CAN ANIMALS BE MORAL?

Mark Rowlands
Department of Philosophy
University of Miami
Coral Gables, FL 33124

1. Grace of the Virtues

Eleanor, the matriarch of her family, is dying and unable to stand. Grace manages to lift her to her feet. She tries to get Eleanor to walk, pushing her gently along. But Eleanor falls again. Grace appears very distressed, and shrieks loudly. She persists in trying to get Eleanor to stand, unsuccessfully. Grace stays by the fallen figure of Eleanor for another hour, while night falls. If Grace were human, we might have little hesitation in attributing to her an emotion of a certain sort: compassion or sympathy. But neither Grace nor Eleanor is human. Eleanor is the matriarch of the First Ladies family of elephants. Grace is a younger member of another family of elephants, the Virtues Family.[1] This is the sort of case cited as evidence for the claim that some non-human animals (henceforth "animals") can be motivated by moral considerations. In this paper I am unable to survey the large and growing body of empirical research that bears on this claim.[2] This work sits in the background – the springboard for a discussion that is rather more abstract and conceptual. I use this case both as a representative example of this research, and as a way of organizing discussion.

---

[1] Ian Douglas Hamilton, Bhalla, S., Wittemyer, G. & Vollrath, F., "Behavioural Reactions of Elephants Towards a Dying and Deceased Matriarch." Applied Animal Behaviour Science 100, (2006), 67–102.

[2] For an excellent summary, see Marc Bekoff and Jessica Pierce, Wild Justice: The Moral Lives of Animals (Chicago: University of Chicago Press, 2009).

In this paper, I shall defend the claim that Grace *might* be a moral subject. I shall argue that there are no logical or conceptual obstacles to thinking of Grace as motivated by moral considerations – specifically by emotions that have identifiable moral content. Grace's actual motivation is, of course, an empirical matter, and I take no stand on this. I defend the possibility claim: it is possible for animals to be motivated by emotional states such as compassion, sympathy, and toleration, and also by negative counterparts such as jealousy, malice, and spite. These are emotions that have moral content – *morally-laden* emotions.

One might think that in defending merely this possibility claim, I am unacceptably stacking the dialectical deck in my favor. In fact, there is still much work to do. The attribution of moral motivation to animals is almost invariably regarded as impossible. In part, philosophers have respectable (though ultimately mistaken) reasons for this. But in part, they are confused. And in part, they have fallen under the spell of a certain kind of magical thinking. But, first, I need to make this paper's principal thesis a little bit more precise.

2. Morally-Laden Emotions

The thesis I am going to defend appeals to the concept of a morally-laden emotion. In this section, I'll explain what I mean by this. Unfortunately, I have a rather complicated theory of emotion-attribution, and here I can only outline, not defend, it.

Emotions, in the sense employed in this paper, are intentional states. They have content in the way that thoughts, beliefs, and desires, have content. We can attribute them by way of a sentence embedded in a that-clause. Smith is indignant *that* Jones snubbed him. That Jones snubbed him is the content of Smith's indignation. The attribution of an emotional state to a

subject, S, by way of a proposition, *p*, does not entail that S entertain *p*, or even be able to entertain *p*. The use of *p* is warranted if *p* entails – in a rather broad sense, which approximates 'makes true in a given context' – another proposition that S does entertain. The idea that an animal can entertain a proposition may seem unlikely. But a proposition is, of course, not a sentence but a content expressed by a sentence. If animals can entertain content – if they can think at all – they can entertain propositions.

Given this framework, there are two ways in which the possession of an emotion might, let us say, *misfire* – which I shall understand as roughly, the analogue of what it is for a belief to be false. The category of misfires is a conjunctive one: an emotion misfires when it is either *misguided* or it is *misplaced*. Smith is indignant because he believes Jones has snubbed him, but he is, in fact, mistaken. Jones didn't snub him at all. Let us say that Smith's indignation is, in this case, *misplaced*. An emotion is misplaced when it is predicated on a factual assertion that is false. If his indignation is not to be misplaced, the factual proposition, 'Jones snubbed me' must be true. If it is not misplaced, then Smith's indignation makes true this factual proposition.

Suppose, however, that Jones did indeed snub Smith. However, she has every right to do so – Smith was obnoxious during their previous encounter. Smith, as we might say, deserved no better from Jones in this case. Let us say that Smith's indignation, in this case, is *misguided*. The emotion is misguided, in this case, because it is based on an assumption of entitlement that is, in fact erroneous. Thus, if an emotion, E, is not to be *misguided*, then there is a certain evaluative proposition, *p*, that must be true. It is not necessary that the subject of E be able to entertain *p*. But the truth of *p* is required if E is not to be misguided: the truth of *p*, we might say, *makes sense* of E. Smith's indignation, if it is not misguided, makes true the evaluative

proposition, 'Jones was wrong to snub me'. An emotion that is not misplaced entails the truth of a factual proposition. An emotion that is not misguided entails the truth of an evaluative proposition.

Armed with these ideas, we can define the concept of a morally-laden emotion as follows:

> An emotion, E, is *morally-laden* if: (1) there exists a proposition, *p*, which expresses a moral claim, and (2) if E is not misguided, then *p* is true.

In this paper, I shall defend the claim that animals can be moral *subjects* in the sense that they are the subjects of morally-laden emotions. The category of the moral subject is a novel category that does not reduce to the more familiar categories of moral agent and moral patient.

3. Agents, Patients, and Subjects

Animals, if they are thought of as having moral standing at all, are almost invariably regarded as moral *patients* rather than moral *agents*, where:

> (1) X is a moral *patient* if and only if X is a legitimate object of moral concern: that is, roughly, X is an entity that has interests that should be taken into consideration when decisions are made concerning it or which otherwise impact on it.
>
> (2) X is a moral *agent* if and only if X is (a) morally responsible for, and so can be (b) morally evaluated (praised or blamed, broadly understood) for, its motives and actions.

Nothing in (1) and (2) rules out one and the same individual being both a moral agent and a patient: most humans are both.

As far as both humans and animals are concerned, it is usually assumed that this is as far as the logical geography extends. Animals can be moral agents or patients, or both. They are not, it is generally supposed, moral agents. Therefore, they must be no more than moral patients. There are, however, dissenting voices. For example, David DeGrazia writes:

> These examples support the attribution of moral agency – specifically, actions manifesting virtues – in cases in which the actions are not plausibly interpreted as instinctive or conditioned. On any reasonable understanding of moral agency, some animals are moral agents.[3]

This claim is echoed by Stephen Clark[4], Steven Sapontzis[5], and also by Evelyn Pluhar, who writes:

> Is it really so clear, however, that the capacity for moral agency has no precedent in any other species? Certain other capacities are required for moral agency, including capacities for emotion, memory, and goal-directed behavior. As we have seen, there is ample evidence for the presence of these capacities, if to a limited degree, in some nonhumans. Not surprisingly, then, evidence has been gathered that indicates that nonhumans are capable of what we would call "moral" or "virtuous" behavior.[6]

---

[3] David DeGrazia, Taking Animals Seriously (New York: Cambridge University Press, 1996). p. 203.

[4] Stephen Clark, The Nature of the Beast: Are Animals Moral? (Oxford:Oxford University Press, 1984)

[5] Steven Sapontzis, Morals, Reasons and Animals (Philadelphia: Temple University Press, 1987)

[6] Evelyn Pluhar, Beyond Prejudice: The Moral Significance of Human and Nonhuman Animals (Durham, NC: Duke University Press, 1995), p. 2.

Similar claims, although in varying forms, can be found in the work of Vicki Hearne, Jeffrey Moussiaeff Masson, Stephen Wise, Frans de Waal, and Marc Bekoff.[7] Indeed, Darwin claimed that animals can be motivated by the "moral sentiments."[8] Both DeGrazia and Pluhar express their claim in the language of agency. This, I think, is regrettable. The concept of agency is inseparable from the concept of responsibility, and hence from the concepts of praise and blame. If animals are moral agents, they are responsible for what they do, and so can be praised or blamed for this. At one time, courts of law – both non-secular and secular – set up to try (and subsequently execute) animals for perceived crimes were not uncommon.[9] I assume few would wish to recommend a return to this practice. At the core of this unwillingness is the thought that animals are not responsible, and so cannot be blamed, for what they do. If this is correct, then their characterization in terms of moral agency should be resisted.

I shall argue that there is another option: while animals are indeed moral patients, and not moral agents, they can also be moral *subjects*, where:

(3) X is a moral *subject* if and only if X is, at least sometimes, motivated to act by moral considerations.

---

[7] Vicki Hearne, Adam's Task: Calling Animals by Name (New York: Vintage Books, 1987); Jeffrey Moussiaeff Masson, Dogs Never Lie About Love: Reflections on the Emotional World of Dogs (New York: Three Rivers Press, 1997); Jeffrey Moussiaeff Masson and Susan McCarthy, When Elephants Weep: The Emotional Lives of Animals (New York: Delacorte, 1995); Stephen Wise, Rattling the Cage: Toward Legal Rights for Animals (Cambridge, MA: Perseus Books, 2000); Frans de Waal, Good Natured: The Origins of Right and Wrong in Humans and Other Animals (Cambridge, MA: Harvard University Press, 1996); Marc Bekoff, The Smile of a Dolphin: Remarkable Accounts of Animal Emotions (New York: Discovery Books, 2000); Marc Bekoff, Minding Animals: Awareness, Emotion, and Heart (Oxford: Oxford University Press, 2002).

[8] Charles Darwin, The Descent of Man (London: John Murray, 1871). Darwin did, however, stop short of claiming that animals are fully "moral beings".

[9] See E.P. Evans, The Criminal Prosecution and Capital Punishment of Animals, London, Heinemann, 1906) for a wealth of examples. See also, P. Dinzelbacher, 'Animal Trials: A Multidisciplinary Approach' in the Journal of Interdisciplinary History, Vol. 32, 2002, 405-21.

The concept of a moral subject has almost invariably been conflated with that of a moral agent: to say that Smith is motivated to act by moral considerations is thought to be equivalent to the claim that Smith is responsible for what he does. In some ways this conflation is odd: for as definitions (2) and (3) make clear, these claims are, logically, quite distinct. (2) is a claim about evaluation; (3) is a claim about motivation. Moral agency and moral subjecthood should be as conceptually distinct as the concept of evaluation is distinct from the concept of motivation. And it is reasonably clear that, *in general*, these are quite different things: the motivation for an action is one thing, the evaluation of the action or the motivation quite another. Indeed, an evaluation is often *of* a motivation.[10]

Nevertheless, there are persuasive (in the sense that almost everyone has been persuaded) reasons for supposing that this general distinction between motivation and evaluation is not applicable in the moral case: that is, in the case of *specifically* moral motivation and moral evaluation. Thus, the standard view is that there is no distinction between a moral agent and a moral subject. The reasons for this are not difficult to discern.

Imagine someone – for entirely obvious reasons, we can call him Sigmund – whose motivations are always hidden from him. The motivational component of his mind is akin to a black box: replete with states that successfully guide Sigmund's behavior, but to which he has no first-person access. There is an obvious sense in which Sigmund is, as we might put it, "at the mercy" of his motivations. He has no idea what motivates him to act in the way he does, and therefore has no control over those motivations. Because of this, Sigmund may not be an agent in the sense that a normal adult human is usually taken to be an agent: he is pulled this

[10] Suppose, for example, that hard determinism is true. In such circumstances, no one could be morally evaluated for what they do, but it would not follow that they were not the subjects of motivational states.

and way that by motivations to which he is blind. And if Sigmund is not an agent then, *a fortiori*, he is not a moral agent. Nevertheless, it is still true that he is motivated to act in various ways, even if he is blind to these motivations. Even if he is not an agent, Sigmund is nevertheless a subject of motivation. Can he also be a subject of specifically *moral* motivation? Control over his motivations might be required for Sigmund to be a moral agent. But why should it be required for him to be a moral subject?

The answer is to be found in the Kantian dictum that *ought implies can*. Since Sigmund is blind to his motivations, he has no control over them. They are, therefore, not the sorts of thing he can embrace or resist. But if he can neither embrace nor resist his motivations, it makes no sense to say that he *should* embrace or resist them. Sigmund's motivations, in this sense, have no normative dimension. They are not the sorts of things he *should* endorse or reject. His motivations make no *normative claim* on Sigmund. However, moral motivations are precisely things that make normative claims on their subjects. Morally good motivations are ones that should be embraced by their subject; morally evil motivations are ones that should be resisted. Therefore, it seems Sigmund's motivations cannot be moral ones.

The connection between normativity and control is one that has decisively shaped the most influential accounts of moral motivation. Moreover, in these accounts one finds a specific conception of control. The essence of control lies in a particular form of self-consciousness: one that, broadly, consists in understanding the principles upon which one is inclined to act. Let us consider two influential developments of this idea.

4. Kant and Aristotle: The Reflection Condition

Christine Korsgaard, that most able contemporary representative of the Kantian approach to moral philosophy, writes:

> Kant believed that human beings have developed a specific form of self-consciousness, namely, the ability to perceive, and therefore to think about, the grounds of our beliefs and actions as grounds. Here's what I mean: an animal who acts from instinct is conscious of the objects of its fear or desire, and conscious of it as fearful or desirable, and so as to-be-avoided or to-be-sought. That is the ground of its action. But a rational animal is, in addition, conscious that she fears or desires the object, and that she is inclined to act in a certain way as a result. That's what I mean by being conscious of the ground as a ground. So as rational beings we are conscious of the principles on which we are inclined to act. Because of this, we have the ability to ask ourselves whether we should act in the way we are instinctively inclined to. We can say to ourselves: "I am inclined to do act-A for the sake of end-E. But should I?"[11]

This ability – to understand the principles on which we are inclined to act – is, according to Korsgaard, part of the essence of morality in the Kantian image:

> [T]he capacity for normative self-government and the deeper level of intentional control that goes with it is probably unique to human beings. And it is in the proper use of this capacity – the ability to form and act on judgments of what we ought to do – that the essence of morality lies, not in altruism or the pursuit of the greater good.[12]

---

[11] Christine Korsgaard, "Fellow Creatures: Kantian Ethics and our Duties to Animals," Tanner Lectures on Human Values, ed., G. Peterson (Salt Lake City: University of Utah Press, 2004), p. 148-9.

[12] Christine Korsgaard, "Fellow Creatures: Kantian Ethics and our Duties to Animals," Tanner lectures on Human Values, ed., G. Peterson (Salt Lake City: University of Utah Press, 2004), p. 140.

Any behavior that is not subject to this sort of normative self-government is not moral behavior. If a creature is unable to reflect on what it does – to ask itself whether this is, in the circumstances, a morally good thing to do, ask itself whether its' motive is a morally good one – is not a moral creature.

Grace, the elephant, is presumably unable to reflect on her motivations in this way. She cannot reflect on her motivations: she cannot ask herself whether these are good or bad, whether she should embrace or resist them. From the Kantian perspective, Grace's motivations fail to qualify as moral ones. At most, her motivations count only as non-moral facsimiles of concern – sentiments that resemble the various forms of concern in their phenomenological profile, but which lack the essential moral core that would convert this phenomenology into a properly moral emotion. Her motivations simply push Grace this way and that – she is the moral equivalent of a cork bobbing on the oceans of motivation. The 'specific form of self-consciousness' that consists in the understanding of the principles on which she in inclined to act would (allegedly) change everything. Grace is lifted out of the ocean, transformed from an individual pushed this way and that by her motivations to one capable of standing outside those motivations – serenely observing and adjudicating between them.

Aristotle paints a picture that is similar, at least in broad outline. In this passage from the *Nichomachean Ethics*, Aristotle emphasizes the psychological complexity of the virtues:

> But for actions in accord with the virtues to be done transparently or justly it does not suffice that they themselves have the right qualities. Rather, the agent must also be in the right state when he does them. First he must know that he is doing virtuous actions;

second, he must decide on them, and decide on them for themselves; and, third, he must also do them from a firm and unchanging state.[13]

For an action to be an expression of a virtue, it must not simply be an example of what would commonly be regarded as a virtuous action (have the "right qualities"). In addition, the agent must (a) know that he is performing a virtuous action, and (b) perform the action because it is a virtuous action ("decide on them for themselves"), and (c) this decision must be an expression of a stable disposition on the part of the agent. Conditions (a) and (b) collectively impose a minimal condition of reflection on the virtuous agent. To satisfy these conditions, the agent must understand what a virtue is, and be motivated by this understanding to perform a certain action because it would be expressive of this virtue. I shall call this the *reflection condition* on possession/expression of a virtue:

> For action ϕ, performed by agent A, to be an expression of virtue, V, it is necessary that A (i) be able to understand that ϕ is an instance of V, and (ii) A must perform ϕ because he understands that ϕ is an instance of V and wishes to be virtuous.

Aristotle closely ties the reflection condition to the *normative grip* of a virtue. Aristotle has two, mutually reinforcing explanations of the normativity of virtue. First, he emphasizes that the virtues are things that must be acquired, and this can be done well or imperfectly.[14] This is why we can be praised for our possession of a virtue, and blamed for our lack of it. If virtues were

---

[13] Aristotle, Nichomachean Ethics, trans. T. Irwin (Indianapolis : Hackett, 1999) 1105a27-35

[14] "The virtues arise in us neither by nature nor against nature. Rather, we are by nature able to acquire them, and we are completed through habit. . . . Virtues, by contrast [to the senses], we acquire, just as we acquire crafts, by having first activated them. For we learn a craft by producing the same product that we must produce when we have learned it; we become builders, for example, by building, and we become harpists by playing the harp. Similarly, then, we become just by doing just actions, temperate by doing temperate actions, brave by doing brave actions." Nichomachean Ethics, 1103a19-03b2.

like the senses, whose possession is a matter of nature and not our efforts, this praise or blame would make no sense.

Second, a virtue is something that can be exercised well or poorly. In a well-known passage from the *Nichomachean Ethics*, Aristotle remarks:

> So also getting angry, or giving and spending money, is easy and everyone can do it; but doing it to the right person, in the right amount, at the right time, for the right end, and in the right way is no longer easy, nor can everyone do it. Hence, doing these things well is rare, praiseworthy, and fine.[15]

These are often difficult matters that call for judgment. This judgment can be executed well or poorly; mistakes are always possible. The possibility of mistake in the application of a virtue underlies the normative status of a virtue.

Therefore, with regard to the question of whether Grace's behavior is virtuous, Aristotle would presumably recommend to us the following questions: (1) Does Grace know that what she is doing is virtuous? (2) Does Grace behave in the way she does because she wants to be virtuous? (3) Does Grace think Eleanor is the appropriate recipient of this sort of behavior? (4) Does Grace think her behavior is proportional to Eleanor's suffering? (And so on, and so forth.) If Grace is unable to entertain these (and related) thoughts, then Grace's behavior does not qualify as the expression of a moral virtue.

Both Kant and Aristotle would agree that animals do not qualify as subjects of moral motivation because, and to the extent, they are unable to critically scrutinize their motivations and the relation between their motivations and their actions.[16] Underlying this emphasis on

---

[15] Aristotle, Nichomachean Ethics, 1109-a27-30

[16] See also B.A. Dixon, Animals, Emotion and Morality (New York: Prometheus Books, 2008)

critical reflection are the closely related ideas of normativity and control. Without the ability to reflect on its motivations, no sense can be given to the idea that a motivation is something an animal *should* have rather than something it merely *does* have. Without this ability, the motivations of animals simply push them in one direction then another. In the absence of the capacity for critical moral reflection, the motivations of animals have no normative status.

Aristotle and Kant are two prominent exemplars of the orthodox view of moral motivation. We might dub this orthodoxy the SCNM view. The ability to critically SCRUTINIZE one's motivations is a necessary condition of having CONTROL over those motivations. Control, in turn, is required for those motivations to have NORMATIVE status – for them to exert a normative grip on their subjects. And normative status is a necessary condition of MORAL status. The SCNM schema is the target of this paper. In particular, I shall argue for two claims. First, we have no workable account of the connection between the SCRUTINY and CONTROL. Second, there is an alternative account of NORMATIVE status, one that does not derive from the control a subject supposedly has over her motivations.

5. The Idiot

Let us introduce the character of Myshkin, named after the prince in Dostoevsky's, *The Idiot*. Prima facie, Myshkin has the soul of a prince: throughout his life, he performs many acts that seem to be kind or compassionate. He performs these acts because he is the subject of sentiments or emotions that – again, at least *prima facie* – seem to be kind or compassionate ones. When he sees another suffering, he feels compelled to act to end or ameliorate that suffering. When he sees another happy, he feels happy because of what he sees. If he can help

someone get what they want without hurting anyone else, he will help because he finds that he enjoys doing it. In short, Myshkin deplores the suffering of others and rejoices in their happiness. His actions reflect, and are caused by, these sentiments. What Myshkin does not do, however, is subject his sentiments and actions to critical moral scrutiny. Thus, he does not ever think to himself things like: "Is what I am feeling the right thing to feel in the current situation – that is, is what I *am* feeling the same thing as what I *should* be feeling?" Nor does he think to himself things like: "Is what I propose to do in this circumstance the (morally) correct thing to do?" He is incapable of doing this. His dealings with others operate on a more visceral level. What we might think of as Myshkin's moral profile looks something like this:

> (M1) (i) Myshkin performs actions that seem to be good, and (ii) Myshkin's motivation for performing these actions consists in sentiments or emotions that seem to be good, but (iii) Myshkin is able to subject neither the actions nor the sentiments to critical moral scrutiny.

Are (i)-(iii) sufficient for Myshkin to qualify as a moral subject – for his motivations to count as moral ones? Aristotle and Kant would deny that they are. But how reasonable is this denial? Myshkin spends his life helping others, and does so because he delights in the happiness of others and deplores their suffering. Perhaps, we might distinguish (a) whether something qualifies as a moral subject, in the sense of being motivated by moral concerns, and (b) whether something qualifies as a moral subject in precisely the same way we do. This, however, merely amounts to an invitation to expand our conception of the moral subject. Aristotle and Kant would likely respond with a firm, "No thanks!" To inconvenience them, the invitation needs to be strengthened into something more like an offer that can't be refused.

As a beginning to this strengthening process, let us focus on the characterization of Myshkin's actions and feelings as ones that *seem* to be good. The motivation for this characterization is fairly clear. Those who reject the idea that Myshkin is a moral subject will also reject the idea that his emotions and actions count as good (e.g. kind or compassionate) – such a description would be true only of a moral subject. However, suppose someone – Marlow, after the skilled scrutinizer of motivations who narrates some of Joseph Conrad's novels – is capable of the sort of critical moral reflection of which Myshkin is incapable. And suppose, on the basis of this reflection, Marlow were to endorse the same sorts of sentiments and actions that Myshkin has and performs. For any given circumstance C, Myshkin has sentiment S and, as a result, performs action A. Marlow, an adept moral scrutinizer of his sentiments and actions, independently comes to the conclusion that in circumstance C, it is morally correct to have sentiment S and perform action A. Suppose, further, that Marlow – who we might think of as a *ideal moral spectator* – invariably reaches the correct moral conclusion. If this were the case, we could strengthen (M1) as follows:

> (M2) (i) Myshkin performs actions that are, in fact, good, and (ii) Myshkin's motivation for performing these actions consists in emotions or sentiments that are, in fact, the morally correct ones to have in the circumstances, but (iii) Myshkin is able to subject neither the actions nor the sentiments to critical moral scrutiny.

The transition from (M1) to (M2) is not insignificant: it means that Myshkin has *external* reasons for his actions.[17] An internal reason for action is one that furthers a certain motive of the agent – whether the agent actually has this motive, or whether she would come to have this

[17] Bernard Williams, 'Internal and external reasons', in his Moral Luck (Cambridge: Cambridge University Press, 1981), 101–13.

motive by following a 'sound deliberative route'.[18] An agent would have an external reason to ϕ, on the other hand, if (i) she has reason to ϕ, and (ii) none of her motives or interests is furthered by ϕ-ing.

In the case of (M1), it can be argued that Myshkin has no moral reasons for his actions: The Myshkin of (M1) is at the mercy of his motivations: he has no control over whether he embraces or resists them, and so they exert only causal, but not normative, pressure on his actions. But normativity is essential to an item's being a reason rather than a cause. Myshkin's motivations, therefore, while they might cause his behavior, are not moral reasons. However, (M2) introduces a new type of moral reason: the Myshkin of (M2) has external moral reasons for his actions. Myshkin cannot entertain those reasons, and therefore they are not internal. Nevertheless, they are moral reasons that exist for Myshkin to do what he does.

Would we want to deny that Myshkin – as characterized by (M2) – is a moral subject? He does things that are, in fact, morally good, and he does this from a motivation that is, in fact, a morally commendable one. There are reasons – moral reasons – that connect his sentiments and actions, although these are external ones. Should we really be so confident in the claim that Myshkin does not qualify as a moral being? This is still an invitation to think of morality in a certain way; but is it still an invitation that can be reasonably declined?

Perhaps: but a skeptical moral philosopher who wishes to decline the invitation is committed to the idea that a necessary condition of being a moral subject is that one possesses internal reasons for one's actions. This idea is, of course, far from unreasonable. The existence

[18] Bernard Williams, Making Sense of Humanity (Cambridge: Cambridge University Press 1995), p. 35.

of external reasons is controversial. Bernard Williams, who introduced them into philosophical discourse, went on to argue that there were no such things.

Indeed, even someone who accepts the existence of external moral reasons might be chary of the claim that Myshkin is a moral subject. There is, one might argue, a certain deficit that accompanies Myshkin's lack of internal reasons. There is a clear sense in which, although Myshkin gets things "right" – that is, in given circumstance C, he performs the morally correct actions and he experiences the morally correct emotions – he gets it right by *accident*. Myshkin might *get things right*, morally speaking, but he does not get things right *for the right reasons*. That is what is required for Myshkin to be a moral subject.

The picture of a reflective moral subject as someone who can get things right – i.e. identify the correct moral conclusion – *for the right reasons* is one that many will find compelling. Indeed, this is one way of developing the connection between normativity and control that is central to the orthodox way of thinking about moral motivation. Therefore, let us try and remove the offending element of contingency in Myshkin's moral decision making. We can do this by, to some extent, internalizing Myshkin's reasons – although, crucially, not in a way that converts him into a fully-fledged reflective moral subject.

Suppose that Myshkin's "getting things right" – that is, having the right sentiments and performing the right actions in given circumstances – is not as accidental as it seems. Suppose Myshkin has good reasons for his action-guiding sentiments. It is just that these reasons are not available to his conscious, rational scrutiny. They are reasons that have been embodied in Myshkin's sub-personal, non-conscious, inferential processing operations. For example, if our tastes ran to the modular (although it is strictly optional to think in this way), we might suppose

that Myshkin has a "moral module," the operations of which are cognitively impenetrable – that is, cannot be penetrated by subsequent belief- and concept-forming operations. Nonetheless, the operations of the module are reliable ones and lead to correct moral sentiments and actions with the same level of success as our ideal, rational spectator, Marlow. In such circumstances, we could amend (M2) to the following:

> (M3) (i) Myshkin performs actions that are good, and Myshkin's motivations for performing these actions consist in feelings or sentiments that are the morally correct ones to have in the circumstances, and (iii) Myshkin's emotions are the product of a reliable mechanism – a "moral module" – that links perceptions of morally salient features with appropriate emotional responses, but (iv) Myshkin cannot critically scrutinize the deliverances of this mechanism.

(M3) removes the element of *contingency* implicit in (M2). Myshkin (M3) does not "get things right" by accident. However, he still, arguably, does not get things right "for the right reason." Myshkin simply acts on the deliverances of his moral module – emotions of various sorts – he is unable to critically scrutinize these deliverances.

Would we want to deny that the Myshkin of (M3) is a moral subject? Would it not be more reasonable to suppose that this Myshkin is a moral subject, merely a somewhat different one to the critical moral appraisers we take ourselves to be? The primary opposition to this claim is likely to derive from the thought that while (M3) might have removed the offending element of contingency from Myshkin's moral profile, it has not done this in the right way. (M3) might mitigate the contingency, but it does nothing to alleviate Myshkin's absence of control over what he feels and how he acts. Myshkin has no idea what is going on in his moral

module. He has no authority over its workings and no control over its outputs. But control, it is thought, is a necessary condition of normativity. If Myshkin has no control over the emotional deliverances of this module then these deliverances have no normative status: they are not the sorts of things Myshkin should embrace or resist. Ultimately, the denial that Myshkin's motivations count as moral ones has its roots in the related ideas of authority and control.

Compare Myshkin's moral profile with that of Marlow:

> (M4) Marlow performs actions that are good, and (ii) Marlow's motivations for performing these actions consist in emotions or sentiments that are the morally correct ones to have in the circumstances, and (iii) Marlow's emotions are the product of a reliable mechanism – a "moral module" – that links perceptions of morally salient features with appropriate emotional responses, and (iv) Marlow can critically scrutinize the deliverances of this mechanism.

Myshkin and Marlow differ only with respect to condition (iv). Marlow's ability to critically scrutinize the deliverances of his "moral module" gives him a control over his motivations that Myshkin lacks. Therefore, Marlow's motivations possess a normative status that Myshkin's motivations do not. Therefore, Marlow's motivations can be moral ones while Myshkin's motivations cannot. I shall now argue that this picture is (a) unworkable, and (b) unnecessary.

8. S-C: From Scrutiny to Control

The reason this picture is unworkable is there is no satisfactory account of how the ability to engage in critical scrutiny of one's motivations is supposed to endow a subject with control

over those motivations. Consider, first, the idea of critical scrutiny of motivations. This is best thought of as open-ended, multi-layered phenomenon:

1. Marlow recognizes that he has certain motivations – states that incline him to act in one way or another. This he can think: I am inclined to Φ because of motivation, M.

On the basis of this recognition, he can ask himself certain questions such as:

2. Is M a motivation I should endorse or reject?

This recognition and interrogation is, of course, only the beginning of Marlow's critical moral scrutiny. Marlow also understands how he should to proceed to answer (2). Thus:

3. Marlow brings to bear moral principles or propositions that he antecedently holds. He assesses M in the light of these principles. Is M compatible with, incompatible with, or entailed by, these principles?

In answering these questions, Marlow can come to decide whether M is permissible, impermissible, or perhaps obligatory. Moreover, if Marlow is an appropriately critical subject, he will realize that further questions need to be asked concerning his antecedent adoption of these moral principles or propositions. Thus:

4. Marlow asks if there is any reason to suppose that his antecedent acceptance of these moral principles has been shaped by irrational or extra-rational factors.

Not only does he ask this question, he can also make a decent fist of answering it – by engaging in (what seem to him) processes of impartial and honest reflection on the type of upbringing and education he has had, and the likely consequences of these on his processes of moral deliberation. Furthermore:

5. Marlow asks if there is any reason to suppose that his ability to properly assess compatibilities and incompatibilities between motivations and principles might be compromised by certain non-rational biases that he has acquired.

Once again, Marlow does not simply ask this question but attempts to answer it – perhaps through a similar examination of his history, making (what seem to him) reasonable inferences concerning what sort of inferential biases this history might have produced, and so on.

If this is what critical scrutiny is, at least in broad outline, the question is this: why – in virtue of which of its properties – does this does this sort of scrutiny endow a subject with control over her motivations? One possibility is that the relevant properties are *phenomenological* ones. There are important differences between the phenomenological character of the experiences undergone by Myshkin and Marlow when they find themselves in situations that call for a moral response. Myshkin's motivational states are conscious ones and so possess associated phenomenology. Marlow, however, brings something else to the phenomenological table. Marlow engages in moral deliberation in a way that Myshkin never could. Myshkin *lives* his motivations: he has them and acts on them and there is a phenomenology associated with this. But Marlow can also agonize over his motivations. There may be nagging questions. I am inclined to do this – but should I? Should I embrace this inclination or resist it? These questions have their own phenomenology. With Marlow also, there is the phenomenology associated with bringing moral principles to bear. If all goes well, Marlow may encounter the phenomenological sense of a satisfactory resolution achieved – or failing that the phenomenology accompanying the suspicion that his deliberations have not yet

reached a successful resolution. Marlow's motivations are, as Sartre once put it, *troubled* in a way Myshkin's can never be.

However, while there are clear phenomenological differences between Myshkin and Marlow, it is unlikely that any meaningful model of control can be wrested from them. Thus, for example, the hard determinist is, typically, at pains to not deny the phenomenology of freedom. He simply denies that our sense of being free adds up to our actually being free. Similarly, that we seem to have control over our motivations, a seeming that is bound up with the phenomenology of moral deliberation, does not entail that we actually do. Marlow's phenomenology is as compatible with there mere illusion of control as it is with the presence of control. Control cannot be engineered from phenomenology.

A *prima facie* more promising approach to understanding control appeals not to the phenomenological status of Marlow's moral deliberation but its *meta-cognitive* status. Marlow can have *higher-order* thoughts about his motivations, and this is the source of his control over them. I shall argue that this approach will not work.

9. Meta-Cognition and Control

It is easy to feel the intuitive pull of the idea that Marlow's ability to meta-cognize could imbue him with control over his motivations. Myshkin is the subject of motivations of various sorts. However, because he cannot reflect on those motivations, but simply act on them, he is, in one fairly clear sense, at their "mercy." These motivations push him this way and that – causing him to act in one way or another. Marlow's meta-cognitive abilities, on the other hand, allow him to survey and critically evaluate his motivations. These abilities intuitively support a picture of

Marlow as an individual quite different from Myshkin. Myshkin is "at the mercy" of his motivations. He has them, and he acts on them – and that is all. He is tossed this way and that – a bobbing cork on a sea of motivations. Marlow's meta-cognitive abilities, on the other hand, allow him to float above this sea. He is able to observe his motivations, and, by following certain evaluative procedures, adjudicate between them.

If this picture is indeed intuitively compelling, this is simply a symptom of a certain kind of magical thinking that often underpins an appeal to the meta-level. To suppose that Marlow's meta-cognitive abilities can confer control over his motivations is to fall victim to a common, but illicit, picture – a type of fallacy – that I shall label the *miracle-of-the-meta*. I am going to introduce this fallacy via a particularly obvious instance: the higher-order thought (HOT) model of consciousness.[19] The idea underlying the HOT model of consciousness is that meta-cognition, in the form of higher-order thoughts, confers consciousness on mental states.[20]

In order to understand HOT models of consciousness, two distinctions are required:

(1) *Creature* versus *state* consciousness. We can ascribe consciousness both to creatures or individuals (for example, Jones is conscious as opposed to asleep) and to mental states (for example, Jones' belief that Ouagadougou is the capital of Burkina Faso is, sometimes, a conscious belief). The HOT model is an attempt to explain state consciousness, not creature consciousness.

(2) *Transitive* versus *intransitive* consciousness. Jones is conscious of the fact that Ouagadougou is the capital of Burkina Faso. This is transitive consciousness. Transitive

[19] See, for example, David Rosenthal, "Two Concepts of Consciousness", *Philosophical Studies* 49, (1986), 329-59.

[20] For extended discussion, see Mark Rowlands, "Consciousness and Higher-Order Thoughts" *Mind and Language* 16, (2001), 290-310, and *The Nature of Consciousness* (Cambridge: Cambridge University Press 2001), chapter 5.

consciousness is a form of creature consciousness. Only creatures are transitively conscious; mental states are not. Intransitive consciousness, however, can be ascribed both to creatures and states. Jones is intransitively conscious when he is conscious, as opposed to asleep, knocked out, or otherwise unconscious. A state – Jones' belief that Ouagadougou is the capital of Burkina Faso – is intransitively conscious when he is consciously entertaining it.

The core idea of HOT models is that intransitive state consciousness can be explained in terms of transitive creature consciousness. First: a mental state M, possessed by creature C, is intransitively conscious if and only if C is transitively conscious of M. Second, a creature, C, is transitively conscious of mental state M if and only if C has a thought to the effect that it has M. Jones' belief that Ouagadougou is the capital of Burkina Faso is intransitively conscious when, and only when, Jones is transitively conscious of this belief. And Jones is transitively conscious of this belief when, and only when, he has a higher-order thought about this belief. Thus, intransitive state consciousness is to be explained in terms of transitive creature consciousness, and transitive creature consciousness is to be explained in terms of a higher-order thought – a thought about a mental state.

The HOT account of faces a reasonably obvious dilemma. Suppose Jones is in pain. According to the HOT model, this pain is intransitively conscious if and only Jones has a higher-order thought about this pain – a thought to the effect that he is in pain. However, either the higher-order thought is itself intransitively conscious or it is not. If the higher-order thought about the pain is itself intransitively conscious, then the HOT account has done nothing to explain intransitive state consciousness. The attempted explanation has cited a state that

possesses the very property that the explanation is supposed to explain. The nature of intransitive state consciousness has not been explained; this explanation has merely been deferred. If intransitive consciousness of a mental state requires a higher-order thought about states, and if the higher-order thought is intransitively conscious, then we would need to postulate another thought – a third-order or "higher-higher-order thought" in order to explain the intransitive consciousness of the higher-order thought. But then, once again, the question arises of whether this third-order thought is or is not intransitively conscious. In other words, supposing that the higher-order thought is intransitively conscious yields an infinite regress.

The way to prevent this regress is to deny that the higher-order thought must be intransitively conscious. Jones' thought to the effect that he is in pain need not be intransitively conscious in order to ground the intransitive consciousness of his pain. This, therefore, is the position typically adopted by HOT theorists. However, it is a position that has crippling difficulties of its own. These difficulties make up the second horn of the advertised dilemma.

Suppose that the higher-order thought that (supposedly) confers intransitive consciousness on Jones' pain is not intransitively conscious. Then Jones will have no idea that he is thinking this. But how can thinking that he is in pain make him aware of his pain if he has no idea that he is thinking he is in pain? The problem is that the HOT account understands intransitive state consciousness via transitive creature consciousness. Jones' pain is intransitively conscious to the extent that he – the creature, Jones – is transitively conscious of it. But an intransitively unconscious thought – a thought that he has no idea he is having – is not the sort of thing that can make him transitively conscious of anything. In general, intransitively unconscious thoughts do not make us aware of anything – that is *precisely what it is for them to*

be intransitively unconscious. That is, intransitively unconscious thoughts are not the sort of thing that can underwrite transitive creature consciousness.

I, like the mythical Jones, believe that Ouagadougou is the capital of Burkina Faso. When I consciously entertain this belief it makes me aware of a fact – the fact that Ouagadougou is the capital of Burkina Faso. However, it is only rarely that I consciously entertain the belief. Most of the time this belief is one of my non-conscious beliefs – which, of course, at any given time number the vast majority of my beliefs. The non-consciousness of a belief consists in the fact that it does not make me aware of what it would make me aware of if it were conscious. My non-conscious belief that Ouagadougou is the capital of Burkina Faso is non-conscious because, and to the extent, it does not make me aware that Ouagadougou is the capital of Burkina Faso. This is precisely what it is for the belief to be non-conscious. More generally: intransitively non-conscious states do not make the creature that has them transitively conscious of anything at all – that is precisely what makes them intransitively non-conscious. Conversely, if they do make the creature that has them transitively conscious of their objects, that is because they are, in fact, intransitively conscious. But the explanatory strategy of HOT models is to explain intransitive states consciousness in terms of transitive creature consciousness. And this means that they cannot allow the relevant higher-order thoughts to be intransitively non-conscious.

The dilemma facing the HOT model is, therefore, a crippling one. If the higher-order thoughts are intransitively non-conscious, then they cannot underwrite transitive creature consciousness and the HOT model fails. If the higher-order thoughts are intransitively

conscious, the HOT model faces an infinite regress. Either way, the prospects for a successful HOT explanation of consciousness are grim.

The idea that a higher-order thought about a mental state could explain the intransitive consciousness of that state is one example of the fallacy I have called the *miracle-of-the-meta*. In essence, the fallacy involves the attribution of miraculous powers to meta-cognition or meta-cognitive abilities. The idea that meta-cognition could explain the normative status of a motivation, via explaining a subject's control over that motivation, is, I shall now argue, another example of this sort of miraculous thinking.

Consider, first, the general form this "miracle" takes. There is a set of first-order states that are taken to be intrinsically lacking in some regard. That is, taken in themselves they seem to lack a property that we routinely accept that they have. This property is then supplied through extrinsic means. These extrinsic means consists in a set of second-order states that take these first-order states as their objects. These second-order states supply the first-order states with the feature they were taken to be lacking. However this strategy is fatally flawed. In particular, the issue that arose with regard to the first-order states – their apparent lack of possession of a given property – is simply reiterated at the second-order level. With regard to HOT models of consciousness, the delinquent property of first-order mental states was consciousness. Second-order states – higher-order thoughts – were invoked to supply this delinquent property to the first-order states. However, the same issue of delinquency that arose at the first-order level is also reiterated at the second-order level. Specifically the second-order states could supply the delinquent property of intransitive consciousness to the first-order states only if they already possess this property. This undermines the pretensions of the HOT model to

provide an explanatory account of what intransitive state consciousness is. The appeal to higher-order states ultimately presupposes the feature it is supposed to explain.

The idea that we can explain a subject's control over her motivations by appeal to meta-cognitive abilities is, I shall now argue, a version of the same fallacy. To see why, let us first return to the predicament of Myshkin. We are tempted to suppose that in the absence of the relevant meta-cognitive abilities – the ability to form higher-order thoughts about his motivations – Myshkin is at the "mercy" of his motivations. They push him in this way and that. Unable to critically scrutinize these motivations, he has no control over what they cause him to do. Meta-cognitive abilities, however, supposedly transform Marlow. Armed with these abilities, he can sit above the motivational fray: observing, judging and evaluating his motivations, coolly deciding the extent to which he will allow them to determine his decisions and actions. This is gives Marlow a control over his motivations that Myshkin lacks and so lends his motivations a normative status that Myshkin's lack.

However, what if Marlow's meta-cognitive abilities were not at all like their portrayal in this account? What if they were, in certain crucial respects, like the motivations that they take as their objects? Marlow can undoubtedly think to himself thoughts such as, "What is it that is motivating me to do this?" and "Is this motivation something that I should act upon, or something that I should reject?" In answering questions such as this, he can bring to bear antecedently accepted moral principles or propositions, and assess the compatibility or lack thereof, between principle and motivation. He can, in turn, question his antecedent commitments, and his ability to accurately and impartially makes judgments of compatibility, and so on. However, suppose the answers he gives to all these questions are ones over which he

has little control: being, for example, crucially sensitive to features of the situation in which he finds himself – features of which he has only a dim awareness and whose influence on him largely escapes his conscious grasp. In other words, while Marlow could still observe and evaluate his motivations this observation and evaluation would be both clouded and shaped by contextual factors of which he has little awareness and over which he has little control. In such a circumstance, it would seem that Marlow is at the "mercy" of his meta-cognitive assessments of his motivations. These assessments pull him in this way or that – make him evaluate his motivations in one way or another – but they work in subterranean ways. He has little grasp, and just as little control, over their workings.

In raising these worries, I am, of course, alluding to recent influential work in moral psychology of a broadly *situationist* orientation.[21] Situationist accounts of the moral subject can be contrasted with *dispositionalist* accounts. According to the latter, very roughly, the moral subject is constituted, in part, by a set of dispositions that are internal to the subject and the subject is responsible for their formation and preservation. These dispositions make up what we might call the *character* of the person. Included in these dispositions will be ones that pertain to the evaluation of motivations. Thus, the character of the moral subject is, in part, made up of dispositions to classify given motivations as ones that should be acted upon or ones that should be resisted.

Situationist accounts, however, see things differently. According to a (probably unrealistically) strong version of situationism, a person's character is as malleable as the situations in which she finds herself. Change the situation, and the subject's dispositions to

[21] See, for example, Philip Zimbardo, *The Lucifer Effect: How Good People Turn Evil* (New York: Random House 2007)

morally evaluate motivations in one way rather than another will also change. Thus we might imagine Marlow as a participant in a Stanford Prison Experiment, or as a guard at Abu Ghraib. If the situationist account is correct, we should expect Marlow's evaluation of his motivations to vary with these variations in circumstances. However, the relevant contextual factors operate at all levels of the process of scrutiny. They apply not simply to Marlow's evaluations of his motivations but also to all the resources he brings to bear on these evaluations: his antecedent adoption of certain moral principles, his judgments of compatibility and lack thereof, and so on.

Therefore, the idea that Marlow's meta-cognitive abilities confer control and, via control, normativity, on his motivations – a normativity that Myshkin's motivations therefore lack – seems, at the very least, a hostage to empirical fortune. If Marlow is, as the situationist argues, at the mercy of his meta-cognitive assessments of his motivations in the way that Myshkin has been portrayed as being at the mercy of his motivations, then it is unclear how possession of meta-cognitive abilities can imbue Marlow with control over his motivations; and therefore it is equally unclear how they can confer normativity on these motivations.

However, any resolution of this debate is, for our purposes, relatively unimportant. Underlying these empirical concerns is a more important conceptual point. The appeal to meta-cognition attempts to explain the normativity of our emotions by way of our control over them. This overlooks the fact that the very issue of control that arises at the level of motivations is also going to be replicated at the (second, third, fourth, and so on order) level of our evaluation of those motivations. This failure to see what is, ultimately, a fairly obvious point is a symptom of the pull exerted on us by the miracle-of-the-meta. Meta-cognition, of the sort embodied in principles (1)-(5), was supposed to allow Marlow to sit above the motivational fray, and calmly

pass judgment on his motivations, thus providing him with control over those motivations and so transforming them into normative items. However, there is no reason to suppose that meta-cognition is above this motivational fray. If first-order motivations can pull Myshkin this way and that, then second-order evaluations of those motivations can do exactly the same to Marlow. If Myshkin is indeed at the "mercy" of his first-order motivations, as the traditional picture would have us believe, then, logically Marlow is similarly at the "mercy" of his second-order evaluations of these motivations. In short, second-order evaluation of our first-order motivations cannot lift us above the motivational fray that we think endemic to the first-order motivations for the simple reason that we can be motivated to evaluate our motivations in one way rather than another. Nothing miraculous happens at the meta-level: whatever features, patterns and, most importantly for our purposes, shortcomings we find at the first-order level, these are merely reiterated at the second-order level.

10. Conditions of Moral Subjecthood

The failure to identify a workable account of the connection between SCRUTINY and CONTROL would be troubling if the only way to explain the normativity of motivations was via a subject's control over them. However, there is another, relatively familiar, account of normativity available: an objective consequentialist, evaluationally externalist account. Such an account will assume a fairly strong sense of ethical objectivity, according to which situations contain features that make them good or bad independently of the subjective states of the agent. The evaluation of a motivation will then, be a function of whether it systematically (as opposed

to accidentally) promotes good- or bad-making features of situations.[22] The normative status of a motivation is, therefore, explained in terms of its relation to external factors rather than the subject's control over it.

This objective consequentialist, evaluationally externalist position can provide the basis for an altogether more modest conception of a moral subject. Myshkin qualifies as a moral subject if he meets three conditions: (1) possession of an appropriate *sensitivity* to pertinent features of his environment, where (ii) this sensitivity is *normative*, and (iii) is grounded in a *reliable mechanism*.

First, Myshkin's emotions *track* the good- or bad-making features of a situation. When Myshkin encounters someone suffering he feels a strong desire to alleviate their suffering. When he encounters someone happy, he rejoices in this, and so on. The notion of *tracking* can be captured via counterfactuals. For example, ceteris paribus, if the situation did not contain a person who was suffering, then Myshkin would not have, specifically, this emotion and the resulting desire, and so on.

Second, Myshkin's sensitivity is normative: it can be assessed as correct or incorrect. Marlow, our imagined ideal spectator, endorses Myshkin's emotional responses to environmental circumstance, and he does so because he is, as much as anyone can be, in possession of the correct moral theory.

[22] See, for example, Julia Driver, Uneasy Virtue (Cambridge University Press, 2000)

Third, Myshkin's normative sensitivity does not happen by accident. This sensitivity is grounded in a mechanism – a moral module if you like – that generates certain emotions in given circumstances.

On this account, at the core of moral motivation we find *normative sensitivity grounded in a reliable mechanism*. Once a subject meets these conditions, that subject is a moral one. The possession of meta-cognitive abilities almost certainly expands the type and variety of moral reasons an individual is capable of entertaining. However, when grounded in a reliable mechanism, normative sensitivity to the good- and bad-making features of a situation is sufficient to have a basic moral reason. It is in this sense that Myshkin possesses moral reasons and so counts as a moral subject.

It is in this sense, too, that Grace might possess moral reasons. If Grace's emotions are appropriately sensitive to the good- and bad-making features of situations, if this sensitivity is normative, and if the implicated emotions are produced by a reliable mechanism, this is all that is required for Grace to qualify as acting for moral reasons. If the empirical evidence currently being amassed by cognitive ethologists is best explained in terms of these three features, then we can legitimately conclude that Grace, and many other social mammals, are moral subjects.