Quality of will and radical value reversals

GUNNAR BJÖRNSSON
STOCKHOLM UNIVERSITY, LUND GOTHENBURG RESPONSIBILITY PROJECT

Al Mele's *Manipulated Agents: A Window to Moral Responsibility* (OUP 2019) is an extraordinarily careful and clear little book. A central recurring element is the use of examples of radical value reversals due to manipulation. In this commentary, I discuss the relevance of these examples to a simple quality of will account of blameworthiness without explicit historical conditions. Such an account, I suggest, can fairly straightforwardly explain how value reversals might mitigate blameworthiness. But I also suggest that the intuition that they *completely* remove blameworthiness should instead be explained away.

I start with a case sharing some key features of Mele's many examples of radical value reversal:

> Billy had been a good parent to his three children, aged 2, 4, and 5. Ignoring their needs for his own personal pleasure had just not been an option. But overnight, manipulators had secretly implanted a set of selfish values, erasing all of Billy's competing prior values in the process. Waking up before the children, he now considers sneaking out to go golfing, leaving them to take care of themselves at home during the day. His internal condition after the manipulation, including his collection of values, is such that he is able take his children's needs into account and act on those needs, though he has very little inclination to do so. However, this ability is not rooted in any preexisting values that survived the overnight change. Instead, it is rooted in the implanted collection of values. Billy is surprised by the fact that he is now seriously considering ignoring his children, but because of his new values he also finds it liberating and exhilarating. He decides to do what was previously unthinkable and sneaks out, leaving the children on their own accord.

Let us stipulate that Billy is not only able to tend to his children's needs, but also satisfies popular internal conditions on moral responsibility: he is able to respond to a wide enough range of reasons, is aware of the morally significant features of his options, and identifies with and takes responsibility for the motives from which he acts, for example. Nevertheless, I take it that Mele and many others will have the intuition that Billy isn't morally blameworthy for leaving his children at home. Going by this intuition, blameworthiness is subject to a historical constraint. Historical constraints can be *positive*, requiring that the agent's relevant internal condition has come about in a certain way. But that is not Mele's suggestion: as far as his intuitions and proposals go, agents coming into the world ready-made might be responsible for their actions. Instead, the constraint is *negative*, requiring that the relevant internal conditions has not come about in certain ways.

If Mele is right, this seems to spell trouble for a variety of accounts of moral responsibility that accords no role to how an agent came to be the way they are, focusing only on how they are and what they do as a result. In what follows, I will ask about the relevance of radical value reversal cases for an account which understands an agent's blameworthiness for something as

grounded in the relation between the agent's quality of will and the object of blame, but operating with no explicit historical constraints. (I focus on blameworthiness rather than responsibility generally for reasons of space.)

Here are some tempting thoughts that motivate such a quality of will account: First, if some decision, action, or outcome isn't bad in ways that matters morally, it makes no sense to blame anyone for it. Second, if an agent has displayed all the responsiveness to reasons that can be morally demanded of her in relation to what makes a certain decision, action, or outcome bad, she cannot be morally blamed for it. Third, if a morally bad thing happens and someone has fallen short of demands on responsiveness to reasons, but the bad thing didn't happen because of the shortcoming, or happened because of it but through some deviant route, the agent is not morally to blame for it. If we think of the quality of an agent's will as a matter of how well she responds to available reasons, these thoughts suggest the following minimal account of blameworthiness:

> EXPLANATORY QUALITY OF WILL (EQW): An agent is to blame for something if and only if it is morally bad and due, in a normal way, to the agent's will falling short of what morality demands of her.[1]

Degrees of blameworthiness might then be, in part, a matter of how bad a will was required to explain why the object of blame came about (Björnsson 2017b: 145–8).

Though EQW does not include a separate historical condition, it can straightforwardly account for some ways in which history matters for blameworthiness. Consider Mele's characters Van and Ike. Each gets into a car and runs over a pedestrian, too intoxicated to control the vehicle, but whereas Van willingly and recklessly got drunk, Ike was force-fed alcohol. Only Van seems blameworthy for running over the pedestrian (pp. 7–10). According to EQW, this is due to the fact that only Van ran over the pedestrian because he didn't care as could be demanded of him about the risks of heavy drinking. Though neither agent's quality of will immediately preceding the accident explains why a pedestrian was hit, Van's *prior* substandard quality of will, while getting drunk, explains why he did.

Historical relations might also be relevant in the case of actions over which the agent has control at the time. Generally, what can be reasonably demanded of someone is a function of what they are capable of, in some relevant sense. Moreover, as Mele often stresses, what *will* one is capable of in that sense at a given time might be highly constrained by one's values at that time. It could thus be that the operative demands on Billy's will are weakened by the selfish values that have been implanted in him. And if the demands operative at the time are weakened, the will at work constitutes less of a deviation from the operative demands, rendering Billy less blameworthy.

I just suggested that selfish values might mitigate demands to respond to altruistic reasons. This doesn't mean that selfish actions by those who have more selfish values are in general less

[1] I've defended and applied aspects of this account in e.g. Björnsson and Persson 2012; Björnsson 2017. I take the account to be very much in line with what Strawson (1962) took to be grounding the appropriateness of reactive attitudes. Cf. Arpaly and Schroeder 2014.

blameworthy. For what one does at a time is also standardly the normal upshot of numerous prior choices and actions, not just in drunk driver cases. Even if one does not fall much short of what can be demanded of one *at the time*, what one does might thus be the upshot of having fallen short of modest demands on one's quality of will on numerous prior occasions, and is more likely to be so if one's values are indeed selfish. And if what happened required for one's will to fall short on numerous prior occasions, it arguably required for one's will to be worse overall, thus grounding more blame than the latest shortcoming itself would suggest.

What all this *does* mean is that someone whose values and quality of will is the upshot of a radical value reversal manipulation might be much less to blame than someone who got there themselves, as it were, even if the manipulated agent is still in enough control to be subject to demands on the will. The manipulation of the agent's values breaks any normal explanatory connection between the agent's quality of will prior to the manipulation and the quality of will from which he acts after the manipulation.

This might go some way to accommodating the intuition that those subject to radical value reversals have their blameworthiness diminished, but clearly not all the way. Mele, and many with him, has the sense that an agent like Billy is *not at all* responsible for his action. What should the quality of will theorist say about that intuition?

The first thing to note is that the intuition that radical value reversal cases undermine blameworthiness is far from univocal. Personally, when focusing on the manipulation and the fact that what Billy did was not previously an option for him, I do have the sense that Billy isn't to blame for what he did. But when I focus on the fact that it remained an option for him not to leave the children at home, it also seems that he is to blame, to some extent. So I'm torn. And judging by a survey Mele reports in an appendix, others are too. Though considerably more subjects were inclined to disagree than agree with the claim that an agent like Billy is morally responsible for his actions, the disagreement was not resounding: the mean response on a scale from 1 ("strongly disagree") to 7 ("strongly agree") was a little over 3.

Why are intuitions not clearer than this? Elsewhere I have argued that, in line with EQW, to see someone as blameworthy for something is to see the object of blame as due to, or explained by, the agent's substandard quality of will. Seeing X as explained by Y in turn involves (i) taking a certain explanatory perspective, employing an explanatory model which represents lawlike relations between variables of interest, and (ii) seeing how X follows from Y in accordance with this model given certain background conditions. One is thus in effect treating Y as the value of an independent variable in the model, and X as the value of a dependent variable. Importantly, what explanatory perspective we take can shift depending on explanatory interests, the salience of various candidate explanantia, and on how straightforwardly these explanantia account for the explanandum. If a house is hit by lightning and burns down, we might naturally think that the house burned down because it was hit by lightning. In doing so, we treat the presence of oxygen, the combustibility of the house, and the absence of a lightning rod as background, ignore whatever atmospheric conditions gave rise to the lightning, and understand the events between the strike of lightning and the final outcome as values of intermediary dependent variables: compared to the background, the lightning is more out of the ordinary; whatever caused the lightning is less salient and straightforwardly connected to

the outcome; and the intermediaries—that a fire started in the attic, say—are comparatively less interesting as they are straightforwardly accounted for by the fact that the lightning hit. A fire engineer might take a different perspective. She might treat the lightning as background and see the outcome as explained by the absence of a lightning rod. Or she might find the fact that the fire started in the attic particularly interesting, as houses might be especially vulnerable to fires starting there: why it started there is now less interesting.

Based on this, I've suggested that considerations of determinism, various forms of luck, and manipulation cases all undermine our sense that the agents involve are responsible for their actions because they provide powerful prompts to change explanatory perspective to one from which the agent's quality of will no longer seems to be what explains the object of responsibility (Björnsson and Persson 2012; 2013; Björnsson 2017b: 155–62). This story applies straightforwardly to cases involving radical value reversals. Such reversals are themselves remarkable events, and they straightforwardly account for the agent's quality of will and the resulting action as well as the remarkable fact that what the agent did had previously not even been an option for her. Given this, we will be prompted to see the reversal as what explains the agent's actions. And when we do, the agent's quality of will at the time of action is only seen as an intermediary explanatory step, rather than as what explains the action: the agent no longer seems blameworthy for it. But, as compatibilists often do, one can resist these prompts by focusing attention on the agent's quality of will and treating it as an independent variable, returning to the sort of normal everyday explanatory perspective from which we standardly assess whether some decision, action, or outcome was due to the agent's quality of will. It is the possibility of these different perspectives that explains why intuitions concerning the responsibility undermining effects of determinism, luck, and forms of manipulation varies and remain ambivalent in many of us.

If this is correct, the question is what the right perspective is for assessing an agent's responsibility and blameworthiness. Here, I have argued, various lines of reasoning suggest that it is the sort of everyday perspective that treats the agent's quality of will as an independent variable (Björnsson and Persson 2012). If this is correct, the sense that agents are not at all responsible for what they do after radical value reversals is illusory.

Bibliography

Arpaly, Nomy and Schroeder, Timothy 2014: *In Praise of Desire.* Oxford University Press.

Björnsson, Gunnar 2017a: 'Explaining (Away) the Epistemic Condition on Moral Responsibility'. In *Responsibility: The Epistemic Condition*. Robichaud, Philip and Wieland, Jan Willem (eds) New York: Oxford University Press pp. 146–62.

Björnsson, Gunnar 2017b: 'Explaining Away Epistemic Skepticism About Culpability'. In *Oxford Studies in Agency and Responsibility*. Shoemaker, David (ed) Oxford University Press pp. 141–64.

Björnsson, Gunnar and Persson, Karl 2012: 'The Explanatory Component of Moral Responsibility'. *Noûs*, 46, pp. 326–54.

Björnsson, Gunnar and Persson, Karl 2013: 'A Unified Empirical Account of Responsibility Judgments'. *Philosophy and Phenomenological Research*, 87, pp. 611–39.

Mele, Alfred R. 2019: *Manipulated Agents: A Window to Moral Responsibility*. Oup Usa.

Strawson, Peter F. 1962: 'Freedom and Resentment'. *Proceedings of the British Academy*, 48, pp. 187–211.