13 Moral Judgment and Volitional Incapacity

Antti Kauppinen

Charles Doyle had let down the Mam and condemned his children to genteel poverty. He had been weak and unmanly, incapable of winning his fight against liquor. Fight? He had barely raised his gloves at the demon.

—Julian Barnes, *Arthur & George*

The central question of the branch of metaethics we may call philosophical moral psychology concerns the nature or essence of moral judgment: what is it to think that something is right or wrong, good or bad, obligatory or forbidden? One datum in this inquiry is that sincerely held moral views appear to influence conduct: on the whole, people do not engage in behaviors they genuinely consider base or evil, sometimes even when they would stand to benefit from it personally. Moral judgments thus appear to be motivationally effective, at least to an extent. This motivational success would be readily explained if they simply *were* motivationally effective psychological states, such as desires. This is what Hobbes seems to do when he claims that "whatsoever is the object of any man's appetite or desire, that is it which he for his part calleth good; and the object of his hate and aversion, evil" (Hobbes 1651/1994, VI, 28). But this is far too quick. We know that moral judgments can also fail to lead to corresponding action. For example, since it is conceptually possible—not to mention all too common in the actual world—to think that something is wrong and yet want to do it, thinking that something is wrong cannot simply consist in aversion toward it, contrary to what Hobbes seems to have thought. In this way, reflection on the various conceivable *failures* of moral motivation rules out possible candidates for explaining *successful* moral motivation. This is not an empirical psychological inquiry. We want to find out what kind of psychological state moral judgment *must* be (and what it *cannot* be) in order to have just the right kind of motivational role. What I will suggest in this essay is that a close examination of the various possible

T1

kinds of motivational failure has important and underappreciated implications for metaethics.

In metaethics, there are two major strategies for explaining the existence and ubiquity of the motivational success of moral judgment. According to moral judgment *internalism*, moral judgments involve some degree of motivation by their very nature.[1] Roughly, internalists believe there is a *conceptual* connection between thinking that one morally ought to φ and being to some degree motivated to φ (or being motivated to take what one believes to be the means to φ-ing). Consequently, an agent who is not at all motivated to φ does not genuinely think that she morally ought to φ.[2] The denial of internalism is known as moral judgment *externalism*. On externalist views, it is not the nature of moral judgment but rather human nature that explains why judgments tend to motivate: as a contingent matter of fact, human beings are often disposed to do what they believe to be morally right. Thinking that something is one's duty does not as such involve motivation, but does lead to action when combined with a *desire to do whatever is right*. Internalism and externalism face opposite challenges: internalists have an easy time with motivational success, but difficulty with motivational failures, whereas externalists make failure easily intelligible at the cost of struggling to explain success.

It seems to me that internalists have the upper hand in this debate. As Michael Smith has recently argued, any externalist strategy makes morality look like a kind of fetish.[3] After all, on the externalist picture, what we care directly about is being moral, so that our concern with the particular things we judge to be right is only derivative. If Smith is correct, only internalists can leave room for nonderivative caring about what we think is good or fair as a result of moral judgment, where "what we think is good" is read *de re*—caring about reducing poverty as such, for example, as opposed to caring about it as a way of doing whatever is good. The externalist explanation for motivational success is thus suspect. In addition, externalists seem to allow for too radical motivational failures. On their picture, someone who was never once in her whole life motivated to act morally could nonetheless make genuine moral judgments. Equally, there could be a community within which a recognizably moral language game was played without the judgments making the slightest difference to people's actions.[4] But neither individual nor communal comprehensive indifference to moral judgments seems conceptually possible. At best, we could describe these people as trying to make moral judgments, but failing to do so.

It is thus some form of internalism that holds the most promise. Traditionally, internalism has been perhaps the strongest reason for adopting

noncognitivist views about the nature of moral judgment. According to them, a moral judgment is a *conative* or *affective* state, and only such states are intrinsically motivating. If this is the case, it is easy to see why internalism is true: moral judgments have an internal connection to motivation, since they consist in some sort of motivational states. Internalist *cognitivists* face a challenge of explaining why an internal connection would obtain between merely *believing* that one ought to do something and being motivated to do it, given that on the standard story in the philosophy of action, belief alone cannot constitute motivation.[5] They have responded to this challenge in two main ways: either denying that all beliefs are motivationally inert, or by arguing for a weaker internalist thesis that beliefs with a suitable content necessarily hook up to motivation in rational agents, though not all moral agents. Here the earlier dialectic reappears: noncognitivist internalists are prima facie better placed to explain motivational success, whereas cognitivist internalists have an easier time with motivational failure.

Sophisticated contemporary forms of noncognitivism, such as Allan Gibbard's norm-expressivism, are capable of explaining certain kinds of motivational failure, like weakness of the will. However, I will argue here that the existence of the sort of radical motivational failure that Gary Watson has labeled *volitional incapacity* cannot be accounted for by *any* expressivist theory that identifies a moral judgment with a conative state. Volitional incapacity, crudely put, is the kind of disorder that Charles Doyle in my epigraph might have suffered from, supposing he genuinely judged that he should stop drinking: it is not that his desire to drink was causally stronger than his will to stop, but that he never really willed to stop in the first place—he "barely raised his gloves" at it. It is not just that he was not able to resist the temptation; he was not even able to *fight* it. This contrasts with another kind of addict, the weak-willed agent who yields to temptation in spite of meaning to stay clean. As Watson points out, this kind of incapacity is a neither a failure of judgment nor one of self-control: Charles Doyle knew what he had to do and, we may suppose, would have been able to do it had he put up a fight. Rather, volitional incapacity forces us to distinguish between *judgment* and *the will*: one is volitionally incapacitated when one's self-directed ought-judgment does not result in a corresponding intention. Just as weakness of the will shows that valuing and desire are (at least) modally separable, volitional incapacity shows that valuing and conative states like planning are modally separable and thus cannot be identical, as expressivists claim. Instead, the internal connection between moral judgment and motivation must

be explained in terms rationally binding belief, as internalist cognitivists have argued. I finish with a sketch of a new kind of cognitivist and rationalist account that draws on recent developments in conceptual role semantics.

1 Expressivism and Failures of Motivation

According to any kind of noncognitivism, moral judgments consist in psychological states[6] with a world-to-mind direction of fit. These states include desires, second-order desires, intentions, plans, and the like—any psychological state whose function it is to make the world fit its content rather than the other way around.[7] The identification of moral judgment with desiderative or conative states is motivated primarily by internalism: if thinking that something is right is adopting some positive attitude toward it, it is easy to see why moral judgments essentially have a relation to motivation. As already noted, the crudest form of noncognitivism, identifying valuing with desiring, is untenable in light of the possibility of motivational failure, since it seems plain we can fail to desire what we think is right.[8] More sophisticated noncognitivists, such as contemporary expressivists, make room for motivational failure by making a distinction between noncognitive attitudes: only *some* of them constitute valuing, and having these more complex attitudes is compatible with lacking occurrent motivation.

I will focus here on Allan Gibbard's recent work, since it provides perhaps the most detailed expressivist account of ought-judgments currently available. In his *Thinking How to Live*, he begins with the observation that moral judgments, and normative judgments in general, are essentially judgments about what to do. Correspondingly, ought-questions and reason-questions are questions about what to do. To judge that something is the thing do, in turn, is on his view to *settle* on doing it or to *decide* to do it—more precisely, to rule out doing anything else (Gibbard 2003, 137). As he puts it, "When I deliberate, I ask myself what to do and I respond with a decision. Thinking what I ought to do amounts to deciding what to do" (ibid., 17). Gibbard's main goal is to argue that seeing normative utterances as expressing "contingency plans" or "hyperstates"[9] explains why they behave in many ways as would Moorean assertions about nonnatural properties, but without the need to commit to queer metaphysical or epistemological theses. He claims that this is the key to solving the Frege–Geach problem about the behavior of normative predicates in unasserted contexts. I do not here take a stand on whether planning states

T1

can play this sort of role; my question is rather whether planning states play the right sort of motivational role to qualify as candidates for moral judgment in the first place.

Suppose, then, that thinking that I ought to take the dog for a walk is a planning state, a matter of adopting a plan to take the dog for a walk. Clearly this makes motivational success intelligible: if I plan to do something, then, at least ceteris paribus, I will have some motivation to do it. But how does Gibbard's view accommodate motivational failures? His first response is to deny that one wholeheartedly judges that one ought to do something if one fails to (try to) do it:

> For a crucial sense of 'ought', I say, the following holds: if you do accept, in every relevant aspect of your mind, that you ought right now to defy the bully, then, you will do it if you can. For if you can do it and don't, then some aspect of your mind accounts for your not doing it—and so you don't now plan with every aspect of your mind to do it right now. . . . And so, it seems to me, there's a part of you that doesn't really think you ought to. (Ibid., 153)

These sort of cases result from "conflicts among motivational systems" (ibid.; see also Gibbard 1990, 56–58)—here between the planning system and the more primitive one that results in fear of the bully. At other times, our plans are defeated by the grip that social norms like those of politeness and cooperation have on us even when we acknowledge other things are more important in the situation, morally speaking—Gibbard points here to Milgram's classic experiments, in which subjects continued to give electric shocks to others in spite of recognizing that it was wrong in virtue of the great pain the shocks were causing (as they were led to believe) (Gibbard 1990, 59). It seems that we can make sense of failures like weakness of the will in these terms, whether we describe it as a case of being of two minds about what one ought to do or being overcome by a desire that is stronger than the motivation provided by the planning system. If we identify the self of an agent with the totality of her higher-order planning states, as Gibbard sometimes is tempted to do,[10] the latter description is more fitting, since the competing motivation will be external to the self: *I* am not of two minds, though there is a force in me that may be causally stronger than my decision when it comes to acting.[11] To accommodate cases of listlessness or accidie, in which the problem is not the strength of a competing desire but lack of *any* desire in spite of willing (see below), Gibbard allows that a state may be a planning state even if it only *normally* issues in action (Gibbard 2003, 154).[12] (Of course, it does not make sense to say of a token state that it "normally" does anything; Gibbard must be

T1

thinking of an instance of a certain *type* of state.) Thus, there is room for certain kinds of motivational failure even if moral judgments are identified with planning states. Sophisticated expressivism can make sense of failures like weakness of will and accidie while at the same time making intelligible the internal connection between judgment and action.

Having provided an expressivist explanation of motivational success and failure, Gibbard goes on the offensive. He argues that the expressivist explanation of motivational success is superior to the internalist cognitivist one, so, given that it can handle failure equally well, we should prefer expressivist moral psychology. In effect, the argument is that once we admit that there is an internal connection between judgment and motivation, there is no point in thinking that the psychological states in question are distinct existences: "[W]hy then think that deciding and thinking what one ought to do are separate activities?" (ibid., 13). Making a distinction here merely incurs an extra explanatory debt (why would there be an internal connection between two separate states with opposed directions of fit?), so expressivism has an advantage here, if Gibbard is right.

2 The Difference between Judgment and the Will

I granted above that expressivists can accommodate certain kinds of failure of motivation. In this section I will draw on recent arguments in the philosophy of action to show that we must distinguish between two basic types of motivational failure and corresponding internal conflict, the latter of which would have no conceptual room if moral judgment were a conative state.

The Gap between the Will and Action

The most obvious failures of motivation occur between what an agent sets out to do and what she in fact does or tries to do. There are at least three possible forms here. First, it may be that the agent sets out to do one thing, but a desire to do something else is causally stronger, leading her to form at different proximal intention (the psychological state that causes and sustains the bodily movements that constitute intentional action) and thus doing something else intentionally, unless prevented by an external obstacle.[13] This is *weakness of the will*, as I understand it. Weakness of will may also consist in giving up one's original prior intention and forming a new one—that is, giving up one plan and adopting another—without changing one's mind about reasons for action. The essential thing is that it makes sense to speak of weakness of the will only when there already is a will

T1

that can be either strong or weak. To be sure, weakness of the will is often understood or even defined in slightly different terms, as acting against one's judgment about what is best. It is readily intelligible why it is commonly thought so, since as a rule what we will is what we judge best. In such cases, the weak-willed person acts *both* against her judgment *and* her will. But as I will argue, we must be careful to distinguish between the two. Consequently, we are better off restricting the term 'weakness of (the) will' to cases where the agent has already *formed* a will, that is, set out to do something, without simply assuming that this is what judgment about what is best consists in.[14]

The second form of failure of willing is accidie or listlessness, where the agent sets out to do something but is not capable of mustering any effort to achieve her goal. This seems to be the condition from which some depressives suffer, as Michael Stocker (1979) notes. Third, it may happen that the agent does what she has set out to do, but this action is in fact caused by some other source of motivation than her setting out to do so.[15] This kind of *deviant causation* is a failure of self-control that can be relevant to assessing the agent, since it raises the question about what she would have done in counterfactual situations where the causally efficacious desire would not have been present.

In each of these cases, we can say that in settling on a course of action, the agent has formed a ("prior" or "distal") *intention* to act in a certain way, a commitment that implies forming a certain causally effective proximal intention at the time of action. This commitment has a world-to-mind direction of fit and a characteristic causal-functional role of leading to proximal intention. On pain of (subjective) irrationality, it excludes commitments to contrary courses of action and can thus play an important role in coordinating and planning activities. In Alfred Mele's terms, having an intention is *having an executive attitude toward a plan* (Mele 2003, 28, 210; cf. Bratman 1987). It is natural to talk about the agent's *will* in this context. An agent who has a strong or resolute will is one who follows through on her intentions. I will use the terms 'will', 'planning state', and '(prior) intention' interchangeably in what follows.

The Gap between Judgment and the Will

It seems possible that sometimes we not only fail to do what we have set out to do, but also fail to set out to do what we think we ought to do. This amounts to a different kind of failure and inner conflict, as many in recent philosophy of action have argued. I will first try to make this difference intuitive by contrasting four different cases.

T1

Brave Michelle

It is wartime, and Michelle is in the Resistance. There's a wireless in her room, and one day she gets an urgent message to be delivered to René, an explosives expert: "The hawk will be in the henhouse at 1900 hours tonight." There is a curfew and the town is swarming with Germans, but Brave Michelle believes it is her duty to go out and deliver the message. Consequently, she settles on leaving. Just as she's about to go, she hears a German armored car rumbling nearby, but maintains her resolve and sneaks out anyway.

Weak Michelle

Weak Michelle's situation is otherwise the same as Brave Michelle's, but when she hears the armored car rumbling nearby, she feels too scared to open the door. Her heart beats loudly as she walks around in small circles, her nerves racked. She envisions what will happen if she doesn't go and tells herself she will do it now and then it will be over, but as she puts her hand again on the handle, there's another violent sound and she can't bring herself to do it.

Listless Michelle

Listless Michelle's situation is otherwise the same as Brave Michelle's, but owing to fatal failures and loss of comrades in previous resistance activity, she has become seriously depressed. She finds it her duty to deliver the message and swears to herself she will do so, but when the time comes, she lies inert in her bed. Try as she might, she can't get her legs to move in the right direction.

Incapacitated Michelle

Incapacitated Michelle's situation is otherwise the same as before, but she has, understandably enough in the exceptional conditions, developed a severe case of agoraphobia, fear of open spaces and the outdoors. The wireless crackles out the message, and Michelle thinks it her duty to deliver it. After all, lives are at stake, and it is all up to her. But delivering the message would require leaving the house, and she can't see herself doing it. The mere thought makes her shudder. She knows she should at least make an effort, but finds herself paralyzed by the idea. Consequently, she can't bring herself to make the decision to go out, though, as we might say, she has already decided that it is what she should do.

As I would describe these cases, Brave Michelle exercises her volitional capacity by willing in accordance with her judgment and strength of will

T1

by following through with her plan in spite of the scary situation. Weak and Listless Michelle also will in accordance with their judgment, but lack strength of will, thus suffering from the type of motivational failure I discussed in the previous section. The interesting case is Incapacitated Michelle, who does not even form a will or intention to do what she judges she ought. This is the same sort of radical failure I suggested the fictional Charles Doyle might have suffered from. In neither case is there an internal struggle between two competing motivational systems. Incapacitated Michelle is, as Watson puts it in a similar case, too terrified even to try (Watson 2004a, 94). Her condition consists precisely in her inability to settle on a course of action that would involve going out.[16] As Watson notes of a similar case, "Quite apart from being unable to see clearly what is to be done, or to do what [she] wills, [she] is sometimes unable to commit herself to implementing [her] judgments."[17] In Mele's terms, we can say that in spite of her judgment, Incapacitated Michelle has a plan to go out as a sort of recipe of how to carry out her duty, but she fails to adopt an *executive attitude* toward it. (In the final section, I will return to what adopting such an attitude might amount to.) In this, she differs from Listless and Weak Michelle, who fail to follow through on a plan they have settled on executing, either through general inertia or the presence of a competing desire. Their conditions, consequently, support different counterfactuals. Were Incapacitated Michelle to form an intention, she would be able to act on it (unless she also suffers from some other motivational disorder). By contrast, Listless Michelle would be unable to act on any or most other intentions as well, and Weak Michelle would act on the intention if she only got rid of the competing desire or fear. To be sure, the difference here is subtle, and we may well be unable to tell whether Michelle is weak willed, listless, or incapacitated on the basis of a single case of inaction. But the difference can come out on the basis of a pattern in her action, or inaction, as it may be. Listless Michelle will not answer the phone either, whereas Incapacitated Michelle may have no problem with it.

The situation of Incapacitated Michelle seems to be a coherent possibility, assuming that conceivability is here a reliable guide to possibility. The lesson is that there is a potential difference between judgment and the will that is distinct from the failure to act on one's plans or intentions (the failure of Weak Michelle), but a failure to adopt an executive attitude to what one takes to be the best supported plan in the first place. This distinct failure is *volitional incapacity*, the inability to will as one judges best.

Distinguishing between ought-judgment and the will is not an ad hoc move to make sense of volitional incapacity. It does other work as well.

Watson argues that it leaves room for the "work of the will" in cases where it is necessary to form an intention even though the balance of reasons is even or unclear, leaving the normative issue open. Perhaps I cannot decide whether it is better to stay or to go, but I must do either, and so I will.[18] Second, Jay Wallace argues that a realistic account of normative motivation must leave room for counternormative intentions. This is manifest in the fact that there doesn't seem to be a Moorean paradox involved in saying "I really ought to do *x*, but I'm going to do *y* instead" or "I've chosen to do *y*, though it's not in fact the best alternative open to me," as one would expect if there were no possible gap between ought-judgment and intention.[19] (There is still something odd about such statements; I will return to this below.) Finally, Al Mele points out that there may be a temporal gap between judgment and decision (understood as settling on an action). For example, we can imagine that Joe deliberates on New Year's Eve about whether or not to continue smoking, and judges it best to stop smoking at midnight; that will be his New Year's resolution (Mele 2003, 199). But he may not *yet* have resolved to stop, Mele argues.

All these cases support distinguishing normative judgment and planning states, though in most situations the two are closely aligned. Both ought-judgment and settling on a plan involve a verdictive, all-things-considered stance on action. That is why it is natural to talk of "decision" in each case. But they are different kinds of decision. The decision involved in reaching a first-personal ought-judgment could be called a *cognitive decision*, a verdict that all things considered, the balance of reasons in the case favors a particular course of action.[20] Forming the will or adopting a plan, by contrast, amounts to a *practical decision*, actually settling on a course of action. Thinking about a court case should make this distinction clear.[21] The jury deliberates whether the defendant is guilty or not, weighing the evidence pro and con. Let us assume that in the end, it *decides that* the defendant is guilty. As a result, the judge *decides to* send the defendant to prison for, say, eight years. Here, the jury's decision is a *belief* that, in the context of legal practice, has practical consequences, calling for action on the part of the judge—that is the whole point of forming the belief. Nonetheless, it can be true or false, depending on whether the defendant actually is guilty or not. The judge's decision, by contrast, is an announcement of an (institutional) *intention* or plan to punish the defendant in a certain way. It cannot be true or false, though we can agree or disagree with it. This process can malfunction in two ways, analogous to the motivational disorders: the judge may fail to issue a verdict corresponding to the one given by the jury (perhaps she is bribed) or the wheels of the institution

T1

may grind to a halt (prison guards don't do their job properly and the prisoner walks free).

3 From Noncognitivist to Cognitivist Internalism

I argued in the previous section that we must separate ought-judgment and volitional states from each other to make volitional incapacity and related phenomena intelligible. This is very bad news for noncognitivist forms of internalism, Gibbard's expressivism included. The argument is quickly stated. Gibbard claims that we should think of ought-judgments as consisting of contingency plans, decisions to act in a certain way in a certain situation. But this cannot be the case: were normative judgments planning states, the sorts of failure I discussed in the previous section would be inconceivable. To borrow a term from Michael Smith, who uses a similarly structured argument to argue against the identification of moral judgments with desires, the existence of volitional incapacity and the like shows that normative judgments and planning states are *modally separable*, so that their relationship cannot be either identity or logical connection. Moral judgment must be different from practical decision or adopting a plan.

The internal connection between ought-judgment and motivation cannot, then, be explained by identifying the judgment with a conative state. It must instead be a cognitive state that is in some way special in terms of giving rise to plans and actions. It cannot be an ordinary belief like a belief that there is milk in the fridge, since the ordinary belief has motivational effects only contingently, given one has an appropriate desire, such as the desire to drink milk. The whole point of the practice of making moral judgments and persuading each other of them seems to be making a difference to action and attitude. (In this respect, a moral judgment is like the jury's verdict of guilt, discussed above.) It is just this difference between moral judgments and ordinary beliefs that externalism misses. What we need, then, is a cognitive state that has a noncontingent but defeasible tie to the will and corresponding action. The story should go something like this. Suppose that I am being threatened by a schoolyard bully who wants me to do his homework. I weigh the practical reasons bearing on the situation and reach the conclusion that all things considered, I ought to just walk away. Insofar as I successfully exercise my volitional capacities—insofar as my will listens to my reason, to put it in Aristotelian terms—I form the intention to walk away.[22] The intention to walk away, in turn, leads me to intentionally walk away (as a causal result

of forming the relevant proximal intention) insofar as I successfully exercise my strength of will. The various motivational disorders and inner conflicts show that the required capacities can fail to function at both junctions. When judgment fails to lead to willing, I suffer from volitional incapacity. My will stands opposed to my reason. On the other hand, if, for example, a causally stronger desire to avoid risk intervenes between my will and action, I either form a different proximal intention than I had intended or go a step back and form a different intention altogether, with the result that I (intentionally) do the bully's work rather than walk away. If my will misfires entirely and I remain passive in spite of my intention, I suffer from accidie. In these cases, some of my desires stand opposed to my will. Acting on a *moral* ought-judgment is simply a special case of normative control, as are failures to do so.[23]

The picture I have presented is supposed to be a weakly internalist one. But how can that be the case, if ought-judgment is at least two steps removed from action? As Michael Smith observes, talk of internalism in the absence of necessary connections makes sense if (and perhaps only if) volitional capacity and strength of will are distinctively *rational* capacities, so that judgment, will, and action go necessarily together *for rational agents* (cf. Smith 1994, 177–180). A verdictive, all-things-considered ought-judgment, after all, is a judgment that I have most reason to do this. This would also explain the pragmatic oddity of claims like "I really ought to do *x*, but I'm going to do *y* instead"—they amount to avowals of practical irrationality on one's own part, admitting that one cannot justify one's behavior to others. This seems to be the most promising *form* of an internalist cognitivist account.

Are volitional capacity and strength of will rational capacities, then? For Smith, the paradigm rational capacity is the logical capacity of improving the *coherence* of our beliefs (Smith 1997, secs. 5–6). Consequently, he argues that what corresponds to volitional capacity (or strength of will—Smith's model does not distinguish between the two) in his Humean model is a rational capacity only if the content of the belief that one ought to φ (that is, one has most reason to φ) is such that when the belief is present, one's psychology is more coherent if it includes the desire to φ than if it lacks it. He claims that this is the case only if the belief itself is a belief about desires—more precisely, the belief that one's ideally informed and perfectly coherent self would desire one to desire to φ. On this analysis, to believe one ought to φ is to believe that one's ideal counterpart would desire one to φ, and forming this belief leads an agent with a disposition toward coherence to desire to φ.

T1

Smith's proposal involves strong assumptions about the content of our moral beliefs (do we really form causally effective beliefs about an ideal counterpart?) and the role of coherence in rationalizing attitudes (could not acting against an ought-judgment cohere better with the rest of one's psychology in some cases?), among other things.[24] These and other problems are now well known, so instead of discussing the virtues and vices of Smith's account, I want to finish with a sketch of a promising alternative model that treats *inference* rather than coherence as the basic rational capacity. According to inferentialism, or more broadly, conceptual role semantics, the content of a belief is to be understood in terms of how it fits into the context of other mental states, perceptual inputs, and behavioral outputs; that is, what it properly follows from and what properly follows from it. The version that seems to me the most promising is developed in detail by Robert Brandom in *Making It Explicit* (1994) and other works.

The key ideas of inferentialism, in short, are the following. First, sounds, inscriptions, and tokenings of mental states acquire content by being caught up in the right way in a social practice of giving and asking for reasons. Second, the distinctive feature of the game of giving and asking for reasons is the deontic score that all participants attribute to each other, a record of which linguistic and nonlinguistic performances each of us is committed and entitled to. On Brandom's social pragmatist picture, these deontic statuses of commitment and entitlement result from deontic attitudes—treating someone as committed or entitled. In simplest terms, to treat someone as *entitled* to φ (that is, to attribute an entitlement to φ -ing) is to *positively* sanction φ -ing—for example, treating a ticket-holder as entitled to enter a concert may consist simply in allowing only ticket-holders in. Treating someone as *committed* to φ, in turn, is a matter of *negatively* sanctioning *not* φ-ing—for example, treating a library user as committed to returning a book in time may consist simply in revoking library privileges in case of failure to return the book in time.[25] Importantly, undertaking a commitment can be understood in these terms as entitling others to negatively sanction failure to perform; I will develop this idea below.

Third, the *pragmatic significance* of speech acts consists in the difference they make to deontic scores—what commitments and entitlements they give rise to, what commitments preclude entitlement to them, and so on. Fourth, the *semantic content* of expressions, their inferential role, is understood in terms of their pragmatic significance. Crudely, to know what something means is to know (in the sense of being able to discriminate in

practice) what entitles one to it, what it commits one to, and what is incompatible with it in terms of precluding entitlement. This is to know the material inferential role that the expression has. Finally, *logical vocabulary is expressive*: its function is to make explicit the material inferences that give content to nonlogical vocabulary. Asserting (or thinking) that Matilda is a cow commits one to thinking that Matilda is an animal, since licensing that (material) inference is part of what gives the concept the content it has. Using the conditional to say "If Matilda is a cow, Matilda is an animal" or the negation to say "Cows are not horses" *makes explicit* part of the concept's content, rather than underwriting the correctness of the inference.[26]

Though inferentialist accounts are inspired by Gentzen-style definitions of logical connectives in terms of introduction and elimination rules,[27] the circumstances and consequences of proper application that define the inferential role of a commitment need not be limited to other commitments. For example, one can acquire entitlement to "That car is red" by being in the right kind of causal commerce with the world and having the right kind of discriminatory abilities. Indeed, this is how concepts come to have empirical content, even broad content—the broad inferential role of "water" is not the same in the actual world and on Twin Earth, since the proper circumstances of application on Twin Earth involve different stuff. Importantly, the same goes for nonlinguistic output. It can be a part of the inferential role of an expression that it commits one to *act* in a certain way. Rational agency, on Brandom's view, is a matter of responding to the acknowledgment of such a commitment with the suitable behavior. For example, "I shall open the door at ten o'clock" expresses a practical commitment of this kind; an agent who undertakes such a commitment will open the door at ten o'clock, insofar as she is practically rational (disposed to respond to acknowledging a commitment with appropriate behavior) and the performance is in her power. To be more precise, there are two kinds of practical commitments, future-directed and present-directed ("I shall open the door *now*"), which correspond to prior and proximal intentions, respectively. The difference is that they are defined in normative rather than causal-functional terms.

What, then, is the distinct inferential role of moral ought-beliefs—what is it to think ought-thoughts? Let us begin with the language that makes these thoughts explicit. Brandom's original and at first sight unlikely idea is that normative vocabulary is akin to logical vocabulary in its function, indeed a species of it: it, too, makes explicit proprieties of inference. What is specific about it is that it makes explicit inferences from doxastic

T1

commitments (beliefs) to practical commitments—that is, in Brandom's terms, proprieties of a kind of *practical reasoning*. To illustrate this with a nonmoral case, let us imagine that Al and Sean volunteer as firemen in their small village. Suppose that Abe treats firemen as committed to rush to the fire station when the fire bell rings—that is, he negatively sanctions them for failing to do so, say, in the form of disapproval and not inviting them to a party, and takes everyone else as being entitled to do the same. This is a matter of Abe taking anyone, himself included, to be entitled to a pattern of inferences like the following: (Al fails to rush to the fire station when the bell rings → I shall not invite Al to the party), (Sean fails to rush to the fire station when the bell rings → I shall not vote for Sean in the election), and so on for any number of (social) sanctions and for anyone who might volunteer to be a fireman. In the pattern, the antecedent is a belief about nonperformance of an action in a situation, and the consequent a practical commitment to a form of sanctioning. At the same time, Abe takes responding to a belief that the fire bell rings with a practical commitment to rushing to the fire station to be an instance of good practical reasoning on Al and Sean's part—that is, he takes them to be entitled, not simply committed, to making this move.[28] One can presumably be disposed to draw such inferences without possessing explicit normative vocabulary. But when one does possess normative vocabulary, one is able to express one's commitment to endorse all these inferences by simply saying "Firemen *ought* to rush to the fire station when the bell rings."

Provided that Sean, for example, possesses the relevant vocabulary, he, too, can say or think "I *ought* to rush to the fire station when the bell rings." On the inferentialist view, to say or think "I ought to φ" amounts to explicitly acknowledging the appropriateness of the kind of pattern of inferences by oneself and others I sketched above. It is *not* itself to undertake the practical commitment—in traditional terms, form the intention—to φ. That would be expressed by "I will φ" (in the committive rather than predictive sense) or "I shall φ." Ought-talk makes explicit *inferential commitments*, not practical ones. These commitments fall into two classes that together make it intelligible why first-personal ought-judgments are *rationally binding* for the will. First, the *rationality* of the binding must be understood in the inferentialist picture in terms of entitlement-preserving inferences. Crudely, to think that the balance of reasons favors φ-ing is to think that anyone in *this* situation is entitled to φ. That is, anyone who is entitled to my doxastic commitments, whatever they are, is also entitled to form the practical commitment to φ—to reason

T1

practically to the conclusion expressed by "I shall φ."[29] But this does not yet capture an ought, since one could be entitled to do many incompatible things. So, second, the *bindingness* of the ought-judgment is manifest in the fact that it excludes entitlement to do anything else (to undertake other practical commitments). To take myself to be committed to φ-ing is to take anyone, myself included, to be entitled (and perhaps committed, too) to negatively sanction my failing to φ.

In my view, this is all we can say about generic ought-judgment as such. What are sometimes called moral or prudential or institutional oughts are just ought-judgments grounded in different reasons or norms, and consequently entitle different kinds of sanctions.[30] The specific motivational role of first-personal *moral* ought-judgments must be understood in terms of the specific nature of moral reasons and moral sanctions. Let us return to the case of Michelle to get a grasp of them. When she forms the belief that she ought to go out and deliver the message to René, she commits herself to a pattern of practical reasoning, both in the first person and in the third person. She takes it that anyone in her situation would be entitled to (make the practical commitment to) deliver the message. This she takes to be licensed by the moral wrongness of not delivering it. The inferentialist account as such leaves open how to think of moral wrongness. Perhaps to think something is morally wrong simply is to think that its nonmoral properties license the inference to a practical commitment of a moral kind—to think that staying in the house is wrong is just to think that one ought not stay in the house, on pain of moral sanctions.[31] This approach would rest all the specificity of moral judgment on the sort of practical commitments licensed. Or maybe thinking that something is wrong is ascribing to it a nonnatural property that grounds an obligation, or a particular sort of natural property, like causing more pain to a creature capable of suffering than other open alternatives. Each of these theories of the proper circumstances of application of a moral ought is compatible with inferentialism.

In making the ought-judgment, Michelle also takes herself to be committed to delivering the message, that is, not entitled to do anything else. As we have seen, that means that she positively sanctions anyone negatively sanctioning her not delivering the message. And in the moral case, this means that anyone is entitled to adopt a particular emotional posture toward her, namely, that of *blaming*. She herself is entitled, and perhaps also committed, to blame herself—in other words, to feel guilt.[32] Moreover, insofar as the commitment is grounded in moral wrongness, she takes herself to have it independently of anything that she herself has done.

T1

Let us, then, suppose that to judge that one (morally) ought to do something is to make such an inferential commitment. How does this view fare with the task of explaining motivational success and motivational failure? To begin with the success, on the inferentialist picture to think that one ought to φ is to think both that one is entitled to form a practical commitment to φ and that one is not entitled not to form a practical commitment to φ. Clearly, commitment to the correctness of this inference gives rise to the practical commitment in question insofar as the agent is a rational one, since an essential part of being rational is just being disposed to match one's psychology to one's commitments. And once the practical commitment is adopted by the agent, the agent will act accordingly, again insofar as she is rational. Thus, motivational success in rational agents is easily explained and predicted by the inferentialist model. What, then, of the various kinds of failure? To understand weakness of will, we must bear in mind the distinction between the two different kinds of practical commitments, future- and present-directed ones. It seems natural to understand weakness of will as the failure of present-directed practical commitments to match the future-directed ones, or failure to maintain a future-directed commitment in the face of a contrary desire. In either case, the result is that one fails to do what one has set out to do. In paradigmatic cases, the explanation for the failure is that a causally stronger desire disrupts the rational causal mechanism that standardly translates practical commitments to actions. Volitional incapacity and the like, in turn, are on this model failures to respond to inferential commitments by adopting the corresponding (future-directed) practical commitment while acknowledging the doxastic commitments that serve as the antecedent.[33] Rationally speaking, this is the same sort of failure as believing that *p*, endorsing the inferential commitment expressed by "*p* implies *q*," and not believing that *q*. It can have various kinds of explanations—in Michelle's case a deep psychological disorder, in Charles Doyle's case his "weak and unmanly" character.[34]

The inferentialist story that I have sketched has some obvious affinities with both Smith's and Gibbard's views. I hope its advantages in naturally accommodating various motivational failures while making intelligible the rationality of the transitions are clear at this point. Still, a few words may be in order to distinguish it from the expressivist model. On the inferentialist picture, to make an ought-judgment is in effect to locate oneself in the space of reasons, understood as a space of inferentially articulated commitments and entitlements. These commitments and entitlements, in turn, are understood in terms of deontic attitudes. Does this undermine

the claim that the account is a *cognitivist* one? Not so, for the deontic metalanguage applies across the board. Surely we can *believe* that cows are not horses, even if the content of that belief is understood in terms of commitments and entitlements,[35] and those in terms of deontic attitudes. Cognitivism is thus compatible with construing ought-judgments as inferential commitments. What is essential is that they are not understood as practical commitments, as expressivists would have it. Moreover, there is nothing to prevent inferential commitments from being *true*—just as it can be true that if *p*, then *q*—nor having the sort of phenomenology that is commonly associated with moral convictions. No doubt further argument is needed, but we have at least a prima facie case for the inferentialist account being a form of cognitivism.

In short, it seems that an inferentialist account of the content of moral beliefs provides a basis for a rationalist, weak internalist cognitivism that leaves room for the different kinds of motivational failure while also making it intelligible why motivational success can be expected from rational agents, since it shows how ought-judgments can be rationally binding. Given that the modal separability of moral judgment and planning states is a fatal problem for expressivism, this version of internalism deserves serious consideration.

Acknowledgments

I have greatly benefited from comments by Sven Nyholm, Lilian O'Brien, Geoffrey Sayre-McCord, James Skidmore, and Jussi Suikkanen.

Notes

1. The term 'internalism' is used for a wide variety of views in metaethics, more than can be covered in a note; I will use it only for moral judgment internalism, as defined in the text.

2. I omit an important intermediate step here, namely, the move from thinking that something is morally obligatory in my situation to thinking that I ought to do it, or conversely from thinking that something is morally wrong to thinking that I ought not do it. I return to this briefly in the final section.

3. Smith 1994, 1997. For a defense of Smith against critics, see Toppinen 2004.

4. For this kind of argument, see Blackburn 1998 and Lenman 1999.

5. This challenge, in effect, is what Michael Smith calls "the moral problem" (Smith 1994).

T1

6. It should be noted that we talk about "judgment" in two different though related senses: judgment as a linguistic expression (an utterance like "Cheating on one's girlfriend is morally wrong") and judgment as a psychological state (the thought that cheating on one's girlfriend is morally wrong). On the standard picture, sincere moral judgments in the former sense conventionally express moral judgments in the latter, constitutive sense. It this constitutive sense of judgment that we are concerned with; it would be a category mistake to say that an utterance is inherently motivating.

7. For a well-known attempt to cash out the notion of direction of fit, see Smith 1987. For persuasive criticism, see Tenenbaum 2006. I will not press the point here.

8. Another cautionary note: though I follow here the convention of speaking loosely about valuing and ought-judgments in the same breath, they are prima facie distinct precisely in terms of their motivational consequences. Moral judgments fall into two basic kinds, evaluative ("Helping the poor is good") and deontic ("I ought to help the poor" or "Not helping the poor is wrong"). Though the relationship between the two is a matter of dispute—it is one of the things at issue in the buck-passing debate, for example—we can say that prima facie, (first-personal) deontic judgments have a more direct relation to motivation and action. They are *verdictive* in nature, amounting to an all-things-considered take on the (moral) reasons present in the situation, whereas value judgments are of a *contributory* nature—they are judgments about which features of the situation weigh in which direction. Although it is debatable whether merely thinking that something would be good necessarily involves motivation (can I not see *some* good in quitting philosophy without having the slightest desire to do so?), it seems that moral verdicts do push toward an action—at least in the sense that if they do not result in corresponding behavior, it makes sense to speak of a motivational *failure*.

9. Hyperstates are idealized states in which one has decided what to do with regard to every possible factual circumstance (Gibbard 2003, 53–59).

10. See Gibbard 1990. In a similar spirit, Michael Bratman argues for understanding the self in terms of a special kind of plans, "self-governing policies," in various papers in Bratman 2007.

11. Indeed, Gibbard's explanation of the bully case is somewhat bizarre from this perspective: if my fear or desire to please leads me to submit against my decision to stand up for myself, it is surely not true that there is a part of me that thinks I *ought* to submit. To be sure, Gibbard does not exactly say this, but only that "there's a part of you that doesn't really think you ought to" (Gibbard 2003, 153). Still, why bring up oughts and plans when the part of me that leads me to submit is neither involved with nor responsive to ought-thoughts and plans?

12. Defining functional roles in terms of statistical normality seems too weak; in the case of the depressive, it may be that it is *normal* for her planning states to fail

to lead to action. In the spirit of Gibbard's earlier work (Gibbard 1990), one might appeal to the reason why the capacity to form planning states was selected for in the course of evolution, which is plausibly intelligent action in pursuit of long-term goals. In this sense, it is "normal" that plans lead to action. But this alone does not suffice for explaining why failures of motivation are *normatively* criticizable—why it is a kind of failure on the *agent's* part not to follow through on plans, something for which *she* is responsible.

13. For the notion of proximal intention, see Mele 1992.

14. For another view according to which weakness of the will is a matter of giving up one's intentions rather than acting against one's best judgment, see Richard Holton 1999.

15. Perhaps I have decided to treat myself to a doughnut, but the very sight of a Dunkin' Donuts sign arouses in me a desire that takes over control of my steps, rendering my original plan irrelevant.

16. If Gilbert Harman (1975) is correct in claiming that intention involves the belief that one will act as one intends (or at least the absence of the belief that one won't), the reason why Incapacitated Michelle fails to intend to leave may be that she believes she won't anyway.

17. Watson 2004a, 95. Watson talks, in fact, about "judgments or prior intentions," but this is misleading in my context.

18. See Watson 2004b, 134–135. Here "deciding what is better" is a *cognitive* decision, as explained below.

19. Wallace 2000, 10–12. I don't agree with the conclusions that Wallace draws from this phenomenon.

20. For this terminology, see Mele 2003, ch. 9.

21. I may owe this analogy to Lilian O'Brien's unpublished work on intention.

22. On this simplified picture, means–end deliberation is part of the process of forming the ought-belief. Some actions may be best represented as resulting from a process where the belief that one ought to φ combines with the means–end belief that *x*-ing is the best (or sufficient) means to φ-ing to lead to an intention to *x*. Smith (2005) takes this to be the standard case (and misleadingly talks about desire for an end rather than intention; see the next note).

23. There's a conspicuous absence from this story: desire as an independent psychological state is nowhere mentioned in the success story. This is no accident. Once we have intentions in the picture, we have no need for desires; indeed, it is hard to see where they would fit in. However we understand desire, it is what we might call a *partial* pro-attitude or stance toward its object, in contrast to intention, which is an all-things-considered attitude that is able to play a central role in

T1

the coordination of action. They are conduct-*controlling*, not merely conduct-influencing attitudes, as desires are (Bratman 1987, 15–16). Intentions do all the work that desires were supposed to do on the Humean picture, and they do it better, since they make irrationality intelligible in virtue of the commitments they embody. To be sure, there is still room for consequential attribution of desire: whenever *A* φs intentionally, we can attribute to *A* a desire to φ, as Nagel famously argued (Nagel 1970, 29). But this is to say nothing of the etiology of the action.

24. For critical discussion of Smith, see Sayre-McCord 1997; Arpaly 2003, ch. 2; and Dancy 2004, ch. 6.

25. As the latter case shows, the sanctions involved need not be external to the practice in question. This point is of fundamental importance for Brandom, since it makes possible understanding conceptual practice as "normative all the way down."

26. So, for Brandom, the inference (Matilda is a cow → Matilda is an animal) is not enthymematic; it does not need the conditional as a premise in order to be valid. See Brandom 2000, ch. 1.

27. See not only Brandom 1994, but also Peacocke 1992 and Wedgwood 2001.

28. By emphasizing the commitment and negative sanctions involved, I depart from Brandom, who defines what he calls an "institutional ought" solely in terms of entitlement: "Taking it that there is such a norm or requirement [in Brandom's example, the bank employees wear neckties] just is endorsing a patter of practical reasoning—namely taking [the inference from going to work in a bank to putting on a necktie] to be an entitlement-preserving inference for anyone who is a bank employee" (Brandom 1994, 250–251). But bank employees are not just *allowed* to wear neckties, but precisely *required* to do so, and talk of entitlement does not capture that essential aspect. More on this below.

29. The relevant doxastic commitments must be picked out demonstratively, as it were, since the thought I ought to rush to the fire station does not as such involve any particular grounds for rushing there. If my house is on fire, the reason why I ought to rush to the fire station may be that I need help putting out the fire. The relevant doxastic commitment, the entitlement to which I take to entitle anyone to rush to the fire station, turns out in this case to be the belief that one's house is on fire.

30. For this sort of view, see also Broome 2004; Wedgwood 2006.

31. This resembles so-called buck-passing theories of value, such as Thomas Scanlon's (1998) view. According to buck-passers, for something to be good is just for it to have nonevaluative properties that give reasons to adopt positive or negative attitudes toward it. For criticism and defense, see, e.g., Väyrynen 2006; Suikkanen 2005.

T1

32. This is obviously a *very* simplistic story about moral emotions, but I lack the space for a proper discussion here. One thing to note is that the inferential commitment to accept blame does not entail that one *will*, in fact, accept it—as it is with commitments in general, it is always possible to fail to act accordingly. Thus you may very well get upset when someone criticizes you for something you yourself acknowledge you should have done.

33. Here my story departs significantly from Brandom, who describes weakness of the will in much the same terms as I discuss volitional incapacity (Brandom 1994, 267–271). This is at least in part because he takes conative, not just normative, vocabulary to express inferential commitments; I disagree, but cannot pursue this here.

34. The nature of the explanation has major implications for whether volitional incapacity excuses from responsibility. We would be far more likely to blame Charles Doyle than Michelle, and perhaps blame him more than the "manly" alcoholic who would at least put up a fight and only then succumb to temptation.

35. Namely, that commitment to something's being a cow precludes entitlement to the commitment that it is a horse.

References

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Barnes, J. 2006. *Arthur & George*. London: Jonathan Cape.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Clarendon Press.

Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Bratman, M. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. 2007. *Structures of Agency*. Oxford: Oxford University Press.

Broome, J. 2004. Reasons. In *Reason and Value*, ed. J. Wallace, M. Smith, S. Scheffler, and P. Pettit. Oxford: Oxford University Press.

Dancy, J. 2004. *Ethics without Principles*. Oxford: Oxford University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

T1

Gibbard, A. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Harman, G. 1975. Practical Reasoning. In *The Philosophy of Action*, ed. A. Mele. Oxford: Oxford University Press.

Hobbes, T. 1651/1994. *Leviathan*. Ed. E. Curley. Indianapolis: Hackett.

Holton, R. 1999. Intention and Weakness of Will. *Journal of Philosophy* 96: 241–262.

Lenman, J. 1999. The Externalist and the Amoralist. *Philosophia* 27:441–457.

Mele, A. 1992. *Springs of Action*. Oxford: Oxford University Press.

Mele, A. 2003. *Motivation and Agency*. Oxford: Oxford University Press.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Oxford University Press.

Peacocke, C. 1992. *A Theory of Concepts*. Cambridge, MA: MIT Press.

Sayre-McCord, G. 1997. Michael Smith's *The Moral Problem*. *Ethics* 108:77–82.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Smith, M. 1987. The Humean Theory of Motivation. *Mind* 96:36–61.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Smith, M. 1997. In Defense of *The Moral Problem*. *Ethics* 108:84–119.

Smith, M. 2005. The Structure of Orthonomy. In *Agency and Action*, ed. J. Hyman and H. Steward. Cambridge: Cambridge University Press.

Suikkanen, J. 2005. Reasons and Value—In Defence of the Buck-Passing Account. *Ethical Theory and Moral Practice* 7:513–535.

Stocker, M. 1979. Desiring the Bad. *Journal of Philosophy* 76:738–753.

Tenenbaum, S. 2006. Direction of Fit and Motivational Cognitivism. *Oxford Studies in Metaethics* 1:235–264.

Toppinen, T. 2004. Moral Fetishism Revisited. *Proceedings of the Aristotelian Society* 104:305–313.

Väyrynen, P. 2006. Resisting the Buck-Passing Account of Value. *Oxford Studies in Metaethics* 1:295–324.

Wallace, R. J. 2000. Normativity, Commitment, and Instrumental Reason. *Philosophers' Imprint* 1. http://www.philosophersimprint.org/.

T1

Watson, G. 2004a. Volitional Necessities. In *Agency and Answerability*. Oxford: Oxford University Press.

Watson, G. 2004b. The Work of the Will. In *Agency and Answerability*. Oxford: Oxford University Press.

Wedgwood, R. 2001. Conceptual Role Semantics for Moral Terms. *Philosophical Review* 110:1–30.

Wedgwood, R. 2006. The Meaning of 'Ought'. *Oxford Studies in Metaethics* 1:127–160.