"What is the difference between your subset objection to Rawls on utilitarianism and T.H. Irwin's commentary?"

Author: Terence Rajivan Edward

Abstract. T.H. Irwin's stimulating commentary on John Rawls anticipates but does not make "the subset objection to Rawls." This term of mine is potentially misleading, but Irwin's commentary is more so: I argue that relevant parts involve dubious commitments.

Draft version: Version 3 (May 23st 2022, "rejection in favour…").

In his monumental book *A Theory of Justice*, John Rawls sets out to argue against utilitarianism. Some of his arguments involve his original position thought experiment, or informal model, which involves self-interested individuals choosing from a menu of principles. Each individual lacks knowledge that would lead them to be biased towards their own case, so that what they choose is fair.[1] Assuming we should implement a fair set of principles, we should implement what they choose. And Rawls argues that they would prefer his proposed principles over utilitarianism. What I call the subset objection to Rawls versus utilitarianism is the following: there are a number of premises involved in his original position argument, or arguments, against utilitarianism (such as that we should implement a fair set of principles, that lacking the knowledge these individuals lack results in a fair procedure, etc.), and only a few of these premises are needed to construct an argument against utilitarianism, so the long argument Rawls makes is pointless. We can just take that small subset of the total set of premises and draw

[1] That is the hope anyway. I am not adding qualifications at this scene setting stage.

the conclusion (Edward 2016: 40-41). I call this "the subset objection" but that is potentially misleading, amongst other reasons because there may be multiple small subsets which enable one to arrive at the conclusion. Anyway, I wish to make a comparison between what I say and what T.H. Irwin says in his commentary on Rawls, within the third volume of his history of Western ethics. Unfortunately, commentary on Rawls is scattered across disciplines and I was not acquainted with this commentary when I wrote on Rawls versus utilitarianism. A reviewer of all three volumes fears the work is destined to be a great unread (Kalderon 2016: 510).

Irwin tells us:

> …the assumptions about justice and fairness that we build into the Original Position must not be so precise that they explicitly reject utilitarianism. For it should still be an open question for us, when we look at the Original Position, whether utilitarian or non-utilitarian principles will come out of it. The conditions that specify the Original position should not embody an explicit commitment to non-utilitarian views about justice (2009: 909)

Thus Irwin has already anticipated a subset objection, which is that when we look into the assumptions built into Rawls's method we can quickly derive a rejection of utilitarianism. What is the point of using his longer method then, at least for the purpose of arguing against utilitarianism? But when we turn over the page to read more of Irwin, he does not actually make the objection himself. So that is one difference, a key difference: *Irwin conceives of the possibility of a subset objection, but he does not make it.* He is not even neutral. His view is that Rawls avoids such an objection, because any inconsistency between utilitarianism and the assumptions built-in to the method is hidden, when they are presented side by side. For Irwin, the

original position is essential, or at least very valuable, for revealing the inconsistency (2009: 910).

It is worth noting that Irwin depends on an unstated commitment when moving from the first sentence quoted above, or sentence fragment, to the second sentence. The commitment is this:

> (*Forwards-inconsistency commitment*) If the assumptions built into the original position are clearly inconsistent with a proposed principle, then prior to applying the original position method, it is a closed question whether individuals in the original position will reject it.

If there were clear inconsistency, we would apparently know in advance to applying the method that individuals in the original position will reject utilitarianism. That brings us to a second difference: *I do not rely on this commitment.* I look at the premises of the whole original position argument and find a quicker route to its conclusion, but I do not examine its beginnings and anticipate in advance where the whole argument will lead. I am neutral on whether this can be done. But I concede here that this is a strange thought: "We can see that utilitarianism is inconsistent with a background assumption of the original position method and yet when we apply that method, the result is an acceptance of utilitarianism." Is it impossible though? It would be a paradox, I suppose.[2]

Let us move onto a third difference. Here is what Irwin says when we turn to the next page of his commentary, after the material quoted:

[2] Nozick sounds as if he is committed to an analogous paradox: the original position is justified on the basis of the separateness of persons but individuals in it choose an option which is inconsistent with that separateness (1974: 228).

> But if Rawls's argument is sound, the initial judgments about fairness cannot be neutral between different theories of justice, and so we cannot consistently accept both these judgments and utilitarianism. He argues that utilitarian principles would not be chosen in the Original Position. Since the Original Position is a legitimate device only in so far as it incorporates our initial considered judgments about fairness, these initial considered judgments must rule out a utilitarian conception of justice.
>
> The Original Position, therefore, must rest on principles that are strong enough to rule out utilitarianism. Since it is designed so that it rules only what fair initial conditions rule out, judgments that define fair conditions must somewhere be inconsistent with utilitarianism. (2009: 910)

Irwin's thought is that, even if there is no clear inconsistency before applying the method, there must be an inconsistency between what he earlier called the assumptions built into the method and what he here calls the judgments defining fair conditions,[3] because the result is that utilitarianism is rejected. A third difference is this: *I argue that there is an inconsistency between these assumptions and utilitarianism, but I don't argue that there must be an inconsistency if individuals in the original position reject utilitarianism.* There seems to be an inconsistency and it is not that hidden, but I don't proceed backwards from the original position result: I am not committed to being able to infer an inconsistency, hidden or not, from the mere fact of rejection.[4]

[3] I presume these are the propositions one uses in justifying the method, for example the proposition that individuals A and B are separate beings and this separateness should be respected.

[4] Irwin's world is that of moving forwards from any clear inconsistency to knowing what the original position individuals will reject, and also moving backwards from rejection to some inconsistency earlier, albeit perhaps not clear then, whereas mine is that of examining the parts and getting the same result from fewer parts – optimization.

Although Irwin focuses on Rawls versus utilitarianism, other options are also rejected using the original position method and Irwin is presumably committed to something more general:

> (*Backwards-inconsistency commitment*) If individuals in the original position reject a proposed principle, we can infer that there is an inconsistency between that principle and the assumptions built-in to the original position method.

One reason for not accepting this commitment has to do with how individuals in the original position choose from a menu of options that we design and if they reject an option in favour of Rawls's option, it is still possible that they would choose that option on another menu, in which Rawls's option is removed – or at least the fact of rejection alone, using the menu Rawls designed, does not exclude this possibility. Inconsistency, classically conceived, is an either-or matter: either there is an inconsistency between the built-in assumptions or there is not. But what we learn from rejection in favour of another principle, without examining the argument for rejection, is merely "They reject this option because there is a better option on the menu they have been offered." If the original position is an inconsistency-detecting device, helping us to detect more hidden inconsistencies as well,[5] then it seems we should learn something more robust than this: "They would never agree to the rejected option, even if one switches to an impoverished menu. It does not matter what the options are: they would always have some problem with agreeing to that option when it is on the menu, which reflects inconsistency with the assumptions built into the method." The mere fact of rejection using the original position and

[5] If it is such a device, I am assuming it detects any inconsistency.

a certain menu[6] does not allow us to conclude this, so I think it is not an inconsistency-detecting device.

Another reason for rejecting the commitment identified has to do with social science knowledge. Individuals in the original position do not know features of themselves, but they possess a general social science knowledge, which they are to use in selecting between principles (1999: 119). It allows them to favour some principles over others on the grounds that these are more likely to be stably implemented. Rejection may not reflect inconsistency with the built-in assumptions, but rather a preference for what is slightly more likely to be stable over what is slightly less likely given their knowledge. Utilitarianism appears to be logically inconsistent with the separateness of person assumption because it recommends producing happiness regardless of the boundaries of individuals and rights for these boundaries to be respected; but a rejected option may not display any logical inconsistency with a background assumption, such as this one, or a combination of them.

References

Edward, T.R. 2016. Rawls versus utilitarianism: the subset objection. *E-Logos* 23 (2): 37-41.

Irwin, T.H. 2009. *The Development of Ethics: A Historical and Critical Study*. *Volume III: From Kant to Rawls*. Oxford: Oxford University Press.

Kalderon, M.E. 2016. Review of Terence Irwin *The Development of Ethics: A Historical and Critical Study* Vol. 1, *From Socrates to the Reformation*; vol. 2, *From Suarez to Rousseau*; vol. 3, *From Kant to Rawls*. *Ethics* 126 (2): 510-513.

[6] The commitment more fully is that rejection from the original position *using a given menu*, such as Rawls's menu, allows us to infer inconsistency with the built-in assumptions.

Nozick, R. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Rawls, J. 1999 (revised edition). *A Theory of Justice.* Cambridge, Massachusetts: Belknap Press.