Equal Rights for Zombies? Phenomenal Consciousness and Responsible Agency

2,988 words (main body)

Abstract (133 words)

Intuitively, moral responsibility requires conscious awareness of what one is doing, and why one is doing it, but what kind of awareness is at issue? Levy (2014) argues that phenomenal consciousness—the qualitative feel of conscious sensations—is unnecessary for moral responsibility. He claims that only access consciousness—the state in which information (e.g., from perception or memory) is available to an array of mental systems (e.g., such that an agent can deliberate and act upon that information)—is relevant to moral responsibility. I argue that a wide class of views entail that the capacity for phenomenal consciousness is necessary for moral responsibility. I focus in particular on considerations inspired by Strawson (1962), who puts a range of qualitative moral emotions—the reactive attitudes—front and center in the analysis of moral responsibility.

1. Introduction

Intuitively, moral responsibility requires conscious awareness of what one is doing, and why one is doing it, but what kind of awareness is at issue? In *Consciousness and Moral Responsibility* (2014), Neil Levy argues that phenomenal consciousness—the qualitative feel of conscious sensations—is not necessary for moral responsibility. He claims that access consciousness—the state in which information (e.g., from perception or memory) is available to an array of mental systems (e.g., such that an agent can deliberate, report, and act upon that information)—is the

only type of consciousness necessary. However, Levy's argument against the necessity of phenomenal consciousness begs the question against the broad class of views inspired by Strawson's "Freedom and Resentment" (1962), which puts a range of qualitative moral emotions—the reactive attitudes—front and center in the analysis of moral responsibility.

2. Levy's Argument

Levy's argument appeals to philosophical zombies, who are functionally identical to "normal"[1] human beings but lack phenomenal consciousness (PC). Let us assume for the moment that zombies are conceivable, although this is controversial. I will return to this controversy in §4, but if it turns out that zombies are inconceivable, that result would suit the broader aims of this essay relatively well. (Also note that Levy's argument will only ask us to conceive of zombies who are *functionally*, rather than *physically*, identical to us; prima facie, the former seems easier to conceive than the latter.) Levy argues that:

> since zombies are functional duplicates of us, there is nothing we can do that they can't. They are able to perform morally significant actions just as we are. They are able to do so after due deliberation. They are able to exercise control over their actions. Indeed, they seem capable of fulfilling almost any proposed sufficient conditions of moral responsibility. Since this seems… to be the case, it also seems as though it cannot be phenomenal consciousness that is required for moral responsibility. (2014, 28)

Levy concludes that access consciousness (AC) is the only type of consciousness necessary for moral responsibility (MR).[2] The part of his argument of interest to me here seems to be the following:

(1) Zombies lack PC.

[1] Throughout this paper, I put the term "normal" in scare quotes. The very idea of "normal" human beings is obscure and arguably ableist, but I don't have the space to address these problematic implications here.
[2] Strictly speaking, Levy is only arguing here that that the debate about the necessity of consciousness for MR is a *debate specifically about AC*: "what is at issue in debates over moral responsibility is whether agents must have a certain kind of access to a certain kind of content in order to be morally responsible" (28n5).

(2) Zombies can do everything we can do.

(3) We can do whatever is sufficient for MR [e.g., we can deliberate over reasons for action; exert self-control over our inclinations, etc.].

(4) Therefore, zombies can do whatever is sufficient for MR. (2 and 3)

(5) Therefore, PC is not necessary for MR. (1 and 4)

Premise (3) assumes that "normal" adults are sometimes morally responsible: that we actually have certain properties or capacities that constitute sufficient conditions for MR for at least some of our actions. Levy remains neutral about what such sufficient conditions might be.[3]

3. Phenomenal Consciousness and Moral Responsibility

Levy's argument, however, is question-begging. All the work is imported into premise (2), that zombies can do everything that we can do. Numerous, relatively independent lines of reasoning from ethics, epistemology, and neuroscience suggest that PC is necessary for MR, and for a variety of other moral and epistemic achievements. Given space constraints, I focus on considerations inspired by Strawson. On a Strawsonian approach, MR essentially involves the capacity to feel a range of moral emotions—the reactive attitudes—and to be concerned with the emotional experiences and reactions of others. The reactive attitudes are (partly) *affect-laden* experiences. They do not consist solely in cold, cognitive moral judgments, but constitutively involve experiences with distinctive qualitative characters, e.g., "what it's like" to feel

[3] Of course, global skeptics about MR would not accept (3). Levy (2011) himself *is* such a skeptic, although for reasons related to luck rather than consciousness. Following Levy's own locution, several premises in this reconstruction suggest that the sufficient conditions for MR consist entirely in "what we can do." Does this locution suggest that Levy assumes that the relevant conditions for MR consist solely in *capacities for action*? Almost certainly not: Levy argues that AC is necessary for MR, and AC is not an action. But if Levy *is* making this assumption, he is begging the question in ways straightforwardly analogous to those I focus on in this essay. See, e.g., Reader (2012).

smoldering resentment when someone expresses ill will toward you, to suffer the sting of another's blame, or to feel the glow of another's praise.

Thus, if you are a Strawsonian of almost any stripe, then you are likely committed a) to affirming that PC is essential to MR and therefore b) to denying that zombies, despite their putative capacities for deliberation and self-control, are capable of MR. Even if zombies have perfect cognitive access to their perceptions, memories, and reasons for action, *ex hypothesi* they do not actually *feel* gratitude, resentment, etc. Nor, for that matter, do they even feel pleasure or pain. Zombies seem to respond appropriately to others' expressions of approval and disapproval, but they do not actually care about the quality of others' wills. They lack qualitatively good or ill wills for others to care about. When a zombie shouts, "I'll never forgive you in a million years!" she is not actually feeling rage or resentment. Such utterances are "full of sound and fury, signifying nothing."[4] It is unclear, therefore, that such affectless utterances could constitute genuine acts of blame. There is, it seems, much we can do that they can't. Strictly speaking, zombies cannot perform morally significant actions at all—or so a Strawsonian would argue.

That Levy is so quick to dismiss PC is striking because he elsewhere writes that, "Like most theorists of moral responsibility, I am concerned with a notion of responsibility that is constitutively linked to the appropriateness of the reactive attitudes…"[5] It is difficult to see how, say, scolding a zombie could be appropriate—or even intelligible as a genuine act of blame—given zombies' in-principle incapacity to experience any affective responses, like shame or indignation, as a result of being scolded. Since Levy countenances a conceptual link between the

[4] Shakespeare (5.5).

[5] He continues: "I am skeptical that any conception of moral responsibility that divorces it from the reactive attitudes concerns anything that is genuinely similar enough to the kind of moral responsibility at issue here to perspicuously be referred to by the same label" (forthcoming). Levy would likely have to deny that the reactive attitudes *constitutively* involve PC. For an argument to this effect about reactive attitudes and the MR of groups, see Tollefsen (2003, 231ff). See §5 (especially note #17) for further discussion. I take the thought experiments introduced in this section to support the constitutive claim.

reactive attitudes and MR, and since the reactive attitudes plausibly involve PC, how could Levy overlook the potential significance of PC to MR?

One possibility is that Levy is thinking of PC in unduly narrow terms, as merely perceptual (almost epiphenomenal) properties:

> An agent is phenomenally conscious of something (a taste, a sensation, a sound) when their mental state has… a qualitative character: the apparently ineffable qualities we feel when we perceive colors, or taste wine, or hear the soft pattering of rain… Why does the redness of a ripe tomato look like *that* and not, say, like the blue of a late afternoon sky (or, for that matter, like the ringing of a church bell)? (2014, 27).

It may, then, simply not have occurred to Levy that PC also includes a host of morally relevant affective experiences.[6]

What if we were to ask, by comparison, why does the sting of blame feel like *that* instead of like the glow of praise? Consider, in this vein, two further thought experiments. First:

> INVERSION: Vera has "inverted" moral qualia. She is functionally identical to "normal" people, and responds in typical ways to praise and blame, but she actually *feels* the glow of praise when she is blamed and the sting of blame when she is praised.

Insofar as this case is conceivable (and I will discuss reasons to question its conceivability in what follows), it brings into sharp relief the moral relevance of affective PC. While we would feel sympathy toward Vera's tragic moral-psychological plight, and should try to help uncross her emotional wires if possible, we would not view her as morally responsible in the same way or to the same extent as a person who feels a more "normal" range of emotional responses. Vera's experience of others' blame as if it were praise significantly diminishes her responsibility for acting in blameworthy ways. So the sheer fact that, at a certain level of description, she

[6] Perhaps Levy is also hostile to PC on anti-dualist grounds? He may take a deflationary or eliminativist view of PC because he takes PC to fit uneasily within scientific theories of the world. But concerns about dualism are a red herring. The question is whether PC matters for MR, regardless what further investigation reveals about the underlying nature of PC. Perhaps it is (empirically or conceptually) necessary that anyone as functionally complex as a "normal" person is capable of PC, in which case: so much the worse for the possibility of zombies, and so much the better for the necessity of PC to MR.

functions like we do does not immediately settle whether she meets any sufficient conditions for MR one might reasonably propose.

Vera's case may be difficult to genuinely conceive, however. How could someone with inverted moral qualia even *develop* into a functioning moral agent? A virtue-ethicist or sentimentalist might emphasize that Vera cannot cultivate virtue because she is incapable of taking the right sort of pleasure in doing the right thing. Even Kant, though often cited for downplaying the importance of feeling to moral agency, argued that the capacities for various moral feelings were preconditions for being susceptible to the moral law.[7] This brings us to a key distinction—which Levy may simply overlook—between the conditions on being a morally responsible agent *at all* and the conditions on being morally responsible *for some particular action or omission*.[8] Plausibly, both PC and AC are necessary for responsible agency in general. Both the capacity to access the contents of one's mind and the capacity to qualitatively experience a range of feelings and reactive attitudes are likely necessary for being capable of MR.[9] On this line, zombies and individuals like Vera would not even be candidates for MR. That said, PC and AC arguably play different roles in MR. An important project for future research is articulating the various roles that different sorts of consciousness play in constituting responsible agency. Different moral theories will likely spell this out in different ways.

However, Levy might be on firmer ground were he to argue that PC is not necessary for MR *in all particular cases*. It is not clear that, for every action, there exists some particular feeling that one must have in order to merit praise or blame for that action. I suspect it depends

[7] For helpful discussion, see Denis and Wilson (2016) and Gasdaglis (ms).
[8] See, e.g., Wallace's (1994, 84) distinction between "accountability conditions" on being an agent capable of MR at all, and "blameworthy conditions" on being blameworthy for particular actions.
[9] Thus, even theorists who deny that awareness is necessary for MR in *particular cases* (Adams 1985, Smith 2005) might accept a different necessity claim: that capacities for PC and AC are preconditions for being the sort of individual to whom MR could ever be appropriately assigned. However, I agree with these theorists (against Levy) that AC is *not* necessary for MR in all particular cases (citation removed).

on the action in question. To see how particular feelings might matter in particular cases, consider another case, involving an individual who, in comparison to zombies, more plausibly satisfies the general background conditions of responsible agency:

> BOUTS OF INZOMBIA: Zed is an otherwise "normal" person who suffers from temporary bouts of "inzombia": brief periods where everything goes "dark inside," but he continues to act fully like himself. During such bouts, Zed would vehemently deny that he was in a zombie state if you asked him. Afterwards, he retains propositional memories of what happened, but the memories are entirely "numb" and non-episodic: they lack any perceptual or affective character.
>
> ZED & FRIENDS: now imagine that you are very close to Zed and have an intense, long-awaited, moral-emotional exchange with him. Zed's eyes well with tears and his voice quivers as he says, "I can't tell you how much I admire you and how grateful I am to count you as a friend" (or "After all these years, I'm finally ready to forgive you," or "Words can never express how deeply sorry I am"). Later, however, you discover that Zed was in a zombie state during the interaction.

Once you learn that Zed was "blacked out" in this way, will you think the interaction retains the full moral significance it seemed to have? Might you feel cheated in any way? Might you perhaps want a "do-over" of the conversation to make sure that Zed *really felt* the feelings you thought he was expressing? Knowing that Zed did not experience the affective states paradigmatically associated with the relevant reactive attitudes influences our moral assessment of his behavior, and our sense of which reactive attitudes it would, in turn, be appropriate to take towards him. If Zed didn't feel deeply sorry when he apologized, does he still deserve gratitude for apologizing? Can we even say that there was an apology if he literally felt nothing when he gave it? More generally, the cases of Zed and Vera undermine the intuition that merely functionally equivalent cognitive-behavioral processes suffice for MR.

Most of the "folk" seem to agree that individuals who *systematically* lacked the capacity for these qualitative experiences could not be morally responsible.[10] When asked to imagine

[10] Shepherd (2015).

humanoid machines that acted just like human beings but lacked all conscious sensations (including pain, emotion, etc.), participants tended to agree that such individuals were conceivable, but that they would lack free will and MR for acting badly. By contrast, they believed that humanoid machines with PC would bear MR. In fact, there are rapidly expanding experimental-philosophical literatures on consciousness, responsibility, and their interconnections, which I cannot fully explore here.[11] One clear upshot from this research, however, is that the capacity for PC plays an incredibly powerful role in the folk's judgments about how individuals ought to be treated. We feel as though individuals entirely lacking PC have no "skin in the game" of morality. There is nothing "at stake" for zombies, which prevents them from being genuine participants in the moral community.

4. Sympathy for the Zombie?

Of course, if actually faced with zombies functionally identical to "normal" human beings, it would be extremely difficult to withhold our ordinary affect-laden reactions toward them. It would, in fact, be extremely difficult to believe (and perhaps even to conceive) that they were zombies. They would seem to be ordinary participants in the moral community, and they would presumably act outraged, distraught, or at least perplexed by the suggestion that they lacked inner mental lives. I suspect that we would (or at least should[12]) sooner doubt whichever scientist or authority figure told us they lacked PC than we would withhold our sympathy or resentment toward them. (This is effectively the plot of countless tales of science fiction and fantasy: a non-human entity—robot, extraterrestrial alien, animal, plant, or even an ecosystem or

[11] See Goodwin (2015) for a helpful overview. See Sytsma and Machery (2012) for studies ostensibly suggesting that participants sometimes judge that individuals with sophisticated cognitive capacities but impoverished experiential capacities deserve significant moral consideration, but I agree with Jack and Robbins (2012, 402) that Sytsma and Machery's cases do not clearly involve the total absence of PC.
[12] Cf. Antony (1996) and Tanney (2004).

planet—displays evidence of feeling, intelligence, and reactive attitudes, such that the perceptive, empathic protagonists of the story fight to defend the interests and rights of the entity, while the villains show the entity a callous disregard.) In other words, insofar as we would feel wholeheartedly compelled to praise or resent a zombie, I predict that to *just that extent* we would also find it difficult to wrap our heads around the idea that she was really a zombie. Seriously imagining ourselves in an interpersonal relationship with such an individual naturally involves imagining that we care about the quality of her will and that she cares about ours, and *thereby* contributes to the difficulty of conceiving that such functionally identical individuals could entirely lack PC.

But what if we stipulate that we could establish, in some way we all agree to be conclusive, that certain individuals really were zombies?[13] In that case, I think we would—and should—shift from the "participant" stance to the "objective" stance toward these individuals. We would—and should—cease to think of them as genuine participants in the moral community, whom we might wholeheartedly resent, and shift toward seeing them more as "objects of social policy… to be managed."[14] It might remain useful to keep praising and blaming them, to keep them in line; the traditional consequentialist defense of MR might still apply. We might also *let ourselves* become emotionally engaged with zombies, much as we become emotionally engaged with fictional characters,[15] but such engagement would be exclusively for *our sake*, not the zombies', i.e., not owed to them by the requirements of morality.

[13] Cf. Putnam (1963).
[14] Strawson (1962).
[15] For further discussion of emotional engagement with fiction, see Gendler (2013, §5.3) and the references therein.

5. What the Zombie Didn't Know

I believe that affectless, nonconscious agents, like zombies, would not only fail to be morally responsible: they would lack the moral status we recognize in far less cognitively sophisticated individuals.[16] To repurpose Bentham's (1789) famous exhortation about non-human animals, the question is not *merely*, "Can they reason? nor, Can they talk? But [*also*], Can they suffer?" But the central question here is not, as is debated about animals, whether the capacity for PC is sufficient for moral status, or whether some more exalted cognitive capacities are also necessary. The question here revolves around entities that ostensibly have higher-order, rational capacities but lack experiential capacities (including "lower-order" capacities for pain and pleasure and "higher-order" moral-emotional capacities, e.g., to react to praise by feeling valued, or, for that matter, by feeling embarrassed). Are the experiential capacities *necessary* for the rational capacities, or necessary for the rational capacities to bear the moral significance often associated with them?[17]

The conceivability of zombies might seem to entail the conceivability of possessing rational capacities without possessing experiential capacities, but this inference would be too quick. Since zombies are entirely unacquainted with affective experience, there is reason to doubt that they can properly be said to *understand* others' feelings, or their moral significance.[18]

---

[16] It might be the case that zombies have some modicum of moral status (i.e., patiency, standing, considerability, etc.). For example, if zombies are living organisms, and if, as some argue, all living organisms deserve some moral consideration, then zombies deserve some moral consideration. It might, then, be about as intrinsically wrong to behead a zombie as it would be to chop down a tree. We would, I hope, nevertheless prohibit zombie "abuse" because, as Kant says about the cruel treatment of animals, doing so would cultivate vicious traits.

[17] Two other types of entity with ostensible possession of rational but not experiential capacities are autonomous robots and group agents. See Sparrow (2007, 71-2) for insightful treatment of robots' potential for MR. Regarding group MR, I believe I can remain neutral here. For example, one could maintain that a group's MR *depends in part* on the experiential capacities of its members without arguing that the group's MR *reduces entirely* to features of its individual members. This would entail that a group comprised solely of zombies could not bear MR in the same way as regular groups, which strikes me as plausible. Alternatively, one might argue that groups can legitimately experience at least some moral emotions (Schmid 2014; cf. Sosa 2009, Tollefsen 2003).

[18] Thanks to (name removed) for discussion of this point.

(And how could a zombie be blameworthy for hurting my feelings if it doesn't understand what feelings are?) Can zombies even understand what it means to be a "reason," moral or otherwise, if they don't know "what it's like" to take something as a reason?[19] Finally, can zombies, per Levy's stipulation, actually exercise self-control over their actions? What would be the nature or force of the "inclinations" they'd have to resist? (And how could a zombie be praiseworthy for overcoming a temptation if it doesn't actually feel tempted?) While I (think I) can imagine that zombies have *functional analogs* to the affective-motivational pushes and pulls of reasons and inclinations, I can't imagine that these analogs figure in genuine exercises of *rational*—or otherwise normatively significant—capacities insofar as they're not even potentially felt.

The capacities to access, integrate, and act upon information thus seem relevant to, but insufficient for, MR. Computers, for example, far outstrip human minds in terms of the capacities to access and integrate (certain sorts of) information, but we don't typically think that computers are *simply thereby* "conscious" of this information in any morally relevant sense (e.g., such that it would be appropriate to praise them for accessing data). As far as human minds go, our rational capacities to access and integrate content are thoroughly dependent on affect. In fact, Levy elsewhere writes at length about how neuroscience suggests that "affect is indispensable to rationality."[20] Citing Antonio Damasio's research, he contests the assumption that "reason" and "emotion" are inherently opposed:

> rather than emotions crowding out reasoning, they might partially *constitute* it… even when we have time to deliberate, we cannot dispense with affect. It makes options salient for us, helping thereby to solve the problem of combinatorial explosion which faces any pure calculating machine. (2009, 76)

[19] For further discussion, see, e.g., Smithies (2012, 2013) and the references therein.
[20] Levy (2009, 76). Cf. Levy (2007, 80-1, 112, 116-20, 187-95, 293-308).

To make such claims consistent with the denial that PC is necessary for MR, Levy would have to develop an account of affect which does not merely *explain* its qualitative character(s) in terms of non-qualitative phenomena, but eliminates reference to its qualitative character altogether. If, however, we're talking about affect that is not qualitatively experienced, it's not clear that we're talking about affect at all.

References

Adams, R. M. (1985). Involuntary sins. *The Philosophical Review, 94*(1), 3-31.

Antony, L. (1996). Equal Rights for Swamp-persons. *Mind & Language*, *11*(1), 70-75.

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. Clarendon Press.

Denis, L., & Wilson, E. Kant and Hume on Morality. *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/fall2016/entries/kant-hume-morality/>.

Gasdaglis, K. L. (Manuscript). Moral Regret and the Psychological Constitution of the Kantian Agent.

Gendler, T. S. Imagination. *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/imagination/>.

Goodwin, G. P. (2015). Experimental Approaches to Moral Standing. *Philosophy Compass*, *10*(12), 914-926.

Jack, A. I., & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, *3*(3), 383-403.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

Levy, N. (Forthcoming.) Implicit Bias and Moral Responsibility: Probing the Data. *Philosophy and Phenomenological Research*.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge University Press.

Levy, N. (2009). Neuroethics: Ethics and the sciences of the mind. *Philosophy Compass*, *4*(1), 69-81.

Putnam, Hilary (1963). Brains and behavior. In Ronald J. Butler (ed.), *Analytical Philosophy: Second Series*. Blackwell.

Reader, S. (2007). The other side of agency. *Philosophy*, *82*(04), 579-604.

Schmid, H. B. (2014). The feeling of being a group: corporate emotions and collective consciousness. *Collective Emotions: Perspectives from Psychology, Philosophy, and Sociology*, 3-16.

Shakespeare, W. (2003). *MacBeth*. B.A. Mowat and P. Werstine (eds.). Simon & Schuster.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical psychology*, *28*(7), 929-946.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115*(2), 236-271.
Smithies, D. (2012). The mental lives of zombies. *Philosophical Perspectives*, *26*(1), 343-372.
Smithies, D. (2013). The significance of cognitive phenomenology. *Philosophy Compass*, *8*(8), 731-743.
Sosa, D. (2009). What is it like to be a group?. *Social Philosophy and Policy*,*26*(01), 212-226.
Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77.
Strawson, P.F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48: 1–25.
Sytsma, J., & Machery, E. (2012). The two sources of moral standing. *Review of Philosophy and Psychology*, *3*(3), 303-324.
Tanney, J. (2004). On the Conceptual, Psychological, and Moral Status of Zombies, Swamp-Beings, and Other 'Behaviourally Indistinguishable' Creatures. *Philosophy and Phenomenological Research*, 69(1), 173-186.
Tollefsen, D.P. (2003). Participant reactive attitudes and collective responsibility. *Philosophical Explorations*, *6*(3), 218-234.
Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.