Part II, Paper 3, Kant's Ethics

Christopher Benzenberg (cpb67)

LT5-8, Fridays, 9:30 - 11:00 am

Seminar Room E, 17 Mill Lane

Course Description

In this lecture, we examine Kant's *Groundwork to the Metaphysics of Morals* (1785). We will especially address the following four questions: (i) In what sense is the good will good without limitation? (ii) What is moral worth and how does it relate to Kant's concept of duty? (iii) What are the different formulas of the categorical imperative, and how do they relate? (iv) Why should we think that the categorical imperative is universally valid? In answering these questions, the lecture seeks to convey the basic elements of Kant's moral thought. However, the lecture also tries to counter the common misconception that Kant's ethics is anaemic or cold. We will see that happiness plays an important and role for Kant.

Format

There will be four lectures in total, one on each of the above questions. At the beginning of every lecture, I will give a 45 to 60min presentation, after which there will be time for questions. It is important that you read the primary text before each lecture. If you cannot read the German original, I suggest that you use the Gregor-Timmermann or the Wood translation. The readings for each week are specified below (in italics). The stated page numbers reference volume 4 of the academy edition; you find them in the margins of your Cambridge edition. If you have any questions about the readings or the general format, please contact me at cpb67@cam.ac.uk.

Required Readings

1. **Lecture** "Good Without Limitation": Preface & Section I (4:387-396)
2. **Lecture** "Inclinations versus Duty": Section I (4:397-405)
3. **Lecture** "The Categorical Imperative(s)": Section II (4:406-445)
4. **Lecture** "Reason, Freedom, and Morality": Section III (4:446-463)

Part II, Paper 3, Kant's Ethics
Christopher Benzenberg (cpb67)
LT5-8, Fridays, 9:30 - 11:00 am
Seminar Room E, 17 Mill Lane

1. Lecture — "Good Without Limitation"

Preface & Section 1 (4:387-396)

1.1 The Project of the Groundwork

The Metaphysics of Morals. Kant first distinguishes between the three main branches of philosophy: physics, ethics, and logic (4:387). Logic abstracts from any specific object of thought and is thus purely formal. Both physics and ethics, however, have determinate objects and are thus material. Physics concerns laws of nature (laws about what happens); ethics deals with laws of freedom (laws about what ought to happen) (4:387f.). Like physics, ethics has an empirical and an a priori/pure part. The empirical part is called practical anthropology; the pure is called metaphysics of morals (4:388). The metaphysics of morals contains all a priori laws of freedom in systematic order.

The Supreme Principle of Morality. The *Groundwork* only deals with the highest law of freedom, the supreme principle of morality, thereby giving a general answer to the question: "What should we do" (A804-5/B832-3). Kant distinguishes two more specific goals of the Groundwork, namely (i) to identify the highest principle of morality, and (ii) to justify the highest principle of morality (4:392). The identification occurs mainly in Section I and is embellished in Section II. The justification, or deduction, is to be found in Section III. Kant further suggests that the identification proceeds by analysing common cognition, whereas the justification proceeds synthetically from the sources of morality (4: 392).

Kant's Notion of a Pure Will. The supreme principle of morality is not a free-floating principle of the form "An action is right iff X (for some X)". Instead, Kant conceptualises the principle of morality as the principle of a faculty, specifically the pure will: "the metaphysics of morals is to investigate the idea and principles of a possible pure will" (4:390). The idea is this: a pure will is "carefully cleansed" by everything empirical, specifically the corrupting influences of sensibility/inclinations (4:388). Because the pure will is not determined by private conditions, it must have a universal character. Yet this universal character is the hallmark of morality. The pure will is therefore a good will.

1.2 Final versus Instrumental Value

The opening sentence. "It is impossible to think of anything at all in the world, or indeed even beyond it, that could be taken to be good without limitation, except the good will" (4:393). Kant contrasts the good will with talents of the mind, qualities of the temperament, and gifts of fortune, especially happiness. Unlike the good will, they are not good without limitation. But how are we to understand this claim? What does it mean to be good with and without limitation? A first plausible suggestion is that something is good without limitation iff it has final value, and it is good with limitation iff it has instrumental value.

Final versus Instrumental Value

(1.) For all x, x has final value iff x is valued for its own sake.
(2.) For all x, x has instrumental value iff x is valued for the sake of something else.

The problem. On the face of it, this reading seems to explain the text. Consider persistence of intend, as one of the talents of the mind. Clearly, persistence is only instrumentally valuable to obtain other goods, such as a well-paying job. Conversely, the good will is not good because it is instrumental to achieve some other good. The good will is valued for its own sake. The problem, however, is that not all things that are good with limitation are instrumentally good. Consider happiness. We clearly value happiness for its own sake, yet, according to Kant, it is only good with limitation.

1.3 Intrinsic versus Extrinsic Value

Korsgaard's distinction. Many people conflate final with intrinsic good and instrumental with extrinsic good. But Korsgaard shows that the two distinctions should be kept apart. The final/instrumental distinction tells us how we value a thing: for its own sake or for the sake of something else. The intrinsic/extrinsic distinction, on the other hand, clarifies how things have value: whether it is good in itself or not. As Korsgaard puts it:

> There are [...] two distinctions in goodness. One is the distinction between things valued for their sakes and things valued for the sake of something else—between ends and means, or final and instrumental goods. The other is the distinction between things which have their value in themselves and things which derive their value from some other source: intrinsically good things versus extrinsically good things. Intrinsic and instrumental good should not be treated as correlatives, because they belong to two different distinctions. (Korsgaard, "Two Distinctions in Goodness", 250)

Intrinsic versus Extrinsic Value

(1.) For all x, x has intrinsic value iff the value of x strongly supervenes only on (or is grounded in) x's intrinsic properties.

(2.) For all x, x has extrinsic value iff the value of x strongly supervenes also on (or is grounded in) x's extrinsic properties.

How this might help. With this two-fold distinction in place, we might reapproach Kant. Being good without limitation could mean that a thing is finally and intrinsically good. And clearly, the good will is both finally and intrinsically good. Happiness, on the other hand, is finally but not intrinsically good. We value happiness for its own sake, but its goodness is at least partially grounded in something extrinsic to it, namely the good will. Only if the will is good, does happiness have value. On this picture, the good will turns out to be the source of the value of happiness.

1.4 Unconditional versus Conditional Value

The problem. While Korsgaard's reading is widely accepted, Bader argues that it faces serious textual challenges. Happiness must have intrinsic value, for Kant, (i) to contribute to the complete good (*bonum consummatum*), the attainment of which is a duty for finite rational beings. What's more, (ii) happiness comes in degrees and so it's sources should also come in degrees, but the good will does not come in degrees. And (iii) it seems that the badness of unhappiness is also conditional on the good will, but it cannot be said to derive from the good will. Bader therefore suggests introducing yet another distinction: between conditional and conditional value.

Unconditional versus Conditional Value

(1.) For all x, x has unconditioned value iff, there is no y, such that y is a condition on x being valuable.

(2.) For all x, x has conditionate value iff, there is a y, such that y is a condition on x being valuable.

Conditions versus sources. Pace Korsgaard, Bader intends the unconditional/conditional distinction to be different from the intrinsic/extrinsic distinction. For this to work, Bader must distinguish between sources, on the one hand, and conditions, on the other hand. He suggests that this distinction is intuitive as we would, for instance, not think that the absence of a disabler is a source of value. But without the absence of a disabler the thing would not be good. Bader suggests that the distinction between conditions/and sources tracks these intuitions. We can then define intrinsic sources in a hyperintensional way, by only looking at duplicates where the condition is satisfied.

How this helps. Bader now has the resources to argue that happiness, for Kant, really is intrinsically valuable, just like the good will, because its goodness is grounded in intrinsic properties. However, unlike the good will, the intrinsic value of happiness is conditional upon the presence of an enabler, namely the good will. On this picture then (which I think is the correct reading of Kant), the good will is good without limitation because it is unconditionally good. And, according to Kant, the good will is the only thing that is unconditionally good.

Part II, Paper 3, Kant's Ethics
Christopher Benzenberg (cpb67)
LT5-8, Fridays, 9:30 - 11:00 am
Seminar Room E, 17 Mill Lane

2. Lecture — "Duty versus Inclination"

Section 1 (4:397-405)

2.1 Analysing the Good Will

Quick recap. Kant has introduced the good will as the only thing that is good without limitations; all other goods, such as happiness, are only good with limitation. We have explored three readings of "good without limitation", meaning either (i) finally good, (ii) intrinsically good, or (iii) unconditionally good. While most people would endorse the second reading, I suggested that the third reading is both textually and systematically adequate. In a next step, Kant now wants to *identify* the principle of the good will, which he thinks is the supreme principle of morality (4:390).

The approach. Kant could identify the principle of morality by analysing the will of a purely rational being that necessarily adheres to morality (4:453). However, Kant wants the Groundwork to help us overcome a "natural dialectic" (4:405) — our human tendency to massage the moral law so that it aligns with our inclinations. To curb this tendency, Kant analyses the human will, i.e., *the will of finite rational beings*, which are driven both by reason ($\sim$ duty) and sensibility ($\sim$ inclination): "we shall examine the concept of duty, which contains the concept of a good will, though under certain subjective limitations and hinderances" (4: 397).

Four cases. Setting aside actions that are contrary to duty (4:397), Kant examines four cases which are in accordance with duty, and thus candidates for good-willed actions with moral worth (4:397-9): (i) Kant's prudent shopkeeper treats his customers fairly to maximise his long-term profit; acting on mediate inclinations, his conduct lacks moral worth. (ii) Preserving one's life only has moral worth if it is done from duty rather than an immediate inclination. (iii) The philanthropist only deserves to be praised for her donations if she did so from duty rather than love or compassion. And finally, (iv) the attainment of happiness lacks moral worth if it is done from a direct inclination rather than duty.

2.2 Kant's Caricature

Two criticisms. Kant is commonly mocked for his view that morality requires that we must act on cold duty, rather than warm sentiments — "he were by temperament cold and indifferent to the suffering of others" (4:398). Schiller ridiculed this view during Kant's lifetime in his *Xenien* (see below). In "Persons, Character, Morality", Williams famously observes that Kant's moral agent has "one thought too many." Take a husband, who saves his drowning wife rather than another person, thinking "It is my wife, and in situations of this kind it is permissible to save one's wife rather than other person." It would have been better had he saved his wife simply because he loves her, or for her own sake.

> I like to serve my friends, but unfortunately I do it by inclination
> And so often I am bothered by the thought that I am not virtuous.
> There is no other way but this! You must seek to despise them
> And do with repugnance what duty bids you.

What's at stake. If the criticism is true, then it seems that Kantian morality is necessarily pitched against happiness. Happiness, for Kant, consists in the total satisfaction of all our inclinations (4:393, 394, 396, 399, 406, A806/B834). Now, those actions which Kant credits with moral worth either lack an inclination or even have an opposing inclination: only the suicidal person deserves moral praise for preserving her life; only the misanthrope's donations have moral worth; and only the gout suffere's efforts to stay healthy are good willed. And indeed, Kant suggests that the good will "in many ways limits [...] the attainment of [...] happiness" (4:396; see also 4:400, 405).

2.3 Three Charitable Readings

Battle citation reading. Henson suggests that Kant intended moral worth to be an exceptional distinction reserved for genuine moral battles. If you win a moral battle, you overcome an opposing inclination, and only then, Henson claims, does your action have *true* moral worth. On this picture, however, moral worth is not something that we should strive for, just as we should try to avoid battles in general. Mundane actions, Henson argues, still deserve to be praised, just not to the same extend. And indeed, Kant writes about the philanthropist that her donations "deserve praise and encouragement" (4:398).

Counterfactual dependence. Herman suggests that actions with moral worth need not be done from duty alone but can also involve inclination. This can be done in two ways: (i) either we have an accompanying inclination when we act from duty, or (ii) we act from inclination, but have duty as second-order motive that kicks in if something goes wrong, like a back-up generator. And indeed, Kant qualifies his claim that we should not act from inclination when he writes that we should only "excludes [them] [...] from

calculations when we make a choice" (4:400) — this leaves open that we can act from inclination.

Epistemic reading. Kant's real reason for excluding inclinations is epistemic. Remember that Kant's main goal is to identify the principle of the good will, by examining our finite rational will. To isolate the contribution of pure reason, Kant must abstract from the effects of sensibility, i.e., inclinations. Kant thus writes: "we shall inspect the concept of *duty*, which contains that of a good will, though under certain subjective limitations and hinderances, which, however, far from concealing it and making it unrecognizable, rather *bring it out by contrast and make it shine all the more brightly*." (4:397; my emphasis) This raises the question: How does Kant derive the moral law from the concept of duty?

2.4 Deriving the Moral Law

The three propositions. Out of the blue, Kant writes "The second proposition is: an action from duty has its moral worth [...] in the maxim according to which it is resolved upon, and thus it does [...] merely [depend] on the *principle of willing*" (4:399-400). He then adds "The third proposition, as the conclusion from the previous ones I would express as follows: *duty is the necessity of an action from respect for the law*." (4:400). This derivation raises one big question: What is the first proposition supposed to be?

Two possible readings. The third proposition contains three elements; that actions from duty are (i) necessary, (ii) involve respect, and (iii) concern the law. The second proposition has contributed the third element, which leaves the first proposition to somehow contribute the first two elements. Either the first proposition already assumes that actions with moral worth are done from respect — understood as an a priori feeling (4:401n) —, and shows that such actions must be necessary; or it assumes that good-willed actions are necessary and shows that only actions from duty involving respect satisfy this condition.

The correct reading. I suggest that the second reading is correct. By establishing that the good will is good without limitation, Kant aims to show that the good will is necessarily good (see also 4:390, 398). The four cases then establish that actions done from inclination are only accidentally good because we can always imagine cases in which they fail to achieve the demanded action. They are hence not expressions of the good will. Only actions from duty are necessarily good. Here then is how I reconstruct the three propositions:

The three propositions

(1) Actions from duty are necessary because they are based on *respect* (not inclinations).
(2) Actions from duty are based on a principle which must be a *law* (not an effect).
(3) Actions from duty are necessary from *respect for the law*.

My revised version

(P1) Good-willed actions are necessarily in accordance with duty.
(P2) Actions are necessarily in accordance with duty only if they are done from duty.
(P3) Actions are done from duty only if they are done from respect for a lawful principle.
(C) *Therefore*, good-willed actions are done from respect for a lawful principle.

The Categorical Imperative. From the third proposition, Kant derives the principle of the good will, which binds all rational beings (4:390-1): the Categorical Imperative. The principle of the good will must have a law-like form. Now, since Kant abstracts from all sensible matter, this principle cannot be any specific law, but must the lawfulness of an action in general: "Since I have robbed the will of all impulses that could arise for it from some particular law, nothing remains but as such the universal conformity of action with law, which alone is to serve the will as its principle, i.e., I ought never to proceed except in such a way *that I could also will that my maxim should become a universal law*." (4:402)

2.5 Non-Accidental Rightness

The pressure point. We could raise many potential objections against Kant's argument. Why, for instance, should we conceptualise the supreme principle of morality as the principle of a faculty? I think the most interesting objection, however, concerns Kant's notion of non-accidental rightness. In my version of Kant's argument, Kant assumes that actions with moral worth are necessarily or non-accidentally right — something that most contemporary ethicists would sign. However, why should we think that only actions from duty are non-accidentally right? How do the four cases support this conclusion?

The first problem. In the four cases, Kant argues that actions from inclination are only accidentally right because we could easily imagine scenarios in which one lacks the inclination and fails to perform the action. But couldn't we likewise imagine scenarios in which we happen to lack the motive of duty? One solution is to fix the motives and see whether they will necessarily lead to the right outcome. Based on Nozick's notion of epistemic tracking, we could formulate the following conditions for moral tracking:

Moral Tracking

S's action φ based on M is non-accidentally right iff
(a) if φ wasn't right and S were to act based of M, S wouldn't perform φ based on M,
(b) if φ was right and S were to act based on M, S would perform φ based on M.

The second problem. While these conditions certainly excludes some inclination-based actions (e.g., Herman's benevolent accomplice), it seems that we can easily construct cases

where the agent acts on inclinations, but the action is non-accidentally right. Inspired by standard counterexamples to epistemic tracking, we could introduce a benevolent demon, which ensures that the inclination-based actions always result in the correct conduct: performing the right actions and omitting the wrong ones. In cases like Kant's shopkeeper, we could similarly imagine scenarios where the invisible hand guarantees that greed and self-interest necessarily result on morally right actions. But can we also construct cases of actions from duty that are only accidentally right?

Part II, Paper 3, Kant's Ethics
Christopher Benzenberg (cpb67)
LT5-8, Fridays, 9:30 - 11:00 am
Seminar Room E, 17 Mill Lane

3. Lecture — "The Categorical Imperative(s)"

Section II (4:406-445)

3.1 Normativity

What happened thus far. In the first section of the Groundwork, Kant has identified the supreme principle of morality, which he conceptualizes as the principle of a pure will. Yet rather than examining the pure will directly, Kant focused on our human will, which is incentivised both by duty and by inclination. If we abstract from inclinations, Kant suggests, we can isolate the supreme principle: "I ought never to proceed in such a way that could also will that my maxim should become a universal law." (4:402) This principle becomes the focus of the second section of the Groundwork.

Sources of normativity. Unlike other forces of nature, a will acts "according to the representation of laws" (4:412). But like all other forces, a will never deviates from its own law without an outside influence (A294/B350). Yet failure is necessary for normativity: I ought to φonly if I could fail to φ. The holy will cannot fail because it is not subject to the corruption of inclinations, and so the moral law describes how it *in fact* operates (4:412-4). Our human will, however, can fail because it is subject to corruption from inclinations, and so the moral law is a categorical imperative for us, telling us how we *ought to* act (4:413).

3.2 Hypothetical vs Categorical Imperatives

Kant's distinction. Hypothetical imperatives, for Kant, "represent the practical necessity of a possible action as a means to achieving something else that one wants" (4:414). They can be either imperatives of skill when the end is problematic/possible, or they can be imperatives of prudence when the end is assertoric/actual, as is the case only with happiness. Categorical imperatives, by contrast, represent "an action as objectively necessary by itself, without reference to another end." (4:414; see also 3:298-300) Categorical imperatives demand apodictic/necessary ends, according to Kant.

Instrumental/categorical versus prudential/moral. It is worth noting that Kant runs together two separate distinctions: the instrumental/categorical distinction and the prudential/moral distinction. Instrumental rationality demands that "If you want the end, you ought to want the means", whereas categorical demands have the form "you ought to want the end". This distinction leaves open whether the end in question is problematic, assertoric, or necessary. Prudential demands only concern problematic/assertoric ends, whereas moral demands concern necessary ends. Kant seems to think that instrumental/categorical and prudential/moral distinctions coincide.

	instrumental	**categorical**
prudential	*hypothetical imperatives*	—
moral	—	*categorical imperatives*

3.3 Universalizability

Formula of Universal Law

(FUL) "Act only according to that maxim through which you can at the same time will that it become a universal law." (4:421; cf. 4:402)

Formula of the Law of Nature

(FLN) "So act as if the maxim of your action were to become by your will a universal law of nature." (4:421; cf. 4:436)

Different duties. Kant distinguishes between duties towards myself and duties towards others. He further distinguishes between perfect and imperfect duties. Perfect duties involve a contradiction in conception (there couldn't be a law to φ), whereas imperfect duties only involve a contradiction in the will (I couldn't will there to be a law to φ). Since both distinctions are orthogonal to each other, we get four classes of duties, specified in the table below. It is worth noting, however, that this is only the tip of the iceberg. In the *Metaphysics of Morals*, Kant attempts an exhaustive characterisation of all our duties.

	towards myself	**towards others**
perfect duties	*Don't commit suicide!*	*Don't lie!*
imperfect duties	*Develop your talents!*	*Be beneficent!*

Example: Don't lie!

> May I not, when I am hard pressed, make a promise with the intention of not keeping it? Here I readily distinguish the two senses which the question can have—is it prudent, or is it right, to make a false promise? [...] Suppose I seek, however, to learn in the quickest way and yet unerringly how to solve the problem 'Does a lying promise accord with duty?' I have then to ask myself 'Should I really be content that my maxim (the maxim of getting out of difficulty by a false promise) should hold as universal law (one valid both for myself and others)? And could I really say to myself that every one may make a false promise if he finds himself in a difficulty from which he can extricate himself in no other way?' I then become aware at once that I can indeed will to lie, but I can by no means will a universal law of lying; for by such a law there could properly be no promises at all [...] my maxim, as soon as it was made a universal law, would be bound to annul itself. (4: 402)

Problems. Kant insists that we are not even permitted to lie about a victim's whereabouts to a would-be murderer (6:429-31, 8:425-30). But is that plausible? Could we not form the maxim "I lie whenever doing so would deter a would-be murderer." This maxim clearly seems to be universalizable. Conversely, there seem to be puzzle maxims of actions that seem perfectly permissible but whose maxims fail the universalizability test. Take the maxim "Every Sunday at 10am, I attend the service at St Edmund's chapel." It is impossible for this maxim to become a universal law: it would get crowded in the chapel if all 8 billion humans attended.

Possible solution. The general way to resolve such puzzle maxims is to formulate some demands on what qualifies as maxims (something Kant never really does). The general move here must be that maxims cannot be too specific. But how can we spell out this notion of "not too specific" in a more precise way? This is genuinely hard. I suggest that we should take seriously the idea of maxims as a law of nature. Natural laws make no reference to a specific place and time: they hold at everyplace and every time. This would at least avoid the above maxim. It is not clear, however, whether this solves the general problem.

3.4 Humanity

Formula of Humanity

(FH) "So act that you use the humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means." (4:429)

Example: Don't lie!

> [T]he one who has it in mind to make a lying promise to another will see right away that he wills to make use of another human being merely as means, without the end also being contained in this other. For the one I want to use for *my ends* through such a promise *cannot possibly agree* with my way of conducting myself toward him and thus cannot *contain in himself the end* of this action. (4:429f.; my emphasis)

Reconstruction. In a lying promise, we treat the other person as a mere means because the other person "cannot possibly agree" with our way of treating her. What Kant arguably has in mind is that the other person, *if she was perfectly rationa*l, could not agree with our way of treating her — we are dealing with perfect rational agreement. Specifically, a rational person could not agree with contradicting ends because rationality requires consistency. Humanity is commonly understood as the ability of rational beings to set their own ends. On this reading then, FH essentially demands that we respect everyone's ability of setting their own ends by not pursuing contradicting ends.

Korsgaard's reading. Korsgaard puts great emphasis on FH. On here account, our ability to set our own ends, i.e., our humanity, confers value upon our specific ends. In other words, the humanity of rational beings is the source of all value in the world. Note that this dovetails with her reading of the good will, on which the good will is the only thing that has intrinsic value (see lecture 1). However, note also that, on Bader's interpretation of the good will, this picture cannot be correct as happiness also has intrinsic value and therefore does not derive its value from some outside source, such as humanity.

3.5 Equivalence

Formula of the Kingdom of Ends

(FKE) "By kingdom, however, I understand the systematic union of several rational beings through common laws. Now, since laws determine ends according to their universal validity, it is possible [...] to conceive a whole of all ends (of rational beings as ends in themselves, as well as the ends of its own that each of them may set for itself) in systematic connection, i.e., a kingdom of ends, which is possible according to the above principle." (4:433)

A moral world. For Kant, every world has (i) a form, (ii) matter, and (iii) constitutes a totality. FUL/FLN demand that we universalise our maxims, guaranteeing that the laws of the moral world are in "systematic union." These laws are the form of the moral world. FH ensures that our ends are consistent, constituting "a whole of all ends." These ends constitute the matter of the moral world. FKE describes a kingdom in which everyone always adheres to FUL/FLN and FH, thus establishing a totality. The kingdom of ends therefore is a moral world. This reading also unlocks the formulas' equivalence:

Equivalence. The above three ways of representing the principle of morality are fundamentally only so many formulae of the salfsame law, one of which of itself unifies the other two within it. However, there is yet a dissimilarity among them, which is indeed subjectively rather than objectively practical, namely to bring the idea of reason closer to intuition [...]. For all maxims have 1) form, which consists in universality [...]; 2)

a matter, namely an end [...]; 3) a complete determination of all maxims by that formula, namely: that all maxims from one's own legislation ought to harmonize into a possible kingdom of ends as a kingdom of nature. (4:436)

My suggestion. Kant thinks that theoretical and practical reason are just two applications of the same faculty. We know from the *Critique of Pure Reason* that theoretical reason demands a systematic unity of all empirical laws. I suggest that practical reason is doing the very same thing in the practical domain, demanding a systematic unity of all maxims, which is done by FUL/FLN. While the matter of nature is guaranteed to be consistent, practical matter (i.e., ends) must be made consistent. This is done by FH. However, since the laws govern the matter, systematising our ends ensures that we systematise our maxims, and vice versa. This establishes the equivalence of the three formulas.

3.6 Universal Happiness

Universal happiness.

> Now in an intelligible world, i.e., in the moral world, in the concept of which we have abstracted from all hindrances to morality (of the inclinations) such a system of happiness proportionately combined with morality can also be thought as necessary, since freedom, partly moved and partly restricted by moral laws, *would itself be the cause of the universal happiness*, and rational beings, under the guidance of such principles, would themselves be the authors of their own enduring welfare and at the same time that of others. But *this system of self-rewarding morality is only an idea, the realization of which rests on the condition that everyone do what he should*, i.e., that all actions of rational beings occur as if they arose from a highest will that comprehends all private choice in or under itself. (A809-10/B837-8; my emphasis)

My reading. A world in which everyone is perfectly moral just is the kingdom of ends. We know from above that everyone's ends in such a kingdom would be consistent. Now, happiness is the total satisfaction of all our inclinations/ends. For this to be possible, however, our ends must be consistent, for otherwise they couldn't be mutually satisfied. The categorical imperative thus aims at universal happiness. What is more we can understand the demands of the categorical imperative as prudential demands of a highest will: suppose that all our ends were the ends of one highest being; this being's prudential demands are our moral demands. We might thus even define morality in terms of

Part II, Paper 3, Kant's Ethics
Christopher Benzenberg (cpb67)
LT5-8, Fridays, 9:30 - 11:00 am
Seminar Room E, 17 Mill Lane

4. Lecture — "Reason, Freedom, and Morality"

Section III (4:446-463)

4.1 Identification vs Justification

A quick recap. In sections I and II of the Groundwork, Kant has identified the supreme principle of morality, i.e., the principle of a pure will. Section I derives a first formulation of the categorical imperative from the concept of duty, "which contains that of the good will, though under certain subjective limitations" (4:397). Section II then develops this first formulation into a full-blown moral theory, distinguishing different formulas of the categorical imperative. Having *identified* the supreme principle of morality, Kant now moves on to *justify* it. But what does it take to justify the supreme principle of morality?

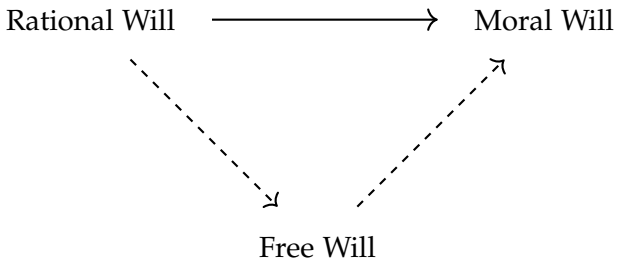
The project of a deduction. Kant claims that he is undertaking a "deduction" of the moral law (4:447, 454, 463). This means he tries to prove that the moral law is universally valid, i.e., valid for everyone. Yet Kant seems to specify this ambition in two different ways: (i) to show that the moral law is valid for all rational beings (4:447), and (ii) that the moral law binds all human beings as a categorical imperative (4:425, 431-2, 445). However, there is no tension. As we will see, Kant is pursuing both of these projects; in a first step, Kant shows that the moral law is valid for all rational beings; in a second step, he then demonstrates that it also binds all humans, i.e., finite rational beings.

4.2 The First Step

Synthetic judgements a priori. Kant claims that the proposition "All rational/good wills are subject to the moral law" is a synthetic judgment a priori (4:447). What does that mean? A priori judgements do not rely on experience and are therefore necessary and strictly universal; empirical judgments are based on experience and are therefore contingent (B2-4). The predicate of analytic judgments is already contained in the subject term, albeit in a "hidden way"; synthetic judgments attribute a predicate that is not already contained in the subject term (B10). Kant thinks that the a priori/empirical and analytic/synthetic distinctions are partially independent of each other, yielding the following table.

	a priori	**empirical**
analytic	*conceptual truths*	—
synthetic	*metaphysical truths*	*empirical truths*

The possibility of synthetic a priori. All three *Critiques* and the *Groundwork* ask the question "How are synthetic judgments a priori possible?" Answering this question is not easy. To establish a synthetic proposition, one cannot simply analyse the subject term but needs a "third thing" (A155/B194), which justifies the synthesis of subject and predicate. But unlike empirical judgements, this third thing cannot be given in experience, but must be given a priori. Kant suggests that we can link "rational willing" and "being under the moral law" by using as a third thing the a priori idea of freedom, or a free will.



The Key Argument

(P1) All free wills are under the moral law, and vice versa (Reciprocity Thesis).
(P2) All rational wills are free.
(C) *Therefore*, all rational wills are under the moral law.

The first premise. Kant establishes the first premise right at the beginning (4:446-7). Freedom, negatively defined, is a type of causation that does not require a preceding external cause. As causation, however, freedom does not mean lawlessness, or arbitrary uncaused choosing. Instead, it must be governed by a law. Since this law cannot be given by nature, it must be given by the will itself. Kant calls this positive conception of freedom autonomy (from the Greek 'αυτο-νομος', literally meaning 'self-law'). Since giving oneself a law is just what the moral law demands, Kant concludes "a free will and a will under moral laws are one and the same." (4:447).

The second premise. Why does Kant think that "Freedom must be presupposed as a quality the will of all rational beings" (4:447)? Kant's answer to this question builds on two premises: (i) everyone who must act under the idea of freedom is really free from a practical point of view; and (ii) rational beings must consider themselves to act under the idea of freedom. Kant concludes that rational beings are really free from a practical point

of view. In other words, all rational wills must also be considered as free wills. Here is the argument in Kant's own words:

> Every being that cannot act except *under the idea of freedom* is just on that account really free from a practical point of view, i.e., that all the laws that are inseparably bound up with freedom hold for it just as if its will were also declared free in itself and in a way valid in theoretical philosophy. Now I assert: That we must necessarily lend the idea of freedom, under which alone it acts, *to every rational being that has a will*. For in such a being we conceive of a reason that is practical, i.e., has causality in respect to its objects. (4:448; my emphasis)

4.3 Possible Problems

The cycle. The above argument raises one big question: Why must rational beings take themselves to act under the idea of freedom? One answer to this question goes like this: We, as rational beings, have pure practical reason, which stands under the moral law. We know that moral actions presuppose that we are free. Hence, we, as practically rational beings, must act under the idea of freedom. This line of argument, Kant suggests, rests on a vicious "cycle" (4:450, 453) because it presupposes that all rational willing is under the moral law, which, however, is what Kant wants to prove.

Solution. To avoid the cycle, Kant argues that freedom is not only presupposed as a necessary condition for practical reasoners, but for reasoners in general, including theoretical reasoners. When we act and judge, we become aware of our noumenal self and realise that we have reason "as pure self-activity" (4:452). To impute our actions and judgements to us, we must think of them as results of a free reason rather than some external natural cause. That's why, it is not just practical reason, but theoretical reason too, which allows us to infer that rational willing is free. This blocks the cycle.

Noumenal Ignorance. The above argument has received a lot of heat in the secondary literature. Kant seems to suggest that, by being attentive to our rational nature, we gain a positive insight into our noumenal existence outside of space and time. Yet in the *Critique of Pure Reason*, Kant explicitly argues against this. We can only ever establish the existence of noumena in the negative sense, but we have no knowledge of noumena positively defined (A252). Kant's argument in the *Groundwork* thus seems to step beyond the boundaries which Kant has so carefully erected in the first *Critique*.

Solution. To be clear, this objection has been a thorny issue in the secondary literature for decades and there won't be an easy fix. That said, I think that Kant does not transgress the boundaries of *knowledge* because he only establishes rational freedom as (doctrinal) *belief*, not as knowledge. We know from the second *Critique* that freedom is a postulate of practical reason; I suggest that, in the Groundwork, Kant introduces freedom as a postulate of reason in general or theoretical reason. Postulates are beliefs and have a weaker epistemic standing than knowledge. That's why they are epistemically legitimate.

4.4 The Second Step

Human obligation. Thus far, Kant has shown that all rational willing stands under the moral law; Kant now argues that we humans, as rational beings, also stand under the moral law. Yet this inference requires more than a simple modus ponens. As humans, we are not just rational, but also finite and thus don't necessarily adhere to the moral law. In particular, Kant thinks that we must overcome two challenges to justify the moral law as a categorical imperative for us: (i) we must show how we can act in accordance with the moral law; and (ii) we must show why the moral law *binds* us.

Kant's compatibilist incompatibilism. As finite beings, we humans are part of the natural world, which is determined by cause and effect. In such a world, freedom seems utterly impossible. Kant, drawing on his discussion of the third antinomy, resolves this tension by arguing that we are citizens of two worlds, the natural and the noumenal world. Only the natural world is thoroughly determined; the noumenal world has place for freedom (455-8). While there is no logical tension in saying *that* we are both determined and free, Kant thinks that we can never understand *how* this is possible (4:459-63).

The normative question. Even if we are able to follow the moral law, we must still answer the sceptic's question, "Why should I subjugate myself to that principle?" (4:449; see also 4:450). Kant answers this question in two steps: (i) even the "most vile villain" must see that, as rational being, he is part of the intelligible world and so that his will is both free and moral (4:454); and (ii) our "authentic self" (4:457, 461) is not our sensible but our intelligible self, because "what belongs to mere appearance is necessarily subordinated to the constitution of the thing in itself by reason" (4:461; see also 4:453-4). We are bound by the moral law, not by inclinations, because our noumenal identity takes priority.

4.5 An Alternative Deduction

The fact of reason. Kant revises the deduction of the moral law in the *Critique of Practical Reason*; while he doesn't explain this change, most think that he tries to avoid the epistemic concerns discussed in 4.3. In the second Critique, Kant simply posits that "the moral law is given, as it were, as a fact of pure reason of which we have a priori consciousness" (5:47). Kant seems to think that we are directly aware of the moral law as universally binding. He further blocks the "cycle" by distinguishing between freedom as the "*ratio essendi*" of morality, and morality as the "*ratio cognescendi*" of freedom (5:4n). The changes suggest that Kant himself was not entirely happy with his argument in section III.