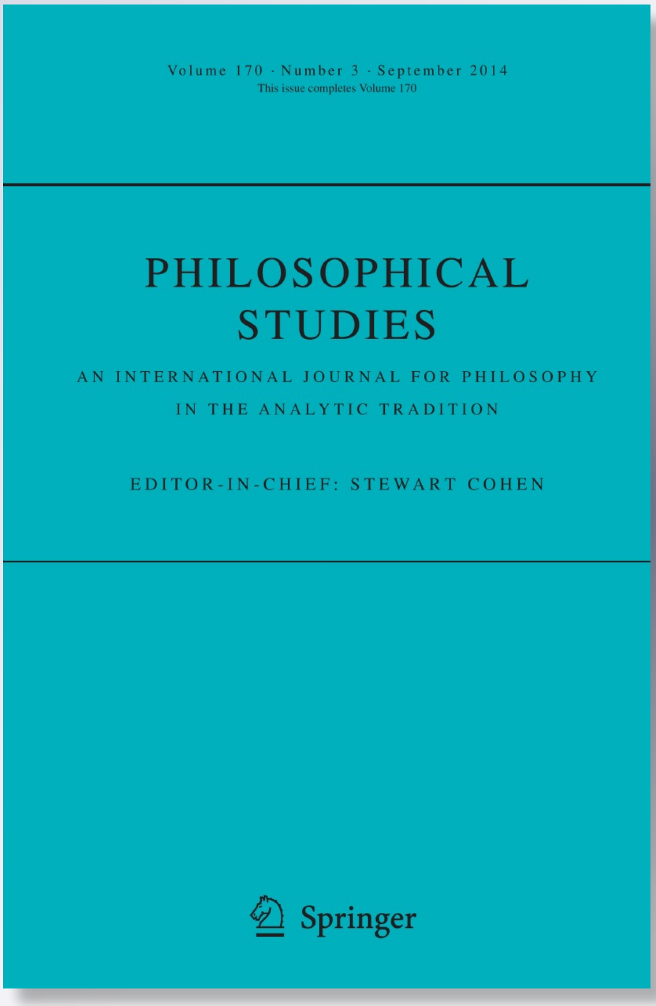
Disagreement behind the veil of ignorance

Ryan Muldoon, Chiara Lisciandra, Mark Colyvan, Carlo Martini, Giacomo Sillari & Jan Sprenger

Philosophical Studies
An International Journal for Philosophy
in the Analytic Tradition

ISSN 0031-8116
Volume 170
Number 3

Philos Stud (2014) 170:377-394
DOI 10.1007/s11098-013-0225-4



Volume 170 · Number 3 · September 2014
This issue completes Volume 170

PHILOSOPHICAL
STUDIES

AN INTERNATIONAL JOURNAL FOR PHILOSOPHY
IN THE ANALYTIC TRADITION

EDITOR-IN-CHIEF: STEWART COHEN

Springer

Philos Stud (2014) 170:377–394
DOI 10.1007/s11098-013-0225-4

Disagreement behind the veil of ignorance

Ryan Muldoon · Chiara Lisciandra · Mark Colyvan · Carlo Martini · Giacomo Sillari · Jan Sprenger

Published online: 22 October 2013

Abstract In this paper we argue that there is a kind of moral disagreement that survives the Rawlsian veil of ignorance. While a veil of ignorance eliminates sources of disagreement stemming from self-interest, it does not do anything to eliminate deeper sources of disagreement. These disagreements not only persist, but transform their structure once behind the veil of ignorance. We consider formal frameworks for exploring these differences in structure between interested and disinterested disagreement, and argue that consensus models offer us a solution concept for disagreements behind the veil of ignorance.

Keywords Rawls · Veil of ignorance · Disagreement · Consensus modeling · Bargaining

R. Muldoon (✉)
Philosophy, Politics and Economics Program, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: rmuldoon@sas.upenn.edu

C. Lisciandra · C. Martini
Department of Political and Economic Studies, Finnish Centre of Excellence in the Philosophy of the Social Sciences, University of Helsinki, PO Box 24, 00014 Helsinki, Finland

M. Colyvan
Department of Philosophy, University of Sydney, A14 Main Quadrangle, Sydney, NSW 2006, Australia

G. Sillari
Department of Political Science, LUISS Guido Carli, Viale Romania 32, 00197 Rome, Italy

J. Sprenger
Tilburg Center for Logic, General Ethics and Philosophy of Science, Tilburg University, 5037 AB Tilburg, The Netherlands

1 Moral disagreement and self-interest

Our moral and political disagreements are often extremely complex. Not only do they involve many competing interests, they also involve many competing principles. Sen (2009, pp. 12–15) illustrated this with the example of three children arguing over who gets to have a flute. Alice argues that the flute should be hers because she spent a great deal of time making it, and this effort should give her rights to the flute. Bob argues that he is the only one who can play the flute, and so everyone would benefit were he to have the flute. Carol argues that since she has no toys of her own, while the other two have many toys, she should get the flute.

Each position describes the competing interests of the children. Not only that, each position represents a legitimate philosophical position: Alice standing in for a libertarian, Bob for a utilitarian, and Carol for an egalitarian. These are all positions for which there are serious arguments in their defense—none is obviously untenable. Given that they are all viable philosophical positions, the mere fact that a child holds a given position is not evidence that they are acting in an interested manner. Crucial to this particular disagreement, however, is that each child has taken on the position that most benefits him or her. A first step in our analysis of this disagreement is whether each child would hold the position that they do were they not to stand to benefit from it. Would Alice continue to want to espouse libertarianism if she were in Carol's position?

In these sorts of moral disagreements, we find that one or more disputants can disguise their self-interest as moral indignation. This may or may not be conscious—individuals can honestly believe that they are fighting for their conception of justice without realizing that they prefer a conception of justice that just so happens to favor people in their position. Sometimes, of course, people knowingly adopt moral language to defend themselves, precisely because others are much more likely to be sympathetic to moral language than the language of raw self-interest. People are naturally influenced by the incentives in front of them: they often find positions that favor their interests more appealing than they would have otherwise, and they may have a tendency to ignore information that might cause them to rethink their position (Babcock and Loewenstein 1997). As a result, individuals have a tendency to choose those accounts of fairness that best suit their self-interest (Babcock et al. 1995; Bicchieri and Mercier 2013). What is more, while they may be able to identify bias in others, they are not able to introspect well enough to see that their own perceptions are similarly biased (Pronin et al. 2002). We encounter this regularly in everyday life. Politicians, judges, and private citizens frequently adopt positions that either advance their larger ideological goals, or their material self-interest, even if they are not consciously doing so. So, whether it is from clouded judgment, or intentional sophistry, our competing interests can inhibit our ability to arrive at moral agreement in a reasoned manner.

Given all this, we might suppose that it is just competing interests that cloud our judgment. If we could simply eliminate our interests, then we would immediately be able to arrive at robust agreement. Rawls offered a powerful version of this in *A Theory of Justice*, by means of the thick veil of ignorance (Rawls 1971, pp. 11,17, Sect. 24). Agents under the thick veil of ignorance do not know who they are in society, whether they are a past, present or future generation, whether they are male or female, rich or poor, in the majority or in a minority, or any other relevant demographic feature. Not only that, but they also do not know the overall demographics of the society in which they live, or any other morally arbitrary facts. Agents are only aware of basic facts about biology, economics, physics, and any other relevant scientific facts that are well-established. In fact, the agents in the Original Position could be thought of as lawyers who do not know whom their clients are. They want to make sure that everyone gets as good of a deal as they can.

Their knowledge about the world is thus constrained to agreed-upon facts that shape the nature of human needs and our ability to provide for them. Rawls writes that:

> [In the original position] the parties must not know the contingencies that set them in opposition. They must choose principles the consequences of which they are prepared to live with whatever generation they turn out to belong to. (Rawls 1971, p. 119)

The veil of ignorance thus blocks the possibility of agents having knowledge of any particular interests that might sway them when deliberating about the basic structure of society. This is an extremely powerful idea—a framework that removes morally irrelevant considerations from our moral decision-making. It removes the possibility of our self-interest getting in the way of our reasoning. However, it is implicitly assumed in Rawls' work and elsewhere that the individuals in the Original Position, thus unburdened from self-interested bias, will then be able to agree.

> To begin with, it is clear that since the differences among the parties are unknown to them, and everyone is equally rational and similarly situated, each is convinced by the same arguments. Therefore, we can view the agreement in the original position from the standpoint of one person selected at random. If anyone after due reflection prefers a conception of justice to another, then they all do, and a unanimous agreement can be reached. (Rawls 1971 p. 120)

Rawls does not discuss the possibility of disagreement *behind* the veil of ignorance. There is no doubt that eliminating self-interest does cut out a major basis of disagreement but as we shall argue in this paper, it does not guarantee that *all* bases of disagreement will be eliminated. More precisely, we wish to argue that the device of the 'veil of ignorance' in moral and political philosophy, does not guarantee that all agents can be effectively reduced to a single agent selected at random. Even if agents behind the veil of ignorance are equally rational and similarly situated, it does not guarantee that they are convinced by the same arguments. In particular, rationality itself, even rationality constrained by massive ignorance and other features of the Original Position, does not fully specify an agent's judgments. Rawls himself was aware of this problem, and spent much time specifying the utility functions of the parties involved in the Original Position (see Rawls 1971, Sect. 25: The Rationality of the Parties). The worry is that the way that rationality as developed in Rawls is too demanding, as it includes a *perspective*. *Rationality* in Rawls is not *rationality* of decision theory, but it is the second, and less demanding, kind of rationality that we will appeal to in this paper. Moreover, applications of ‘veil of ignorance’ arguments go beyond the carefully constructed Original Position that Rawls developed. Our primary interest is considering the form of argument more generally, using Rawls’ work as an important illustration of this method. There is an important tension that we uncover: the more careful the theorist is in blocking the problems that we uncover with ‘veil of ignorance’ strategies, the less that theorist will be able to deal with the fact of pluralism. As we will see, this is because agents’ epistemic states will have to be highly circumscribed. This is in conflict with liberal goals more generally, insofar as liberal theories seek to accommodate a diversity of viewpoints, at least insofar as they are compatible with liberal social contracts. Thus, ‘veil of ignorance’ arguments are much more problematic than they initially appear to be, particularly when they serve as justifications for political claims that apply to pluralistic societies.

In the rest of paper, we develop a systematic account of the consequences of pluralism for veil of ignorance strategies in moral and political philosophy. Our main aim is to show two things: first, that disagreement can easily persist even in ideal implementations of veils of ignorance, and second, that the disagreement changes its form. We show that formal models of consensus formation, fruitfully used elsewhere, can shed light on how such disagreements might be resolved.

We will proceed as follows. First, we will survey the structure of the sorts of moral disagreements we are concerned with, and how they should be resolved (Sect. 2). To better understand the different kinds of disagreement we encounter, we present two formal disagreement-resolution procedures, namely bargaining (Sect. 3) and consensus through mutual respect (Sect. 4). Finally, we offer a brief conclusion (Sect. 5).

2 Disinterested disagreement

Rawls’ development of the veil of ignorance has had an enormous impact in both philosophy and economics. Huge numbers of papers in both fields rely on employing a veil of ignorance to investigate what our choices would be when we eliminate morally superfluous considerations. The appeal of relying on disinterested agents to resolve moral/political disputes is quite widely shared. We focus on Rawls here, but the basic structure of the argument extends well beyond Rawls. Our critique is primarily with the deployment of veil of ignorance strategies, and not at all directed at any particular principles of justice, Rawlsian or otherwise.

As we have seen, Rawls has suggested that we should think of agents behind the thick veil of ignorance as similar to lawyers who do not know who their clients are. These agents, since they are similarly situated and are equally rational, will end up reasoning as if they were a single agent. So, a broad understanding of Rawls’ account argues that once we eliminate divergent interests, we can eliminate disagreement. We aim to show, however, that the combination of rationality and the lack of interests is not sufficient for eliminating potential bases of disagreement. As we mentioned above, Rawls went beyond this, by further specifying the set of considerations that each agent has, arguing that they are more or less identical. So in Rawls’ case, we find a tension between preventing the possibility of disagreement behind the veil of ignorance, and making preconditions for entering into the veil of ignorance implausible for pluralistic societies. Even in ideal theory, agents in pluralistic societies would have differences in perspective that would introduce the possibility of disagreement. To see why this is the case, we will generalize to the effects of a veil of ignorance on a specific distributional question, and then return to the particulars of the Original Position.

Recall Sen's flute example. In the described situation, Alice, Bob and Carol are not simply arguing about *who* should get the flute, but *why*, on the basis of the other goods each of them possesses. Even under the veil of ignorance, they might still genuinely disagree and support different moral principles, regardless of any incentive or return they might expect from them. In other words, each child, rather than having particular interests in the moral principles that they wish to choose, can be understood as having different underlying categorizations for the world (e.g., in terms of primary goods, liberties, and capabilities). If so, the three children propose three different *principles* for assigning the flute to a particular agent and each of them legitimately believes that their view is the right one, with no bias forcing them in this direction. If one of them uses capabilities as an evaluative criterion, an argument that requires an acceptance of primary goods as the appropriate evaluative criterion is not going to be immediately compelling. This will similarly hold true for liberties, or any basic principle that can be employed as an evaluative criterion. Rationality itself does not resolve this dilemma, as it says nothing relevant with respect to how to categorize the world, or which evaluative criterion to employ. At best, rationality can be used to eliminate self-contradicting evaluative criterion, but that will not take us very far. The veil of ignorance does not eliminate all bases of disagreement, even if it does cut out a primary source of disagreement.

If we return to the specifics of the Original Position, though, one might think that Rawls can simply avoid the basic thrust of this objection. After all, Rawls stipulates that the agents behind the thick veil of ignorance will simply be making pairwise comparisons between a limited set of candidates for the basic structure of society. If it is just this, then we can point out that rational agents could well choose different measures of 'well-off' to ground their comparisons. In the Sen story above, we might find that some agents would prefer capabilities, others liberties, still others primary goods. If this is possible, then it is extremely straightforward to see how disagreements could arise. Everyone could be trying to make the worst-off best-off, but have different accounts of what those terms mean. We could end up with substantially different results from pairwise comparison in this instance.

One might well respond that a key element of the Original Position is that agents accept that primary goods are the appropriate measure for determining the ability of free and equal persons to formulate and carry out their life plans. As such, suggesting that an agent could instead choose to rely on capabilities as a measure would be changing the conditions of the Original Position too much. This returns us to the tension that Rawls faces: he may able to largely eliminate a significant source of disagreement, but he does so at a rather high cost. Because on this account, it is not the veil of ignorance that is doing the justificatory work for the two principles of justice—instead, it is the additional assumptions that go into structuring the Original Position. That all agents share a common perspective on how we ought to evaluate individual and social outcomes is extremely demanding, and not one that falls out of the veil of ignorance, but is rather imposed onto it. Requiring that everyone agrees on the same set of measures and evaluative criterion is even more demanding than requiring people to agree on preferences over outcomes. Even worse, there is no moment of agreement in the choice of primary goods—the agents are simply stipulated to all share the same perspective. There is no reason to believe that we would be able to achieve this without a great deal of political dispute. Disagreement in the Original Position is made much harder when such a stringent constraint is put into place, but it comes at a large cost: it is exceedingly rare that members of a polity would all agree on evaluative criterion upfront. That Rawls is working in ideal theory does not particularly protect him from this criticism, since if ideal theory is supposed to serve as a regulative ideal and have any normative force over us, it has to be able to capture important features of who we are. As is evidenced by the many different branches of political philosophy, even people who have dedicated their careers to thinking about what justice comprises in persist in having multiple accounts of what the best perspective is. Why would we suppose that ordinary individuals would be able to do this without any debate?

Just to see how deep this problem can be, let us suppose that there is nothing objectionable about constraining individuals in the Original Position such that they all share the same perspective, and agree that primary goods are the appropriate measure for distributional concerns. But then we are still left with possibilities for disagreement: many of the primary goods—particularly primary social goods—are not precisely defined. Take, for instance, the social bases for self-respect. This could very easily mean different things to different agents. Does one's ability to have nice dress clothes count, as Adam Smith argued in the Wealth of Nations? The answer to this question, and many others like it, would shift the nature of the basket of goods that we need to use as a measurement device. Differences in each of these questions could potentially help shift one's preferences away from the two principles of justice toward the principle of average utility combined with a social minimum, for example. So, even with heavily constrained agents, all by stipulation sharing the same goals and evaluative criteria, and ignorant of their interests, we can still find possible sources of disagreement. If we were to try and avoid this objection by simply further specifying the details of each primary social good, all we would be doing is further ratcheting up the demandingness of the assumption, and decreasing its power as a form of justification for whatever comes out of the Original Position. Just as we would not treat a theory that both demanded and hinged on the fact that we all live to be 150 years old as an ample source of justification, we may find an approach that demands something equally implausible to also lack in some justificatory force, regardless of what we think of the conclusions of the method.

The same problem affects urgent societal challenges, such as assessing the consequences of climate change. For example, is it acceptable to allow some little-inhabited islands to be submerged by rising sea-levels? Is it acceptable to allow some animal species inhabiting the world's coral reefs to go extinct due to ocean acidification? Whatever moral judgments one makes on the extent of acceptable climate change, these judgments are not only contentious, but also based on different categorizations of social, economic and environmental goods. Even when they abstract from their personal position in the world, agents may disagree on which quantities should be included in a moral assessment of climate change, and what counts as a social good in the first place. Is it only about increasing frequency of natural disasters and lower crop yields, or also about decreasing quality of life due to temperature duress in (sub)tropical regions, and decreasing biodiversity on Earth? Such judgments evidently depend on the agents' personal value systems, and they nicely illustrate deliberation in a sort of Original Position without invoking the entire climate change debate.

If this is right, then the device of the thick veil of ignorance can only accomplish part of the task that Rawls wishes for it. Eliminating interests can eliminate a significant source of potential moral conflicts, but it does not eliminate all of them. There is no guarantee that the agents behind the thick veil of ignorance will decide which account of justice to adopt, since there is nothing *epistemically* irrational about the persistence of disagreement. Bob's adherence to utilitarianism may lead me to identify evidence that Alice does not see as salient as a libertarian. It is not because either of them fails to properly include evidence in our judgments, but rather they disagree about what counts as evidence. Each perspective—each choice of an evaluative criterion—imposes a categorization on possible states of the world. The libertarian will reason in terms of gains or losses in personal liberties, whereas the utilitarian reason in terms of the amount of utility gained or lost in society. This necessarily requires them to understand situations in different ways, precisely because different aspects of the same situation will be relevant to how they evaluate it. Taking in 'all' of the aspects of a situation is simply cognitively infeasible. We necessarily limit what evidence we can take in, and how we evaluate shapes what we respond to.

This is an issue that's prior to rationality: how we categorize the world is prior to the axioms of rationality, not something that rationality decides for us. Since Rawls conceives of the thought experiment as political rather than metaphysical, we have no reason to believe that the agents are omniscient, or have privileged access to some 'best' evaluative criterion from the beginning. By design, legitimacy stems from the hypothetical contract agreement amongst ideally-situated agents. But there is no politically neutral way of eliminating differences in categorizations amongst agents. We cannot declare that, for example, primary goods are the only rational evaluative criterion, because there is nothing that we can provide to ground that claim without relying on political or moral judgments that may themselves be reasonably contested.

In the larger project of *A Theory of Justice*, this matters a great deal for when we consider the stability conditions of a Well-Ordered Society. The justificatory strategy that Rawls has for the Original Position is to show that the reasonable pluralism of different liberal conceptions could lead to an universal endorsement of the two principles of justice under idealized conditions. But, if this is not right, then the privileged position that the two principles have will be undermined.[1] If we are

[1] Of course, Rawls saw a version of this problem, which he called the 'burdens of judgment.' This motivated him to switch to the idea of the overlapping consensus in *Political Liberalism* (Rawls 2005).

constraining agents so heavily in the Original Position, and then eventually deal with unconstrained agents, why would we think that they would continue to share the same evaluative criterion? For those with different evaluative criterion, the choices made in the Original Position would have very little binding force. Even disinterested parties may not be able to accommodate my legitimate social and political interests if their judgments of salient features of justice are divorced from my own. Bob would not find reason to be bound by Alice's judgment, even if she were reasoning behind a veil of ignorance, precisely because Bob simply disagrees with how Alice understands the evaluative criteria for a distribution.

So, a veil of ignorance, unless it is stipulated to not only remove agents' particular interests, but also to force agents to conform to a particular categorization scheme for the (political) world, leaves open the possibility of agents finding themselves in a state of meta-disagreement. The Original Position, as we have seen, does in fact make both stipulations. This does resolve many issues with meta-disagreement, but not all of them. However, this comes at a cost: the stipulation of a particular categorization dramatically narrows the scope of justification.

Even while we argue that the veil of ignorance does not quite accomplish what Rawls claims for it, the general strategy does do something extremely important for our moral and political reasoning. If we think about the flute example once more, we note that once behind a veil of ignorance, the situation changes. No longer are we talking about individuals who have moral arguments that they might have changed had they been in a different position—we are not in a situation where moral arguments are mere veneers for self-interest. The veil of ignorance can safely shield us from the intrusion of self-interest. Instead, we are looking at principled disagreements about the nature of desert. It is not that we need to satisfy the individuals' competing interests, but rather we need to have them find a way of aligning their senses of justice in such a way that there is a mutually acceptable outcome. The veil of ignorance has shifted us from an object-level dispute (who gets what) to a meta-level dispute (what are the appropriate evaluative criteria to determine a system of allocation). Since the disagreeing agents have no special epistemic access to the world, this meta-level dispute is occurring amongst individuals who are on an equal footing. Everyone has their beliefs about the appropriate principles to use in moral disagreements, but given that there is disagreement, anyone could be mistaken. This should give us pause and introduce a little moral humility.

The veil of ignorance eliminates moral disagreement driven by idiosyncratic individual interests. However, it is not clear that the agents have adopted a common perspective on the world and a joint categorization moral goods that would allow them to apply the veil of ignorance in the first place. Thus, we need a *solution concept* for resolving disinterested disagreements. Without having universal scope, these solution concepts will help us to elaborate the conceptual differences between these two kinds of disagreements, and contribute to a better understanding of how agents might attempt to resolve tensions about the assignments of goods, duties and rights. We focus on two concepts that are particularly prominent in social science and philosophy: the Nash bargaining solution and the Lehrer–Wagner model of consensus. The latter is, in our view, particularly well-suited to helping resolve disinterested disagreement.

3 Modeling interested disagreement: bargaining games

The previous sections have argued that whenever we have a reasonably diverse set of agents in the Original Position, our moral/political disagreements do not simply disappear but instead change in character. What was once an object-level dispute becomes a meta-level dispute. Because of this change in the character of the disagreement, we argue that different forms of disagreement require different resolution procedures. To make this as clear as we can, we will investigate some formal solution concepts for moral disagreement.

Notably, moral arguments sometimes depend on the personal preferences of those who advance them. One may imagine that in Sen's example, Alice would no longer muster libertarian arguments if she had not made the flute herself.[2] Therefore, we will review whether the *Nash bargaining solution*—the most popular solution concept for overcoming interested disagreement in economics and politics—also implies a good strategy for resolving moral disagreement, be it interested or disinterested.

Assume that Alice, Bob and Carol are allowed to communicate, negotiate and enter binding agreements on how to divide a common good X. For instance, X may stand for the number of hours that the flute is available to the children per week. Alice, Bob and Carol then bargain for a *solution*, represented by the vector $x = (x_A, x_B, x_C)$, where the variables x_A, x_B and x_C represent the amount of time that Alice, Bob and Carol may use the flute individually. Typically, it is assumed that the total payoff to all players, $X = x_A + x_B + x_C$, is constant. This corresponds to our intuition that the bargaining problem is about dividing a cake, or another common good, and that no resources are wasted.

For the bargaining procedure, the *disagreement point* $d = (d_A, d_B, d_C)$ is crucial: it represents the outcome that would occur if negotiations failed (Nash 1950). Typically, $d_A + d_B + d_C$ is smaller than the total good. For instance, the parents of Alice, Bob and Carol might decide to withhold the flute if the children fail to resolve their disagreement. In that case, none of the children would have access to the flute and the disagreement point would be $d = (0, 0, 0)$. Players thus have an incentive to negotiate and to find a solution where resources are used efficiently and the entire good is divided.

Under these constraints, John Nash proposed as the solution for a bargaining game the point $x \in X$ in which the product $N(x)$ is maximized:[3]

[2] Note that these personal preferences need not be selfish: an altruist may fiercely pursue the well-being of those who are more disadvantaged than her. In such situations, an agreement on moral questions is transformed into a settlement of competing, position-dependent interests.

[3] The definition below is easily generalizable to more than three players.

$$N(x) = (x_A - d_A)(x_B - d_B)(x_C - d_C) \tag{1}$$

This solution concept, well-entrenched in the economic literature, has an interesting consequence: those agents that can more easily afford a disagreement (and a resulting lower payoff) have greater *bargaining power* and, as a result, receive more than those agents that can less easily afford a disagreement. For instance, instead of assuming that none of the children starts out with a flute, we may imagine that Alice starts out with it, since she made it. We might expect then that Alice will end up with greater access to the flute at the end of the bargaining process. For two players, we may conceptualize this as $X = 10$, $d_A = 4$ and $d_B = 0$. Then the Nash solution will be the rather unbalanced allocation $x_A = 7$, $x_B = 3$, as illustrated by Fig. 1.

As a result, the Nash solution tends to favor power relations, that need not display any fairness content, over classical utilitarian considerations in defining the solution of a bargaining game. If, in the above example, the good were more valuable to player B than to player A, one could argue that the ideal utilitarian solution should be tilted toward B, in order to maximize total happiness. However, according to the Nash solution, player A, who has the greater bargaining power, would obtain a greater share of the good, even if it he took less pleasure from it than B does. The implication for our example is that the bargaining solution will lean closer and closer to the outcome favored by the more powerful bargainer, remaining entirely unconcerned with issues of fairness towards other players, or towards future generations. In spite of the existence of a large literature on fairness in bargaining (Gauthier 1986; Binmore 1994, 1998), it thus seems that the Nash solution can only apply to resolving interested disagreements, not to resolving disinterested disagreement.

To this one could object that the moral reasoners in the Original Position form judgments over the appropriate distributive schemes in a society (e.g., whether desert, merit or liberty is crucial). They have preferences over the set of candidate schemes, and we can represent these preferences by means of a utility function that is unique up to positive affine transformation. Clearly, these agents need not have personal interests in the agreed upon scheme. So why should we not invoke the Nash bargaining solution in order to resolve the disagreement?

This objection sounds persuasive, but it neglects some crucial features of the Nash bargaining solution. First, there has to be a *default outcome* that will be implemented if we fail to reach agreement. This may be a realistic assumption in the context of real world deliberations where the status quo may play this role. But taking the status quo as a point of reference seems inappropriate for the abstract moral reasoning employed in the Original Position. Second, the agents need to *care* for the outcome, according to their personal utility scheme: if they were not concerned that failure of negotiations led to an inferior distributive scheme, they would have no reason to make concessions. Fear is one of the main driving forces for reaching agreement in a Nash bargaining game, and the location of the disagreement point makes some accept much greater concessions than others. Clearly, such a conceptualization is inconsistent with a view of the Nash bargaining game as a procedure for resolving disinterested moral disagreement. On the other

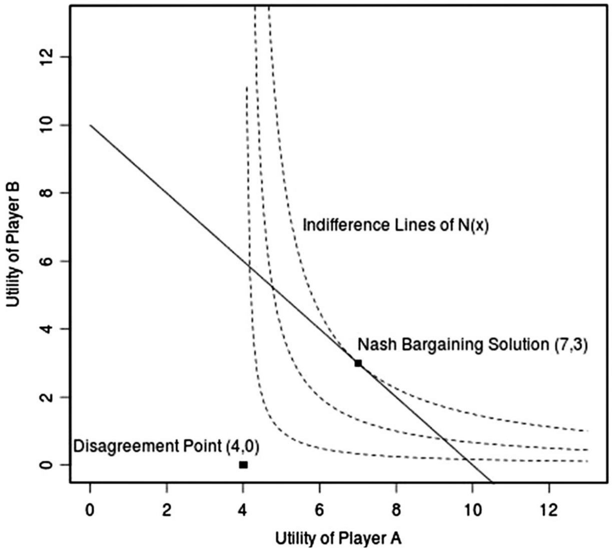


Fig. 1 The Nash Solution for the constraint $x_A + x_B = 10$ (*full line*) and disagreement point $d_A = 4$, $d_B = 0$. The *dashed lines* represent, in increasing order, the *indifference lines* of the $N(x_A, x_B)$-function. The Nash bargaining solution is the point where the constraint $x_A + x_B = 10$ lies tangential to the *indifference lines*

hand, they may be suitable tools for representing interested moral reasoning of real-world agents, e.g., in the case of climate negotiations.

The reasons that bargaining models can be so effective at capturing the nature of interested moral disagreements are those why they fail at capturing key features of disinterested disagreements. What bargaining brings to the fore—caring for the outcome, relativity to a default state of affairs—is precisely what the Original Position seeks to eliminate. As we have seen, reasoning in the Original Position does not eliminate all disagreement, so we need to find an alternative framework that will help us understand the nature this residual disagreement. To get a handle on this, we examine alternative models where the disagreement resolution is based on *peership* and *mutual respect* between the parties involved. The Lehrer–Wagner model, introduced in the next section, is the most prominent representative of those models.

4 Modeling disagreement: consensus through mutual respect

One promising approach to disinterested disagreement is *consensus through mutual respect*. When we disagree on an issue on which we have no stakes, what effectively prevents a consensus is our conviction that our position is the correct one. But on the other hand, we typically regard our own arguments and positions as fallible. That is, we admit that other positions are acceptable, although we may give less weight to them in our considerations. We have respect for the reasoning of other agents who reach a different conclusion. This is the concept of *mutual respect* on which several prominent consensus models are based.[4] Disagreement behind the veil of ignorance is translated as disagreement about the appropriate moral reasoning, where my belief that the other group members are sincere moral reasoners, and my respect for their competence, has me agree to a process that aims at a consensus. Applying this procedure to disagreement in the Original Position, we have individuals apply a procedure that does not eliminate their initial conflicting views about the appropriate ways of categorizing the world but, rather, combines them in a process of mutual aggregation.

There are two minimal elements required for modeling the elements of a consensus: *individual opinions* and *degrees of mutual respect*. Reaching a consensus by combining opinions and respect was first implemented by French (1956) and developed further by DeGroot (1974), Lehrer (1976), Wagner (1978) and Lehrer and Wagner (1981). The idea is that an agent averages her own judgments with the judgment of her peers, according to the respect weights she assigns to her peers, and their competence as moral reasoners. By repeating this procedure, consensus will eventually be achieved. As we will show, the importance of analyzing a situation of disinterested disagreement with a formal model is that we can highlight exactly those aspects of the problem that mark the difference with the interested disagreement manifested in the bargaining case. Moreover, the theory developed by Lehrer and Wagner allows us to understand the precise conditions needed for resolving disinterested disagreements. While we leave the technical details of the model to the interested reader (see Lehrer and Wagner 1981), a few remarks on how the process of convergence works are in order. A summary of the technical details is given in Appendix.

Let us return to Sen's flute-example, although this time under the veil of ignorance. Alice, Bob, and Carol are now three neutral agents who 'do not know certain kinds of particular facts' (e.g., which skills they possess, how wealthy they are, etc.) but only have general information, for instance, on 'political affairs', 'principles of economic theory', 'basis of social organization', 'laws of human psychology', and so forth (Rawls 1971, p. 119). In this case, as we argued in the previous sections, there can still be disagreement, namely, disagreement on the appropriate moral principles for assigning the flute to an individual.

Firstly, the formalism of the Lehrer–Wagner gathers all relevant information on the agents in a compact format: each agent is assigned a *profile* that contains her opinion. For example, Alice thinks that egalitarianism should be used to resolve the disagreement and this will be reflected in her determination of the time allocations with the flute (i.e., the time Alice would have each person spending with the flute). In the formal treatment, the opinions of the N agents is summarised in a vector $(v_1, \ldots, v_N)$. The Lehrer–Wagner formalism also works with a $N \times N$-matrix of *weights of respect* W, whose entries w_{ij} mirror agent i's weight of respect for agent j. For example, Alice may assign herself twice as much weight as she assigns to Bob

[4] Consensus models were originally developed with factual disagreements in mind, but recent work on the topic has suggested that these models have wider applications (Steele et al. 2007; Martini et al. 2012).

and Carol, reflecting her view that their moral reasoning is not as trustworthy as her own. This would then correspond to the weight assignment (1/2, 1/4, 1/4). Bob and Carol's thoughts about the others' competence would then complete the weight matrix *W*.

Secondly, the agents in the group accept that the consensual result will be an aggregation of the information contained in each profile, so that an agent's updated opinion, after the first round of deliberation, will be a weighted average over my initial opinion and the opinions of the other members.[5] Lehrer and Wagner argue that the rationale for agreeing to this aggregation procedure is implicit in the fact that the agent is prepared to enter into deliberation:

> If a person refuses to aggregate, though he does assign positive weight to other members, he is acting as though he assigned a weight of one to himself and a weight of zero to every other member of the group. [...] One justification for aggregation is consistency, since refusing to aggregate is equivalent to assigning everyone else a weight of zero and aggregating. (Lehrer and Wagner 1981, p. 43)

That is, if we refuse to aggregate our opinions with the rest of the group even to the smallest degree, we make a very implausible assumption: namely that we find our own opinion infinitely more important than the opinion of the other group members. Not only that, but when we apply this to Rawls' Original Position, agents have no basis for granting a weight of zero to others. They do not know anyone's backgrounds, special skills, or position in society. All they know is that they disagree on how to categorize features of the world.

Given this ignorance, there isn't any rational ground on which an agent could choose to ignore everyone else.[6]

> Actual disagreement among experts must result either from an incomplete exchange of information, individual dogmatism, or a failure to grasp the mathematical implications of their initial stage. What is impossible is that the members of some community of inquiry should grasp the mathematical implications of their initial state and yet disagree. (Lehrer 1976, p. 331)

Thirdly, Lehrer and Wagner show that the process leads to convergence of opinions in the long run. For the purposes of this paper it will be enough to point out the fact that the conditions for convergence are not particularly demanding ones: as long as the positions that the agents in the model hold can be somehow numerically aggregated, and as long as the agents in the model are willing to give an (even minimal) degree of respect to the other agents, then the model will allow the agents to converge to a common view. What is to be noted, however, is that the model does not suppose that there is a single best principle, which can be individuated under the veil of ignorance; but rather, that such principle can be individuated in the first place

[5] Formally, this amounts to matrix multiplication of the opinion vector by the weights matrix.

[6] Besides, an agent who refused to be rational and listen to her peers in the Original Position would surely violate the spirit of Rawls' thought experiment in a fundamental way.

is the result of adopting the Lehrer–Wagner model, and the uniqueness is guaranteed by the model itself. What is assumed by the Lehrer–Wagner model is the mathematical function (a matrix of weights multiplied by a vector of opinions) that represents the situation of disagreement. Even this, though, is not a presupposition if, as Lehrer and Wagner (1981) do, we take the function to be a somewhat idealized but realistic representation of the situation of disagreement.

It remains to explain how the weights are to be interpreted under the veil of Ignorance. Initially, we suggested that the weights could be thought of as measures of respect or trust. This could be understood as the simplest case. If we think of these measures as one of respect we might reasonably assume that all agents have equal respect under the veil of ignorance, and so they would find themselves respecting every opinion equally. However, an advantage of the Lehrer–Wagner model is that it can also accommodate more complex scenarios. This can be advantageous if we wish to enrich our understanding of moral disagreements.

These observations are independent of whether we are concerned with factual or with non-factual disagreements (Steele et al. 2007; Martini et al., 2012). While the epistemic justification of repeated pooling is problematic for the case of weights that express ascriptions of truth-tracking abilities, there is a consistent and intuitive justification for resolving non-factual disagreements through mutual respect, namely as a gradual discovery of mutual strength of preferences. Martini et al. (2012) argue that in the non-factual case, repeating the differential weighting procedure can be better defended than in the factual, epistemic case. Thus, the Lehrer–Wagner model not only helps capture several key features of disinterested disagreement, but does so in a way that allows us to see the possibilities for many different disagreement scenarios.

Another natural interpretation of the weights in the case of non-factual disagreement is in terms of *coherence of values*. On this interpretation, the agent is motivated to put more weight in those opinions that are closer to her own, given that they are more likely to cohere with the outcome of reflective equilibrium. This interpretation does not only protect the group view against an excessive influence of radical outsider views—it can also be motivated by prominent models of group decisions developed in social psychology, such as Davis' (1973) social decision scheme (SDS). To wrap up, the possibility of modeling non-trivial respect relations in a potentially large group of deliberating agents is one of the main assets of the Lehrer–Wagner model. Notably, it can be applied independently of whether the subject matter of the disagreement is factual or non-factual: the only thing that changes is the interpretation of the respect weights (as degrees of competence, degrees of care or coherence of values).[7]

We have already seen that a canonical model of interested disagreement—the Nash bargaining model—fails to capture the nature of disinterested disagreement. Why might we think that the Lehrer–Wagner model does any better? Crucial here is

[7] These weights need not depend on how close the agents are to each other with regard to their respective estimates. Using distance between estimates as a measure of respect was advocated by Regan et al. (2006) for pragmatic reasons and also by Hegselmann and Krause (2002). The model permits us to assign relative weights as a function of the difference of opinion, if this is desired, but it is not required.

that the nature of the model's outcome is rather different. Agents in a Lehrer–Wagner model come to a *consensus*. They come to this consensus through a procedure based on mutual recognition and respect. Agents in the Original Position need to be able to resolve disagreements over evaluative criteria, categorizations, and other such basic elements that are pre-requisites of the kind of reasoning that the Original Position requires. Since these social determinations will effect the outcomes of later decisions made behind the thick veil of ignorance, agents must focus on gaining consensus. If any parties are left unconvinced by the choices made at this stage, the use of the Original Position as a source of political justification becomes questionable. So it would seem that consensus is the appropriate standard for agreement. Not only that, but since the thick veil of ignorance does not suppose an adversarial relationship amongst the agents, we might expect that the appropriate attitude between agents is one of recognition and respect. The Lehrer–Wagner model captures these assumptions very well, and unlike the Nash bargaining solution, it refers neither to a default state, nor for caring for the outcome (but it cares for the moral reasoning abilities of the other group members).

A consequence of adopting the Lehrer–Wagner model for understanding disinterested disagreement is that the resolution of this disagreement is contingent on both the opinions of each individual, as well as their levels of respect for each other. We might suppose that given the setup of a veil of ignorance that respect weights are importantly constrained: it's inconsistent with the veil of ignorance for us to have differential weighting amongst agents: Bob couldn't respect Alice more than Carol simply because Bob has nothing beyond the fact of disagreement to rely on. By assumption, all individuals are equal in esteem. It is an open question as to whether individuals should weigh their own opinion more than others in such a setting—something that could be fruitfully investigated, but is beyond the scope of this paper. Once both opinions and respect measures are fixed, there is a unique solution that describes the point of eventual consensus. But the Lehrer–Wagner model reminds us that even in an ideal theory, the output is dependent on the inputs: we can only know what a pluralistic society would choose when we establish which perspectives are represented in the society. This is both mundane and deeply important: of course, what we can reasonably come to a consensus on depends on the makeup of the group coming to consensus. But if this is true, then it suggests that we ought to be suspicious of veil of ignorance arguments that make claims to universalism. Not only does diversity matter, but the composition of our diversity matters to what we are able to justify using this form of argument.

Where the Lehrer–Wagner model truly shows its worth is that it helps us to see how disinterested, well-meaning agents could still disagree. And given this disagreement, how they might come to a consensus. It also reveals the limitations of this form of justification, particularly for theorists that aim to justify universalist claims. We do not claim that the Lehrer–Wagner model is the final word on how we might resolve disagreements behind the veil of ignorance, but it does help reveal the structure of these disagreements and how they are different in kind from interested disagreements. Given this, we are better placed to understand the dynamics of resolving these disagreements and how the dynamics too look quite different in the two cases.

5 Conclusions

In this paper, we have argued that even under Rawls' thick veil of ignorance, there can be disagreements about the right moral account, and the right division of goods. Rawls' veil of ignorance very elegantly provides a framework for eliminating interest from our moral reasoning. What we have argued, however, is that simply eliminating interest does not eliminate disagreement. Good-faith disagreement amongst rational agents can arise behind the veil of ignorance.

This is an important insight for two reasons. Of primary importance, it highlights the challenges that we can face in a pluralistic society. Competing interests are not the only source of disagreement. Many disagreements, and in fact, many of our most fundamental disagreements, come not (just) from conflicting interests, but from deep-seated theoretical commitments. Our perspectives on the world have a much larger effect on our reasoning than we might appreciate. These perspectives do not get filtered out in the process of stepping behind the thick veil of ignorance. Rationality alone does not tell us how to see the world, how to measure things in it, or which things we should measure. When we are stripped to our bare rationality, we still have to choose how to answer these questions. Pluralistic societies ought to expect a wide variety of answers to these questions. Simply idealizing this diversity away does not help us address the problem. When we take diversity seriously, we find that unique solutions that are independent of the composition of the population are no longer plausible outcomes.

A second reason why this is important is that it illustrates the value of model-based reasoning in moral and political philosophy. Thought experiments such as Rawls' Original Position are true achievements in their ability to identify crucial issues, and help us reason about them, by placing complex ideas in situations that we better understand. Formal models can offer an additional resource. They help us simplify a problem and get at the core issues, but since we can explore these issues mathematically or algorithmically, we are sometimes able to uncover surprising nuances. If we think of disagreement as stemming from conflicts of interest, then we would naturally suppose that it would disappear when we remove interest. But if we model each case, we can see how disagreement might still be present, albeit in a different form than before. Modeling offers us an extra tool that can help us reason about complex situations. In this instance, it enabled us to see how assumptions about homogeneity do at least as much work as the structure of the veil of ignorance itself in justifying particular outcomes.

We do not take ourselves to have exhausted the ways in which moral agents could seek to resolve interested disagreements or disinterested disagreements. What we have shown, however, is that disinterested disagreement is possible—and in fact likely in pluralistic societies. If anything, we should expect it in multiple aspects of the debate. Not only that, but disinterested disagreement is different in kind from interested disagreement. This not only enriches our understanding of *A Theory of Justice*, but opens up new possibilities for moral and political philosophy more generally. With an enriched idea of the nature of disagreement, we may turn our attention to new and better ways of resolving deep-seated conflict in pluralistic societies.

Acknowledgements The authors wish to thank Samuel Freeman, Douglas Paletta, Gerald Gaus, and the audience of the Conference 'New Directions in the Philosophy of Science' held at the Bertinoro Conference Centre (University of Bologna) for their useful comments on earlier versions of the manuscript. The usual exculpations apply. Part of this work was funded by a Dutch Science Foundation Internationalisation Grant.

Appendix

Summary of the Lehrer–Wagner model

Let $G = \{1, \ldots N\}$ be a group of agents. The Lehrer–Wagner model is concerned with the problem of estimating an unknown quantity x from the individual estimates x_i of every group member i. This quantity x is thought of as objective and independent of the group members' cognitive states.

Lehrer and Wagner's central idea consists in ascribing the agents beliefs about each other's expertise, or in other words, mutual degrees of respect for the issue at hand. These weights w_{ij} represent the respect that agent i has for agent j, relative to the subject matter in question, and describe the proportion to which j's opinion affects i's revised opinion. The mutual respect assignments are in an $N \times N$ matrix W:

$$W = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1N} \\ w_{21} & w_{22} & \ldots & w_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ w_{N1} & w_{N2} & \ldots & w_{NN} \end{pmatrix}.$$

An important mathematical constraint is that the values in each row are nonnegative and normalized so as to sum to 1: $\sum_{j=1}^{N} w_{ij} = 1$. Then, W is multiplied with a vector $\vec{x} = (x_1, \ldots, x_N)$ that contains the agents' individual estimates of x, obtaining a novel updated value for $\vec{x}$:

$$W \cdot \vec{x} = \begin{pmatrix} w_{11}x_1 + w_{12}x_2 + \ldots + w_{1N}x_N \\ w_{21}x_1 + w_{22}x_2 + \ldots + w_{2N}x_N \\ \ldots \\ w_{N1}x_1 + w_{N2}x_2 + \ldots + w_{NN}x_N \end{pmatrix}.$$

In general, however, this procedure will not directly lead to consensus, since the entries of $W \cdot \vec{x}$ differ: $(W\vec{x})_i \neq (W\vec{x})_j$. However, it can be shown that the *iterated* application of the pooling procedure represented by W, W^n, converges to the so-called "consensus matrix" W^∞. It can also be shown that the individual entries of $W^\infty \cdot \vec{x}$ are equal to each other. Thus, repeating the pooling procedure leads to (rational) consensus.

References

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review, 85*(5), 1337–1343.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives, 11*(1), 109–126.

Bicchieri, C., & Mercier, H. (2013). Self-serving biases and public justifications in trust games. *Synthese, 190*(5), 909–922.

Binmore, K. (1994). *Game theory and the social contract, Vol. 1: Playing fair*. Cambridge: MIT Press.

Binmore, K. (1998). *Game theory and the social contract, Vol. 2: Just playing*. Cambridge: MIT Press.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review, 80*, 97–125.

DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association, 69*, 118–121.

French, J. R. P. Jr. (1956). A formal theory of social power. *Psychological Review, 63*(3), 181–194.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Hegselmann, R., & Krause, U. (2002). Opinion Dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3). http://jasss.soc.surrey.ac.uk/5/3/2.html.

Lehrer, K. (1976). When rational disagreement is impossible. *Noûs, 10*(3), 327–332.

Lehrer, K., & Wagner, C. (1981). *Rational consensus in science and society*. Dordrecht: Reidel.

Martini, C., Sprenger, J., & Colyvan, M. (2012). Resolving disagreement through mutual respect. *Erkenntnis*.

Nash, J. (1950). Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48–49.

Pronin, E., Lin, D., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381.

Rawls, J. (2005). *Political liberalism*. New York: Columbia University Press.

Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.

Regan, H. M., Colyvan M., & Markovchick-Nicholls, L. (2006). A formal model for consensus and negotiation in environmental management. *Journal of Environmental Management, 80*(2), 167–176.

Sen, A. (2009). *The idea of justice*. Cambridge: Belknap Press.

Steele, K., Regan, H. M., Colyvan, M., & Burgman, M. A. (2007). Right decisions or happy decision makers?. *Social Epistemology, 21*(4), 349–368.

Wagner, C. (1978). Consensus through respect: A model of rational group decision-making. *Philosophical Studies, 34*, 335–349.