IV*—THE VOICE OF CONSCIENCE[1]

by J. David Velleman

ABSTRACT I reconstruct Kant's derivation of the Categorical Imperative (CI) as an argument that deduces what the voice of conscience must say from how it must sound—that is, from the authority that is metaphorically attributed to conscience in the form of a resounding voice. The idea of imagining the CI as the voice of conscience comes from Freud; and the present reconstruction is part of a larger project that aims to reconcile Kant's moral psychology with Freud's theory of moral development. As I reconstruct it, Kant's argument yields an imperative commanding us to act for reasons whose validity we can consistently will to be common knowledge among all agents. Universalizing a maxim thus turns out to consist in willing, not that there be some universally quantified rule of conduct, but rather that a principle of practical reasoning be common knowledge—as a principle of reasoning ought to be.

How do you recognize the voice of your conscience? One possibility is that you recognize this voice by what it talks about—namely, your moral obligations, what you morally ought or ought not to do. Yet if the dictates of conscience were recognizable by their subject matter, you wouldn't need to think of them as issuing from a distinct faculty or in a distinctive voice. You wouldn't need the concept of a conscience, any more than you need concepts of distinct mental faculties for politics or etiquette. Talk of conscience and its dictates would be like talk of the mince-pie syllogism, in that it would needlessly elevate a definable subject matter to the status of a form or faculty of reasoning.[2]

1. In writing this paper I have drawn on conversations and correspondence with Marcia Baron, Jennifer Church, Stephen Darwall, David Hills, David Phillips, and Connie Rosati. Work on this paper has been supported by a sabbatical leave from the College of Literature, Science, and the Arts, University of Michigan; and by a fellowship from the National Endowment for the Humanities.

2. The mince pie syllogism was the ironic invention of Elizabeth Anscombe. Anscombe objected to the notion that the practical syllogism was merely a syllogism on a practical topic, such as what one ought to do. She argued that if there were a distinct logical form for reasoning about what one ought to do, then there might as well be distinct forms for reasoning about every definable topic, including mince pies. (*Intention* [Ithaca, NY: Cornell University Press, 1957], 58.)

*Meeting of the Aristotelian Society, held in Senate House, University of London, on Monday, 23rd November, 1998 at 8.15 p.m.

Our having the concept of a conscience suggests, on the contrary, that ordinary practical thought does not contain a distinct, moral sense of 'ought' that lends a distinct, moral content to some practical conclusions. The point of talking about the conscience and its voice is precisely to mark a distinction among thoughts that are not initially distinguishable in content. Among the many conclusions we draw about what we ought or ought not to do, some but not others resonate in a particular way that marks them as dictates of conscience. The phrase 'morally ought' is a philosophical coinage that introduces a difference of sense where ordinary thought has only a difference of voice—whatever that is.

But what is it? Conscience doesn't literally speak. The idea of its addressing you in a voice is thus an image, albeit an image that may infiltrate your experience of moral thought and not just your descriptions of it. Yet whether the dictates of conscience are somehow experienced as spoken or are just described as such after the fact, this image must represent something significant about them, or it wouldn't be used to identify them as a distinctive mode of thought. The question is what literal feature of these thoughts is represented by the image of their being delivered in a voice.

The answer, I think, is that the dictates of conscience carry an authority that distinguishes them from other thoughts about what you ought or ought not to do.[3] The voice of conscience is, metaphorically speaking, the voice of this authority. To recognize an 'ought' as delivered in the voice of conscience is to recognize it as carrying a different degree or kind of authority from the ordinary 'ought', and hence as due a different degree or kind of deference.

If the voice of conscience does represent a distinctive authority that accompanies some practical conclusions, then it is more than a curiosity of moral psychology: it symbolizes a fundamental feature of morality, regarded by some philosophers as *the* fundamental feature. Kant, in particular, thought that what morality requires can be deduced from the authority that must accompany its requirements. If Kant had written in the imagery of conscience, he might have put it like this: by reflecting on how the voice of

3. The authority of conscience is the central theme of Butler's *Sermons*. For a recent discussion of Butler, see Stephen Darwall, *The British Moralists and the Internal 'Ought' 1640–1740* (Cambridge: Cambridge University Press, 1995), Chapter 9.

conscience must *sound,* you can deduce what it must *say*—whereupon you will have heard it speak.

Of course, Kant didn't formulate his moral theory in these terms, but I think that they can be substituted for terms such as 'duty' and 'moral law' in Kant's own formulations, with some gain in clarity and persuasiveness for modern readers. My goal is to reconstruct Kant's categorical imperative in the terms of conscience and its voice.[4]

The idea of reconstructing the categorical imperative as the voice of conscience originated with Freud. Freud was interested in the voice of conscience because he thought that it could explain why paranoiacs heard voices commenting on their behaviour;[5] and that it could in turn be explained by the psychological origins of conscience in parental discipline 'conveyed... by the medium of the voice'.[6] In tracing conscience to the voice of parental discipline, Freud also thought that he could explain why its power 'manifests itself in the form of a categorical imperative.'[7] This explanation showed, according to Freud, that 'Kant's Categorical Imperative is... the direct heir of the Oedipus complex'.[8]

My view, which I cannot defend here,[9] is that the categorical imperative can indeed be identified with the super-ego, at least in one of its guises. For I think that the categorical imperative is what Freud would call an ego ideal. The ego ideal, in Freudian theory, is that aspect of the super-ego which represents the excellences of parental figures whom the subject loved and consequently

4. There is at least one passage in which Kant uses the word 'conscience' in reference to the activity of applying the categorical imperative: *Groundwork of the Metaphysic of Morals,* trans. H.J. Paton (New York: Harper and Row, 1964), 89 (422). (Page numbers in parentheses refer to the Prussian Academy Edition.)

5. 'On Narcissism: An Introduction', in *The Standard Edition of the Complete Psychological Works of Sigmund Freud,* Vol. 14, ed. James Strachey (London: The Hogarth Press, 1957), 69–102, at 95. See also *Group Psychology and the Analysis of the Ego, S.E.* 18: 67–143, at 110 [53]; *New Introductory Lectures on Psycho-Analysis, S.E.* 22: 3–182, at 59 [74]. (Page numbers in brackets refer to the Norton paperback volumes of individual works from the Standard Edition.)

6. 'On Narcissism', 14: 96.

7. *The Ego and the Id, S.E.* 19: 3–66, at 35, 48 [31, 49]. Freud also uses this phrase in *Totem and Taboo, S.E.* 13: ix–162, at 22.

8. 'The Economic Problem of Masochism', *S.E.* 19: 156–70, at 167. Freud also identified the super-ego with the Kantian 'moral law within us' (*New Introductory Lectures,* 22: 61, 163 [77, 202]).

9. But see 'The Direct Heir of the Oedipus Complex' (MS).

idealized when he was a child.[10] Although Kant often framed the categorical imperative as a rule for the will to follow, I think that it is better understood as an ideal for the will to emulate, in that it describes an ideal configuration of the will itself. And I think that this ideal could indeed be internalized from parental figures as they appear to the eyes of a loving child.

This conception of the categorical imperative as an ego ideal will reappear at the end of this paper, but it is not my immediate concern. What concerns me here is Freud's suggestion that the categorical imperative can be identified with the voice of conscience.

The image of conscience as having a voice is potentially misleading in one respect. Taken literally, the image may lead us to think of conscience as an external intelligence whispering in our ears, like Socrates's *daimon.* Even when taken figuratively, the image still suggests that the dictates of conscience occur to us unbidden, as thoughts that we don't actively think for ourselves, and hence as external to us, in the sense made familiar by the work of Harry Frankfurt.[11]

Conscience is most likely to seem external in this sense when it opposes temptation: conscience and temptation can seem like parties to a dispute on which we sit as independent adjudicators. Yet even this judicial image is misleading, since the disputing parties do not appear as distinct from ourselves. We ourselves play each role in the mental courtroom, now advocating the case of temptation, now that of conscience, representing each side *in propria persona.* In short, we vacillate—which entails speaking in different voices, not just hearing them.

Thus, hearing the voice of our conscience is not really a matter of hearing voices. It's rather a matter of recognizing a voice in which we sometimes speak to ourselves.

Freud's theory of the super-ego might seem to favour the image of conscience as an independent agency, distinct from and in opposition to the self. Freud certainly thought that in cases of

10. Freud's views on the relation between super-ego and ego ideal are clearly summarized in Joseph Sandler, Alex Holder, and Dale Meers, 'The Ego Ideal and the Ideal Self', 18 *The Psychoanalytic Study of the Child* 139–58 (1963). See also Joseph Sandler, 'On the Concept of the Superego', 15 *The Psychoanalytic Study of the Child* 128–62 (1960).

11. *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), especially chapters 2, 5, 7, and 12.

mental illness, the super-ego could become the source of voices heard involuntarily, and hence from outside the self in Frankfurt's sense.[12] Yet in the normal subject, the super-ego bears an ambiguous relation to the self. It is 'a differentiating grade in the ego',[13] and the process of introjection by which it is formed is a way of identifying with other people, which is necessarily a deployment of the self. So another description of what happens when the super-ego addresses the ego is that the self identifies with others in addressing itself.

Although Kant doesn't tend to speak of the conscience *per se,* his moral philosophy also reflects the complexity of its relation to the self. On the one hand, Kant says that the moral law is necessary and inescapable; on the other hand, he describes it as a law that we give to ourselves. For Kant, giving ourselves the moral law represents both our exercise of an autonomous will and our subjection to a necessity larger than ourselves; just as, for Freud, conscience is the ego addressing itself in the voice of external authority.[14]

This analogy reveals what is right about Freud's claim that the voice of the super-ego is the voice of Kant's categorical imperative. The necessity to which we submit in the law that we give to ourselves can be imagined as the authority we recognize in a voice with which we address ourselves—namely, the voice of conscience. I want to show that Kant's attempt to derive the content of the moral law from the very concept of its practical necessity can be restaged as an attempt to derive the words of conscience from the authoritative sound of its voice.

An example of rational authority. The first step in this reconstruction of Kantian ethics is to analyze the authority that Kant would attribute to the conscience. Whereas Freud thought of the conscience as the seat of internalized parental authority, Kant would think of it—if he thought in such terms at all—as a seat of rational authority. But what sort of authority is that?

12. See the passages cited in note 5, above.

13. This is the title of Chapter XI of *Group Psychology.*

14. Kant seems to reject the image of an external voice of conscience at *Groundwork* 93 (425–26), where he insists that moral philosophy cannot serve 'as the mouthpiece of laws whispered to her by some implanted sense or by who knows what tutelary nature...'.

Consider, by way of analogy, the authority of cognitive judgments whose propositional content is self-evidently true. You make such a judgment, for example, when you confirm for yourself that 2+2=4. To say that such a judgment is authoritative is to say that it merits deference. But why should anyone defer to your judgment on matters of elementary arithmetic?

The answer is not that you're especially well positioned to think about such matters. When it comes to adding 2 and 2, all thinkers are in the same position. But for that very reason, a computation performed by you here and now can take the place of anyone's, including your own on future occasions. That is, you can compute the sum of 2 and 2 *once and for all,* in that you would only compute it similarly in the future; and you can also compute it *one for all,* in that others would only compute it similarly, too. Your judgment is thus authoritative because it can serve as proxy for anyone's, including your later selves'. To see yourself as judging authoritatively is to see yourself as judging for all in this sense—in the sense, that is, of judging as anyone would.

But what if your judging as anyone would were, in turn, a matter on which judgments might differ? In that case, your arithmetic judgment might only seem authoritative to you. Surely, however, you recognize your judgment as having an authority that anyone would recognize. You must therefore see yourself as judging, not just as anyone would, but as anyone would judge that anyone would.

And now an infinite regress rears its head. For what if judgments could differ as to whether you were judging as anyone would judge that anyone would—and so on? Fortunately, there is independent reason to expect such a regress in the present context, and also to regard it as benign.

The reason is that the facts of elementary arithmetic are common knowledge among those who consider them, and common knowledge involves a regress of the present form. Anyone who adds 2 and 2 sees, not just that it's 4, but also that anyone who added 2 and 2 would see that it's 4, and that such a person would see this, too, and so on. The facts of elementary arithmetic are like objects in a public space, where everyone sees whatever everyone else sees, and everyone sees everyone else seeing it. Unlike publicly visible objects, however, the facts of arithmetic are common knowledge

among all possible thinkers rather than a finite population of actual viewers.

As a participant in this common knowledge, you have higher-order knowledge about the judgments of all other thinkers, and about their judgments about the judgments of all. This higher-order knowledge constitutes a perception of authority in your own judgment that 2+2=4, since it represents this judgment as that which anyone would think, and would think that anyone would think, and so on.

So it's just as we might have expected: the voice of authority is the one with the reverb. But now we know the source of the reverberations. A judgment resounds with authority when it is perceived as echoing and re-echoing in the minds of all other thinkers, as it does when its content is a matter of common knowledge.

This authority attaches, as we have seen, to items of *a priori* knowledge, such as the judgment that 2+2=4. Items of *a priori* knowledge would seem to be the only bearers of this authority, in fact, since only the *a priori* can be regarded as what anyone would think, or be thought to think, and so on.

The authority of the moral law. I suspect that the form of common knowledge among all thinkers—of that which anyone would think, and would think that anyone would think, and so on—is the form that Kant attributes to the moral law in calling it universal. Of course, Kant thinks that the moral law is universal in the sense that it applies to all rational creatures; and the most economical way of representing a universally applicable law is with a universal quantifier, as in 'All rational creatures must keep their promises' or 'No rational creature may lie'. But serious problems, both textual and philosophical, stand in the way of reading Kant's talk of universal law as referring to universally quantified rules.

Consider, to begin with, these two passages from the *Groundwork:*

> Everyone must admit that a law, if it is to hold morally—that is, as the ground of an obligation—must carry with it absolute necessity; that the command 'Thou shalt not lie' does not hold just for men, without other rational beings having to heed it, and similarly with all the other genuine moral laws; and that consequently the ground

> of obligation here must be sought, not in the nature of man or in the circumstances of the world where he is located, but solely *a priori* in the concepts of pure reason.[15]
>
> It may be added that unless we wish to deny to the concept of morality all truth and all relation to a possible object, we cannot dispute that its law is of such widespread significance as to hold, not merely for men, but for all *rational beings as such*—not merely subject to contingent conditions and exceptions, but *with absolute necessity*.... And how could laws for determining *our* will be taken as laws for determining the will of a rational being as such—and only because of this for determining ours—if these laws were merely empirical and did not have their source completely *a priori* in pure, but practical reason?[16]

These passages are central to the *Groundwork,* because they introduce the conceptual connections among morality, universality, and the *a priori*—the connections through which Kant hopes to derive the content of the categorical imperative from the very concept of morality. The passages argue that the concept of morality entails that its laws carry 'absolute necessity'; which entails that they hold not only for men but for all rational creatures; which entails that they hold *a priori.*

Suppose that we interpret this argument as using the word 'laws' to denote general rules, and as contrasting rules that quantify over men with rules that quantify over rational creatures. We must then wonder why the former rules are any less necessary than the latter, since the former apply necessarily to anything insofar as it is a man, just as the latter apply necessarily to anything insofar as it is rational, and either represent some conduct as necessary for the relevant agents. 'All men must keep their promises' and 'All rational creatures must keep their promises' would seem to be equally necessary, each within its specified domain. We may also wonder why the concept of morality calls for laws of the latter form. Couldn't there be a distinctively human morality, in which 'All men must keep their promises' would count as a law? Finally, we may wonder why such a law could not follow *a priori* from the concept of a man, just as a rule quantifying over rational creatures might follow from the concepts of reason and rationality.

15. *Groundwork* vi (389), my translation. For reasons that will be explained below, I have brought this passage into conformity with Paton's translation of the following passage, in which 'gelten für' is translated as 'hold for'.

16. *Groundwork* 76 (408).

Note, however, that Kant's example of absolute necessity is not a general rule that quantifies over all rational creatures. His example is rather a second-person command, 'Thou shalt not lie'. And what Kant says about such a requirement is not that it must refer to all rational creatures but that it must 'hold for' them—an expression that he repeats throughout the *Groundwork,* as we shall see.

Of course, the pronoun in 'Thou shalt not lie' might be standing in for a universal quantifier, and what's at issue could be the domain of that implicit quantifier. Yet if the issue were whether 'thou' referred to all men or to all rational creatures, then Kant wouldn't ask for whom the rule holds. The rule, fully spelled out, would be either '(All) thou (men) shall not lie' or '(All) thou (rational creatures) shall not lie', and in either case it would have to hold or not hold, without limitation. 'All men shall not lie' cannot hold only locally or selectively, any more than 'All men are mortal'.

Suppose, however, that 'Thou shalt not lie' were a type of which various tokens were addressed to various agents, with corresponding variance in the reference of the pronoun. Commands of this type could be said to 'hold for' particular agents in two related senses: they might be authoritative from the perspectives of particular agents as addressees, and they might consequently be valid in application to those agents. To ask for whom the rule holds would be to ask who finds himself addressed by an authoritative command of this type.

According to this interpretation, Kant isn't thinking of moral requirements as universally quantified rules; he's thinking of them as personally addressed practical thoughts, of the form 'Thou shalt not lie'. We can now extend the interpretation so as to explain Kant's chain of inferences.

For suppose, next, that when Kant insists on the 'absolute necessity' of moral requirements, he means that the corresponding thought must be absolutely authoritative from the perspective of the addressee: an agent should not be able to exempt himself from the force of such a thought. Absolute necessity, so understood, can indeed be said to follow from the very concept of a moral requirement. So we have accounted for the first link in Kant's chain.

Now suppose that 'Thou shalt not lie' would be absolutely authoritative, in the requisite sense, if and only if it were what any agent would think to himself upon considering whether to lie, and

would think that any agent would think, and so on. If it were such a thought, then an agent considering whether to lie would not only think to himself 'Thou shalt not lie' but would also think of himself as *having nothing else to think,* because this thought would strike him as what anyone would think on the subject, including himself on other occasions. He would therefore think of the question as having been settled once and for all—or, in other words, authoritatively. By contrast, if 'Thou shalt not lie' weren't such a thought, then even an agent who thought it would regard it as optional, there being other things that anyone, including himself, might think on the subject. He would therefore find it lacking in authority. Here is a sense in which the absolute authority entailed in the very concept of moral requirements can be seen to consist in their 'holding for' all rational agents—that is, by constituting what anyone would think, or would think that anyone would think, and so on. We have now accounted for the second link in Kant's chain.[17]

The third link follows without further suppositions. The form of what anyone would think, and would think that anyone would think, and so on—the form, if you like, of that than which there is nothing else *to* think—is the form of *a priori* knowledge. When it attaches to a thought such as 'Thou shalt not lie', it yields a thought that is simultaneously *a priori* and practical. Hence the very concept of a moral requirement can be seen to entail an absolute authority that is found only in *a priori* practical thought.[18] Kant's argument is now complete.

I have embroidered this interpretive hypothesis on two mere swatches of text. How it will look against the broader fabric of Kantian ethics remains to be seen. First, however, I want to register an important qualification.

My hypothesis is that moral laws, for Kant, are not universally quantified rules but rather personally addressed practical thoughts, whose universality and authority both consist in their being what anyone would think, and would think that anyone would think, and so on. Yet if 'Don't lie' is universal in this sense, then everyone in the relevant circumstances will find himself with nothing else to

17. See also *Groundwork* 92–3 (425): '[D]uty has to be a practical, unconditioned necessity of action; it must therefore hold for all rational beings...'.

18. See *Groundwork* 93 (426): 'These principles must have an origin entirely and completely *a priori* and must at the same time derive from this their sovereign authority....'

think; and if everyone in the relevant circumstances finds himself with nothing else to think but 'Don't lie', then there will, in effect, be a universal rule of not lying.

For this reason, my hypothesis cannot be that moral laws, for Kant, aren't universally quantified rules at all; it must be that they aren't universally quantified rules in the first instance. Moral laws, as I understand them, can be expressed in universally quantified rules, provided that those rules are understood as expressing the authority of personal practical thoughts, whose authority just consists in their being what anyone would think that anyone would think.

Let me emphasize, then, that I do not mean to ignore or dismiss the many passages in which Kant himself enunciates laws as universally quantified rules of behaviour. I merely suggest that the universal rules enunciated by Kant should be understood as summaries of something more complex, or as the outer surfaces of something deeper—namely, a state of affairs in which practical thoughts, in personal form, are common knowledge among all agents.

How universalization works. With this qualification in mind, I want to apply my interpretive hypothesis to Kant's account of universalization, the procedure by which maxims are tested under the categorical imperative. Here, too, the hypothesis helps to resolve both textual and philosophical problems.

Consider this instance of universalization:[19]

> [A person] finds himself driven to borrowing money because of need. He well knows that he will not be able to pay it back but he sees too that he will get no loan unless he gives a firm promise to pay it back within a fixed time. He is inclined to make such a promise; but he has still enough conscience to ask 'Is it not unlawful and contrary to duty to get out of difficulties in this way?' Supposing, however, he did resolve to do so, the maxim of his action would run thus: 'Whenever I believe myself short of money, I will borrow money and promise to pack it back, though I know that this will never be done.' Now this principle of self-love or personal advantage is perhaps quite compatible with my own entire future welfare; only there remains the question 'Is it right?' I therefore transform the demand of self-love into a universal law and frame

19. *Groundwork* 90 (422). I have brought Paton's version of this passage into conformity with his translation of the preceding passage, by rendering 'gelten' as 'to hold'. (See note 15, above.)

> my question thus: 'How would things stand if my maxim became a universal law?' I then see straight away that this maxim can never hold as a universal law of nature and be self-consistent, but must necessarily contradict itself. For the universality of a law that every one believing himself to be in need can make any promise he pleases with the intention not to keep it would make promising, and the very purpose of promising, itself impossible, since no one would believe he was being promised anything, but would laugh at utterances of this kind as empty shams.

The target of universalization in this passage is what Kant calls a maxim of action: 'Whenever I believe myself short of money, I will borrow money and promise to pay it back, though I know that this will never be done.' We might think that the way to make this maxim universal is to replace the first-person pronoun with quantified variables ranging over all rational creatures.[20] Kant seems to suggest such a procedure when he refers to 'the universality of a law that every one believing himself to be in need can make any promise he pleases...'. But Kant also suggests a different procedure, when he considers whether his maxim itself 'can... hold as a universal law'. Kant's maxim is framed in the first person, and so it—the maxim itself—can 'hold' as a universal law only if first-personal thoughts can somehow be universal.

Kant's framing his maxim in the first person is no accident. He could not have restated it, for example, as 'Immanuel Kant will make lying promises when he is in need'. Such a third-personal thought would not be a maxim of action, since it could not be acted upon by the thinker until he reformulated it reflexively, in the first person. Insofar as the target of universalization is a practical thought, it is essentially first-personal.[21]

20. For an interpretation of universalization along these lines, see e.g. Onora O'Neill, *Acting on Principle: an Essay on Kantian Ethics* (New York: Columbia University Press, 1975), esp. Chapter Five, 59–93; and 'Consistency in Action', in *Constructions of Reason; Explorations of Kant's Practical Philosophy* (Cambridge: Cambridge University Press, 1989), 81–104. See also Christine Korsgaard, 'Kant's Formula of Universal Law', in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 77–105. According to Korsgaard, universalization 'is carried out by imagining, in effect, that the action you propose to perform in order to carry out your purpose is the standard procedure for carrying out that purpose' (92). In the present case, then, the agent 'imagines a world in which everyone who needs money makes a lying promise and he imagines that, at the same time, he is part of that world, willing his maxim' ('Kant's Analysis of Obligation: the Argument of *Groundwork I*', in *ibid.*, 43–76, at 63). Finally, see Roger J. Sullivan, *Kant's Moral Theory* (Cambridge: Cambridge University Press, 1989), 168–69.

21. On this topic, see John Perry, *The Problem of the Essential Indexical and Other Essays* (New York: Oxford University Press, 1993).

This first-personal thought should remind us of the second-personal injunction considered above, 'Thou shalt not lie', which was there regarded as being addressed by the agent to himself. So regarded, 'Thou shalt not lie' was couched in what might be called the reflexive second-person—the second-person of talking to oneself. And when it is thus addressed to oneself, 'Thou shalt not lie' is just the contradictory of 'I shall lie', the maxim that is currently up for universalization. Our earlier reflections on how the second-personal injunction could be a universal law are thus directly relevant to the universalization of the first-personal maxim.

As before, we might consider transforming the maxim into a universal law, by substitution of a quantifier for the first-person pronoun. But Kant speaks more often of maxims' *being* laws themselves than of their being *transformed into* laws. In addition to asking whether a maxim can 'hold as a universal law',[22] he asks: whether maxims can 'serve as universal laws',[23] whether they have 'universal validity... as laws'[24] or 'the universality of a law';[25] whether a maxim 'at the same time contains in itself its own universal validity for every rational being'[26] or is constrained 'by the condition that it should be universally valid as a law for every subject';[27] whether it 'can have for its object itself as at the same time a universal law'[28] or can 'have as its content itself considered as a universal law'.[29] All of these expressions call for a single thought to be regarded simultaneously as the maxim of one agent and as a law for all.

According to my interpretation, however, a single thought can simultaneously be a first-personal maxim and a universal law, if it is what anyone would think in response to the relevant practical question, and would think that anyone would think, and so on. It is then a type of thought whose tokens would be authoritative for any agent. And imagining that 'I will make false promises' would be

22. Also at 103–4 (438).
23. 94 (426).
24. 126 (458); see also 129 (461).
25. 128 (460).
26. 105 (437–38).
27. 105 (438).
28. 114 (447).
29. 115 (447).

authoritative for anyone is a way of imagining a universal law of making false promises.

This interpretation explains how an individual maxim can 'have as its content itself considered as a universal law'[30] or 'contain in itself its own universal validity for every rational being'.[31] Universalizing a first-personal maxim ('I will make false promises') is not, in the first instance, a process of conjoining it with some universally quantified variant of itself ('Everyone will make false promises'). Universalizing this maxim is rather a matter of regarding the maxim itself as what anyone would think, or would think that anyone would think, and so on. The universalized maxim is more like this—'Obviously, I will make false promises'—where 'obviously' indicates that the following thought would occur to anyone, as would occur to anyone, and so on. That's how a first-personal maxim can contain its own universal validity within itself.

Kant says that a universal law of making false promises would have the result that 'no one would believe he was being promised anything, but would laugh at utterances of this kind as empty shams'. If we think of this law as a universally quantified rule, to the effect that everyone may or will make false promises when in need, then we shall have to wonder why it would have the results predicted.

The answer might be that people's adherence to such a law would entail the issuance of so many false promises that everyone would eventually learn to distrust everyone else.[32] But this answer would be a piece of empirical reasoning, about how social interactions would evolve in response to a particular pattern of conduct; whereas Kant says that the requirements of morality must be derivable *a*

30. 115 (447).

31. 105 (437–38).

32. For this interpretation, see O'Neill, 'Universal Laws and Ends-in-Themselves', in *Constructions of Reason*, 126–44, at 132: 'The project of deceit requires a world with sufficient trust for deceivers to get others to believe them; the results of universal deception would be a world in which such trust was lacking, and the deceiver's project was impossible.' See also Korsgaard, 'Kant's Formula of Universal Law', 92: 'The efficacy of the false promise as a means of securing the money depends on the fact that not everyone uses promises this way. Promises are efficacious in securing loans only because they are believed, and they are believed only if they are normally true.' Finally, see Sullivan, *Kant's Moral Theory*, 171: 'Truthful assertions cannot survive any universal violation of the essential point of such speech. Once everyone lies for what each considers a "good" reason, we can never know when any verbal behavior counts as "telling the truth".'

priori. This piece of empirical reasoning would therefore be out of place in the process of universalization, by which the specific requirements of morality are derived.

What's more, the same empirical reasoning wouldn't apply to a law licensing promises whose falsity would go undetected, since the proliferation of undetectably false promises would not undermine people's trust; yet Kant reaches the same conclusion about a law of undetectable falsehoods. He imagines a case in which 'I have in my possession a deposit, the owner of which has died without leaving any record of it'. Moral reflection in these circumstances raises the question 'whether I could... make the law that every man is allowed to deny that a deposit has been made when no one can prove the contrary'. Kant's conclusion is 'that taking such a principle as a law would annihilate itself, because its result would be that no one would make a deposit'.[33] This conclusion cannot be an empirical prediction of what would happen under a universally quantified rule of denying unrecorded deposits. General adherence to such a rule would not in fact discourage prospective depositors, precisely because there would be no record of the deposits involved.

In my view, however, the way to imagine a universal law of denying unrecorded deposits is to imagine that the maxim 'I will deny unrecorded deposits' is authoritative, in that it is what anyone would think, and would think that anyone would think, and so on. This law would indeed undermine the faith of prospective depositors—not empirically, through the pattern of conduct it produced; but rationally, through the *a priori* practical thinking that it embodied, which would be common knowledge among all agents. *No one would make unrecorded deposits if stealing them were all there was to think of doing with them.*

If the maxim of denying unrecorded deposits were a law in this sense, then the authority of that maxim would be evident to prospective depositors no less than it was to their intended trustee, since the maxim would be what anyone would think that anyone

33. *Critique of Practical Reason*, trans. by Lewis White Beck (Indianapolis: Bobbs Merrill, 1956), 27 (27). The same case appears, with embellishments, in the essay 'On the Proverb: That may be True in Theory, But Is of No Practical Use', in *Perpetual Peace and Other Essays*, trans. Ted Humphrey (Indianapolis: Hackett Publishing, 1983), 61–92, at 69–70 (286–287).

would think. Depositors would only have to reason about the case from the perspective of their trustee in order to see what his maxim for dealing with their deposits would be, since there would be nothing else to think of doing with them. That the trustee would deny having received their deposits isn't something that depositors would have learned from past experience of his or anyone else's behaviour; it's something that would be evident to them through their own practical reasoning, as proxy for his. They would consequently be deterred from making unrecorded deposits.

This interpretation simply assumes that the connections fundamental to Kant's conception of morality—the connections among universality, necessity, and the *a priori*—hold for all of the laws involved in universalization, including: (1) the categorical imperative, in which the procedure of universalization is prescribed; (2) the specific requirements derived by means of that procedure; and, finally but crucially, (3) the laws imagined within it. In this last instance, imagining one's maxim to be a universal law must entail imagining it to have all three connected properties—that is, to be universally inescapable *a priori*. Hence universalization is a procedure of imagining one's maxim to constitute practical but *a priori* and hence common knowledge.

The nature of maxims. Thus far I have avoided inquiring into the nature of maxims, choosing instead to work with simple expressions of intent, such as 'I'll make false promises' or 'I'll deny unrecorded deposits'. Now that I have offered an hypothesis as to how maxims are universalized, however, I can no longer avoid the question of what they are and, more importantly, why they might be subject to such a procedure. And I don't think that maxims are simply intentions or expressions of intent.

Kant says that maxims are 'principles of volition'.[34] Many interpreters have noted that Kant usually formulates maxims of action so as to specify both a type of behaviour and a purpose to be served by it—or, in other words, an end as well as a means.[35] I think that maxims so often connect end and means, and do so in the form

34. *Groundwork* 68 (400).

35. See O'Neill, *Acting on Principle*, 37–38; Korsgaard, 'Kant's Analysis of Obligation', 57–58, and *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), 108.

of general principles, because they state the connection between reasons and action.[36]

Consider again the maxim of a lying promise: 'Whenever I believe myself short of money, I will borrow money and promise to pay it back, though I know that this will never be done'. I interpret this maxim to mean that financial need is a reason for promising to return a loan, and that this reason outweighs the countervailing consideration that the promise would be false. The maxim is thus a principle of volition in the sense that it licenses a practical inference, from the premises 'I need money' and 'I'd be lying if I promised to repay a loan', to the conclusion 'I'll promise to repay a loan'. The license for this inference is framed as a general principle because the validity of an inference-type cannot vary from one token to another.

More importantly, the validity of an inference is a logical relation that must be recognizable *a priori*. That's why a maxim is naturally subject to the test of universalization. If there is a valid inference from 'I need money' to 'I'll make a false promise', then the validity of that inference must be such as anyone would recognize, and would recognize that anyone would recognize, and so on. The validity of a practical inference, like the validity of *modus ponens*, must hold for—and be common knowledge among—all thinkers.

In this case, the inference can't be valid, precisely because its validity would have to be common knowledge, which would undermine a presupposition of the inference itself—namely, that making false promises is a means of getting money.[37] If it were common knowledge that a decision to make false promises followed from a need for money, then nobody would lend on the basis of promises; promises wouldn't be a means of getting money; and a decision to make them would no longer follow. Thus, if 'I'll make false promises' did follow from 'I need money', then it

36. See Korsgaard, 'An Introduction to the Ethical, Political, and Religious Thought of Kant', in *Creating the Kingdom of Ends*, 3–42, at 13–14: 'Your maxim must contain your reason for action: it must say what you are going to do, and why'; 'Kant's Analysis of Obligation', 57: 'Your maxim thus expresses what you take to be a reason for action.' I am inclined to put a slightly finer point on this claim, by saying that the maxim states the rule of practical inference, from reason to action.

37. Here I follow what Korsgaard calls 'the practical contradiction interpretation' ('Kant's Formula of Universal Law', 92). I differ from Korsgaard, however, in tracing the practical contradiction to an imagined piece of common knowledge rather than an imagined standard practice. (See note 20, above.)

wouldn't follow, after all; and so it doesn't follow, to begin with. A desire for money isn't a valid reason for making false promises.

Its not being a reason is also *a priori.* And this point provides the most challenging twist in Kant's argument. Kant thought that we cannot wait passively to receive practical dictates with *a priori* authority, and hence that we cannot wait for the voice of conscience to speak.[38] We have to propose our own practical dictates and ask whether they could possibly carry *a priori* authority. And sometimes, when the answer is no, *that answer* turns out to carry the sought-for authority: *it* resounds with the voice of conscience.

The practical dictate in the present example is the maxim that making a false promise follows from circumstances of financial need. That the validity of this inference must be *a priori* is itself *a priori,* since validity is a matter of rationality, which is common to all thinkers. From the *a priori* requirement that the validity of an inference must be *a priori,* the impossibility of a valid inference from financial need to false promises follows *a priori* as well. Anyone can see, and can see that anyone can see, that the validity of this inference would have to be *a priori,* but that one of the inference's presuppositions would then be false, so that the inference wouldn't be valid, after all. The fact that the validity of such an inference would have to be common knowledge, which would invalidate the inference—this fact is itself common knowledge among all who care to reflect on the matter. So when the question is whether a need for money is a reason for making false promises, anyone can see that the answer is no, and that anyone can see it, and so on.

Here, finally, is a dictate of conscience, reverberating with the appropriate authority. Conscience tells us that the reasons we thought we had for doing something couldn't be reasons for doing it; and it tells us authoritatively, once and for all. They couldn't be reasons for doing it, conscience tells us, because their being reasons couldn't be seen, and be seen to be seen, by all. And what conscience here points out to us is something that can be seen, and seen to be seen, by all. Thus, conscience authoritatively reveals that our proposed reasons for acting couldn't be authoritative and consequently couldn't be reasons.

38. See again the passage quoted in note 14, above.

The role of autonomy. But isn't conscience supposed to forbid us from doing things rather than merely inform us that we don't have reason for doing them?

Kant's answer, I think, would be that by informing us of the absence of reasons for doing things, conscience rules out the possibility of our doing them for reasons and, with it, the possibility of our doing them autonomously—or, indeed, the possibility of *our* doing them, since we are truly the agents of the things we do only when we do them for reasons. And ruling out the possibility of our being the agents of the things we do is the way that conscience forbids us from doing them at all.

Kant says:[39]

> *[M]orality* lies in the relation of actions to the autonomy of the will.... An action which is compatible with the autonomy of the will is *permitted;* one which does not harmonize with it is *forbidden.*

Kant could have put his point differently. An action that is incompatible with the autonomy of the will isn't, properly speaking, an action at all: it's a piece of behaviour unattributable to an agent, a bodily movement in which there is nobody home. So put, of course, the point seems to be that we *won't* do the forbidden thing—or, at least, that *we* won't do it. Yet this point is compatible with the recognition that we might still do the forbidden thing in the weaker sense of 'do' that includes nonautonomous behaviour. As I interpret Kant, the recognition that we could do something only nonautonomously deters us from doing the thing even in this weaker sense. The deterrent force of this recognition derives from our reverence for the idea of ourselves as rational and autonomous beings.

Kant speaks of a 'paradox' with the following content: 'that without any further end or advantage to be attained[,] the mere dignity of humanity, that is, of rational nature in man—and consequently that reverence for a mere idea—should function as an inflexible precept for the will.'[40] In other words, the prescriptive force of moral dictates is a force registered in our reverence for the idea of ourselves as rational and autonomous beings. Conscience tells us that if we do something, we shall have

39. *Groundwork* 107 (439).

40. 106 (439).

to do it nonautonomously, without reason; and conscience thereby appeals to our reverence for this self-ideal as a motive against doing the thing at all.

The Kantian ego ideal. I have now returned to the idea that Kant resembles Freud in positing an ego ideal. This ideal is necessary to motivate our adherence to the conclusions that result from applying the categorical imperative—the conclusions that I have identified with the dictates of conscience. These conclusions authoritatively refute our proposed reasons for acting; but in order to deter us from acting, they must engage our respect for the conception of ourselves as acting only for reasons. Moral requirements thus motivate us via an ideal image of our obeying them.

I believe that the ego ideal plays a similar role in Freudian theory.[41] Freud sometimes speaks as if the commands of the super-ego are backed by threats and obeyed by the ego solely out of fear. In fact, however, his descriptions of the relations between ego and super-ego depend heavily on the ego's admiration for the super-ego, as an internalized object of love. And it is in this latter capacity that the super-ego is described by Freud as being, or as including, an ego ideal.

I believe that Freud's theory of the ego ideal can help us to humanize Kant's ideal of ourselves as rationally autonomous. It can help us to see that what Kant called 'reverence for a mere idea'—reverence, that is, for 'the mere dignity of humanity'[42]—is in fact our response to something that we have internalized from real people in the course of our moral development. More specifically, I believe that the object of this reverence, the ideal of ourselves as rationally autonomous, is an ideal that we acquire in the course of loving our parents, in the manner described by Freud. But my reasons for this belief will have to wait for another occasion.

Department of Philosophy
University of Michigan
Ann Arbor, Michigan 48109, USA
velleman@umich.edu

41. The claims made in this paragraph are defended in 'The Direct Heir of the Oedipus Complex' (MS).

42. Quoted at note 40, above.