Attributionism and Counterfactual Robustness

Rutger van Oeveren & Jan Willem Wieland

In this journal, Vishnu Sridharan presents a novel objection to attributionism, the view according to which agents are responsible for their conduct when it reflects who they are or what they value. The key to Sridharan's objection is that agents can fulfil all attributionist conditions for responsibility while being under the control of a manipulator. In this paper, we show that Sridharan's objection falls prey to a dilemma—either his manipulator is counterfactually robust, or she is not—and that neither of its horns undermine attributionism.

AJP, 2016

1. Introduction

There's a heated debate between volitionists and attributionists on the nature of moral responsibility.[1] Basically, volitionists say that agent S is responsible (blameworthy or praiseworthy) for an action or omission X *only if* X is the result of a conscious choice. Attributionists, in contrast, say that S is responsible for X *if* X reflects who S is or what she values, whether or not X is the result of a conscious choice. Sridharan [2016] presents a challenging objection to attributionism. In the following, we'll challenge this objection. If our reply succeeds, then attributionism—an umbrella term for many current accounts of moral responsibility—isn't defeated by it. As always, this doesn't mean that attributionism is true. But it does mean that if attributionism incorporates our strategy, the debate between attributionism and volitionism can't be decided on the basis of Sridharan's objection.

2. Sridharan's objection

Basically, Sridharan's objection is that when a manipulator arranges the world in such a way that the attributionist conditions are satisfied, the given agents are not responsible even though they meet all the alleged conditions. Sridharan applies this strategy to three main attributionist accounts, namely the accounts by Sher [2009], Smith [2005], and Arpaly and Schroeder [2014].[2] Sher's

[1] The labels are due to Levy [2005]. For a recent overview of the debate, see Wieland [2016].

[2] As Sher [2008] points out, his account differs from the other accounts. On Sher's view, what matters is whether one's constitutive attitudes stand in a suitable causal relation to one's conduct, and not whether one's attitudes are also judgment-sensitive.

account receives most attention, and we'll explain Sridharan's objection in terms of it. Consider Sher's well-known Hot Dog case:

> (1) Alessandra, a soccer mom, has gone to pick up her children at their elementary school. As usual, Alessandra is accompanied by the family's border collie, Bathsheba, who rides in the back of the van. Although it is very hot, the pick-up has never taken long, so Alessandra leaves Sheba in the van while she goes to gather her children. This time, however, Alessandra is greeted by a tangled tale of misbehavior, ill-considered punishment, and administrative bungling which requires several hours of indignant sorting out. During that time, Sheba languishes, forgotten, in the locked car. When Alessandra and her children finally make it to the parking lot, they find Sheba unconscious from heat prostration. [Sher 2009: 24]

On Sher's account, S is blameworthy for unwitting X if relevant non-epistemic conditions are met, and S's failure to realize that X is wrong falls below a relevant standard, and is caused by dispositions constitutive of S [ibid.: 88]. Roughly, S's failure falls below a relevant standard if S has the cognitive capacities to meet that standard, and does meet it in counterfactual situations with slightly different conditions [ibid.: 109]. In terms of (1), for example, Alessandra is blameworthy for leaving her dog Sheba in the car partly because her memory is not impaired, and she does remember that she shouldn't leave her dog in the car in counterfactual situations where she is not detained. Alessandra is blameworthy, then, even though she didn't choose to forget about Sheba.

Sridharan presents the following counterexample to Sher's account:

> (2) Alessandra is the same person with the same character as in (1). In this case, however, her children do not misbehave, and she is on track to preserve the fragile health of her dog. However, her ex-wife Parvati, who hates Sheba, has been plotting her revenge for years. Although Parvati is not a skilled neuroscientist, she possesses a brilliant understanding of how Alessandra's brain functions. She uses this specific knowledge to develop a smell that has the power to cause Alessandra to forget Sheba in the car. The smell is so personalized that, if Alessandra's character were different, it would have no effect. The success of Parvati's scheme is far from guaranteed, however, as the day is quite windy. As Alessandra is on her way to pick up her kids, Parvati releases her smell and, as fate would have it, successfully causes Alessandra to forget her canine companion. With her mind cleared of Sheba, Alessandra takes her kids to get ice cream. When they finally return to the parking lot, they find Sheba unconscious from heat prostration. [Sridharan 2016: 467]

Hence the counterexample: Alessandra appears blameless for leaving Sheba in the car even though she satisfies all Sher's conditions. In particular, her failure to be aware of Sheba falls below a relevant standard, since she would have this awareness in counterfactual situations where the wind blew differently, and her failure is the result of dispositions constitutive of her, since Parvati's smell is adapted to them, and wouldn't work for different agents.[3]

The same kind of counterexample, Sridharan proposes, can be generated to other attributionist accounts.[4] In the following, we will defend attributionism from Sridharan's objection by showing that his counterexample is ambiguous, and falls prey to a dilemma. Neither of its horns, we will argue, yields an objection to attributionism.

Basically, our response is a version of the origination response anticipated by Sridharan: "In (1), Sher might argue, Alessandra's actions originate in herself, whereas in [(2)] Alessandra's actions originate in Parvati." [2016: 468] Sridharan ultimately rejects this response because in (2) Alessandra's make-up does play a role in why she forgets Sheba (and that is why Alessandra's actions can be said to originate in herself). To a certain extent, we agree. Yet, as we'll argue, it just fails to play a relevant role.

3. Counterfactual robustness

Let us say that a manipulator M is *counterfactually robust* when M arranges things differently in worlds where S's make-up is different. Henceforth, we will use 'make-up' as shorthand for any attributionist term (such as Sher's "constitutive dispositions").[5] More precisely, M is counterfactually robust with respect to an agent S and a wrong act X when M intervenes in such a way that, if relevant circumstances obtain, S fails to realize that X is wrong, even in worlds where S's make-up differs.[6] For example, Parvati is counterfactually robust with respect to Alessandra and the wrong act of leaving Sheba in the car if Parvati adjusts her manipulative operations to the specific sensitivity of Alessandra's (that is, over worlds where the latter's make-up varies) in such a way that, if the wind is right, Alessandra will forget about Sheba.

[3] The cases are not fully parallel. First, the cause of Alessandra's conduct is different (school issues and the smell, respectively), and, second, her conduct itself is different (sorting out issues and getting ice cream). Since Sher's conditions are fulfilled in both cases, however, we believe this presents no problem for Sridharan's counterexample.

[4] Whereas familiar manipulation arguments challenge compatibilism generally (see Pereboom [1995]), Sridharan's argument, on which we are focusing here, targets attributionism.

[5] Or "judgment-sensitive attitudes" from Scanlon [1998] and Smith [2005], or "intrinsic desires" from Arpaly and Schroeder [2014].

[6] We'll focus on blameworthiness for wrong acts and discuss praiseworthiness for right acts in section 4.

In the following, we will spell out Sridharan's counterexample (2) in two quite different ways: one where the manipulator is counterfactually robust, and one where she isn't. We will show that this makes all the difference. Here is a variant of (2) where Parvati is counterfactually robust:

> (3) As in (2), Parvati developed a smell that is so personalized that, if Alessandra's character were different, it would have no effect. Had Alessandra's character been different, Parvati would have developed a different smell to achieve the same effect. As Alessandra is on her way to pick up her kids, Parvati releases her smell, and causes Alessandra to forget her canine companion. With her mind cleared of Sheba, Alessandra takes her kids to get ice cream. When they finally return to the parking lot, they find Sheba unconscious from heat prostration.

In this case, we'd agree with Sridharan that there's a strong intuition that Alessandra isn't responsible. However, we also think that attributionism can provide a straightforward and non-volitionist answer in such cases. Basically, we don't think that in (3) Alessandra's conduct reflects who she is or what she values.

To see this, a distinction needs to be made between *causal* and *explanatory* factors of her conduct. In (3), we concede that the specific features of Alessandra's make-up play a causal role in her forgetting: the smell affects the specifics of her make-up. If Alessandra's features were otherwise, however, the outcome would still be the same. For in that case the robust manipulator would devise a different smell, and Alessandra would still forget about Sheba. Therefore, the specific features of Alessandra's make-up play no explanatory role, and that's why she's not responsible (on attributionist terms).[7]

To account for cases like these, we propose, attributionists should adopt an extra clause in their accounts. Namely:

> S is responsible for X iff the given attributionist condition A obtains, *and any cause of X, which is non-redundant and external to S, is not adjusted to variations in A*.[8]

If Alessandra doesn't satisfy this additional clause, as is the case in (3), her conduct is not explained by *her* make-up. According to this clause, then, in (3), Alessandra is not responsible for what happens to Sheba. In this way,

[7] So long as the manipulator is robust, we think it doesn't really matter whether the manipulator is a person who acts on the basis of a nasty plan, or whether it's a mere cosmic coincidence that the interventions in all different worlds fit Alessandra's make-up (for, again, in such a case Alessandra's conduct is not explained by her make-up).

[8] We add 'external to S' because, trivially, S herself is a cause of X, and is adjusted to variations in A. We add 'non-redundant' because S can still be responsible for X if her conduct were entirely the same without the external cause.

attributionism can account for counterfactually robust manipulators.[9] Non-robust manipulators may appear more problematic, but they also pose no problem for attributionism, as we will discuss next. Let us say that a manipulator M is *not* counterfactually robust when M does not intervene effectively in other worlds, or does not intervene at all. Consider the following example:

> (4) The story is almost the same as in Sridharan's case (2) except that, this time, Parvati's knowledge of Alessandra's brain is not so brilliant. As a result, Parvati does not really know whether the smell she developed will work.[10] Given that she wants to take revenge, she releases it, and the story ends as in (2): the wind carries the smell to Alessandra, the smell proves to be effective, and causes Alessandra to forget Sheba. In all nearby possible worlds, however, the story ends differently. Even where the wind is blowing in the right direction, and so does carry the smell to Alessandra, it doesn't match Alessandra's make-up, and she does not forget her dog.

If Sridharan is right, we should think that Alessandra is not responsible here. But this is far from clear. In this case, Parvati's interventions function in exactly the same way as other circumstantial factors (such as the wind, or events in the school): they obtain in the actual world, but fail to obtain in many nearby possible worlds. Moreover, such circumstantial factors don't pose specific problems for Sher's account.[11] There are many circumstantial factors that sustain Sher's original case (1), but they do not render Alessandra any less responsible. (4) is analogous to (1) in that she is not detained in many nearby possible worlds (and does remember Sheba), and if you share Sher's intuition that Alessandra is responsible in (1), then your intuition that she is responsible in (4) shouldn't be any different. Crucially, in (4), and unlike in (3), Alessandra's specific features *do* play an explanatory role. If Alessandra had had a different make-up than she does—if, for example, she were more thoughtful or caring—the smell wouldn't have distracted her. In (4), then, who Alessandra is and what she values is relevant in explaining why she forgot Sheba.[12]

[9] Arguably, robustness comes in degrees. To account for this, one might say that the *more* nearby worlds in which causes of X are adjusted to variations in A, the *less* one is responsible for X, and vice versa.

[10] Does Parvati still qualify as a manipulator in this case? If one doesn't recognize this case as an interesting variant of Sridharan's objection, we can simply set it aside.

[11] Of course, circumstantial factors pose a general sceptical worry, but they afflict *all* non-sceptical accounts of responsibility, not just Sher's.

[12] If you disagree with this argument and think that Alessandra is not responsible when Parvati is non-robust, you may find our solution to the first horn of the dilemma promising as well. For you may think that, in this case too, and contrary to our claim, Alessandra's specific features don't explain her forgetting.

Sridharan might respond that in (4) the intervention is *designed* for Alessandra, whereas in (1) there's no agential intervention. In (4) Alessandra seems to be 'a puppet of another person', and this may mitigate her blameworthiness. But though Parvati might *think* Alessandra is her puppet, it is important to see that her intervention works only in the actual world, and not in any nearby world where all circumstantial factors are held fixed (except for Alessandra's make-up). If this is so, it seems clear that Alessandra is no more Parvati's puppet than she is the puppet of the problems at the school (for example).

4. Generalization

Given that our account is formulated in general attributionist terms, our response to Sridharan's objection can be generalized. Next, we'll briefly show this for Arpaly and Schroeder's attributionist take on praiseworthiness (and leave Smith's case and further attributionist accounts, for the reader to consider).[13] According to Arpaly and Schroeder, S is praiseworthy for X, roughly, if S does X with an intrinsic desire for the right or the good. Here is Sridharan's counterexample (slightly edited from [2016: 473]):

> A college student named Roark is enthralled with the philosophy of Ayn Rand that, to his understanding, obligates people to be selfish and to leave others to fend for themselves. However, his intrinsic desires favour the less privileged. One day, on the campus quad, Roark comes across an unknown person in distress. In line with his beliefs, he thinks that the person should help herself, and his conscious reasoning is able to overpower his intrinsic desires to help. However, as he is about to turn away, his ethics instructor Professor Gutierrez—who knows Roark well enough to push his buttons—releases a high-pitched acute noise. This blast of noise is designed to bypass Roark's beliefs and to stimulate his intrinsic desires. After hearing the noise, Roark is overwhelmed by his intrinsic desires, and he helps the stranger.

According to Sridharan, this case poses a counterexample to Arpaly and Schroeder's account, because intuitively Roark is not praiseworthy even though he does help the distressed person with an intrinsic desire for the right. Again, we think the case is ambiguous, and that the objection can be countered by considering the robustness of Gutierrez's interventions.

The dilemma is as before: either the manipulator is robust, or she is not. In the former case, Gutierrez adjusts the noise in the nearby possible

[13] Regarding praiseworthiness, we may assume: M is counterfactually robust with respect to S and a right act X when M intervenes in such a way that, if relevant circumstances obtain, *S realizes that X is right*, no matter S's specific make-up.

worlds where Roark has weaker intrinsic desires, and Roark still performs the right action. So whatever the strength of his intrinsic desires, Roark would have helped the stranger anyway. In this case, we wouldn't say Roark is praiseworthy for helping the stranger. After all, his good action isn't explained by the strength of his intrinsic desires. Our clause rightly excludes a praiseworthiness verdict in this case.

In contrast, if the manipulator is not robust, and does not push buttons (or not effectively) in nearby possible worlds, then it is not the case that Roark will perform the good action regardless of the strength of his intrinsic desires. This time, the fact that Roark has strong rather than weak intrinsic desires explains why he helps the stranger, and the manipulator is just one among many circumstantial factors (factors that obtain in the actual world, but not in many nearby possible worlds). Roark is to be held praiseworthy in this scenario, and the account by Arpaly and Schroeder, extended with our clause, yields this verdict.[14]

5. Conclusion

According to attributionism, agents are responsible for their conduct when it reflects who they are or what they value. As we see it, what Sridharan's challenge shows is the possibility of cases where this reflection is merely superficial. In such cases, the fact that S is *this* person with *this* set of values does not explain why her action came about. This is excluded in our account: S is responsible for X only when X is explained by S's make-up, which requires, as we have suggested, that no cause of X is adjusted to S's make-up.[15]

References

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry Into Moral Agency*, New York: Oxford University Press.

Arpaly, N. and T.A. Schroeder 2014. *In Praise of Desire*, New York: Oxford University Press.

Levy, N. 2005. The Good, the Bad and the Blameworthy, *Journal of Ethics and Social Philosophy* 1/2: 1–16.

Pereboom, D. 1995. Determinism Al Dente, *Noûs* 29/1: 21–45.

Scanlon, T.M. 1998. *What We Owe to Each Other*, Cambridge, Massachusetts: Belknap Press of Harvard University Press.

[14] The case description is silent about Roark's conduct in nearby worlds where the manipulator isn't active, and we would need more information about this in order to make a more definite statement about Roark's praiseworthiness (see Arpaly [2003: 85–6]).

[15] Thanks to Phil Robichaud, Matthé Scholten, and two referees for comments, and to the Netherlands Organization for Scientific Research for their support through a Veni grant.

Smith, A.M. 2005. Responsibility for Attitudes: Activity and Passivity in Mental Life, *Ethics* 115/2: 236–71.

Sher, G. 2008. Who's In Charge Here?: Reply to Neil Levy, *Philosophia* 36/2: 223–6.

Sher, G. 2009. *Who Knew? Responsibility Without Awareness*, New York: Oxford University Press.

Sridharan, V. 2016. When Manipulation Gets Personal, *Australasian Journal of Philosophy* 94/3: 464–78.

Wieland, J.W. 2016. The Epistemic Condition, in *Responsibility: The Epistemic Condition*, eds P. Robichaud and J.W. Wieland, Oxford: Oxford University Press.