PRINCIPIA **28**(1): 175–185 (2024) doi: 10.5007/1808-1711.2024.e96739
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

ILLOCUTIONARY LOGIC AS A TOOL FOR RECONSTRUCTING KANT'S DERIVATION OF THE FORMULA OF THE CATEGORICAL IMPERATIVE FROM ITS MERE CONCEPT

DIRK GREIMANN
Universidade Federal Fluminense, BRASIL
dirk.greimann@gmail.com

Abstract. This paper aims to reconstruct Kant's derivation of the formula of the categorical imperative from its mere concept with the help of the resources of Searle's and Vanderveken's illocutionary logic. The main exegetical hypothesis is that the derivation envisaged by Kant consists in deriving the formula from the success conditions of categorical imperatives. These conditions, which are analogous to the success conditions of ordinary orders, contain restrictions for the successful construction of a system of moral laws that determine what the content of the categorical imperative must be.

Keywords: Kant • categorical imperative • illocutionary logic • self-defeating speech act

RECEIVED: 30/12/2022 REVISED: 21/01/2023 ACCEPTED: 02/05/2023

Introduction

The fundamental step in the construction of Kant's ethics is the "identification" (*Aufsuchung*) of the supreme principle of morality. To carry it out, Kant uses three different procedures: first, the Socratic method of making the standard of morality explicit that implicitly underlies our ethical judgments; second, the derivation of the formula of the categorical imperative from its mere concept; and third, the regressive method of analyzing the validity conditions of categorical norms. In what follows, my aim is to reconstruct the second method. The characteristic of the reconstruction proposed here in comparison with the approaches available in the literature is the use of speech-act theoretical means.[1] It is argued that the notions of illocutionary logic, and in particular the notion of a "self-defeating" speech act, can be used to reconstruct Kant's derivation in a plausible way. The paper has four parts. In part 1, the problem of identifying the supreme principle of morality is briefly recapitulated. In part 2, the components of the concept of moral imperative that Kant makes use of in the derivation are explained. In part 3, the derivation is reconstructed in detail. Finally, in part 4, the question whether the derivation is correct is briefly addressed.

1. The Problem of Identifying the Categorical Imperative

To develop the framework for our reconstruction, we must first clarify what Kant understands by a "foundation" (*Grundlegung*) of the metaphysics of morals. The key to this lies in Kant's classification of the ethical disciplines. It is based on the distinction between "natural science" (*Naturlehre*) and "moral science" (*Sittenlehre*) and this in turn on the distinction between is and ought.[2] Kant defines natural science as the science of the laws according to which everything happens and moral science as the science of the laws according to which everything ought to happen. Both natural and moral science divide into an empirical and an a priori part. Kant calls the a priori part of natural science the "metaphysics of nature" and the a priori part of morals the "metaphysics of morals" or "pure moral philosophy". The metaphysics of nature is the science of the a priori principles of nature. These principles differ from the empirical ones mainly in two ways: they are universal and necessary. Analogously, the metaphysics of morals is the science of the a priori moral principles, which – in a derivative sense – are also characterized by their universality and necessity. Such principles are universal in the sense that they apply not only to a special group of people, but to all people, and not just to humans, but to all possible rational beings. If a person x in the situation S is required to fulfill a universally valid norm N, then the same also applies to every other person y who is in the same situation. A norm is an "unconditional" or "necessary" or "categorical" norm if it is obligatory under all circumstances, regardless of our interests.[3]

"Conditional" or "hypothetical" norms, on the other hand, are only obligatory if there is a corresponding interest. The norm to save for old age in youth, for example, is conditional because, first, it applies only to persons who have an interest in having savings in old age, and second, its normative force can be removed by giving up this interest. Conditional norms do not bind in virtue of a moral obligation, but in virtue of end-means relationships. For example, a person who does not save for old age when s/he is young is acting unwise because s/he is violating her/his own interests, but s/he is not acting immorally because s/he is not violating a moral obligation. Kant identifies proper ethics with the a priori part of moral science, that is, with the metaphysics of morals understood as the science of the universally and unconditionally mandatory norms of morality. He understands the empirical part of moral science as a mere doctrine of prudence, which does not inform us about what our duty is, but only gives "advice for the purpose of our desires".[4]

By the "supreme principle of morality" Kant understands a moral norm that is superior to all others, both in the normative and in the logical hierarchy of norms. That norm A is superior to norm B in the *normative* hierarchy means that A ought to be followed even if B would be violated as a result. A supreme norm in this hierarchy is characterized by the normative property that it overrides the binding force

of all norms with which it conflicts. That norm A is superior to norm B in the *logical* hierarchy means that B can be derived from A. The supreme principle of this logical hierarchy is characterized by the logical property that all other norms of the hierarchy can be derived from it.

The "identification" (*Aufsuchung*) of the supreme principle of morality is the first step in Kant's foundation (*Grundlegung*) of the metaphysics of morals.[5] To characterize this step more closely, we must determine how the foundation of the metaphysics of morals relates to the metaphysics of morals itself. At first glance, the foundation appears to be a normative discipline forming part of the metaphysics of morals.[6] It contains many normative statements such as the categorical imperative. However, this conflicts with the fact that Kant considers the foundation as a "business that is complete in its purpose and to be separated from every other moral investigation" which, in the order of the ethical disciplines, precedes the metaphysics of morals themselves.[7] The "synthetic use of pure practical reason", which characterizes normative ethics, is dispensed with in the foundation, if only because such a use would first have to be legitimized by a critique of pure practical reason. The assertion that the categorical imperative is actually valid can therefore only be made after the foundation has been laid. On this point, Kant is quite clear: whether the categorical imperative actually "takes place" is left open in the context of its identification.[8]

To take this into account, we must understand the foundation as a *metaethical* discipline whose subject is the metaphysics of morals. The question of the foundation is not what we ought to do, but what the principles of a metaphysics of morals would be, if such a science exists at all. Accordingly, the question of the "supreme principle of morality" is to be understood as a metaethical question that can be formulated as follows: If there is a system of universally and absolutely obliging norms – what would be the supreme principle of these norms? Or, in Kant's terminology: if morality "is something real, and not just a chimerical idea without truth" – what principle of morality must then be conceded?[9] In what follows, I shall assume that it is this metaethical question that Kant seeks to answer by means of the identification of the supreme principle of morality.[10]

2. Kant's Notion of a Moral Imperative

The derivation of the formula of the categorical imperative from its mere concept is intended to show that this question can be answered by means of a mere conceptual analysis. In order to reconstruct this derivation, we must first make the components of the concept of a moral imperative explicit that Kant implicitly uses in the derivation. In addition to the formal characteristics of generality and unconditionality, this also includes the conception of moral laws as "laws of freedom".[11] To put this aspect in

perspective, it is helpful to place Kant's conception of the moral imperatives in the grid of the following three basic views: ethical positivism, subjectivism, and objectivism.[12]

Ethical positivism conceives of ethical norms as "positive" norms, that is, as norms that are constituted by our social interactions. If, for example, the employer orders his subordinate: "Leave the room!", then the employee ought to leave the room because the employer successfully commands him or her to do so. But not all positive norms are constituted by such speech acts. For example, the rule that men ought to wear black shoes after 6 p.m. is a positive norm that is based on well-established social conventions and ultimately on the expectations of a social group.

The main strands of ethical positivism are ethical conventionalism and the non-cognitivism advocated by Hare in his early phase, according to which the answer to the basic ethical question, "What should I do?" is to be given by issuing a command, not by making an assertion.[13] Because of his principle of autonomy, also Kant considers the ethical norms as positive norms; I will explain this further below.

Hypothetical norms, such as the rule that one should save for old age when one is young, are not constituted by social interactions but by interests and ends-means relationships. The notion of ought that is used here is not the positive one, but the hypothetical one. In this case, the fact that x ought to follow the rule R does not mean that x has received the order from a person authorized to issue instructions to follow R, or that x is socially expected to follow R, but that x is required to do so in view of his interests.

Ethical subjectivism is the doctrine that the moral norms are to be construed as hypothetical ones. According to Kant, this conception is based on a category mistake. In his view, moral ought and subjective desire are two different categories that are in principle independent of one another. He justifies this distinction with the linguistic insight that already the common use of language distinguishes between good (*gut*) and evil (*böse*) on the one hand, and well-being (*Wohl*) and woe (*Wehe*), on the other hand, so that "so that there are two quite different judgments according to whether in an action we take into consideration its good and evil or our well-being and woe (bad)" (Kant [1788], p.80-81).

Finally, the norm that one ought to treat each person as an end in itself is an interest-independent, categorical norm. Since such norms have the character of absolutely valid laws, it would seem natural to interpret them platonistically, as part of the objective (though non-empirical) world. According to this realistic view, the world itself is not morally neutral, but it rather has a normative structure consisting of moral facts that constitute categorical norms. This realism is not shared by Kant. His understanding of ethical norms is rather guided by the *constructivist* view that moral persons themselves write the laws to which they are subject.[14] The guiding principle for this conception is the principle of *autonomy*, according to which moral norms are constituted by the self-legislation of autonomously acting subjects. Moral norms are

therefore understood by him as "laws of freedom", that is, as positive norms that are posited by acts of self-legislation. A pure recipient of orders, who is not also a legislator at the same time, but is passively subject to a ready-made system of norms or to the authority of an absolute ruler, would not be a moral subject in Kant's sense due to the lack of autonomy.

If one understands *moral objectivism* as the view that moral norms are to be conceived as categorical norms that do not depend on subjective interests, then the brand of moral objectivism defended by Kant is characterized by the positivist claim that moral norms are "normative posits". According to this hybrid conception, categorical norms are constituted by the act of self-legislation, just as assertions are constituted by the act of making an assertion.

3. The Derivation of the Formula of the Categorical Imperative

In the first section of the *Groundwork*, Kant identifies the supreme principle of morality by means of an explication of the criteria (*Richtmass*) for morality that are implicitly applied in the ethical judgments of common sense. This Socratic method is based on the assumption that knowledge of the supreme principle of morality is already contained in everyday ethical knowledge.[15] The criteria given by Kant are: 1. the usefulness or futility of an action can neither add to nor diminish its moral worth; 2. an action has moral value only if it is not done out of inclination but out of duty; 3. an act of duty has its moral value not in the effect that the act is intended to achieve, but in the maxim it follows; 4. whether it is permissible to follow a maxim does not depend on its matter, but only on its formal properties; 5. it is only permissible to follow a maxim if one can will that the maxim become a general law, i.e., if one can will that all other persons also follow the maxim.

Because of Kant's rationalistic conviction that "all moral concepts have their seat and origin completely a priori in reason" ([1785], p.51), this empirical procedure is unsatisfactory in his view, however. A "pure moral philosophy" must draw its concepts and laws from pure reason. He does justice to this requirement in the second section of the *Groundwork* by identifying the supreme principle of morality by means of an analysis of the concept of moral law. The starting point here is the question "whether the mere concept of a categorical imperative may perhaps also furnish its formula, which contains the proposition that alone can be a categorical imperative" ([1785], p. 69). Kant affirms this, and he sketches the following derivation of the formula of the categorical imperative from his concept:

> When I think of a *hypothetical* imperative as such I do not know in advance what it will contain, until I am given the condition. But when I think of a

> *categorical* imperative I know at once what it contains. For since besides the law the imperative contains only the necessity of the maxim to conform with this law, nothing is left but the universality of a law as such, with which the maxim of the action ought to conform, and it is this conformity alone that the imperative actually represents as necessary. There is therefore only a single categorical imperative and it is this: *act only according to that maxim through which you can at the same time will that it becomes a universal law*. (Kant [1785], p.69-71)

Kant's thesis here is that a moral restriction for action can already be derived from the mere concept of a categorical imperative, namely the prohibition to follow a maxim if one cannot want the maxim to be generally followed.[16] To reconstruct this derivation, we must show that from the premise

(1) There is a rule R such that we are categorically obliged to follow R,

we can draw the conclusion

(2) R is the second-order rule not to follow a rule R' if it is not possible that one can will that all persons follow R',

where R is a "maxim" in Kant's sense. Since we are categorically obliged to follow a rule R if and only if we are under no circumstances permitted not to follow the rule R, the categorical imperative can also be formulated as follows:

(CI) It is morally permitted to follow a rule R if and only if it is possible that one can will that all persons follow R.[17]

To derive (CI) from the mere notion of the categorical imperative, two fundamentally different approaches come into consideration. The first is to consider (CI) as an analytic definition of the concept of permission in terms of the notions of possibility and intention. Although this approach achieves its goal, it is exegetically unsatisfactory because it involves a reduction of the moral notion of permission to the non-moral notions of possibility and intention that Kant certainly would reject. In his view, the moral notion of permission is closely connected to the notions of good and evil, and this does not also apply to the non-moral notions.

The second method avoids this difficulty. It is based on the idea of deriving the restriction formulated in (CI) from the conditions of success for the construction of a moral legislation (a coherent system of moral laws). To this end, we can take illocutionary logic as a framework for the reconstruction of Kant's derivation. Let me explain. The successful performance of linguistic acts is linked to certain conditions, which are called "success conditions" in speech act theory. Since every speech act

(such as asserting, asking, commanding, etc.) is associated with a set of success conditions that uniquely characterize it, we can define these acts in terms of their success conditions.

In the case of orders and similar speech acts, the success conditions include certain restrictions on the propositional content of these speech acts. Thus, an order is successful only if its propositional content refers to the future. The order "Clean your shoes yesterday!" fails because its propositional content refers to the past. This success condition is obviously not of an empirical kind, but it already follows from the mere notion of giving an order.

Analogously, the moral legislation of a person is successful only if it meets certain conditions that derive from the notion of a categorical imperative. By definition, a categorical imperative is universally valid. This feature implies:

(3) If a person x, considered as a moral legislator, permits her/himself to follow rule a R, then x also permits any other person to follow R.

In other words, it is analytically true that a morally acting person would not allow her/himself to follow a rule R if s/he did not allow all other persons to follow R as well. Moreover, by definition, a categorical imperative is a "law of freedom". A free person is the author of the categorical imperatives s/he must obey. This feature implies:

(4) It is permitted to a person x to follow a rule R if and only if x, considered as a moral legislator, permits to follow R.

From (3) and (4) we can already derive the following variant of *quod tibi fieri non vis, alteri ne feceris*:[18]

(5) Act in such a way that you allow yourself only what you, as a legislator, would also allow anyone else to do.

However, this imperative is weaker than Kant's categorical imperative (CI). The prohibition of a false promise, for example, cannot be derived from it. Someone could, in accordance with (5), allow her/himself to make a false promise and accept that everyone else is also allowed to do this. The problem is that we cannot derive from (5) the following stronger conclusion:

(6) If a person x, considered as a moral legislator, allows her/himself to follow a rule R, then it is possible that x may want that all other persons also follow R.

To fill this gap, we have to account for Kant's claim that a person acting according to moral laws does not allow her/himself to follow a rule R if it is not possible that s/he may want all other persons also follow R. As can be seen from Kant's explanatory

examples of the categorical imperative, this claim is closely related to his concept of a "self-destructive" law (*sich selbst zerstörendes oder vernichtendes Gesetz*).[19] It can be explicated in terms of the concept of a *self-defeating* speech act, which is defined in illocutionary logic as follows: a speech act is self-defeating if its success conditions are inconsistent.[20] Thus, the order "I order you to disobey any order" is a self-defeating speech act because it is inconsistent to order to disobey any order. It presupposes two intentions that contradict each other: first, the intention to get the hearer to disobey any order, and second, the intention to get the hearer to obey this order. If the speaker does not have the second intention, the order fails because it is not sincere.[21] Similarly, "I promise you not to keep any promise" is a self-defeating promise because, to keep it, it must not be kept. This speech act also presupposes two conflicting intentions: first, the intention not to keep any promise, and second, the intention to keep this promise. The promise fails if the speaker does not have the second intention, because in this case the promise is not sincere. In a broader sense, the assertion "I assert that snow is black, but I don't believe this" ("Moore's Paradox") also defeats itself because the assertion that snow is sincere only if the speaker believes that snow is black.[22]

Kant's concept of a "self-destructive law" can be analogously understood as the notion of a law whose positing has inconsistent success conditions. It is illustrated by Kant's example of the false promise, in which a person x allows her/himself to follow the rule of getting out of a predicament by making a false promise.[23] This permission presupposes two incompatible intentions, too: first, the intention to get out of a predicament by making a false premise, and second, the intention that everyone else should be allowed to do the same. It is, however, impossible to make a false promise when everyone else is allowed to do the same. The problem is that, in this case, it is impossible to make the hearer believe that the promise is sincere:

> For the universality of a law that everyone, once he believes himself to be in need, could promise whatever he fancies with the intention not to keep it, would make the promise and the end one may pursue with itself impossible, as no one would believe he was being promised anything, but would laugh about any such utterance, as a vein pretence. (Kant [1785], p.73)

It is consequently a success condition of the permission that x grants to her/himself that other persons do not grant themselves the same permission. On the other hand, due to the universal validity of a moral legislation, a moral permission is successful only if it is valid for all persons: if x allows her/himself to follow R, x must allow that all other persons also follow R. Hence, the permission that x grants to her/himself is self-defeating; it has contradictory success conditions. From this we can finally derive (6) and hence also the categorical imperative (CI).

4. Is Kant's Derivation Correct?

We have seen that the fundamental rule *quod tibi fieri non vis, alteri ne feceris* can actually be derived from the categorical imperative; the sentence "If a person x acts according to moral laws, then x only allows her/himself what x would also allow all other persons" is a conceptual truth. Whether this also applies to the stronger imperative (CI) set up by Kant is questionable. He claims that (6) is not only a necessary condition for acting morally, but also a sufficient one. According to (CI), it is morally permitted to follow a rule R *if and only if* it is possible that one can will that all persons follow R. But (6) implies only the following weaker version of the categorical imperative:

(CI′) It is morally permitted to follow a rule R *only if* it is possible that one can will that all persons follow R.

Moreover, Kant assumes that the permission to follow a rule R is self-defeating if it is impossible that all persons make use of this permission. This seems to be wrong. The permission to make a false promise in a certain situation fails only when a larger number of people *actually* make use of it. As long as this is not the case, there is no reason to doubt the sincerity of such promises. No one would laugh about any such utterance, as a vein pretence. As far as I can see, Kant offers no solution to either of these problems.

References

Allison, H. 1991. On a Presumed Gap in the Derivation of the Categorial Imperative. *Philosophical Topics* **19**: 1–15.

Beck, L. W. 1960. *A Commentary on Kant's Critique of Practical Reason*. Chicago: Chicago University Press.

Bittner, R. 1993. Das Unternehmen einer Grundlegung zur Metaphysik der Sitten. In: O. Höffe (ed.), *Grundlegung zur Metaphysik der Sitten. Ein kooperativer Kommentar*, 13–30. Frankfurt am Main: Klostermann.

Greimann, D. 2004. Ist die Ethik Kants ontologisch unschuldig? *Kant-Studien* **95**: 107–127.

Greimann, D. 2003. Kants Ableitung der Formel des Kategorischen Imperativs aus seinem blossen Begriff. In: U. Meixner & A. Neven (ed.), *Philosophiegeschichte und logische Analyse*, vol. 6, Geschichte der Ethik, 97-111, Paderborn: Mentis.

Habermas, J. 1999. *Wahrheit und Rechtfertigung*. Frankfurt am Main: Suhrkamp.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Henrich, D. 1975. Die Deduktion des Sittengesetzes. In: A. Schwan (ed.), *Denken im Schatten des Nihilismus, Festschrift für W. Weischedel*, 55-112. Darmstadt: Wissenschaftliche Buchgesellschaft.

Höffe, O. 1993. Kants nichtempirische Verallgemeinerung: zum Rechtsbeispiel des falschen Versprechens. In: O. Höffe (ed.), *Grundlegung zur Metaphysik der Sitten. Ein kooperativer Kommentar*, 206–233. Frankfurt am Main: Klostermann.

Kant, I. 2011 [1785]. *Groundwork of the Metaphysics of Morals*. German-English edition, edited and translated by M. Gregor and J. Timmermann, Cambridge: Cambridge University Press.

Kant, I. 2002 [1788]. *The Critique of Practical Reason*, translated by W. Pluhar and introduced by S. Engstro, Indianapolis: Hackett Publishing Company.

Kutschera, F. von. 1999. *Grundlagen der Ethik*, second edition. Berlin: de Gruyter.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*, London: Penguin Books.

Marina, J. 1998. Kant's Derivation of the Formula of the Categorical Imperative: How to Get it Right. *Kant-Studien* **89**: 167–178.

Onof, Ch. 1998. A Framework for the Derivation and Reconstruction of the Categorical Imperative. *Kant-Studien* 89: 410-427.

Rawls, J. 1980. Kantian Constructivism in Moral Theory. Rational and Full Autonomy. *Journal of Philosophy* **77**: 515–572.

Rawls, J. 1989. Themes in Kant's Moral Philosophy. In: E. Förster (ed.), *Kant's Transzendental Deductions. The Three Critiques and the Opus Posthumum*, 81-113. Stanford: Stanford University Press.

Vanderveken, D. 1980. Illocutionary Logic and Self-Defeating Speech Acts. In: J. R. Searle; F. Kiefer; M. Bierwisch (ed.), *Speech Act Theory and Pragmatics*, 247–272. Dordrecht: D. Reidel.

Wood, A. 1999. *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Notes

[1] See, for example, (Allison 1991), (Marina 1998), (Onof 1998), and (Wood 1999). The reconstruction presented here is a strongly revised version of the reconstruction in my German.

[2] Cf. (Kant [1785], p.3 f.).

[3]Cf. (Kant [1785], p.69.).

[4] See, for instance, (Kant [1785], p.65).

[5] Cf. (Kant [1785], p.13).

[6] See also (Bittner 1993, p.22) and (Höffe 1993, p.206 f.).

[7] Cf. (Kant [1785], p.13).

[8] Cf. (Kant [1785], p.71, p.79). Whether Kant's intention in the third section, which provides the transition to the Critique of Pure Practical Reason, is to prove the actual validity of the categorical imperative is unclear. His assertion in the "Concluding Note" that an unconditional moral law cannot be made intelligible according to its absolute necessity indicates that he considers such a proof impossible. On the other hand, Kant announces at several places that he intends to carry out a deduction of the moral law. Cf. also (Henrich 1975, p.62 ff.).

[9] Cf. (Kant [1785], p.119).

[10] Cf. also (Beck 1960).

[11] Cf. (Kant [1785], p.3).

[12] These termini technici are not used consistently in the literature. My explanations have therefore a stipulative component.

[13] Cf. (Hare 1952, p.46). For an account of ethical conventionalism, see (von Kutschera 1999, p. 126-137).

[14] I adopt here the constructivist interpretation of Kant's ethics defended in (Rawls 1980) and (Rawls 1989) and also in (Habermas 1999). A more Platonist interpretation can be found in (Mackie 1981). In (Greimann 2004), these different interpretative approaches are contrasted and discussed.

[15] Kant himself draws the parallel to Socrates in (Kant [1785], p.37).

[16]Since, according to Kant, there is a multiplicity of categorical imperatives, his talk of "the" categorical imperative in the singular is terminologically slightly inconsistent. By "the" categorical imperative, Kant means the supreme categorical imperative, which subsumes all other imperatives of this kind. For the sake of brevity and simplicity, I follow this usage here.

[17]See also the similar reconstruction in (von Kutschera 1999, p.330 ff.).

[18]See also (Kant [1785], p.89, footnote).

[19]Cf. (Kant [1785], p.35) and (Kant [1788], p.41).

[20]Cf. (Vanderveken 1980, p.249 f.).

[21]For a more complete account of the contradictory success conditions of such speech acts, see (Vanderveken 1980, p.274-271).

[22]Cf. (Vanderveken 1980, p.264 ff.).

[23]Cf. (Kant [1785], p.73).