Prolegomena 21 (1) 2022: 41-58
doi: https://doi.org/10.26362/20220104

Was Rawls a Kantian?

NENAD CEKIĆ
Department of Philosophy, Faculty of Philosophy, University of Belgrade
Čika Ljubina 18-20, 11000 Belgrade
ncekic@f.bg.ac.rs

ORIGINAL SCIENTIFIC ARTICLE – RECEIVED: 14/03/2022 ACCEPTED: 12/05/2022

ABSTRACT: In this article, the author evaluates whether Rawls rightly believed that his theory of justice could be interpreted as Kantian. Rawls' Kantianism is primarily treated as the general ethical foundation of his theory of justice. Providing insight into the debate conducted on Kantian's interpretation in the 1970s and early 1980s, the author explains the first doubts about Rawls' Kantianism, and how Rawls' defenders subsequently responded to them. At the center of the debate was the question of whether Rawls' principles of justice were essentially the fruit of heteronomous rather than autonomous motivation, and whether they could be treated as "categorical imperatives." Noting the significant differences in the conception of moral motivation between Kant and Rawls, the author explains how and why Rawls' Kantianism is essential to the history of moral and political philosophy. Finally, the author suggests that Rawls' Kantianism should be treated more like Kantian inspiration rather than Kantian interpretation in a literal sense.

KEY WORDS: Rawls, justice, Kantian interpretation, autonomy, heteronomy, the categorical imperative.

It is not always entirely clear why precisely John Rawls' theory of justice is nowadays characterized as "Kantian" and, in some contexts, why this label is still being used. Of course, it is common knowledge that Rawls declared himself Kantian in his famous *A Theory of Justice*. The entire section 40 of this book bears the name "Kantian interpretation of justice as fairness." However, it is often forgotten that Rawls' declaration of the Kantian approach came with an explicit warning that his theory "has departed from Kant's views in several respects" (Rawls 1971: 256). Questions about the status of these departures, and how important they are for assessing whether Rawls was really a Kantian, require clarification. Given Rawls' numerous references to Kant's ideas, shedding light

on all the connections between Rawls and Kant would require much more space than what is at our disposal. Therefore, our interest here is deliberately limited to an initial discussion of Rawls' Kantianism.

Furthermore, there is a good historical reason for this approach. Namely, the vivid debate about the normative (Kantian) assumptions of Rawls' political theory was conducted in open form only for slightly more than a decade after the publication of *A Theory of Justice* (1971). Subsequently, the discussion fell almost silent, not so much because of the loss of philosophical interest in Rawls' core ethical arguments, but because Rawls continued his research in the domain of political philosophy. Philosophical interest then naturally shifted to his more recent works. At the same time, *A Theory of Justice* gained the status of a classical philosophical work, and its ethical framework had already widely been considered theoretically established.

Nevertheless, the issues from the debate in the 1970s and early 1980s regarding Rawls' Kantianism still make much sense today. Our focus is on the original fundamental objections intended to show that Rawls' theory did not actually follow a Kantian path in its normative background. We will limit ourselves to clarifying Rawls' use of Kantian ethical concepts of autonomy and the categorical imperative. Before these clarifications, we will present a general conceptual framework within which Rawls had to move when formulating his Kantian interpretation.

1. The roots of Rawls' Kantianism

The philosophers who write about Rawls today barely consider his work on normative ethics and meta-ethics that preceded *A Theory of Justice*, therefore it might be understood that Rawls has always been a Kantian. If we look at his works from the 1950s and 1960s, we will see that he relied on Kant only when he realized that even sophisticated versions of utilitarianism could not cope with the demands imposed by fairness in a broad sense. It has been almost forgotten that his famous article "Two concepts of rules" (Rawls 1955) is part of the history regarding both normative ethics and meta-ethics. In this article, Rawls tries to explain how utilitarian theory could be used for establishing – at the metaethical level – the priority of deontological moral rules. His considerations at the time indirectly touched on the problem of fairness. Namely, during the debate on utilitarianism, the most potent anti-utilitarian arguments were based on cases that showed that the utilitarian point of view produces unfair and counterintuitive consequences, such as allowing

the breaking of promises or the punishing of the innocent. It turns out that the utilitarian logic of maximization contradicts the usual notions of fairness, and does not recognize the separateness of individuals, or a separable notion of duty.

Inside the debate on utilitarianism, Rawls encountered the famous "priority problem." The question is: "Which fundamental moral notion – Good or Right – must have primacy in a moral or political theory?" The question is crucial because the demands imposed by Right, in the form of an obligation or a duty, can very easily be at odds with the consequences of augmenting Good. That is the background of the problems arising from the standard concept of "fairness." Rawls himself has taken the stance that the priority of Right is a central feature of Kant's ethics (Rawls 1971: 31n). Rawls appears to have thought that Kant's moral philosophy offered a frame in which the non-utilitarian "priority of Right" (and freedom alone) could still be philosophically preserved (see especially Rawls 1971: §6, §8).

Rawls situates his conception of justice as fairness in the "contractarian" tradition. For him, this tradition is the best option for avoiding the normative problems brought by competitive theories of utilitarianism and (deontological) "intuitionism." Rawls declares: "My aim is to present a conception of justice which generalizes and carries to a higher level of abstraction the familiar theory of the social contract as found, say, in Locke, Rousseau, and Kant" (Rawls 1971: 11). Rawls is also convinced that his analysis of justice as fairness provides a "procedural rendering" of Kant's key concepts, such as universality, autonomy, and the categorical imperative (Rawls 1971: 264). Kantianism, then, looks to him to be a way out of a fundamental normative problem of grounding a political theory. Indeed, Rawls abandoned "pure" ethics, but his *A Theory of Justice* deals with the problems he had already faced in the previous discussion of utilitarianism during the 1950s and 1960s. To facilitate a better understanding of the discussion, it is helpful to point out that Rawls never has completely abandoned a consequentialist (utilitarian in a broad sense) conception of the nature of general human motivation. That is why Rawls' attempt to balance the egoistic–utilitarian motivation and the Kantian normative framework was the subject of the initial criticism of his proclaimed Kantianism (e.g., Nagel 1973; Levine 1974; Johnson 1974; Wolff 1977). The common thread of initial criticism was the belief that Rawls had radically misinterpreted Kant's theory to fit his views into Kant's moral philosophy framework.

2. Grounding the morality: A note on two methods

We should note that Rawls' account of justice is explicitly an account of human, not abstract, universal justice. Rawls readily admits that his argument depends on contingent assumptions about humans and their particular situations: "moral theory must be free to use contingent assumptions and general facts as it pleases" (Rawls 1971: 44). This claim reflects a crucial methodological difference between Kant and Rawls. Contrary to Rawls' view, Kant insists on the claim that ethics must be "formal," meaning the source of morality can be found only in pure concepts of Reason (*Vernunft*) rather than in anything "material" (empirical).[1] Kant emphatically insisted on the requirement that moral principles can only be established independently of any contingent assumptions about human nature or the human situation in the empirical "world."[2]

In the Kantian view, the only moral principle (which is the "categorical imperative") is openly revealed to us through the a priori concept of duty, which necessarily presents itself as "universal" (a duty that would allow exceptions would not be a duty at all). This universality would imply that the account of morality and justice applies to all possible moral agents (not just humans) and in all possible worlds.

Contrary to Kant's insistence on a claim that the universality required by (the power of) Reason is a moral demand itself, the famous "Kantian interpretation" starts with the assertion that the central notion of Kantian moral theory is the idea of autonomy and not universality (Rawls 1971: 251). This conviction seems odd because it looks like Rawls does not even take into consideration the interpretations of Kant's thought that have been most prominent from the late 1940s (Paton 1947) until today (e.g., O'Neill 2018), which were focused on the importance of universality and the need to distinguish logical universality and empirical generality.

[1] To clarify: "Practical principles are *formal* when they abstract from all subjective ends; but they are *material* when they are grounded on… incentives" (*G* 4:427). The pagination of Kant's works is given in a standard fashion according to the *Academy Edition* (the *Akademie Ausgabe*) of Kant's writings. Abbreviations: G = *Groundwork of the Metaphysics of Morals* (Kant 2002), KpV = *Critique of Practical Reason* (Kant 2015), MM = *Metaphysics of Morals* (Kant 1991).

[2] Kant said: "it is of the utmost necessity to work out once a pure moral philosophy which is fully cleansed of everything that might be in any way empirical and belong to anthropology; for that there must be such is self-evident from the common idea of duty and of moral laws" (G 4:389).

3. Kantian autonomy

Let us assume that Rawls was right, and the concept of autonomy has some primacy over universality or the categorical imperative. One of the difficulties of this approach lies in the fact that even nowadays, there is still room for disagreement regarding the best complete interpretation of Kant's notion of autonomy. However, we will put this difficulty aside because Kant's basic idea is straightforward, and initially does not create a theoretical problem. When Kant speaks of autonomy or its counterpart, heteronomy, his considerations are aimed at the notion of the Will. His exploration shows that two contrasting motives can lead agents to act. If the Will is entirely determined by "incentives" or "inclinations" (based on sensuality), then the agent acts heteronomously. Heteronomy is opposed to (genuine moral) "autonomy", and is an origin of "all ungenuine principles of morality" (G 4:441). Kant sums up the difference between autonomy and heteronomy as follows:

> Autonomy of the will is the property of the will through which it is a law to itself (independently of all properties of the objects of volition). The principle of autonomy is thus: "Not to choose otherwise than so that the maxims of one's choice are at the same time comprehended with it in the same volition as universal law"[3]... [However], If the will seeks that which should determine it *anywhere else* than in the suitability of its maxims for its own universal legislation, hence if it, insofar as it advances beyond itself, seeks the law in the constitution of any of its objects, then *heteronomy* always comes out of this. (G 4:440-441)

For somewhat unclear reasons, different depictions of Rawls' Kantianism circumvent the fact that autonomy, universality, duty, and the categorical imperative in Kant's ethics are inextricably linked. For instance, in acting autonomously, we act not only according to duty, but also "from duty." Those actions, then, are not only "legal" but also "moral" (KpV 5:81).

When Kant says "law" or "pure form of lawfulness," he has in mind a principle that applies with a logical necessity, which is just another name for "universality." On the other hand, empirical *generality* does allow exceptions. Kantian "universalization"[4] is a straightforward and purely formal test: If a proposed principle of action ("maxim") is in any way self-contradictory, it cannot be "universalized" and is *forbidden*.

[3] This is one of the forms of the most known "universality formula" of Kant's categorical imperative. See Section 5 below.

[4] "Universalization" is a technical term often used to interpret Kant's moral philosophy, but it does not belong to Kant's vocabulary.

Autonomy itself is nothing more than acting on principles that can be universal. Thus, Rawls' intentionally dismissing of universality as "slander bases" that lead to "triviality" (Rawls 1971: 251) in favor of supposedly substantial "autonomy", has no support in Kant's original works.

4. Original position and Kantian autonomy

In comparing Rawls' Kantianism and the ethical views of Kant himself, what kind of justice Rawls, using Kantian language, actually had in mind is often forgotten. For Rawls, "the justice is the first virtue of social institutions, as truth is of systems of thought." Therefore, "a theory, however elegant and economical, must be rejected or revised if it is untrue; likewise, laws and institutions, no matter how efficient and well-arranged, must be reformed or abolished if they are unjust" (Rawls 1971: 3). It is clear that Rawls' justice is social justice or, more accurately, distributive justice. Rawls' justice is meant for social institutions; Kant's morality is for individuals.

The process of establishing the principles of social justice is presented by Rawls as a "bargaining game." Rawls starts from the fact that people are naturally biased by their situations. These biases can be eliminated by redefining the initial situation in which fundamental social choices are made. Rawls calls the initial situation of fairness "original position." It is the artificially designed ideal and "appropriate initial status quo" (Rawls 1971: 12).

The critical feature of the original position is "the veil of ignorance." The veil is meant to ensure the impartiality of judgment. Briefly, the veil postulates that individuals reaching an agreement do so in ignorance of most particular facts about themselves, such as their place in society, class position, and overall social status, as well as their natural abilities, the distinctive features of their individual psychology, and their conception of the good. The veil deprives the bargaining parties (individuals) of all knowledge about themselves, each other, and even of their society and history. Still, they are aware of their fundamental interests and general facts about psychology, economics, biology, and other fundamental social and natural sciences. Parties are still aware of the desirability of the "primary social goods" that anyone needs for a good life. These essential goods are rights, liberties, opportunities, income, wealth, and the social basis of self-respect. Individuals "normally prefer more primary goods rather than less" (Rawls 1971: 123). This preference is the only motivation ascribed to the parties. In pursuing the primary goods, the

parties are supposed to be "mutually disinterested." This indifference allegedly coincides with Kant's notion of autonomy (Rawls 1971: 253). When choosing the principles of justice, people in the original position are motivated solely by the desire to protect their own interests, not by benevolence or envy. Then, the preference for primary goods is derived from only the most general assumptions about rationality and human life's general conditions.

What are Rawls' reasons for arguing that the original position is actually a position of individual autonomy? Let us look at precisely what he says:

> Kant held, I believe, that a person is acting autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being. The principles he acts upon are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things he happens to want. To act on such principles is to act heteronomously. Now the veil of ignorance deprives the persons in the original position of the knowledge that would enable them to choose heteronomous principles. The parties arrive at their choice together as free and equal rational persons knowing only that those circumstances obtain which give rise to the need for principles of justice. (Rawls 1971: 252)

We should stop here for a moment.

Principally, it is unclear why Rawls thinks the veil of ignorance fits the genuine Kantian concept of autonomy. Let us remember that Kant defines autonomy as "the property of the will through which it is a law to itself (independent of all properties of the objects of volition)" (G 4:440). The expression "independent of the objects of volition" indicates a possible problem in Rawls' Kantian interpretation. Kant does not think that when deciding, agents are deprived of relevant knowledge. Moreover, in opposition to the situation in the "original position," Kantian autonomy *presupposes* relevant knowledge of persons' "inclinations" and "incentives." In his most famous examples from *Groundwork*, Kant shows that morality is demonstrated when we are free to counteract sensuality's impetus when duty based on Reason dictates.[5] Moreover, within the general practical

[5] Kant's most famous illustration of duty is the resolution of the dilemma of whether it is permissible to obtain economic benefits through a false promise. Providing the awareness of the a priori notion of duty, Reason instructs the agent to move away from the natural tendency to enlarge the estate. (G 4:402-403; 419-20). It is important to add that duties are best observed when they have a clear form of prohibition. Thus, the immorality of making a false promise can easily be presented as a "maxim" in the form of a prohibition: "Under no circumstances, make any false promise." In this case, it is obvious that the agent is not deprived of knowledge about incentives, but freely chooses to act morally.

field, sensual incentives are a necessary condition of morality because some "material goal" is needed for *any* acting. Finally, a heteronomous act does not have to be morally *wrong*; it could be "legal" but worthless; allowed, but without moral value.[6]

Initially, Rawls' interpretations of Kant's key terms had faced harsh criticisms, followed by adequate defenses. A good proportion of initial objections to Kantian interpretation in the 1970s consisted of accusations that Rawls simply misinterpreted Kant's theory of moral motivation related to the concepts of autonomy and, indirectly, the categorical imperative. The questions raised by the early critics of Kantian interpretation are essentially anthropological, and they concern Kant's ultimate question of what it means to be autonomous in pursuing ends. While Rawls centers his interpretation on autonomy, early critics and defenders of Rawls' Kantian interpretation turned toward the explanation of rational agency.

In a 1974 article, Andrew Levine argues that Rawls' Kantian interpretation "rests on a systematic confusion of an anthropological understanding of Kant's notion of rational agency (replete with contingent assumptions about human nature) and Kant's own non-anthropological understanding" (Levine 1974: 48). Analyzing Rawls' "original position," the role of which is to free our choice of fundamental principles of justice from what Kant would call "empirical" or heteronomous inclinations, Levine concludes that the considerations we take into account in the original position are not what Kant would call "pure." He thinks that Rawls tries to combine Hobbesian egoistic rationality with the Kantian concept of universality, leading to incoherence. Then Levine argues that instrumental rationality, the one Rawls is in fact using, is empirical, and therefore heteronomous in Kant's sense. In addition, Levine suggests that to account for autonomy, a different, non-instrumental notion of "reasonableness" must be employed, which would fall in the domain of Kantian Reason. In the Rawlsian original position, "we express our nature as bundles of appetites for primary goods endowed with a capacity for instrumental rationality; not as bearers of pure practical reason" (Levine 1974: 57). This understanding of human nature necessarily invokes heteronomous motivation.

Now, Levine continues, we should recall that the whole burden of Kant's moral philosophy – and the point on which it must ultimately be evaluated – centers on the attempt to conceive motivation for the moral

[6] For a distinction of morality and legality see e. g. KpV 5:72 or 5:81.

life, independent of specifically human (as distinct from reasonable) nature. For the proposed Kantian interpretation to be viable, the empirically pure motivation provided by pure Reason would have to be identical to the motivation arising from the contingent assumptions about human nature presupposed in the original position (Levine 1974: 52).

In his 1974 paper, Oliver Johnson offered a picture similar to Levine's. Johnson, like Levine, notes that individuals under the "veil of ignorance" are still motivated by what Kant would call heteronomous inclinations: "An action originally heteronomous is not rendered autonomous, even though performed under a veil of ignorance if the nature of motivation is unchanged" (Johnson 1974: 62).

Robert Paul Wolff follows the line of Levine's and Johnson's arguments. He believes that it is very unusual to interpret Kant as saying that any goods (or the Good) are the source of moral motivation, but Rawls, willingly or unwillingly, does. On the other hand, Kant has never been vague in this regard: A material end cannot be morally significant. The fact that Rawls' "primary goods" are highly general and not specifically adapted to individual wishes does not change too much. That fact does not affect the nature of the chosen principle of justice:

> [The] veil of ignorance, in fact, only guarantees that the principles will be… generally heteronomous rather than particularly heteronomous. The choice of principles is motivated by self-interest, rather than by the Idea of Good. (Wolff 1977, 115)

To these objections from the 1970s, we could add another one. The confusion of the two kinds of rationality (instrumental and "pure"), first observed by Levine, is not involuntary. Rawls sometimes deliberately toys with Kant's terminology. We have already seen that the original position is, according to Rawls, set up so that the parties reflect human nature as "reasonable and rational." This expression ("reasonable and rational") is a dual representation of Kant's single adjective *vernünftig*, and covers both the "pure" and instrumental use of practical Reason (Richardson 2022). However, moral acting in Kant's thought is not ambiguous in that conflating way. Instrumental rationality that deals with means–end relations can and must be strictly separated from pure moral reasoning. Motivational conflicts between "acting from duty" (autonomously) and "acting from inclinations" (heteronomously) reveal the possibility and necessity of the strict separation of two different uses of practical Reason.

After the first doubts regarding Rawls' Kantianism, came the first reactions to those doubts. Stephen Darwall's response to objections

concerning Rawls' misconception of autonomy is now considered classic. His main argument is that, although the decisions in the original position could be construed as heteronomous, later decisions to follow the principles of justice in ordinary life are autonomous in the Kantian sense:

> It may well be the case that the choice of principles in the original position is a heteronomous choice because it is an interested choice and still be true that the decision of actual rational beings, not in the original position to act under such principles, is an autonomous decision, and hence, that action on such principles is autonomous. Even if it is true that if one were under the constraints of the original position (most importantly, the veil of ignorance), one would want a particular principle adopted in one's own interest, it by no means follows that all, or even any, rational beings as they are actually placed in the world would want that same principle adopted in their interest. Thus, if a rational being chooses to act on principles, which would be acceptable to him if he were under the veil (on the grounds that they would be acceptable to him under the veil), such a choice is by no means a choice on the basis of his interests and thus is not, on those grounds, a heteronomous choice. (Darwall 1976: 166)

Elsewhere, Darwall extends this argument in a rather unusual way:

> The complaint that the parties are assumed to be self-interested is a red herring in any case. Because of the veil of ignorance, the original position is not a perspective of self-interest but rather of an interest in selves or individuals as such. The assumption of self-interested motivation plays no essential role. The same principles would be chosen, and the same arguments for them found convincing, were the parties not assumed to be self-interested, but to be completely other-interested. (Darwall 1980: 340)

How could Darwall's arguments be evaluated? As has already been seen, Levine and Johnson suggest that Rawls' general attitude that the agent can be moral (autonomous) *through* the (heteronomous) pursuit of happiness is wrong. Darwall's defense here really feels more like philosophical gymnastics than a philosophical argument. Specifically, the plausibility of Rawls' theory of justice lies precisely in the fact that self-interest is a starting point that we all understand. However, in the first quote, Darwall says we can be autonomous if we stubbornly and repeatedly adhere to heteronomous principles. According to another quote, it does not matter if the agent is self-interested or other-interested because the results (principles of justice) are the same in both cases. As for the first defense, the autonomous decision to adhere to the heteronomous principles is not even conceivable within the Kantian worldview. In the second argument, it is unclear what the murky other-interested motivations might be, and how they might even become general.

Darwall's defense, however, draws our attention to one crucial issue. As it stands, his observations are implicitly based on one of the most common misinterpretations of Kant's position, the allegation that he has no regard for human happiness. Some interpreters go further, and depict Kant's position as "rigorous" and hostile to happiness. To give Kant's moral theory a "human face," many, including Rawls and Darwall, have struggled to "reconcile" Kant with happiness.

It looks like this kind of reconciliation is not necessary at all. Namely, Kant claims that happiness is a material end that "everyone has."[7] He explicitly says it is "safe" to suppose it as real for all rational beings "in accordance with natural necessity" (G 4:415, 4:430; cf. KpV 5:25). In other words, it is the "natural end." Even the doubters would be reassured after a careful reading of Kant's *Critique of Practical Reason*, in which "virtue" (morality, which is "supreme" but not "complete" Good), and "happiness" converge in the idea of "highest" or "complete" Good (KpV 5:110–111). However, we can only hope for this convergence rationally, but without any possible warranty. Whether moral acting is rewarded with "happiness" depends on the transcendent ideas of God and the soul's immortality. Though "regulative," they are outside the realm of possible knowledge.

Finally, sometimes it is overlooked that Kant shows a very profound respect for the pursuit of happiness, but in a non-Rawlsian way. In *Metaphysics of Morals*, he asks, "What are the ends that are also duties?" He replies, "They are *one's own perfection* and *the happiness of others*. Perfection and happiness cannot be interchanged here" (MM 6:385). In a moral sense, one should be interested only in someone else's happiness, but not one's own. The Rawlsian concept of a self-interested individual as the center of morality is out of the question. From Kant's remarks, it is easy to conclude why autonomy as acting "from duty" can have nothing to do with "mutual disinterest," derived from the veil of ignorance. Briefly, other people's happiness is a legitimate moral goal not because of any empirical incentive, but because it can be constituted as a duty. Following previous observations that suggest that duties are best seen in their negative form, we could say that it is forbidden *never* to do anything to improve one's own perfection and someone else's happiness.

[7] Kant's definition of happiness is nicely formulated: "Happiness is the state of a rational being in the world in the whole of whose existence everything goes according to his wish and Will" (KpV: 5.127).

5. Principles of justice as "categorical imperatives"

Another critical point that the first critics of Rawls' Kantian interpretation have focused on is the notion of the categorical imperative. This attention was initially sparked by Rawls' explicitly stated belief that his principles of justice were "categorical imperatives":

> The principles of justice are also categorical imperatives in Kant's sense. For by a categorical imperative Kant understands a principle of conduct that applies to a person in virtue of his nature as a free and equal rational being. The validity of the principles does not presuppose that one has a particular desire or aim. Whereas a hypothetical imperative by contrast does assume this: it directs us to take certain steps as effective means to achieve a specific end… Its applicability depends upon one's having an aim which one need not have as a condition of being a rational human individual. The argument for the two principles of justice does not assume that the parties have particular ends, but only that they desire certain primary goods…To act from the principles of justice is to act from categorical imperatives in the sense that they apply to us whatever in particular our aims are. (Rawls 1971: 253)

To even understand exactly what Rawls is trying to say here, we must first clarify Kant's basic idea of the categorical imperative as moral law. His language is notoriously complicated, leading to obscurity, but the primary idea is simple: The notion of categorical imperatives can be grasped only in contrast to the so-called "hypothetical imperative":

> If the action were good merely as a means to *something else*, then the imperative is *hypothetical*; if it is represented as good *in itself*, hence necessary, as the principle of the Will, in a Will that in itself accords with reason, then it is *categorical*. (G 4:414)

> Hypothetical imperative says, "I ought to do something *because I will something else*." By contrast, the moral, hence categorical, imperative says: "I ought to act thus-and-so even if I did not will anything else." That is, the former one says: "I ought not to lie, if I want to retain my honorable reputation," but the latter says: "I ought not to lie, even if I did not incur the least disgrace." (G 4:441)

Hypothetical imperatives are not moral judgments; they are "analytic" and "conditional" propositions about means–end relations. On the other hand, the concept of the categorical imperative (or *unconditional command*) is directly connected with Kant's conception of autonomy. Namely, as we have already seen, in autonomous acting, the agent *gives itself* the moral law regardless of any "object." Here, Reason "presupposes only itself, because a rule is objectively and universally valid only when it holds without the contingent, subjective conditions that distinguish one rational being from another" (KpV 5:21). The principle of Will's immediate (with no "incentives") self-determination, therefore, must

be a *categorical* imperative (G 4:440), and it alone is worthy of the title "imperative of morality" (G 4:416).

Now let us take a look at how Kant sees the operation of the categorical imperative, and then we will revisit Rawls' view of that concept. Above all, Kant's ethics is referred to as "universalistic" with good reason. It is no coincidence that the most famous wording of the categorical imperative is "Formula of the Universal Law." Without this, it is impossible to interpret Kant's ethics. What is Kant's basic idea here?

First, the categorical imperative is a *single* moral criterion, and can only be one because Reason's demand is one and only – universality (logical prohibition of exceptions). The formula of the Universal Law explicitly expresses this demand: "Act only in accordance with that maxim through which you can at the same time will that it become a *universal* law" (G 4:421; cf. G 4:402). In the minimal technical context, we must not forget that the categorical imperative is a "synthetic judgment a priori," just as are judgments like "Every consequence has a cause" that belong to pure natural science.

Second, it should not be forgotten that the way the categorical imperative works is best seen when it has a form of prohibition. We can clarify this claim based on Kant's key notion of duty that is best observed when working *via negativa*. Namely, the human Will does not *by its nature* fulfill Reason's order to respect the very "pure form of lawfulness" (the form of *universality*)[8] because the agents are inclined to make exceptions in their own favor. So, Will must, as a subjective (fallible) principle of volition ("maxim"), conform to the objective law of morality by the following prohibition: "I ought never to act except in such a way that I could also will that my maxim should become a universal law" (G 4:402).

This negative formulation of the categorical imperative is rarely quoted, but it clearly illustrates Kant's general line of reasoning. Reason prohibits the adoption of subjective principles (maxims) that cannot be universal. This prohibition stops heteronomous maxims because they are based *solely* on empirical inclinations, which paves the way for the notion of morally destructive exceptions. By logically forbidding morally wrong maxims, Reason allows morally right ones. Some of them lead to the previously mentioned "happiness of others" as an end that is itself a duty.

[8] An imagined rational being whose Will has no "subjective conditions," for humans these are sensibility or natural inclinations, has no *duty* to fulfill the moral law because their God-like Will is purely determined by the law (rationality itself). Those beings have a "holy Will" (KpV 5:32).

Bearing in mind this technical explanation, what can be said about Rawls' assertion that the principles of justice in the original position are also categorical imperatives? The attack on this assertion was based on the same arguments challenging the view that individuals in Rawls' original position are autonomous. Let us remember, this objection emphasizes the confusion of two senses of rationality: instrumental and pure. Thus, Levine agrees with Rawls that a categorical imperative is an expression of a person's nature "as a free and equal rational being."

> However, he then adds that this freedom and rationality are transcendental and unconditioned by any merely contingent end, no matter how universally entertained. This is the sense in which the categorical imperative commands "categorically." On Rawls' account, however, the desire for primary goods is part of being rational. In that sense of "rationality," principles of conduct that apply to us in virtue of our "nature as free and equal rational beings" are conditioned by merely contingent ends; namely, the set of primary goods. It is only by confusing these two quite distinct senses of "rationality" that Rawls can go on to conclude that these principles command categorically. (Levine 1974: 55)

Let us take another look at Rawls' use of the phrase "categorical imperatives" (plural) for *two* principles of justice. His contention that the desire for primary goods extends to all human beings, whatever their particular conceptions of the Good, does not really modify their hypothetical character. At best, it is a contingent fact that all human beings have some common aims. Thus, claiming that "wanting the primary goods is part of being rational" entirely opposes Kant's conception of rational agency. For the principles of justice to be categorical imperatives, they would have to determine Will independently of anything that does not come from Reason. Rawls' principles of justice do not and cannot do it. As already noticed, imperatives derived from a desire for "primary goods" would be hypothetical in Kant's sense, not categorical. Despite their indefiniteness, they are still empirical "incentives" (see Nagel 1973: 223 n3; cf. G 4:415–416). Rawls' principles of justice, as a categorical imperative(s), resemble Kant's "counsels of prudence" that are a specific kind of hypothetical imperative.

Now let us recall Rawls' claim that the motivational assumption of *mutual disinterest* in the original position accords with Kant's autonomy (Rawls 1971: 253). This assertion seems very strange when viewed in the light of another formulation of the categorical imperative – the famous "Formula of Humanity": "Act so that you use humanity, as much in your

own person as in the person of every other, always at the same time as end and never *merely*[9] as means" (G 4:429; cf. G 4:436).

The very wording of this "formula" casts into doubt Rawls' idea of an individual's absolute "mutual disinterest" as a part of Kantian morality. The empirical "mutual disinterest" of individuals in the original position is the consequence of the veil of ignorance. As no one knows anything about others, no one can be "interested in others" in the sensory and empirical sense. This disinterest is consistent with Kant's views only in relation to one point: His ethical theory does imply that moral decisions exclude empirical interest. However, Rawls again conflates two sources of rationality and motivation–instrumental and pure Reason. Kantian empirically disinterested individuals as "noumenal selves" and "end-in-themselves" are still fundamentally, transcendentally interested in one other's essential feature, namely, Reason, which is a priori familiar to all of them.

Rawls continues to read Kant in a very exotic way. His view on "treating other as means" in the Humanity Formula is oddly rigid compared to Kant's understanding. For unclear reasons, he thinks that principles of justice give an even more vital and characteristic interpretation of Kant's intentions: "They rule out even the tendency to regard men as means to one another's welfare. In the design of the social system, we must treat persons solely as ends and *not in any way* as means" (Rawls 1971: 183).

Why does Rawls' understanding of "treating humanity as means" contradict Kant's Formula of Humanity? We have to remember that Kant, in this formula, does not forbid every "use" of another, but merely requires that in those cases, the latter will not be treated *only* as means. This kind of conviction is quite close to common sense: contracts, trade, friendship, love, and so on. These are activities in which people "use" each other, and the largest part of our lives consists of such activities. Kant forbids extortion and violence, but not "using other" with prior consent. Rawls' particular formulation is chiefly understood as mistaken in classical Kantian literature. A well-known interpreter of Kant's work, Allen Wood, warns us:

> It is possible to treat persons as end in themselves and also as means, as long as you respect their… dignity. This is not only possible, but Kantian ethics positively enjoins it. (...) In realm of ends, every rational being would… be treated as end in itself and at the same time as a means to the system of shared ends. (Wood 2008: 87)

[9] Emphasis added.

6. Kantian interpretation or Kantian inspiration?

After reading the literature on Rawls' Kantianism, one can easily feel that many philosophers asked whether Rawls could have been more Kantian than he was. To answer this question, we must check if Kant and Rawls' theories fit the same normative approach within moral philosophy. Regarding that, Ronald Dworkin notes that any "deep" political theory must be based on the goal, rights, or duty. In duty-based theories, unlike the consequentialist and rights-based theories, individual actions and decisions are viewed as fundamentally significant. On the other hand, rights-based theories are more interested in personal independence and protecting the value of individual opinion and choice. Rights- and duty-based theories conceive of moral rules or laws as independent of selfish interests. However, a difference exists. The duty-based theories consider these rules the essence. Theories based on rights consider moral laws instrumental (Dworkin 1977: 169–176).

Kant's ethics is emphatically "ethics of duty." Kant's famous saying provides the best illustration: "The majesty of duty has nothing to do with the enjoyment of life; it has its own law and also its own court" (KpV 5:89). However, Kant presents the notion of duty as "popular," available to everybody with common sense. It contains the essential quality of *any* (not just moral) law: being absolute, universal, and necessary (G 4:389). On the other hand, Rawls's political theory is based on the notion of rights. "Rights-based" theories treat morality as instrumental: "The man who is in the center of rights-based theories is a man who benefits from someone else's respect for the law, not a man who leads a life of virtue by himself by respecting the law" (Dworkin 1977: 172).

Thus, the answer to the original question "Could Rawls have been more Kantian?" is "No, he could not because the normative framework he chose would not allow it." However, this assessment does not necessarily mean Rawls "made a mistake" or "abused Kant." It seems now that he simply took concepts from Kant's theory that fit his basic idea and then adapted them. After all, Rawls did not hide his departures from Kant; he announced them. Therefore, the real question for the end of our analysis is not could Rawls have been more Kantian, but whether there a valuable heritage emerges from Rawls' Kantian interpretation.

In response to these questions, we can only say that Rawls' Kantian interpretation probably provided a lasting philosophical life of some of Kant's ideas for at least two reasons.

First, Rawls' idea of the inviolability of individuals, on which he grounds the concept of "the priority of the Right," certainly has a Kantian flavor. This idea found a place in contemporary ethics and political philosophy as the idea of "deontological constraints." Of course, Kant and Rawls' notions of inviolability are based on diverse approaches. Kant's is based on duty, Rawls' on the term of rights. Still, there is almost no doubt that Kant would agree with the following widely quoted remark by Rawls:

> Each person possesses an inviolability founded on justice that even the welfare of society cannot override. For this reason, justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many… The rights secured by justice are not subject to political bargaining or to the calculus of social interests. (Rawls 1971: 2–3)

Second, Rawls' attempt to reconcile a concept of naturally self-centered "material" motivation with Kantian "formalism" is also significant. Specifically, even some philosophers who jointly challenge Kant's philosophy and Rawls' Kantianism give Rawls credit for trying somehow to fill Kant's abstract and formal "skeleton" with theoretical "flesh". Rawls' Kantian interpretation could be taken as a philosophical tactic to supply Kant's theoretical position with tangible "content", and not as mere interpretation (see Wolf 1977: 115–116). This belief is still widespread among philosophers.

What can we say regarding the debate on Rawls' Kantianism from today's perspective? We would probably conclude there is "something in it after all," and the debate was not futile. Let us end by playing with words a bit, by changing "Rawls' Kantian interpretation" to "Rawls' Kantian inspiration." Are we wrong? One thing is for sure: By changing Rawls' "Kantian interpretation" to his "Kantian inspiration," we would not lose anything. Moreover, some things might be more apparent to future generations of readers of *A Theory*.

Acknowledgments

This article is partially based on the presentation "Rawls, Nozick and Kantian conceptions of personality" at the conference "*A Theory of Justice*: Fifty Years After" held on November 16th and 17th 2021 in Zagreb (Croatia).

References

Darwall, S. 1976. "A defense of the Kantian interpretation", *Ethics* 86(2), 164–170.

Darwall, S. 1980. "Is there a Kantian foundation of Rawlsian justice", in: G. Blocker and E. H. Smith (eds.), *John Rawls' Theory of Justice* (Athens: Ohio University Press), 311–345.

Dworkin, R. 1977. *Taking Rights Seriously* (Cambridge Mass.: Harvard University Press).

Johnson, O. 1974. "The Kantian interpretation", *Ethics* 85(1), 58–66.

Kant, I. 1991. *Metaphysics of Morals*, transl. by M. Gregor 1991 (Cambridge: Cambridge University Press).

Kant, I. 2002. *Groundwork of the Metaphysics of Morals*, transl. by A. W. Wood (New Haven and London: Yale university Press).

Kant, I. 2015. *Critique of Practical Reason*, transl. by M. Gregor (Cambridge: Cambridge University Press).

Levine, A. 1974. "Rawls' Kantianism", *Social Theory and Practice* 3(1), 47–63.

Nagel, T. 1973. "Rawls on justice", *The Philosophical Review* 82(2), 220–234.

O'Neill, O. 2018. *From Principles to Practice* (Cambridge: Cambridge University Press).

Paton, H. J. 1947. *Categorical Imperative* (Philadelphia: University of Pennsylvania Press).

Rawls, J. 1955. "Two concepts of rules", *The Philosophical Review* 64(1), 3–32.

Rawls, J. 1971. *A Theory of Justice* (Cambridge Mass.: Harvard University Press).

Richardson, H. S. 2022. "John Rawls (1921-2002)", *Internet Encyclopedia of Philosophy*, https://iep.utm.edu/rawls [accessed 15 January 2022]

Wolff, R. P. 1977. *Understanding Rawls* (Cambridge: Cambridge University Press).

Wood, A. W. 2008. *Kantian Ethics* (Cambridge: Cambridge University Press).