The fallibility objection to the original position

Author: Terence Rajivan Edward

Abstract. Do individuals in John Rawls's original position take into account the fallibility of human nature? Some notable commentators on Rawls say that they do or that they should. But this enables us to say that individuals in the original position would not come to an agreement at all.

Different people have different views about what the rules of society should be, at least in the country where I live. Some people think that the rules should preserve tradition; other people think that the rules should give people as much liberty as possible; yet other people have other views. John Rawls thinks the rules should be fair (1971: 3-4). But when are the rules fair?

Rawls has devised a method for evaluating sets of rules, in order to determine when one set is fairer than another. The underlying idea behind Rawls's method is that one set of rules would be fairer if self-interested individuals would prefer that set, so long as each individual does not rely on information about their own specific situation (1971: 136). To illustrate this possibility: a person with a university degree may, if given the opportunity, try to ensure that the rules of society favour people with this qualification, e.g. by recommending the rule that only people who have this qualification are allowed to be members of government. If self-interested individuals come to an agreement about the societal rules

without any individual relying on information about their own specific case, the result would be a fair agreement, Rawls thinks.

On the basis of this idea, Rawls asks us to imagine some self-interested individuals coming together and forming an agreement on the rules of society, but each individual lacks information about their own specific case. Amongst other things, they do not know their occupation, gender, class position, natural endowments, or conception of what a good life would be (1971: 137).

Rawls calls the conditions these individuals are in "the original position." In these conditions, there are some things they do know. They know that they are human beings and they know facts about human nature (1971: 137). They also know that some agreements are harder to keep than others. They are not to make an agreement that, given appropriate circumstances, they would not keep. A further thing that they know is that the agreement they make will be final. Finality is a special condition that Rawls incorporates into the original position. He writes:

> They cannot enter into agreements that may have consequences they cannot accept. They will avoid those that they can adhere to only with great difficulty. Since the original agreement is final and made in perpetuity, there is no second chance. (1971: 176)

However, what Rawls says here leads to a contradiction. Rawls thinks that it is possible for individuals in the conditions he proposes to come to an agreement. But they know facts about human nature. One fact about human nature is that human beings are fallible. There is a chance of making mistakes. In which case, individuals in the original position may be making

a mistake when agreeing on certain rules. In which case, given that any agreement they enter into is final, a consequence of any agreement is that they may have to live under a mistaken agreement, with no opportunity to revise it. Or at least they must be open to this possibility. In which case, because they will not enter into agreements with consequences that they cannot accept, they will not enter into any agreement.

The question of whether individuals in the original position take into account the fallibility of human nature has been addressed before. To my knowledge, Allen Buchanan was the first to explicitly address the question. He proposes that that they will take fallibility into account. He thinks that this proposal blocks certain objections to Rawls (1975: 182-183). He does not notice the argument above. Joseph Raz says that if they do not take human fallibility into account, then the original position will justify a constitution with no room for revision if a mistake is uncovered (1986: 126).[1] He implies that they should take it into account. If humans are indeed fallible, Rawls implies this as well (1971: 137-138). But then they will not form any agreement at all.

References

Buchanan, A. 1975. Revisability and Rational Choice. *Canadian Journal of Philosophy* 5: 179-192.

Rawls, J. 1971. *A Theory of Justice.* Cambridge, Massachusetts: The Belknap Press.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

[1] Raz makes this point in a footnote. I originally tried to provide a more detailed interpretation of his concern about fallibility and eventually arrived at the argument made in this paper.