E(w,i) + u_k^E(z,j)$ and therefore $u_k^E(x,i) - u_k^E(w,i) \geq u_k^E(z,j) - u_k^E(y,j)$, i.e., the utility *difference* between (x, i) and (w, i) is higher than or equal to the utility *difference* between (z, j) and (y, j).[6] This quantitative information notably opens the door for a myriad of utilitarian-based social welfare functions.

The formalism of the preceding paragraphs has left two related questions unanswered. The first concerns the meaning of the binary relations R_i^E and more generally how extended preferences should be interpreted. The second is about the extent to which extended preferences can be expected to be uniform across a whole population. The two issues are dependent since the answer to the first one will presumably make a difference with respect to the answer to the second issue.

[6] Note that the fact that the two pairs of 'extended alternatives' involve the same persons is irrelevant. We could substitute any persons i' and j' for i and j in (w, i) and (z, j) without that making any difference.

Regarding the interpretational issue, virtually all social choice theorists have suggested that extended preferences are obtained through a process of *empathetic* identification. It is especially clear in Harsanyi's writings:[7]

> Value judgments in social welfare [...] may still be interpreted as an expression of what sort of society one would prefer if one had an equal chance to be 'put in the place of' of any member of the society. (Harsanyi 1953, 435)

> We have assumed that *i* will attempt to assess these utilities $u_j(x)$ by some process of *imaginative empathy*, i.e. by imagining himself to be *put in the place of individual j* in social situation *x*. This must obviously involve his imagining himself to be placed in individual *j*'s *objective position*, i.e. to be placed in the objective conditions (e.g. income, wealth, consumption level, state of health, social position) that *j* would face in social situation *x*. But it must also involve assessing these objective conditions in terms of *j*'s own *subjective attitudes* and *personal preferences* (as expressed by *j*'s own utility function u_j). (Harsanyi 1977, 51–52; emphasis in original)

Empathetic identification, or 'imaginative empathy', is achieved by putting oneself in others' shoes, i.e., by identifying oneself with all the objective and subjective features constitutive of other individuals' social position, and personal identities. This interpretation almost requires accepting what can be called a 'sovereignty principle' according to which individual *i*'s extended preferences must respect individual *j*'s preferences over any pair of social alternatives, i.e., $R_i^E = R_j$ over the restricted domain $X \times \{j\}$. As noted by Mongin (2001) and other commentators, the interpretation of extended preferences in terms of empathetic identification does not save this account from difficult ambiguities. Two are worth noting. A first difficulty is related to the concept of preferences. Welfare economists have generally identified preferences and welfare on the basis of a—mostly intuitive—consumer sovereignty principle. It has been pointed out many times, however, that preference satisfaction cannot be constitutive of welfare, especially if preferences are understood in terms of actual or hypothetical choices (e.g., Hausman and McPherson 2006; Sen

[7] As it may create some confusion with the terminology I am using in this paper, it is worth noting that several authors, such as Arrow (1978) and Sen (1970), use the term 'extended sympathy' to refer rather to the identification mechanism underlying extended preferences. But in this context, they use the term 'sympathy' in its old meaning, which effectively makes it synonymous with the modern meaning of empathy.

1973). Harsanyi (1996) himself recognized that not all preferences are conducive of personal welfare and that antisocial and self-detrimental preferences should be ignored by the welfare analysis. This casts doubts on the normative strength of the sovereignty principle highlighted above. The second difficulty is even more significant. The empathetic reading of extended preferences seems to indicate that an individual *i* evaluating social alternatives from *j*'s point of view should completely identify with *j*'s preferences. However, if asked to compare two extended alternatives featuring two different individuals *j* and *k*, it is not clear how *i* should proceed. Endorsing successively *j*'s and *k*'s preferences is presumably not sufficient because, as such, it does not make them comparable. Hence, it seems clear that extended preferences cannot merely replicate each individual's preferences. They are the preferences of the individual who is making an assessment between two extended alternatives. Therefore, it is not possible to avoid the following question: Where do extended preferences come from?

The difficulties that surface with respect to the second issue regarding the uniformity of extended preferences across a population are directly related to the impossibility to answer this question in a satisfactory way. As a first step, it is useful to remark that the uniformity of extended preferences (formally, $R_i^E = R_j^E$ for all pairs of individuals i and j in the population) has been sometimes assumed to derive interesting formal results from a social choice perspective.[8] What is at stakes here, however, is not its theoretical usefulness but its normative relevance. The reason why uniformity is generally thought to be required is that, in its absence, individuals' comparative assessments of extended alternatives would differ, leading to several extended utility functions u_i^E. But then, we would be back to square one as we would be devoid of any way to compare these extended utility functions: each individual would make her own interpersonal comparisons, but disagreement over the right way to make them would ensue. If it is assumed that a *collective* choice cannot be the choice of one person, there are only two possibilities: either everyone agrees on a common standard to compare utilities or only social welfare functions not requiring interpersonal comparisons are acceptable. However, though several arguments have been offered in the social literature to ground the uniformity principle (e.g., Arrow 1978; Binmore 1998; Harsanyi 1977),

[8] For instance, Sen (1970, Chapter 9*) makes the uniformity assumption to establish a formal relationship between Suppes' grading principles of justice and aggregative and Rawlsian social welfare functions.

none of them has generally been regarded as convincing for several reasons.[9] The problems with the uniformity principle emphasize that the quest of grounding interpersonal comparisons and extended preferences from an objective, valueless point of view is hopeless. Disagreement over the standards for making interpersonal comparisons seems inescapable, putting the social choice approach to equity and justice issues under pressure.

III. INTERPERSONAL COMPARISONS AND 'TELEOLOGICAL CONTRACTUALISM'

The discussion of the previous section therefore calls for an essentially negative conclusion: there seems to be no convincing argument establishing that extended preferences must be uniform while keeping their normative meaning. This makes the prospects of a pure social choice approach to equity and justice unlikely, because without interpersonal comparisons almost nothing can be said about equity and justice within this framework. This conclusion may leave social contract theorists indifferent: so much the worse for the social choice approach, but as a rival tradition social contract theory is unaffected.

I shall argue in this section that this conclusion is too quick. The idea that social contract theorists may spare themselves the need to deal with interpersonal comparisons is due to the conflation of two related but still distinct philosophical underpinnings of the social contract tradition. On the one hand, the social contract tradition is constituted by a (set of) first-order moral doctrine(s) that are loosely referred to as 'contractualism'. Though there is no widely agreed definition, contractualism may be characterized as the general view that morality is based on contract or agreement (Ashford and Mulgan 2018; Gauthier 1977). On the other hand, the social contract tradition may be viewed as a (set of) normative account(s)

[9] Harsanyi's so-called causal argument is the one that has been given the most attention and has attracted most of the criticisms. The most forceful criticism comes from Broome (1993) who shows that Harsanyi's argument confuses the cause for preferences (the $C(i)$ variables) with the content of preferences. While impartial observers may agree over the causes for different individuals' assessments of social alternatives, this does not logically imply that they must agree over their impartial assessments of extended alternatives. Mongin (2001) also highlights that Harsanyi's argument would at best support the claim that all impartial observers agree over their *predictions* of which extended alternative brings the highest degree of preference satisfaction. Finally, Pattanaik (1968) observes that there is absolutely no reason to expect that impartial observers' attitude toward risk should be identical. As the utility values of extended alternatives are directly derived from preferences over *lotteries*, that implies that two impartial observers may ascribe different utility values to extended alternatives, even if the causal argument is true.

of *justification*, especially of moral justification. Most social contract accounts will then be characterized as subscribing to a 'deontological' account of justification.[10] Each of these labels has its natural nemesis, 'consequentialism' and 'teleology' respectively. I will not have much to say about the contractualism/consequentialism opposition, except that social choice theory is generally classified within the consequentialist category. This is not only or foremost due to the fact that social choice theorists tend to study equity and justice issues in terms of *choice* rather than in terms of *contract* (or another procedure leading to an agreement).[11] It is mostly related to the fact that within social choice theory the evaluation of collective choices is fully dependent on a more or less broad characterization of *outcomes* rather than on the fact of instantiating or following rules and principles. This point is, however, not relevant for the deontological/teleological distinction. I shall indeed argue that even in the realm of contractualism as a first-order moral doctrine, a teleological form of justification may be relevant and that this calls for the possibility of making comparative judgments of goodness. In this regard, Gaus—following Sandel (2010, 3)—characterizes teleology as:

> A form of justification in which first principles are derived in a way that presupposes final human purposes and ends. […] Principles of right, or public morality, are justified through appeal to values of actual people to whom the morality is to apply. (Gaus 1990, 331)

[10] To my knowledge, the explicit distinction between contractualism as a first-order moral doctrine and deontology as a theory of moral justification is due to Sandel (2010). I shall point out that the characterization of the deontology/teleology distinction as competing accounts of justification I use in the text is not the most common in the literature. In particular, it is clearly not how Rawls defined these terms (see Freeman 1994). Consequentialism and teleology indeed tend to be used as quasi-synonymous (though, according to Freeman, 'mixed' forms of consequentialism are more appropriately seen as belonging to deontology (2007, Chapter 3)). In the same way, almost all forms of contractualism are generally regarded as belonging to the domain of deontology. Though the terminology may be unorthodox, I still think that distinguishing between normative accounts of justification (Sandel's and my deontology/teleology distinction) on the one hand, and first-order moral theories (the contractualism/consequentialism distinction) on the other hand, is enlightening because it helps to separate two related but still different issues: first, what is it for an act or a state of affairs to be good or right? Second, what is the relationship between the good and the right in the justificatory endeavor?

[11] On the choice/contract distinction, see Hampton (1980). Though her article focuses on Rawls and the possible interpretations of his theory of justice in terms of the choice paradigm or the contract paradigm, she also notes that Harsanyi's 'contractualism' belongs to the choice paradigm. As noted by Gaus and Thrasher (2015), the fact that Rawls' contractualism depends on a choice from an 'Archimedean point' rather than a proper contract is not peculiar to it, as other forms of contractualism like Gauthier's actually share this same feature.

As we shall see below, teleological justification can take many forms. By contrast, deontology is then a mode of justification of first-order normative principles that does not presuppose ultimate human purposes and ends, i.e., no definite conception of the good.

The deontological underpinnings of the social contract approach are well exemplified by Rawls (1971) in his theory of justice. Rawls repeatedly emphasizes that a theory of justice must recognize the priority of the 'right' over the 'good'. That is, principles of justice determining what is right and just in the society can and must be determined independently of any conception of goodness and thus of any view regarding what a 'good' life is. In this sense, the right not only has priority over the good, the former also constrains the latter: only conceptions of goodness compatible with the chosen principles of justice will be able to prosper within the society. On the other hand, this 'axiological neutrality' characteristic of a deontological theory of justice is also deemed to be compatible and even to favor a 'reasonable' moral pluralism. This feature is especially developed by Rawls in his later writings which emphasize that his theory is foremost 'political' rather than 'moral' (Rawls 1993, 2001). In a 'well-ordered society', an overlapping consensus will prevail through which citizens affirm a unique political conception of justice, while entertaining conflicting religious, philosophical, and moral views. In particular, the agreed upon principles of justice are endorsed *from within* competing comprehensive moral doctrines:

> We say that in a well-ordered society the political conception is affirmed by what we refer to as a reasonable overlapping consensus. By this we mean that the political conception is supported by the reasonable though opposing religious, philosophical, and moral doctrines that gain a significant body of adherents and endure over time from one generation to the next. This is, I believe, the most reasonable basis of political and social unity available to citizens of a democratic society. (Rawls 2001, 32)

There is absolutely no doubt that the priority of the right over the good is an enduring feature of Rawls' account of justice, thus establishing its deontological character. That said, Rawls' later writings also indicate a growing concern for establishing that his political theory of justice is compatible with the *fact* of moral pluralism that is constitutive of modern societies. It is true that that this concern is essentially due to empirical, rather than normative reasons. Moral pluralism and hence moral disagreement are facts with which we have to live and any practically relevant

theory of justice must not only recognize it but also be compatible with it. With respect to the logic of justification, the right still comes first: we must determine and justify political principles of justice without appealing to any feature of a comprehensive doctrine, being moral, religious or anything else. But from an empirical point of view, principles of justice cannot be but supported from within comprehensive doctrines, hence the requirement for a well-ordered society of establishing an overlapping consensus.

The growing recognition by Rawls of the practical importance of reasonable pluralism and the related turn of his justice as fairness account from a 'metaphysical' or 'comprehensive' to a 'political' understanding also involves a change in the interpretation of his thin theory of the good. Contra Sandel (2010) and communitarian critics of Rawls, Rawls' contractualism in *A Theory of Justice* is not fully deontological as it builds on a conception of goodness as rationality that is attributed to the parties in the original position. As Rawls (1971, 396) made it clear, this notion of goodness *precedes* the establishment of the principles of justice by individuals put behind a veil of ignorance. Its role is to ground the assumptions made about the primary goods that all individuals are assumed to pursue to realize their rational plans and *full* conceptions of the good. As a consequence, though 'thin', this account of the good introduces a teleological feature in the otherwise deontological and constructivist procedure of justification developed by Rawls. The thin theory of the good is pivotal in Rawls' account at least at two levels. First, it allows for the creation of an index of primary goods which itself provides a basis for interpersonal comparisons (Rawls 1971, 92). The latter are indeed required to make the first part of Rawls' second principle of justice (the difference principle) operational and meaningful. Second, it plays an essential role in Rawls' complicated account of stability developed in Part III of *Theory of Justice*. Rawls' congruence argument between the right and the good indeed builds on the claim that maintaining a sense of justice is a good in the sense of the thin theory. Rawls however, later rejected the congruence argument as being incompatible with a political theory of justice and accounted for stability in terms of an overlapping consensus. The interpretation of the thin theory of the good has subsequently changed. Goodness as rationality is no longer a plausible account of a person's objective good. It rather refers to a pluralist conception of value compatible with political principles of justice as agreed between persons that mutually regard themselves as free and equal citizens (Freeman 2007, 97–98).

This brief survey of the evolution of Rawls' contractualism indicates that teleological considerations were present from the start to justify and operationalize principles of justice in the context of reasonable pluralism. It also serves as an intermediary step to establish the *normative* relevance of a teleological form of justification for contractualism. Rawls emphasized the fact of moral pluralism essentially for empirical and political reasons. I now want to argue that contractualists may want to go further and deal with interpersonal comparisons in the context of teleological justification for more foundational reasons. Rawls' contractualism has at least three components that account for the secondary role played by teleological justification: first, a particular kind of constructivism that materializes through the original position; second, the specific account of justice consisting in its two principles; third, its 'formal' conception of the person. Modifying one or several of these components may make room for a more important role for teleological justification in contractualism.[12] In the following, I will focus on the last component, but first I comment on the former two.

The derivation of the two principles of justice through the device of the original position is characterized in terms of a 'procedure of construction' establishing a link between a particular (political) conception of the concept and principles of justice (Rawls 1980, 304). This construction builds on the conception of moral persons as free and equal citizens who, because of this very conception, are committed to search for principles of justice while ignoring their own conception of the good—except for the thin theory of the good which justifies the reasoning in terms of primary good. Now, there is room for disagreement regarding the content of the principles of justice, something which Rawls increasingly emphasized in the last part of his career. Constructivism is also itself not the only way to justify principles of justice in a contractualist framework. Even within constructivism, we may imagine a different procedure of construction where persons are aware of their conceptions of the good. Ultimately, I submit that the issue of justification is tightly related to the *theory of the person* that one sees as appropriate. To defend a full theory of this kind is of course a daunting task that I cannot undertake here. Let me, however, sketch an argument indicating that even within a contractualist

[12] These three components are of course tightly related in Rawls' writings and so, changing one may affect the other two. Rawls' article on Kantian constructivism (Rawls 1980) provides the clearest statement of the relationship between his conception of the person and the constructivist nature of his contractualism.

account of morality there is scope for a theory of the person, one which demands a more significant role for teleological justification.

Though somewhat rough, we may distinguish between two broad accounts of the person in a contractualist perspective. A first account characterizes personhood independently of the ends, goods, and values that particular persons may contingently endorse or pursue. This 'formal view' attributes to persons agency powers and capacities that make them rational and reasonable beings, but takes no stance with respect to how these powers and capacities are actually used. This is not to mean that individuals do not endorse values or do not pursue ends, but rather that what makes these individuals foremost moral persons endowed with a particular normative status is not their values or ends, but their *ability* to pursue ends and endorse values. This formal view finds notably its expression in Rawls' *Theory of Justice*. The normative relevance of the original position is fully dependent on the possibility of decoupling persons from their contingent conceptions of the good life. This postulate remains at the core of Rawls' later writings. For instance, Rawls repeatedly emphasizes that his conception of the person as citizens is a 'political' one that gives persons two "moral powers":

i. One such power is the capacity for a sense of justice: it is the capacity to understand, to apply, and to act from (and not merely in accordance with) principles of political justice that specify the fair terms of cooperation.
ii. The other moral power is a capacity for a conception of the good: it is the capacity to have, to revise, and rationally to pursue a conception of the good. (Rawls 2001, 18–19)

As it is well known, Rawls also emphasizes the 'separateness of persons' as a normatively relevant feature. This separateness is precisely grounded in the fact that each person has their *own* capacity for a sense of justice and for a conception of the good. But this separated identity is not grounded on the *content* of this sense of justice and this conception of the good.

Sandel (2010), among others, has emphasized that such a formal view of the person is almost needed by deontologists. This is by decoupling persons from their ends and values that the claim of the priority of the right obtains its normative force. As I have argued, that does not imply, at least in the case Rawls' contractualism, a complete lack of teleological elements. However, there is at least an alternative plausible view of the person that puts the priority claim in jeopardy. What can be called the

'substantive view' conceives the person, both in her rational agency and her identity, as being fundamentally constituted by the values she endorses and the ends she pursues. Another way to characterize this view is that what makes persons morally separated and relevant is that values and ends are *theirs:* they are able to act and to justify their actions on the basis of values that they recognize as their own. Obviously, the substantive view also regards Rawls' moral powers as normatively important. But it goes further: we *also* give importance to persons as bearers of particular values constitutive of a plurality of conceptions of goodness. On the substantive view, the rationality of persons cannot be characterized independently of the values that justify their intentional attitudes. Moreover, the identity of persons is tightly related to their (possibly) evolving conception of goodness. This implies that the *content* of conceptions of the good cannot be arbitrary from a normative perspective: what makes a person a rational or reasonable being is their ability to act on the basis of *some* values; what makes a person a continuous being is their ability to endorse a *particular* conception of goodness. That implies that some persons may lose their particular normative status if their conceptions of goodness are constituted by what is judged to be unacceptable values or if it is impossible to ascribe to them some continuous and consistent conception of the good.[13]

The formal and the substantive views have quite different implications regarding the relationship between what can be called the 'personal point of view' and the 'moral/political point of view'. On both accounts, a person's personal point of view is constituted by their conception of the good. But as far as the moral/political point of view is concerned, the formal view insists that the choice of principles of justice should be detached from particular conceptions of the good going beyond the thin theory, resulting in the priority of the right over the good. The substantive view not only indicates that the moral/political should not be expected to be completely independent from conceptions of the good, but that it ought not to be. The point is not only that, as a matter of practical rationality, it is impossible to choose a set of political principles in *complete ignorance* of one's own values; it is also that such a choice would be normatively irrelevant. Now, what sets the moral/political point of view apart is that when reflecting on the appropriate principles that should regulate

[13] That would not mean that these persons have no normative importance, but rather that this importance would be grounded on a different normative reason (e.g., as sensible beings capable of enduring pain).

a society, each person must acknowledge that she has to find an agreement with other persons bearing their own conceptions of the good. That implies at least two things: first, a substantive assessment of competing conceptions of the good may be needed; second, establishing a minimal common conception of the good or a compromise between competing conceptions may be required by invoking values (e.g., equality, impartiality) that may not transpire in personal conceptions.

Hence, *on the substantive view of personhood*, agreement over a social contract will require one form or another of *teleological justification* building on a compromise between competing conceptions of the good and/or on a common conception of the good. As pointed out by Gaus (1990), this teleology will still be constrained by deontological principles and values that may find their justification in the formal view of personhood.[14] But the point is that, under this conception of personhood, the social contract cannot avoid any form of teleological justification. Once this point is granted, another question arises: How could this justification be achieved? This question marks a point of departure in the social contract tradition between *contractarian* and *contractualist* (in a narrow sense) approaches. The former argues that the social contract is ultimately the product of a compromise between competing conceptions of goodness. Significant instances of the contractarian approach such as Gauthier's (1987) make use of axiomatic bargaining theory to ground the compromise on bargaining principles. Interestingly, they tend to use bargaining solutions that eschew the need to make interpersonal comparisons of goodness. The contractarian approach is, however, notoriously controversial, and I will not discuss it there. Apart from the contractarian approach to teleological justification in terms of compromise, another possibility is a contractualist approach working through the identification of a common conception of goodness. The next section establishes that such an identification must rely on interpersonal comparisons and develops a particular account in terms of 'extended goodness'.

IV. SYMPATHETIC IDENTIFICATION AND EXTENDED CONCEPTIONS OF THE GOOD

Reflecting on the role of teleological justification in a contractualist framework, Gaus (1990) appeals to Gauthier's (1987) contractarianism

[14] As I have pointed out, the substantive view encompasses the formal view, i.e., regarding a person as being constituted by her conception of goodness obviously implies the consideration that a person has the *ability* to form a conception of goodness.

which, with respect to justification, has two key features: on the one hand, values appealed to are *agent-relative values*, on the other hand the justification takes the form of a *compromise* between competing conceptions of goodness. As I indicate in the preceding section, this compromise is shaped by a bargain that—at least in Gauthier's case—can be characterized through axiomatic bargaining game theory. Quite a different form of teleological justification is provided by some variants of utilitarianism, including Harsanyi's. These variants appeal to *agent-neutral* values from which a *community of valuing* is derived. The teleological justification of this brand of utilitarianism can be seen as an instance of Nagel's (1989) 'view from nowhere' account, according to which the objectivity and neutrality of the moral point of view is achieved by adopting a perspective free from any individual contingencies and idiosyncrasies. Harsanyi's impartial observer theorem is indeed an instance of such an account: by ignoring who and where they will end up in the society, individuals behind the veil of ignorance are forced to adopt a set of values that, though they reflect the values (i.e., preferences) of everyone, are the values of nobody in particular. These values are captured by the observers' extended preferences. As we have seen, Harsanyi claims—wrongly—that these extended preferences must be uniform across the population, thus achieving a community of valuing.

I shall make the case for another form of teleological justification. While the values constitutive of the persons' conceptions of goodness are the ones appealed to in the justificatory endeavor, I am agnostic with respect to their agent-relativity or neutrality.[15] Presumably, individuals' conceptions of the good will essentially be constituted of (prudential and non-prudential) values providing agent-relative reasons for action (e.g., personal welfare, dignity, autonomy). Yet, we cannot exclude the possibility that people value things leading to agent-neutral reasons for action (e.g., overall welfare). Whether or not such values should be considered in the justificatory endeavor is an issue that I leave aside. On the other hand, neither Gauthier's compromise nor Harsanyi's community of valuing are satisfactory views of teleological justification. The former because it makes the resulting compromise depend on bargaining factors whose normative relevance are doubtful from the moral point of view; the latter because it depends on the mistaken assertion that extended preferences must be uniform. Instead, I suggest that teleological justification should

[15] For a discussion of the distinction between agent-neutral and agent-relative values and reasons, see Parfit (1984).

proceed on the basis of the identification of a *minimal* common conception of the good. In other words, a form of community of valuing is needed but it cannot be expected to be a complete overlapping of personal conceptions of goodness. Crucially, this minimal common conception of the good requires the use of interpersonal comparisons. The rest of this section is dedicated to developing an account of interpersonal comparisons of goodness in this context.

The formal framework of extended preferences presented in section II will be useful here. However, while the formalism remains, its interpretation must be quite radically changed. First, in the framework of extended preferences, the binary relations R_i and R_i^E and the related utility functions u_i and u_i^E were thought to correspond to and represent, respectively, a *preference* relation. This implies a commitment to preferentialism which may be regarded as problematic given the already mentioned difficulties surrounding the preference concept in a normative context. We may, however, ignore such commitment by taking the binary relations and the functions to capture personal conceptions of goodness. To avoid any confusion, I will denote B_i the binary relation 'as better or as good as' and g_i the related 'goodness function' of individual *i*. This entails a set of assumptions about the formal properties of personal conceptions of goodness. In essence, I am assuming that individuals can rank social alternatives on the basis of their conceptions of goodness. Call it the 'ordering property of goodness'. This first change entails a second one: we no longer deal with extended preferences but rather with 'extended goodness', or more precisely *extended personal conceptions of the good*. They are captured by the binary relations B_i^E and the 'extended goodness functions' g_i^E. As in the case of extended preferences, the relations B_i^E are defined over the Cartesian product $X \times N$. It remains to establish the precise meaning of these relations. A third and final change is the nature of the process through which extended conceptions of goodness are arrived at by individuals. As I note in section II, social choice theorists have interpreted extended preferences in terms of empathetic identification. I shall suggest instead that extended personal conceptions of the good are formed through a process of *sympathetic identification*. I will detail and argue for these three major changes, starting with the last one.

As I explain in section II, empathetic identification consists in taking another person's point of view, i.e., to adopt all her affective and evaluative states and attributes in some given situation. This approach has been argued to be problematic, in particular with respect to the very feasibility

of empathetic identification in some cases. Adler (2014) makes this objection, which he labels the 'essential-attribute problem'. This problem can be briefly put in the following way. Suppose that any social alternative $x \in X$ includes information about individuals' attributes that, under some theory of personal identity, may be referred to as 'essential', i.e., as being constitutive of a person's identity. For instance, plausible theories of personal identity may hold that gender or memories are such essential attributes.[16] Now, the essential-attribute problem is simply the point that when social alternatives include such information, it might be impossible—in quasi-phenomenal terms—for an observer i to assess an extended alternative (x, j) because 'being j' depends on attributes that i cannot even imagine owning. In other words, empathetic identification confronts the problem that two persons i and j may not be able to fully *share* their respective points of view. Adler makes this objection in a preferentialist framework, but it is not hard to see that it applies at least equally strongly here. In the preceding section, I remarked that under a substantive view of personhood, one's conception of the good is constitutive of one's personal identity. That does not mean that conceptions of the good are not shareable in total, but that imaginative projections of the kind 'I assess social alternative x taking j's values as mine' may in some cases be simply impossible.

Instead of empathetic projection, Adler (2014) suggests a 'sympathy-based conception of extended preferences'. Following Darwall (2002), Adler defines sympathy as an attitude of care and concern for others. In particular, to fully sympathize with someone is to be "motivated to pursue what you believe lies in the interests of the sympathy target" (Adler 2014, 146). This definition implies an important connection between sympathy and *welfare*. Take the specific case where person i is asked to take a moral point of view (i.e., to endorse the position of an 'impartial observer') and to compare two extended social alternatives (x, i) and (y, i). i is thus asked to 'self-sympathize' with herself. Under Adler's and Darwall's account, that means that i must assess each extended alternative uniquely in terms of her self-interest, i.e., her welfare. Similarly, when comparing two extended alternatives concerning the *same* person j who is not her, the observer who fully sympathizes should base her evaluation entirely on j's welfare. Finally, when comparing two extended alternatives

[16] Alternatively, we may assume that these essential attributes are attached to the components of the individuals set N. What should be avoided is to postulate that essential attributes can be found both in X and N, as that would make some extended alternatives (x, i) implausible if not metaphysically impossible.

concerning two different persons *j* and *k*, the sympathetic observer should form a welfare judgment and determine whose welfare is higher. This approach has two advantages but also two implications that may be regarded as undesirable. The first advantage is that the sympathy-based account of extended preferences eschews a general problem with the 'view from nowhere' approach to morality and more generally to normativity. It is clear here that the observer's welfare judgment cannot be based but on her *own conception of welfare as part of the good*. Because presumably most if not all individuals value welfare, sympathetic identification is at least well defined. This leads to the second advantage: this account is not confronted to the essential-attribute problem. The sympathetic observer is not asked to put herself in others' shoes but to make comparative welfare assessments *from her own perspective*. Shareability is thus not an issue.

Two other implications should also be considered as they may be judged problematic. On the one hand, there is obviously no guarantee that all observers will agree on their comparative welfare judgments. This is due to several factors: they may not have all relevant information at their disposal and some may be better informed than others, they may not share the same conception of welfare, or they may disagree about the welfare-effects of certain essential attributes related to personal identity. We can assume that perfectly informed observers may eliminate the first factor; but the second and third ones seem unavoidable. However, as neither we (nor Adler for that matter) are looking for uniformity of extended preferences, this problem is not relevant here. The second potentially problematic implication is that the sympathetic approach gives too much priority to welfare over any other value (prudential or not prudential) when persons take the moral point of view. There is no doubt that welfare is an important value and that any normative endeavor has to acknowledge its importance. But from the perspective of moral and political philosophical theories, it may be argued that teleological justification cannot be based *fully* on it.[17] Take the following plausible case: Emma is a young adult who takes pride in being an ecological activist, even if participating in street protests may occasionally result in her being arrested or hurt. John, a forty-something bank employee and father may judge, in wholehearted sympathy with Emma, that Emma would be better—in

[17] It should be acknowledged that Adler (2014) develops his account mostly as a contribution to welfare economics, not to moral and political philosophy. The objection discussed in the text cannot thus be directly addressed to him.

terms of welfare—in a social alternative x where she renounces her ecological activism than in a social alternative y where she ends up being severely hurt. Suppose that, from Emma's conception of goodness, the converse judgment applies. Formally, we thus have $yB_{Emma}x$ but $(x, Emma)B^{E}_{John}(y, Emma)$. It is clearly debatable that John's judgment should have priority over Emma's. The point is that, except for the mostly unlikely case where Emma does not value welfare at all, her goodness judgment already encapsulates a tradeoff between welfare and other values, including non-prudential ones. With regard to providing Emma (among others) a teleological justification to a set of normative assessments about what is good, just, permissible, or obligatory at the collective level, it is not clear that John or other observers should be authorized to simply disregard Emma's tradeoff. There are only two possibilities here: either it is established that Emma is *wrong* in her goodness judgment, either because she is not correctly informed or she has not properly reasoned about the issue, or sympathetic identification should not simply consist in taking care of others' welfare but also of other's good as a whole. The former option leads to paternalistic considerations which should not be straightforwardly rejected on 'naïve' libertarian grounds. But whatever one may think of paternalism, the latter option should also be given due consideration.

The revision of the concept of sympathetic identification in terms of taking care of others' good entails what I call a concept of 'extended goodness', or more precisely, of *extended personal conceptions of the good*. Consider the meaning of the statement $(x, j)B^{E}_{i}(y, j)$. Literally, it reads as 'i judges x to be better than or as good as y for j'. I have just argued that i's goodness judgment cannot be made only on the basis of a concern for j's welfare. Other values should presumably also be taken into consideration. But the statement remains ambiguous: Is i judging that x is better than or as good as y for j on the basis of i's conception of the good? Or is i fully endorsing j's conception of the good in making his goodness judgment? I would argue that neither is acceptable. The former interpretation may be compatible with sympathetic identification but only on the formal view of personhood where personal identity and personal conceptions of the good are decoupled. But under the substantive view, the revised concept of sympathetic identification renders it implausible. The latter interpretation is more plausible but would imply that i invariably defers to j's conception of goodness. Even if it is accepted, this interpretation is not available in comparative judgments involving two different persons, i.e.,

statements of the form $(x, j)B_i^E(y, k)$. I thus submit a third interpretation: in forming her extended goodness judgment through sympathetic identification, the observer should take others' conceptions of the good *as far as they include values that the observer herself endorses in her own conception*. The observer and other individuals may still differ with respect to the *weight* they give to these values, but they agree that these values matter. Thus, they could and should be appealed to in the justificatory endeavor. In the case of statements of the kind $(x, j)B_i^E(y, j)$, *i*'s extended goodness judgment should thus be based on those values that have a positive weight in *i*'s and *j*'s respective conceptions of goodness. Obviously, *j*'s personal goodness judgment and *i*'s extended goodness judgment may differ, in particular if their respective conceptions of goodness do not fully overlap. The same idea holds, perhaps even more convincingly, in the case of a statement of the form $(x, j)B_i^E(y, k)$. Unless j and k share the same personal conceptions of the good, i as an observer may have to ignore some values. The observer is more likely to justify her judgment to j and k by grounding it on values that j and k actually share. At the same time, because i proceeds through sympathetic rather than empathetic identification, she may consider taking into account only those values that she is herself actually endorsing.

Consider a population of n persons, each endowed with a personal conception of the good. These n individuals also are n potential observers who may form extensive goodness judgments. Suppose that we may identify a set V of (prudential and non-prudential) values v that are weighted positively by at least one person. Denote V^* the (maximal) subset of V such that any $v \in V^*$ is common to *all* personal conceptions of the good. V^* forms the basis on which extended goodness judgments can take shape, though values which are only partially common (i.e., they are shared only by a subset of the population) may also be used in specific cases. Two issues then arise. First, what is the relationship between binary relations B_i and B_i^E capturing respectively personal and extended goodness judgments? The two binary relations should obviously be identical when the observer i is comparing a pair of extended alternatives (x, i) and (y, i). In the more general case where a deliberator k is comparing a pair of extended alternatives (x, i) and (y, j), the deliberator's personal goodness judgments captured by the binary relation B_k is partially relevant because by assumption it restricts the set of values on the basis of which she forms her extended goodness judgments. But among this restricted set, only a (possibly empty) subset of values will be shared by *i*'s and *j*'s

personal conceptions of the good. The relation B_i^k compares extended alternatives on the basis of this subset of values that are shared by all protagonists. How then will a deliberator rank social alternatives from the moral point of view? The deliberator must *justify* her ranking to everyone on the basis of a sequence of extended comparative goodness judgments captured by B_i^E. But this is obviously not sufficient because the deliberator must also find a way to aggregate these extended goodness judgments. A person's all-things-considered moral ranking is then captured by a binary relation B_i^* defined over X (and *not* the Cartesian product $X \times N$).[18] B_i^* is determined both by B_i (since it restricts the range of values that i uses to form her judgment) and by B_i^E. The reason why the latter should be an input in the moral ranking is twofold. First, it is an essential part of the justificatory endeavor. They permit each person to make interpersonal comparative judgments which may be used as an input in the justificatory endeavor. Indeed, they guarantee that all B_i^* are based on a minimal common conception of the good. Second, I would suggest that, at least in what Rawls (1971) calls a 'well-ordered society' based on relationships of reciprocity and mutual respect, persons have strong normative reasons to respect others' conceptions of goodness and to engage into sympathetic identification. Indeed, it might be argued that reciprocity, respect, and tolerance are core values of a well-ordered society that tend to be shared by all the members of the population. This provides reasons to consider and even to value others' conceptions of the good, at least as far as issues requiring public justification are concerned.

That said, and this is the second issue that may be mentioned, there is no reason to expect that everyone will agree on their moral ranking, i.e., that $B_i^* = B_j^*$ for all i, j. On the one hand, the fact that the binary relations B_i^* build on a common set of values does not imply that all persons will make the same tradeoffs between these values. This is just another version of the claim that persons' extended goodness judgments B_i^E will differ across the population. However, though not identical, the B_i^E relations should be regarded as partially comparable. This is due to the fact that they build on the same set of values and the same process of sympathetic identification. If this set is sufficiently large, we may expect an agreement

[18] The binary relation B_i^* corresponds to what Harsanyi (1977) called a person's *moral preferences*. In Harsanyi's account, moral preferences are formed by aggregating extended preferences through a utilitarian formula that is itself derived on the basis of particular rationality and epistemic assumptions. I do not presume here any specific aggregation rule. Indeed, as I indicate below, the fact that deliberators may use different aggregation rules is one source of moral disagreement.

over a significant number of pairs of social alternatives. This minimal agreement over goodness judgments shares conceptual similarities with Rawls' thin theory of the good. There are differences though. The most significant one is that the content of this agreement is left undetermined and so that the goodness judgments may differ from one population to another. The second difference is that this agreement does not follow from a practical or empirical necessity but is more foundational. In this sense, teleological contractualism is fully 'comprehensive' rather political in Rawls' sense.

On the other hand, even if all deliberators were to agree on their extended goodness judgments, there is no reason to expect that they would necessarily agree on the way to aggregate these judgments to form their moral rankings. The likelihood of an agreement on this issue and the principles that would serve as a basis for it is another important aspect of the problem of moral disagreement that is, however, beyond the scope of this paper. This point underlines, however, the interest of partially building contractualism on teleological foundations. Sources of moral disagreement are manifold and can result both from teleological and deontological considerations. To be able to account for this fact seems to be a valuable asset within the social contract tradition.[19]

V. CONCLUSION: THE MANY WAYS OF DISAGREEING OVER A MORAL CODE

In conclusion, I would like to insist on the point made just above. I have claimed in this paper that the need and the difficulties surrounding interpersonal comparisons are not an artefact of social choice theory applied to issues of justice and equity. It is true that social choice theorists *do need* to make such interpersonal comparisons to show that individuals agree on such comparisons. The framework of extended preferences, though useful to account for the origins and the meaning of interpersonal comparisons, cannot establish what the social choice theorist needs, i.e., uniformity of extended preferences across a population. Now, I have argued that social contract theorists, especially those who wanted to avoid contractarianism, must also be able to account for interpersonal comparisons. This is due to the fact that a form a teleological justification is

[19] While initially emphasizing the importance of teleology, Gaus' (2012) contractualism has now given up any explicit reference to the role of competing conceptions of goodness in moral disagreement. Though I cannot defend this claim here, I think that on this aspect, Gausian contractualism—while otherwise extremely valuable and important—is making a step backward as compared to Rawlsian political liberalism, which emphasizes and builds on the disagreement over the good.

needed, even within a contractualist approach. I have suggested an approach in terms of sympathetic identification and extended goodness that falls short on establishing that individuals should agree on their judgments. Disagreement over judgments of goodness is, however, not problematic in a social contract perspective as long as individuals are in general agreement over the rules to make collective choices.

Teleological justification is indeed only a part of the justificatory endeavor. Moral pluralism, i.e., the fact that individuals disagree over their conceptions of goodness, can coexist with an overlapping consensus over core values. But individuals should also agree over principles that, *given prevailing conceptions of the good*, rule collective decision-making. One reason Rawls gave priority to the right over the good was precisely his belief that 'reasonable pluralism' was possible as far as conceptions of the good are concerned, but not in the case of the right. But disagreement over the right is also an important fact of modern societies. I have assumed in this paper that the two forms of justification (teleological and deontological) can be tackled quite independently from each other. This is a simplification, however. Principles about the right also build on values that should be endorsed by individuals taking the moral point of view. Further reflections on the relationship between teleological and deontological justification are required to account for the fact of moral disagreement in modern societies.

REFERENCES

Adler, Matthew D. 2014. "Extended Preferences and Interpersonal Comparisons: A New Account." *Economics and Philosophy* 30 (2): 123–162.

Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. New Haven, CT: Yale University Press.

Arrow, Kenneth J. 1978. "Extended Sympathy and the Possibility of Social Choice." *Philosophia* 7 (2): 223–237.

Ashford, Elizabeth, and Tim Mulgan. 2018. "Contractualism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Article published August 30, 2007; last modified April 20, 2018. https://plato.stanford.edu/entries/contractualism/.

Binmore, Kenneth G. 1998. *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: The MIT Press.

Broome, John. 1993. "A Cause of Preference Is Not an Object of Preference." *Social Choice and Welfare* 10 (1): 57–68.

Darwall, Stephen. 2002. *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.

Fleurbaey, Marc. 2003. "On the Informational Basis of Social Choice." *Social Choice and Welfare* 21 (2): 347–384.

Freeman, Samuel. 1994. "Utilitarianism, Deontology, and the Priority of Right." *Philosophy & Public Affairs* 23 (4): 313–349.

Freeman, Samuel. 2007. *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*. New York, NY: Oxford University Press.

Gaus, Gerald F. 1990. *Value and Justification: The Foundations of Liberal Theory*. New York, NY: Cambridge University Press.

Gaus, Gerald F. 2011. "Social Contract and Social Choice." *Rutgers Law Journal* 43: 243–276.

Gaus, Gerald F. 2012. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Reprint edition. New York, NY: Cambridge University Press.

Gaus, Gerald F., and John Thrasher. 2015. "Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi." *In The Original Position*, edited by Timothy Hinton, 39–58. Cambridge: Cambridge University Press.

Gauthier, David. 1977. "The Social Contract as Ideology." *Philosophy & Public Affairs* 6 (2): 130–164.

Gauthier, David. 1987. *Morals by Agreement*. New York, NY: Oxford University Press.

Hampton, Jean. 1980. "Contracts and Choices: Does Rawls Have a Social Contract Theory?" *The Journal of Philosophy* 77 (6): 315–338.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (5): 434–435.

Harsanyi, John C. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Harsanyi, John C. 1996. "Utilities, Preferences, and Substantive Goods." *Social Choice and Welfare* 14 (1): 129–145.

Hausman, Daniel M., and Michael S. McPherson. 2006. *Economic Analysis, Moral Philosophy and Public Policy*. 2nd edition. New York, NY: Cambridge University Press.

Mongin, Philippe. 2001. "The Impartial Observer Theorem of Social Ethics." *Economics & Philosophy* 17 (2): 147–179.

Nagel, Thomas. 1989. *The View From Nowhere*. New York, NY: Oxford University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Pattanaik, Prasanta K. 1968. "Risk, Impersonality, and the Social Welfare Function." *Journal of Political Economy* 76 (6): 1152–1169.

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): 515–572.

Rawls, John. 1993. *Political Liberalism*. New York, NY: Columbia University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*. 2nd edition. Edited by Erin Kelly. Cambridge, MA: Harvard University Press.

Robbins, Lionel. 1938. "Interpersonal Comparisons of Utility: A Comment." *The Economic Journal* 48 (192): 635–641.

Sandel, Michael. 2010. *Liberalism and the Limits of Justice*. 2nd edition. New York, NY: Cambridge University Press.

Sen, Amartya. 1970. *Collective Choice and Social Welfare*. San Fransisco, CA: Holden-Day.

Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica* 40 (159): 241–259.

Sen, Amartya. 1979. "Utilitarianism and Welfarism." *The Journal of Philosophy* 76 (9): 463–489.

Sen, Amartya. 2017. *Collective Choice and Social Welfare: Expanded Edition*. London: Penguin Books.

Suppes, Patrick. 1966. "Some Formal Models of Grading Principles." *Synthese* 16 (3/4): 284–306.

Weymark, John. 2016. "Social Welfare Functions." In *The Oxford Handbook of Well-Being and Public Policy*, edited by Matthew D. Adler and Marc Fleurbaey, 126–158. New York, NY: Oxford University Press.

Cyril Hédoin is Professor of Economics at the University of Reims Champagne-Ardenne (France). He works at the intersection of economics and philosophy, and more particularly on issues related to rationality, rules, and institutions. He has more recently written on topics articulating normative economics with political and moral philosophy. His work has been published in many philosophy and economics journals, such as *Economics and Philosophy*, *Erkenntnis*, and *Social Epistemology*.
Contact e-mail: <cyril.hedoin@univ-reims.fr>