Abstract: This Element defends a reading of Kant’s formulas of the moral law in Groundwork for the Metaphysics of Morals. It disputes a long tradition accepted both by Kant’s critics and defenders, concerning what the first formula (Universal Law/Law of Nature) attempts to do. But the reading proposed here is the only one that agrees with both what Kant says about these formulas and with what he actually does with them in his discussion of the four famous examples – which are also discussed here at some length. The Element also expounds the Formulas of Humanity, Autonomy and the Realm of Ends, arguing that it is only the Formula of Humanity from which Kant derives general duties, and that it is only the third formula (Autonomy/Realm of Ends) that represents a complete and definitive statement of the moral principle as Kant derives it in the Groundwork. The Element also disputes the common claim that the various formulas are “equivalent,” arguing that this claim is either false or else nonsensical because grounded on a false premise about what Kant thinks a moral principle is for.

Keywords: Kant’s ethics, categorical imperative, formula of the moral law, universal law, humanity as end in itself, autonomy

Kant’s ethical theory has traditionally been received, both by its sympathizers and its critics, largely through his short treatise *Groundwork for the Metaphysics of Morals* (1785). In the *Groundwork*, Kant formulates the supreme principle of morality in a number of different ways. Most prominent in the reception of his theory has been his first formulation: the Formula of
Universal Law (FUL). From the beginning, both the interpretation of this formula and its purpose in moral reasoning have been controversial. I believe they have also been generally misunderstood.

The resurgence of interest in Kant’s ethics in the late twentieth century was led by John Rawls and Onora O’Neill, who identified FUL with “the Categorical Imperative,” and interpreted it as a strict “CI-Procedure” providing a general discursive criterion of moral right and wrong (O’Neill, 2013 [1975], Chapter 2; cf. Rawls, 1989, 2000, pp. 162–181). From the start, however, Kant’s critics – beginning with Gottlob August Tittel (1739–1816) (Tittel, 2012 [1786]) and later such famous philosophers as Hegel, Mill, and Sidgwick – questioned whether FUL can provide any such procedural criterion of right and wrong. I do not pretend to know the best interpretation of FUL when it is read with these aims. Nor do I know whether, so interpreted, it could successfully provide a general moral criterion or procedure – though I doubt it. What I am certain of is that both sides in this dispute are misreading Kant. This essay aims to explain Kant’s actual use of this and the Groundwork’s other formulas of the moral law.

First Part: The System of Formulas

§1. The Aims of the Groundwork

It took Kant a long time to formulate his mature moral theory. By the mid-1760s, after a flirtation with Hutcheson’s moral sense theory, Kant decided to reject feeling as the foundation of ethics and identify his position with the phrase ‘metaphysics of morals.’ By this he meant that he took the foundations of ethics to lie in what he called ‘metaphysics’ – that is, synthetic a priori cognitions from concepts. It was not until nearly twenty years later that he published even the first, cautious ‘laying of the foundations’ (Grundlegung) for a metaphysics of morals (1785). The aims of Kant’s little book were very limited: “a
search for and establishment of the supreme principle of morality” (G 4:392). The search occupies the first two sections, and the establishment (or “deduction”) of the principle occurs only in the Third Section. A fuller presentation of Kant’s system of duties had to wait until the very end of his career: the *Metaphysics of Morals* (1798).

**Kant’s search for the supreme principle of morality.** The first section of the *Groundwork* derives a formula of the moral law from the presuppositions of “common rational moral cognition” (G 4:393). This phrase denotes the knowledge of moral truth that Kant takes the ordinary moral agent to possess and to display in everyday judgments, feelings, and actions when these conform to morality. Kant elicits this everyday cognition in the form of our anticipated assent to certain propositions about what it is that we most value as moral agents, and of the kind of value that we regard as most central to morality itself. Kant expects us to assent to the claim that nothing is good without limitation except a good will (G 4:393–396), and that what is central to morality is a special case of the good will – the case where someone must act with self-constraint on moral grounds. Then the action is not merely “in conformity with duty” (*pflichtmäßig*) but also “from duty” (*aus Pflicht*) (G 4:397–399). This constraint exhibits awareness of an obligating rational law that is universally valid for all rational agents (G 4:401). That thought leads to Kant’s first formulation of the moral principle, the Formula of Universal Law (or FUL): “I ought never to conduct myself except so that I could also will that my maxim become a universal law” (G 4:402). This is the only formula developed out of common rational moral cognition; it is therefore apparently the only formula we need to employ in everyday life. Even it, however, is not thought by the ordinary moral agent in this abstract form, but is an implicit standard of moral judgment (G 4:403).

Kant undertakes a fuller search for the supreme principle of morality in the Second Section, based on a properly *philosophical* investigation of the principles of rational volition (G 4:412). Kant distinguishes three such standards. The *first* is that of instrumental or *technical* reason: that we rationally should constrain ourselves to
take the necessary means to an end we have adopted at our discretion (G 4:415). The second is that of prudential or pragmatic reason: that we rationally should form the idea of a sum-total of achievable satisfactions of our empirical desires or inclinations (under the name of ‘welfare’ or ‘happiness’), and rationally should give the pursuit of this end priority over any other ends of inclination with which it might conflict (G 4:416). The third is that of moral reason, which commands us unconditionally to do or omit certain actions, and to set certain ends, irrespective of any ends or desires we may have that are extrinsic to the moral command (G 4:416).

Kant represents these rational standards in the form of imperatives. This term does not refer to a grammatical form, but instead to a ground or reason for rational self-constraint. Kant distinguishes hypothetical imperatives, grounded on an independently adopted end, from categorical imperatives, which presuppose no independently given end as a condition of their rational validity. The instrumental imperative and the imperative of prudence are both hypothetical: the former is grounded on some end we have set arbitrarily or at our discretion; the latter on the end of happiness, which every finite rational agent does set. The rational validity of hypothetical imperatives, Kant argues, is analytic – that is, it is grounded solely on the content of the mere concept of the imperative itself. More specifically, it is contained in the very concept of setting an end that insofar as one’s actions are governed by that end, one rationally ought to take the necessary means to it. Of course specific technical rules and counsels of prudence are not analytic; they are not even practical propositions, but typically contingent and empirical theoretical truths (FI 20:197–200). The rational self-constraint of hypothetical imperatives, however, is grounded analytically on the very concept of setting an end, which includes in it the normative requirement that we should take the necessary means to the end we have set.

Categorical imperatives, however, are not analytic. Even their possibility will have to be established. Kant does this only in the Third Section, where a deduction of the moral law is to be provided. Until this happens, as Kant reminds us repeatedly (even
In the First and Second Sections, we must regard the whole of morality as possibly an illusion or figment of the brain (G 4:394, 404, 407, 419–420, 423, 425, 429, 431, 445).

Kant calls the three (or five) statements of the moral law, formulas of the categorical imperative. In a later work he says he means this term in the sense mathematicians use it. A ‘formula’ is something that “determines quite precisely what must be done to solve a problem” (CPPr 5:8 n). Each formula addresses a different problem. We will see in due course what each of the specific problems is, and we must be careful not to make hasty assumptions about what Kant thinks a moral principle is for.

Kant’s search for the moral principle, as we learn after the three main formulas of the moral law have been presented, is intended to be systematic (see Wood, 2001). It is organized according to a triad drawn from Kant’s theory of concepts, when that theory is applied to the concept of a practical principle or maxim (G 4:436). A maxim is a subjective norm imposed by a rational agent on its actions (G 4:421, cf. CPPr 5:19). It is something like an intention, if we think of intentions as subjective norms we impose on our actions. A maxim, however, would not be the intention of a particular action at a particular time and place, but rather a generalized intention: to do a certain kind of action whenever certain specifiable conditions obtain. A maxim is therefore a subjective norm – one we impose on ourselves at our discretion. A maxim typically has a form: the kind of action; also a matter: an end for the sake of which the subject acts. According to Kant’s theory of concepts, the complete concept of an individual always involves a complete determination of the individual in respect of its properties. In the case of the formulas, this refers to the complete formula, in which form and matter have been combined (G 4:436).

The first and third formulas, though not the second, have both a more abstract and a more “intuitive” variant. The latter variants are said by Kant to be “closer to intuition” and hence “nearer to feeling,” and therefore better able to help the moral principle gain “access” or “entry” (Eingang) to human beings through intuition and feeling (G 4:389, 405, 409, 436–437). The intuitive variant of the
first formula adds to the abstract idea of universal law the more concrete thought of a possible nature as embodying a system of laws. The intuitive variant of the third formula adds to the idea of rational will regarded as legislating universally the more concrete thought of an entire community of rational beings, and the shared system of their collective ends that would result from the perfect obedience to such a universal legislation.

The entire system of these formulas is this:

**First formula**

FUL  *Formula of Universal Law*: “Act only in accordance with that maxim through which you at the same time can will that it become a universal law” (G 4:421),

with its more intuitive variant,

FLN  *Formula of the Law of Nature*: “So act, as if the maxim of your action were to become through your will a universal law of nature” (G 4:421).

**Second formula**

FH  *Formula of Humanity as End in Itself*: “So act that you use humanity, as much in your own person as in the person of every other, always at the same time as an end and never merely as a means” (G 4:429).

**Third formula**

FA  *Formula of Autonomy*: “the idea of the will of every rational being as a will giving universal law” (G 4:431; cf. 4:432), or “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” (G 4:440; cf. 4:432, 434, 438),

with its more intuitive variant,

FRE  *Formula of the Realm of Ends*: “Act in accordance with maxims of a universally legislative member for a merely possible realm of ends” (G 4:439; cf. 4:433, 437–439).
Each of the formulas is derived from the very concept of a categorical imperative: the first from its form, the second from its matter; the third combines the first formula with the second. The first two formulas are thus in a sense one-sided (form without matter, matter without form). The form of law, Kant says, is the law considered objectively (in terms of its universal obligation), while the matter is the law considered subjectively (in terms of the rational ground or motive for the subject’s obedience to it). The third formula, FA (or its variant, FRE), combines the first two formulas (G 4:431). It is this complete determination form of the law that is later used in the deduction (G 4:446–455). It is also FA that is used as the definitive statement of the moral law in Kant’s later ethical works: the Critique of Practical Reason (CPrR 5:30) and the Metaphysics of Morals (MM 6:225–226).

Kant also proposes to systematize his presentation of the three formulas according to his categories of quantity: unity, plurality, totality (G 4:436). In Kant’s table of categories, the category of unity corresponds to the universal form of judgment (∀ S is P); the category of plurality corresponds to the particular form (∃some S is P); and the category of totality corresponds to the singular form (The S is P – representing the singular or completely determined concept of an individual thing) (CPR A70/B95, A80/B106). FUL/FLN thus represents the universal form of moral laws, FH the plurality of particular ends in themselves that are the matter of the law; likewise, FA/FRE results from combining FUL/FLN with FH. FA/FRE thus neither requires nor receives any grounding except from the way it combines FUL/FLN and FH. FA/FRE is the complete presentation of the supreme principle of morality.

§2. Derivation of the Formulas from the Concept of a Categorical Imperative

This is not an essay about Kant’s derivation of his formulas, but about the formulas themselves – what they mean, and the specific problems they are meant to solve. But the meaning of any philosophical thesis is determined by the arguments for it, so I need to introduce the
formulas by saying something briefly about the way Kant proposes to derive each from the mere concept of a categorical imperative.

**Universal law.** A categorical imperative is a principle that rationally constrains any moral agent irrespective of any contingent desires or discretionary ends the agent may have. It therefore applies to all rational agents universally and in the same way. From this point of view, therefore, the only content thought in the concept of such an imperative is conformity of the will to the idea of universal legislation itself – to those principles by which all rational beings might rationally will to constrain themselves and all others (G 4:402). We should note that this requirement of universal validity is not for Kant a specifically moral or ethical requirement at all. That’s what it means to say that it is purely formal. It expresses only the minimal standard of all rationality, whether theoretical, practical, or even aesthetic. The concept of an objectively valid judgment is simply one that is valid universally for all rational beings (P 4:298). Thus it is a rational constraint on all thinking that we ought to accept only that whose ground or rule of acceptance could be made a universal principle for the use of reason (WOT 8:146 n). Even a correct judgment of taste is one considered universally valid (CJ 5:283–284). An action is right (not ethically, but merely in the sense required for everyone’s external freedom) if it or its maxim are consistent with the freedom of all according to universal law (MM 6:230).

**Humanity as end in itself.** The mere form of a categorical imperative, therefore, is not sufficient to get us to a genuine metaphysics of morals. For that we need not only the form but also the matter of a categorical imperative, namely the way it is “bound up (fully a priori) with the concept of the will of a rational being in general.” By this Kant means: the rational ground that would motivate obedience to a categorical imperative. It is to obtain the concept of such a ground that we must “take one step beyond, namely to a metaphysics [of morals]” (G 4:426–427). In other words, we have not reached the principle for which we are searching in the *Groundwork* until we have derived not only FUL/FLN but also FH.

Some philosophers have thought that practical rationality consists solely in choosing the correct instrumental means to ends
given by desire independently of reason. They have therefore
sometimes denied that there could be any ground for obedience
to a Kantian categorical imperative. Kant’s direct response to their
objections is his derivation of FH. The ground of obedience to
a categorical imperative, he argues, must be an end, but end
must be necessarily connected with the concept of a rational will
and equally valid for all rational beings. That is, it must be an end
which is objective rather than subjective: an end that rationally
binds us not because we have set it at our discretion, but because
it is necessarily binding on us simply as rational beings. Therefore,
it also cannot be some possible effect of our actions, since such an
effect is regarded as an end only because we have set it as an end at
our discretion (G 4:427). The end in question, Kant argues, must be
something whose existence itself has absolute worth. Kant calls this
an “independent” or “self-sufficient” end, and an end having
objective worth (G 4:427, cf. 4:437). The concept of such an end is
what Kant calls an “end in itself” (Zweck an sich selbst) (G 4:428).

What could be such an end? In three successive paragraphs, Kant first supposes, then asserts, and then finally argues, that the end in itself is humanity or rational nature in persons (G 4:428–429).

In this context, ‘humanity’ does not refer to the special empirical
nature of human beings. It refers to the capacity to set ends
according to reason (Rel 6:26–27). This includes the technical pre-
disposition to choose or invent means to the ends a rational being
freely sets, and also the pragmatic disposition to combine these
ends into an idea of one’s welfare or happiness. Thus the concept
of humanity is distinct from and does not include the concept of
personality or the moral predisposition (Rel 6:26 n). But Kant
argues that any being having humanity also necessarily has per-
sonality (Anth 7:322–324; cf. CB 8:111–115). Of course it is
a contingent, empirical fact that there exist any beings with such
predispositions and capacities. But it is true a priori that the
rational nature of any such being must be an end in itself.

Kant’s argument that humanity is the sole end in itself is brief and
notoriously controversial. (My own attempts to interpret the argu-
ment can be found at Wood, 1999, pp. 118–139, 2008, pp. 85–93, and
Schönecker and Wood, 2015, pp. 142–149.) For our present purposes, I will present it this way: A categorical imperative, by its concept, cannot be based on anything regarded as valued contingently or based on the desires or discretionary ends of any rational agent. Once these ends are excluded, there is no conceivable value remaining on which it could be based except that of rational agency itself: specifically, that of the rational beings who are supposed to have this ground for obeying a categorical imperative. Kant thinks that in exercising rational agency we implicitly represent ourselves as having such a value. This is what Kant means by the crucial premise in his argument for FH: “The human being necessarily represents his own existence [as an existing end in itself]” (G 4:429).

A natural objection is that there seem to be people who do not do this. Some people regard themselves as worthless and their existence as having no value. But this objection fails because it misunderstands the claim being made. That claim is not one about how people actually think of their existence. The crucial word ‘necessarily’ means that this is a way we must represent ourselves in exercising any rational agency at all – even if we also (incoherently) represent ourselves and our existence as worthless. The essence of Kant’s argument was later put in an arresting form by Nietzsche: “He who despises himself still esteems the despiser within himself” (Nietzsche, 2002, Part Four, §78). Kant would understand this not as an empirical psychological claim, but a philosophical one about what we presuppose in exercising any form of rational agency. In setting any end, choosing any action, esteeming or despising anything, even ourselves – we thereby claim authority over our own volitions. We presuppose a worth that belongs to ourselves, our volition, even an objective worth which, necessarily and a priori, occupies a fundamental place in any system of valuation we might ever adopt on the basis of our volitions. Philosophies, religions, and neuroses that deny humanity this worth may be psychologically, socially, and historically powerful and persistent, but they are always rationally self-undermining. The pathology, lies, and self-deceptions involved in them are usually more social than individual. The way out of these is the long, difficult, and problematic and still incomplete historical