Philos Stud
https://doi.org/10.1007/s11098-021-01744-8

Situationism, capacities and culpability

Adam Piovarchy[1]

Accepted: 27 September 2021

Abstract The situationist experiments demonstrate that most people's behaviour is influenced by environmental factors much more than we expect, and that ordinary people can be led to behave very immorally. A number of philosophers have investigated whether these experiments demonstrate that subjects' responsibility-relevant capacities are impeded. This paper considers how, in practice, we can assess when agents have a reduced capacity to avoid wrongdoing. It critiques some previously offered strategies including appeals to the reasonable person standard, appeals to counterfactuals and understandability of behaviour, and appeals to base rates of wrongdoing. It then proposes we should think a certain factor impeded capacities when this is the best explanation of a change in patterns of responses. With this approach in hand, I then argue that subjects in many of the situationist experiments are (mostly) excused for their actions.

Keywords Blame · Situationism · Moral responsibility · Capacities · Milgram

1 Introduction

The situationist experiments are a suite of studies demonstrating that most people's behaviour is influenced by seemingly irrelevant environmental factors much more than expected. These studies first received philosophical attention in arguments regarding whether we have empirical evidence for the kinds of global character traits proposed by many theories of virtue (Doris, 2002; Harman, 2000; Kamtekar, 2004; Sabini & Silver, 2005). A more recent area of interest has concerned whether

✉ Adam Piovarchy
adam.piovarchy@nd.edu.au

[1] Institute for Ethics and Society, The University of Notre Dame, Australia, Sydney, NSW 2007, Australia

Published online: 22 October 2021

the subjects who act wrongly in these experiments are morally responsible for doing so (McKenna & Warmke, 2017; Nelkin, 2005; Talbert, 2009). Given the subjects seem to be normal agents, and given their actions are obviously wrong, some philosophers have wondered whether these subjects are in some way prevented from doing the right thing, which could mitigate their blameworthiness.

Answering this requires that we grapple with a more basic question: in practice, how can we tell when wrongdoers possessed the capacities required for moral responsibility? A number of philosophers have examined the capacities of these subjects and come to different conclusions regarding their culpability. Though I ultimately side with those who argue that subjects' blameworthiness is significantly mitigated, I believe the ways in which parties to this debate are making inferences about agential capacities require closer attention.

This paper considers how we can tell when an agent's blameworthiness for wrongdoing is excused or mitigated because they had a reduced capacity to avoid wrongdoing. It considers some existing proposals including appeals to the reasonable person standard, appeals to relevant counterfactuals, appeals to base rates of wrongdoing, and appeals to the understandability of responses, identifying problems with each. I argue that we should think subjects lacked a certain capacity when this is the best explanation of changes in their performance across a range of settings. Relying on evidence from a number of other studies, I then argue that many subjects in the situationist experiments had a decreased capacity to avoid wrongdoing.

2 The situationist experiments

Allow me to remind readers of some notable situationist experiments. Given they are well-known and have been covered many times in the literature, I shall be brief:

- *Dime*: Subjects who had just exited a phonebooth encountered a confederate who dropped some papers. In the control condition, only 4% of subjects helped the confederate. But when subjects in the experimental condition first found a dime in the phonebooth, 87.5% provided assistance (Isen & Levin, 1972).
- *Bystander*: A number of experiments demonstrate that an agent's likelihood of helping a stranger is reduced significantly by the presence of other people. For example, when by themselves, 70% of subjects will check on someone in another room who is heard to cry out in pain after a large crashing sound. But if unresponsive confederates are present, only 7% will help (Latané & Rodin, 1969).
- *Fire* (a variant of *Bystander*): 75% of subjects in a room by themselves left within four minutes of smoke beginning to billow in. However, when in the presence of two confederates who ignored the smoke, 90% of subjects remained in the room. Had the smoke been coming from a real fire, subjects would have remained long enough to die of asphyxiation (Latané & Darley, 1970).
- *Samaritan*: Seminarians were asked to give a speech on the parable of The Good Samaritan. Upon arriving, they were told their talk had been moved elsewhere.

On their way to the second location, seminarians passed by someone who appeared to be unconscious in the middle of the day. The chances of any seminarian stopping to help fell dramatically if they were told they were running slightly late (10%), compared to if they were told they were early (63%) (Darley & Batson, 1973).

- *Obedience*: Milgram (1963) had subjects administer shocks to a stranger as part of a purported memory task. The subjects or 'teachers' would administer increasing shocks to a 'learner' (actually a confederate who wasn't receiving shocks) every time the learner got an answer wrong. Even when the learner demanded to be let out, acted as if they'd had a heart attack, and ceased responding altogether, 65% of subjects continued to administer the dangerous shocks when directed.

These experiments are notable not only because many subjects act much worse than expected, but because they seem to be caused to act wrongly by factors that seem normatively irrelevant (or at least, only weakly relevant). The number of people nearby, a dime, or someone in a lab coat, do not give us strong moral or prudential reasons to act any differently, and yet they produce dramatic changes in people's behaviour. As Nelkin (2005) puts it: "the subjects seem to be acting for bad reasons, or at least not acting for good reasons, and they seem stuck doing so" (p. 199).

This unusual 'stuck' behaviour of the subjects leads some to think that perhaps the subjects who act wrongly might not be morally responsible for their actions, or at least, are less responsible than they would be for other choices they make in their lives. To understand why, we need to examine what it means to be morally responsible for something, and what conditions agents must meet to qualify as responsible.

3 Capacities and moral responsibility

What does it mean to be morally responsible for an action? A common answer, which I will adopt, is that to be morally responsible for one's actions is to be connected to one's actions in such a way that makes one an appropriate target of blame (praise) for wrong (right) actions. Blame is in turn understood as a negative reactive attitude that expresses disapproval of the target's conduct. Being blameworthy for wrongdoing is commonly taken to require that one possessed a certain kind of control over one's actions. This in turn requires the possession of certain capacities and opportunities, typically understood as the capacity to perceive moral reasons, the capacity to then act on those reasons, and the opportunity to exercise those capacities (Brink & Nelkin, 2013; Fischer & Ravizza, 1998; Franklin, 2013; Vargas, 2013). If I fail to save a child from drowning in a lake because I am not capable of swimming, or have been tied up by a rope, then I am not morally responsible for failing to save the child precisely because I lacked the capacities or opportunities necessary to do so. Capacities and opportunities can come in degrees,

and agents can have a reduced capacity or opportunity even if they do not lack it altogether, which will result in reduced blameworthiness in the event of wrongdoing (Nelkin, 2016).

Some care must be taken identifying capacities for the purposes of assessing blameworthiness. For example, some philosophers say that an agent who is tied up lacks the opportunity to exercise their swimming capacities, while others will say that the agent has the general capacity to swim but lacks the specific capacity to swim. Jaster (forthcoming) points out that the distinction between general capacities, specific capacities, and opportunities is really just a matter of what factors we hold fixed. Some capacities depend on factors external to the agent (the capacity to baptise depends on being ordained), some opportunities are lacking in virtue of features internal to the agent (being bed-ridden prevents one from exercising their capacity to play soccer), and the distinction between general and specific is more of a gradient that depends on the number of factors we want to hold fixed. Given it is natural to talk about the capacity to *do* the right thing for the right reasons, for our purposes it will be fine to, at times, individuate capacities in a way that includes opportunities. An agent who has a duty to swim to save someone, and who has the general capacity to swim but is tied up, lacks the capacity to do the right thing in this situation.

A further complication is that when it comes to assessing blameworthiness, establishing that an agent has a specific capacity in one sense might not establish that they have the specific capacity in the sense required for moral responsibility. Care must be taken when individuating what 'the right thing' is. Consider Austin's (1956) putt that he could have made, but didn't. Austin lines up the putt, tries his hardest to sink the ball, and misses. All the same, we would say that he had the specific capacity to make that putt. But were something moral at stake such that failing to sink the putt would lead to some terrible outcome, we would not move from the observation that Austin had the specific capacity to make the putt to the conclusion that he was therefore morally blameworthy for missing. Intuitively, he would not be a fitting target of moral blame.

A tempting way to account for our intuitions here is to say that, in this situation, 'the right thing' to do is simply to *try* and sink the putt, which Austin did. However, this move comes with a risk because trying itself often looks a lot like a capacity. And arguing that one is culpable for not Φ-ing only if one had the capacity to try to Φ risks a regress, as agents would then first need to have the capacity to try to try to Φ, and so on (Jaster, 2020; Small, 2017).[1] Instead, Jaster (2020) points out that we need to distinguish between agentive capacities, which are capacities to perform actions, and non-agentive capacities, which are triggered by certain stimuli but are still bona fide capacities. Wanting, trying, intending, and recognising, among others, are all non-agentive capacities, as is the capacity to try to sink a putt as best one

[1] See Jaster (2020) for objections to attempts to understand capacities in terms of conditional analysis or appeals to nearby possible worlds. This paper's question and argument is still relevant to proponents of such views, as in practice we can't directly look into nearby possible worlds, nor do we often know what conditionals are true of an agent.

can.[2] Although we are interested in whether Austin has the capacity to try to sink the putt, in other cases we may be interested in whether the agent had e.g. the capacity to recognise that a certain moral fact obtains, not the capacity to try to recognise that a fact obtains.[3]

To answer whether the wrongdoing subjects in the situationist experiments had the capacities necessary for moral responsibility, we need an investigation into how we know when agents lack certain capacities and the structure of our reasoning when making such assessments. Let's now consider four existing proposals for assessing when an agent lacks a particular capacity.

4 How do we assess capacities?

4.1 The reasonable person standard

Michelle (2013) and Doris (2007) argue that the number of people who act a certain way in a particular situation can help us determine what can be expected of 'a reasonable person' in that situation. In everyday life and in our legal system, we often make culpability judgements according to such a standard, rather than assessing each person's capacities on an individual basis. We can't access other people's minds to assess their quality of will or reasons-responsiveness, so we require some other standard by which to judge their actions. Ciurria argues that how most people behave in a situation can be an indicator of an individual's capacities, as most people are reasonable. If the majority of people fail to act a certain way in a particular setting, this is evidence that we could not reasonably expect most people to have done otherwise in that setting, and this may be grounds for excusing the obedient subjects since a majority of all subjects obeyed. Doris points out that, although we can't determine excuses entirely by the rates at which people act, "still, reflection on base rates helps determine what can fairly be expected of a particular individual in particular circumstances; surely it is partly because most people yield under torture that it seems unfair to hold victims responsible for failing to resist it" (2007, p. 527).

The reasonable person standard does not assess responsibility on the basis of base rates alone. It allows us to rely on 'other evidence' of an individual's capacities. While one could object that without further explication of what this other evidence is, the standard risks being unsystematic and unhelpful for adjudicating between disagreeing assessments, I want to note a greater problem with it. This is a problem which is easy to miss when reasoning about the capacities of individuals generally, and which we need to be careful to avoid in our current investigation.

Many philosophers make a point of noting that in no situationist experiment do 100% of subjects act wrongly, including Adams (2006, p. 149), Badhwar (2009,

[2] See Hieronymi (2007) and McGeer and Pettit (2015) on how explanations for why agents act in one way rather than another run out at some point, and how this can pose problems for some conceptions of capacity.

[3] I will consider worries about unwitting wrongdoing and volitionist arguments below.

p. 261), Brink (2013, p. 129), Solomon (2003, p. 56; 2005, p. 653), and Wielenberg (2006, p. 448). The purpose of drawing our attention to this fact often seems to be to suggest that since some subjects did the right thing, this is evidence that subjects who acted wrongly had the capacity to do the right thing too. This move is curious though, given that, upon reflection, even if 100% of subjects did the wrong thing in some experiments, this wouldn't prove their capacities had been affected in some way. Nelkin (2005) points out that if 100% of customers who enter a bank wait in line, this does not show that they were unable to do otherwise or somehow forced to act as they did by features of the banking environment. This may seem to favour thinking that the subjects in fact had the relevant capacities. But this reasoning cuts both ways: if the base rate of wrongdoing is immaterial to showing that wrongdoing subjects lacked certain capacities, it also does not show that they possessed certain capacities either.

Both the reasonable person standard and aforementioned appeals to right-doing subjects involve making inferences about a particular agent's capacities by referring to the observed behaviour of other agents. The problem is that this move relies on first having identified a relevant reference class for said comparison, and the way this is chosen may fail to notice relevant differences between agents. For example, if we applied a reasonable person standard to assessing the capacities of 'all 6th graders' on a reading task, we would observe that there are no constraints on behaviour and that most 6th graders could pass the task. We would thus be at risk of concluding the children with dyslexia are merely being inattentive. It is only once we change our reference class from 'all 6th graders' to 'all 6th graders with dyslexia' that we can then apply the reasonable person standard to get the correct assessments. Situationist settings may be the kind of thing that have a negative—albeit local and temporary—effect on some people's capacities, but which we just haven't yet identified.

To show how this worry applies to our current investigation, consider one of the variations of *Obedience* in which only 20% of subjects obeyed. Although most subjects were raised in the same country and within a similar time period, there might still be tremendous variation *between* subjects regarding their upbringing and previous experiences which could affect their present capacities. It is true that not every subject obeyed and so, at first glance, when we apply the reasonable person standard, a-reasonable-person-placed-in-this-setting would not administer shocks at 450 V. But it could also turn out that a-reasonable-person-placed-in-this-setting-who-was-not-taught-how-to-question-authority would in fact obey. For example, Adrian Dimow, one of the subjects who disobeyed very early in the most famous variant of *Obedience*, reports that he disobeyed because he didn't trust the experimenters, rather than out of any concern for the learner. He had been raised to distrust authority in a "socialist-oriented family steeped in a class struggle view of society [which] taught me that authorities would often have a different view of right and wrong than mine" (Dimow, 2004; Perry, 2013a, p. 127). He had also been harassed by the FBI while a member of the Communist Party. These experiences were clearly not had by the majority of the other subjects, and acted as a relevant difference which led to him finding it much easier to resist the influence of the authority figure. His interpretation of the setting, his perception of the experimenter,

and what considerations he took to be present were all very different to those of other subjects. That there was likely to be variation in subjects' interpretation of their setting, the pressure they felt, their perception of the experimenter, and understanding of what was at stake has been noted several times by philosophers arguing that the studies do not rule out the existence of global character traits (e.g. Kamtekar, 2004; Sreenivasan, 2002). The relevance for us is that such factors also plausibly affect the extent to which each subject had the capacity to disobey.

4.2 Counterfactuals

A second means of assessing capacities comes from Brink (2013), and is later endorsed by McKenna and Warmke (2017). Brink argues that we should assess an agent's capacities by examining how the agent would act in relevantly similar counterfactual situations. Regarding *Obedience*, he argues that the obedient subjects are responsible agents, because there are relevantly similar counterfactual scenarios in which subjects would successfully disobey, e.g., "if they had been given more time to consider their options, if they had been asked to justify the imposition of apparent harm to innocent parties, or if they had known the learners" (p. 141). Whereas Ciurria and Doris assessed an agent's capacities by looking at how relevantly similar agents would act in a particular setting, Brink looks at the particular agent across relevantly similar settings. This enables Brink to avoid the reference class problem, since we do not have to categorise an agent into a particular class before making our assessment.

The problem for this approach regards *which* counterfactuals are relevant, and how we interpret the outcomes of those counterfactuals. Grant Brink's supposition that an obedient subject would have disobeyed had they been given more time to consider their options. This could be taken to show the subject had the capacity to disobey in the actual sequence all along. But it could also be taken to show that the actual sequence of events was relevantly different to a setting where that subject had more time. The counterfactual could instead show that something we assumed had no effect on an agent's capacities, namely time pressure, was in fact significant.[4] It could instead show that subjects' capacity to disobey is reduced *in novel situations*, and that doing the right thing in unfamiliar situations without certain information (i.e. what has happened when one has acted in certain ways in this situation in the past) is more difficult than we typically assume. By analogy, the fact that an agent can quickly learn to reliably make successful putts can show they possessed the capacity to make a putt on their first failed attempt, but it can also show that they gained a capacity they initially lacked.

One particular type of counterfactual is relevant to assessing the capacities of the obedient subjects. Since we're assessing the capacity of subjects to resist-strong-social-pressures-to-harm-others-in-novel-situations, Brink treats the relevant counterfactuals as ones where the source of social pressures and the overall task remain fixed. But if we're truly interested in *this* capacity, it matters how subjects would

[4] See Quine (1951) for the *locus classicus* on the underdetermination of theory by data.

behave in other, very different situations that still involve strong social pressures and causing harm to others. For example, we might be interested in how the obedient subjects would behave if they were placed in a Stanford Prison Experiment-like setting with the original pressures (Haney et al., 1973), or various military settings featuring strong pressures and gradually escalating demands (Browning, 2001).

4.3 Predictive likelihoods

Another means of assessing capacities comes from Rudy-Hiller (2020). Rudy-Hiller argues that although it is commonplace to think there is a distinction between what we can reasonably expect of people in a normative sense (i.e. how people should behave; what it is reasonable to demand of people) and what we can reasonably expect of people in a predictive sense (i.e. how they will in fact behave), the latter can help us determine the former. More specifically, the expected likelihood of wrongdoing can affect what we can reasonably demand of one another because it helps us identify when agents have a reduced capacity or opportunity to avoid wrongdoing. Once we have an explanation for why a particular environmental factor increases wrongdoing, and that factor plausibly reduces an agent's quality of opportunity (the fairness of their opportunity to avoid wrongdoing by exercising their capacities), a high likelihood of wrongdoing in the presence of that factor enables us to conclude that avoiding wrongdoing is very difficult, which in turn "*make it the case* that normative expectations in these situations are less reasonable" (p. 2956, emphasis original).[5] Since most people fail to do the right thing in *Bystander*, this, along with plausible explanations of why the bystander effect occurs (e.g. pluralistic ignorance), demonstrates that it is very difficult to do the right thing in these settings. This gives us evidence that subjects in fact have a reduced opportunity to do the right thing, entailing that their culpability is mitigated or possibly even exculpated.

While I am sympathetic to this argument, I have a few concerns.[6] First, the extent to which a proposed explanation indicates that agents' capacities or opportunities are diminished seems to depend on already assuming quite a bit about capacities and situational aptness. Rudy-Hiller takes pluralistic ignorance to decrease situational aptness because its effect on interpretation happens automatically and without subjects being aware of its occurrence, meaning they have little chance to overcome it. But one can question whether automaticity and lack of awareness necessarily reduce quality of opportunity. Automaticity is arguably a feature of many of our

[5] Since quality of opportunity tracks fairness of demands, it can include factors beyond capacity and situational features that might impede their execution. Nelkin (2016) points out that one might have a poor quality of opportunity to Φ because Φ-ing will impose high costs or require sacrifice. Even if one is very willing and able to Φ (in that it will be easy to Φ should they choose to), they can still have a poor quality of opportunity to Φ due to said costs, such that they will not be blameworthy for failing to Φ.

[6] As we will see later, my account is compatible with what Rudy-Hiller is getting at. But we first need to establish some foundation from which to identify when agents have a lack of capacity rather than just an unexercised capacity, and this is what I will provide in Sect. 5.

cognitive processes and thus capacities (particularly non-agentive ones), and unless one thinks that unwitting wrongdoing is always excused (discussed further below), appeals to a lack of awareness need to show that the agent's lack of awareness wasn't culpable.[7]

Second, the explanation seems to do most of the work in these capacity assessments. Rudy-Hiller takes predictive likelihood to be relevant in the following way: with our explanation of the factor's (capacity-compromising) effect on behaviour in hand, predictive expectations allow us to measure the pervasiveness of the factor (how frequently said factor has effects) on agency. And when the relevant factor is capacity-compromising, pervasiveness is a good proxy for the potency of that factor. Potency in turn is a measure of how difficult it is for people to overcome a factor's negative effects on their agency. A decrease in likelihood of doing the right thing thus entails that doing the right thing is more difficult. But the problem with this reasoning concerns how we know that a factor affects behaviour *by* increasing difficulty. For example, large changes in the base rates of wrongdoing might result from only minor changes in difficulty, if people's subjective reasons only barely favour doing right prior to the minor increase in difficulty. Additionally, it's possible that situationist factors could decrease our capacities slightly, while also acting as a form of circumstantial moral luck that causes us to act on different reasons (Herdova & Kearns, 2015) but does not impede our capacities. Given circumstantial moral luck doesn't necessarily undermine blameworthiness or capacities in other settings, agents affected by the bystander effect might still be mostly blameworthy. To the extent that one may want to argue against these possibilities, one would be drawn back to considering the original explanation of the behaviour.[8]

Finally, without saying more about what difficulty is, appealing to base rates as evidence introduces a potential for circularity.[9] Rudy-Hiller sets aside the question of what difficulty is and its relation to blameworthiness, taking us to have a sufficiently intuitive grasp of what it amounts to and what role it should play in mitigating culpability. We can see why this move is problematic by considering existing analyses of difficulty within the literature. To say that X is difficult could mean:

[7] While Rudy-Hiller (2017) himself does not think ignorance is always excusing, he seems to think it isn't reasonable to expect subjects to correct for the effects of pluralistic ignorance given they don't know about it, and this is the most reliable way of overcoming it (fn. 38). But it seems that subjects who provide help don't do so by knowing about the bystander effect and correcting for it; rather, they just have a more accurate perception of their reasons to help, or they just avoid succumbing to the effect in the first place. And this is the kind of thing which the non-helping subjects may still have sufficient capacity to do.

[8] In response to the worry that pluralistic ignorance might only *somewhat* degrade quality of opportunity, Rudy-Hiller appeals "to the nature of the social interpretative process that leads to this kind of ignorance—its automaticity and its proceeding undetected by those who are subject to it—and how this process undermines agential control" (p. 2964). But this is just appealing back to the force of the original explanation, and someone like Brink might not take this to yet demonstrate a lack of control or capacity. Similarly, to the extent that base rates "confirm" (p. 2957) the difficulty, we may need to already think the task is difficult, which takes us back to considering evidence for the explanation.

[9] To emphasise, it is not in fact circular, nor is Guerrero's characterisation. However, we first need to do some more work in Sect. 5 to understand why.

- X requires sacrifice (Nelkin, 2016)
- X requires effort (Bradford, 2015; Nelkin, 2016)
- X requires skill (Guerrero, 2017)
- X has a low probability of success (Guerrero, 2017; von Kriegstein, 2019).[10]

Clearly, the first sense of difficulty is not relevant to the present discussion—none of the situationist experiments require subjects to make substantial sacrifices. And Rudy-Hiller agrees that effort is not the relevant sense of difficulty (p. 2961). Though the fact that moving furniture would require effort may make a refusal to help less blameworthy than if it required little effort, doing the right thing in the situationist experiments typically doesn't require much effort at all, and the factors which increase the rates of wrongdoing don't do so by making avoiding wrongdoing more effortful. The third sense is plausible, as there is significant overlap between things that are skills and things that are capacities, and responding to moral reasons could be understood as a skill. But since we are investigating difficulty in order to assess quality of opportunity, we would need to understand how the claim that 'base rates suggest overcoming pluralistic ignorance requires skill' is distinct from the claim that 'base rates suggest overcoming pluralistic ignorance requires certain kinds of capacities to a particular degree', given the latter is precisely what is at issue.

This leaves us with difficulty as a low probability of success, which Von Kriegstein (2019) argues the other senses of difficulty typically boil down to.[11] But now reconsider our reasoning when we take a proposed explanation of the subjects' behaviour—pluralistic ignorance—to be a plausible mitigating factor: the fact that so many subjects who experience pluralistic ignorance failed to do the right thing suggests they should be excused, because the low success rate is evidence that doing the right thing was difficult. But to say that 'doing the right thing in these settings is difficult' *just is* to say that there is a low probability that one will succeed in doing the right thing when experiencing pluralistic ignorance. This is clearly unsatisfying, and so we need to continue our search for a principled method of assessing the capacities of the subjects in the situationist experiments.

4.4 Patterns of success

A final method involving counterfactuals comes from Herdova and Kearns (2017), who modify Fischer and Ravizza's (1998) account of reasons-responsiveness. Like many others, Herdova and Kearns take two capacities to be essential to moral responsibility: reasons-receptivity and reasons-reactivity. An agent is more reasons-receptive to the extent that they recognise patterns of moral reasons. An agent is more reasons-reactive to the extent that, after recognising a moral reason, they act on said moral reason. Herdova and Kearns's novel contribution is to propose that an agents' degree of reasons-responsiveness is determined by the *understandability* of

[10] 'Difficulty in trying', which Rudy-Hiller (2020) takes to apply to overcoming pluralistic ignorance, is cashed out by Guerrero (2017, p. 204) simply in terms of it being unlikely that an agent will try.

[11] Relative to some contextually-defined reference class of agents.

the agent's pattern of behaviour across a range of counterfactual situations. Understandability is here understood from a rational standpoint, rather than a psychological or scientific standpoint. Agents are more responsible for their actions the more that they display an understandable reaction to patterns of reasons.

Herdova and Kearns argue that the situationist experiments do not show that the subjects lack reasons-responsiveness to any significant degree. This is because across a range of counterfactual situations, subjects would recognise and react to moral reasons. Indeed, many do just that in the control conditions. Because subjects' pattern of responsiveness is *overall* understandable, they retain the relevant capacities necessary for responsibility. Herdova and Kearns admit that the subjects' responsiveness is slightly less understandable than it otherwise would be if the situationist factors had no impact on subjects' behaviour, and so subjects may be slightly less blameworthy than if they had committed wrongdoing in the absence of those situationist factors.[12]

Though I don't have any objections to this proposed account of blameworthiness, there is a problem with its application to particular settings: assessing whether a pattern of responses is understandable tacitly relies on us already having a certain conception of what capacities and opportunities agents have in what settings. For example, suppose that I fail to help someone in need whenever a '*pared*' is between me and the person, but regularly help when a pared is absent. Without knowing what a pared is or how it affects agents, we could conclude that, given my overall pattern of responses across a range of counterfactuals is mostly understandable, I remain responsible for my failure. But suppose we now find out that 'pared' is the Spanish word for 'wall'. Now my behaviour becomes perfectly understandable—I can't see through walls. But we are only able to make this assessment because we already know something about human perceptual capacities and how walls inhibit the visual detection of things that are behind them. We need to assess my responsibility for failing to notice my reasons to give aid *given* there is a large wall and *given* I cannot see through walls. Without this knowledge, my failure to help when a pared is present looks simply like an odd bug in my responsiveness, and we will erroneously conclude I am blameworthy because I manage to give aid in a number of other possible circumstances when walls are absent.

This point generalises in a way that justifies individuating capacities in more fine-grained ways than we typically do. Consider agents who, as a result of brain damage, can recognise objects when held at some orientations but not others, or can recognise what an object is but not what it is used for, or can recognise moving objects but not still objects (De Renzi, 2000). Unless we already know to examine

[12] It is worth noting that Fischer (with Tognazzini, 2011) takes blameworthiness to not be decided by reasons-responsiveness alone. Instead, once we have established that an agent is sufficiently reasons-responsive (understood as an all-or-nothing property), it is a further question whether the agent is blameworthy or excused due to things like ignorance. In contrast, Herdova and Kearns take reasons-responsiveness to determine blameworthiness directly, though they are also open to reasons-responsive agents being excused because they lack the opportunity to exercise their capacities. This seems to leave it unclear how they would excuse the kinds of agents that Fischer takes to be morally responsible, but not blameworthy e.g. a mother who wrongly saves her own child over multiple strangers, or wrongdoers from poor formative circumstances.

their performance in contexts where the objects are e.g. held at a certain angle, answering 'Do these agents have the capacity to recognise objects?' based on whether they successfully recognise objects in most situations will give us the incorrect answer. Even if understandability is the correct criterion for assessing capacities, we can't identify understandability by looking at responses to various settings alone. We need to know something about those settings and about agents generally.

Herdova and Kearns imply that their account is relativised to the evidence that agents have (p. 164) and that they accept that agents can be excused because they lack the opportunity to exercise any general capacities they do have (fn. 27).[13] However, they think a particular consideration counts against taking the wrongdoing subjects in the situationist experiments to be excused. This is that a certain kind of counterfactual seems true of the subjects, which is not true of people behind walls or with the aforementioned kinds of brain damage: were the subjects to stop to reflect on their reasons, or were they to expend more mental effort, they would have recognised them.[14] And it seems implausible to say that an expenditure of mental effort causes subjects' capacity to avoid wrongdoing to be restored; rather, subjects would simply be exercising a capacity they already possessed.

These kinds of counterfactuals face a well-known problem: they give incorrect verdicts for cases where something is inhibiting someone's capacity to try or to expend effort. It might be true that someone in a coma would recognise their reasons were they to expend mental effort doing so, but all the same, they lack the capacity to recognise their reasons (Jaster, 2020). Herdova and Kearns acknowledge the counterfactual only 'typically' gives the right verdict, but given the surprisingness of the subjects' behaviour, it is a live option that some situationist experiments are exceptions. This point applies particularly in cases where the capacity to reflect and think clearly, or to translate one's judgments into action, is precisely the thing that is being inhibited, and below I'll argue that this is the case for some of the experiments.[15] We need to hold certain factors fixed when we make our capacity assessments.

Herdova and Kearns worry that giving up this counterfactual may generalise in problematic ways. Won't we have to conclude, for example, that the man who regularly gets into fights after having alcohol, but never when a police officer is present, thereby lacks the capacity to avoid fights but gains this capacity when police officers are nearby? Not necessarily. If this person always gets into fights when drunk (including all cases where doing so is extremely costly, or if they were

[13] Herdova and Kearns take agents to have a specific capacity when they have a general capacity and opportunity to exercise said capacity. They take themselves to be assessing whether the situationist subjects have the general capacity to do the right thing (fn. 27), presumably because there are no visible external factors which could impede their general capacities. But as noted earlier, opportunities can be lacking due to factors internal to agents, so we need some prior way of identifying what things reduce an agent's opportunity to exercise their general capacities.

[14] 'Effort' is understood in terms of energy expenditure. See Herdova and Kearns (2019).

[15] Additionally, once we have an explanation of why some subjects do manage to do the right thing, we will see that this doesn't seem to be attributable to them simply exerting more effort.

to e.g. unknowingly have their drink spiked at a ballet recital) and never when in the presence of a police officer (including an officer from another country, with no jurisdiction to arrest anyone) then we would think that such a person really does lose their capacity. Such a pattern of behaviour would be positively bizarre.[16] But it's because fight-prone agents don't display such reasons-insensitivity that we think they are responsible, and in any case, experience should have taught them to not drink in the first place.

Still, there is a legitimate worry here: if we can't assess an agent's capacity to do the right thing by asking whether an agent would do the right thing were they to reflect on their reasons or expend effort, which counterfactuals do we use? And how do we assess the understandability of a pattern of responses within said counterfactuals without presupposing some conception of capacities and opportunities, and without saying that agents who regularly commit wrongdoing in some settings because they just don't care are excused?

5 Capacities and inferences

Let's take stock. Suppose we accept that an agent has a general capacity to Φ (evidenced by the fact that they often Φ in many ordinary settings), and we now want to assess whether they are blameworthy for failing to Φ in a particular setting. Our basic challenge is that we need a way to tell whether they had the specific capacity (in the sense relevant to blameworthiness) to Φ and just didn't exercise it, or whether they lacked this capacity. We have seen that philosophers typically try to assess whether an agent has said specific capacity by looking at how the agent would respond in various settings. But this requires us to already have some account of which settings are relevant—which counterfactuals to apply, which possible worlds to examine, or which factors to hold fixed when examining patterns of responses.

I believe we can find a foundation from which to make our assessments. The approach I would like to take exploits our status as beings generally capable of acting according to reasons. Here's the proposal: rather than looking at a particular agent in a particular setting, and observing how they (or other agents) behave when we modify various features of said situation, we can instead find cases where certain regular patterns of behaviour are best explained as the result of a particular factor reducing agents' capacities, given we already know something about agents' motivations and incentives. This allows us to build up knowledge about what kinds of factors generally degrade capacities, and we can work from there to make assessments of particular agents in particular settings.

[16] Herdova and Kearns express some sympathy for thinking that agents who inexplicably fail to respond to reasons in some circumstances are excused precisely because their regular success in responding to reasons in other situations is evidence that something has gone amiss in their functioning (fn. 22). But this seems to be how we ought to describe the subjects in the situationist experiments. The inexplicableness is what makes these studies so interesting the first time one hears about them.

For example, suppose we observe that I regularly succeed at completing a Stroop test without error in under 30 s when offered financial incentives, but regularly fail to complete the test in under 10 s when offered equal or greater financial incentives. Given I generally seem pretty motivated to complete quick tasks for money, and given reducing the time of the test from 30 to 10 s doesn't introduce any factors likely to reduce my motivation to try my hardest, the best explanation for my continued failure seems to be that I lack the capacity to complete this Stroop test in under 10 s.

This approach seems promising, but here's one potential worry: agents can sometimes succeed in more challenging circumstances, without this thereby showing that they also had the specific capacity in less challenging circumstances. Consider the following cases:

> *Sleepy*: The longer that Neil stays awake, the less money will be taken from a charity. Neil stays awake for 50 hours.
> *Fearful*: If Neil falls asleep, his children will be maimed. The more time that passes, the higher the chance that his child's abductors will be caught before the maiming. Neil stays awake for 100 hours.

Suppose someone reasoned in the following way: we know Neil cares about charities, and was motivated to stay awake for 50 h. And this suggests that he lacked the capacity to stay awake for longer. But he did stay awake longer when he was more motivated to do so in *Fearful*. So Neil actually had the capacity to stay awake for more than 50 h in *Sleepy* all along. He simply didn't try hard enough.

This reasoning seems mistaken. Rather than thinking Neil just wasn't motivated enough in *Sleepy*, most of us would conclude that Neil had a greater capacity to stay awake in *Fearful*. But how do we show this, beyond reflecting on our own experience with trying to stay awake? We would say something like his appraisal of the second situation as dangerous to his children triggered certain responses in his brain and body e.g. greater arousal and adrenaline which reduce the drive for sleep. Importantly, these factors are not under Neil's control, and were absent in *Sleepy*. So it makes sense to say that *Fearful* itself is not proof that Neil had the capacity to stay awake (in the sense required for blameworthiness) for longer than 50 h in Sleepy, even though *Fearful* is similar to *Sleepy* and Neil in fact stayed awake for longer in *Fearful*.

This case illustrates how we tacitly rely on a host of background beliefs when making capacity assessments. This is important to acknowledge because if our background beliefs are mistaken, we may overestimate agents' capacity to perform certain actions in certain circumstances. But how do we know that factors like adrenaline and physiological arousal affect one's capacity to stay awake in the first place? We've already seen that we can't assess the subjects' capacities by simply investigating whether the subjects would have succeeded at doing the right thing if they'd been given more time, or if they were placed in an identical setting again some time later. But counterfactuals like these are useful in another way. By assessing performance in a range of circumstances where agents' motivations (or at least, mechanisms capable of generating action) should be constant, we can learn

what sorts of factors generally reduce people's capacities, and then see if those factors are present in the setting we're interested in.

This is how we gain our background knowledge regarding things like fatigue and adrenaline which informed our assessment of Neil in *Fearful*. One can readily find evidence that sleep deprivation impedes cognitive capacities by observing that people who are sleep-deprived regularly perform much worse at a range of tasks. While it is possible that sleep-deprived people are simply much less motivated to perform those tasks, it is implausible that they are more likely to crash their car while driving because their sleep-deprivation makes them simply care less about living. One can find similar evidence that adrenaline enhances physical capacities, by noting that it increases physical performance in a range of ways, even though performance did not previously increase when the agent was given additional incentives. This background knowledge then allows us to make inferences about more complex cases.[17]

Once we've identified that severe sleep deprivation significantly reduces an agent's capacity to e.g. pay attention to the road, we can work backwards and infer that slight sleep deprivation will somewhat reduce an agent's capacity to pay attention. Assessing whether it reduces it enough to excuse any particular agent for failing to pay attention may be difficult.[18] But roughly, since we take it that the agent has very strong reasons to pay attention which they experience as motivating, the more that their behaviour deviates from paying attention in the manner that they normally would, the more likely it is that sleep deprivation is degrading their capacity to pay attention.

A complication is that some factors greatly decrease an agents' chances of success, but we typically don't take this to show that the agent lacked the relevant capacity. For example, forming an intention and resolution to do wrong is likely to be the kind of thing which greatly decreases the chances that an agent avoids wrongdoing. But we would not say that agents with such intentions therefore lack the capacity to avoid wrongdoing. Instead, we naturally tend to assess whether agents have the capacity to avoid wrongdoing prior to forming any such intention. To frame it another way, we can grant that intentions and resolutions do impede capacities, but then ask whether the agent had the capacity to avoid forming the intention or resolution which impeded their capacity. Since the answer is often yes, ordinary agents who form an intention or resolution to do wrong retain the capacity to avoid wrongdoing.[19]

[17] We have to be careful to watch out for interaction effects. Injuring my left hand won't greatly affect my capacity to open a door, injuring my right hand won't either, but it would be incorrect to infer from this that injuring both hands wouldn't greatly affect me.

[18] Stipulating that the agent can't pull over to rest, and is not culpable for being sleep deprived in the first place.

[19] Note too that while a very high proportion of agents who experience weakness of will or akrasia fail to act according to the balance of reasons, treating these as relevant factors for investigation risks gerrymandering the set of relevant cases. These terms pick out cases where the agent/s *already failed* to act according to their better judgement (which often is the correct moral judgment, Huck Finn cases being an exception), and exclude cases where the agent/s succeeded. Instead, when asking whether such agents had the capacity to avoid wrongdoing, we need to treat the relevant factor as something like 'experiencing

This move allows us to handle another potential worry: if we're assessing the capacity of agents by looking at their success rates given various incentives, won't this entail that agents who commit unwitting wrongdoing (wrongdoing due to e.g. ignorance, forgetting, failing to notice something, misjudging) are always excused? After all, if anything reduces the chance of doing the right thing, being unaware of what the right thing to do is certainly will. Such an argument has been used to argue for volitionism, the position that agents can only be blameworthy for actions or omissions which are the result of some act of volition, and so agents can only be directly blameworthy for knowingly committing wrongdoing (Levy, 2011; Rosen, 2004; Zimmerman, 1997). Most capacitarians, however, accept that agents can be blameworthy for unwitting wrongdoing given they could have avoided their ignorance in some sense (Amaya & Doris, 2015; Murray & Vargas, 2020; Rudy-Hiller, 2017).[20] Unwittingness itself is not the kind of feature that we hold fixed and then ask whether the agent had the capacity to do otherwise. Just as we ask whether someone had the capacity to avoid intending to do wrong, we can ask whether they had the capacity to avoid their lack of awareness. If the agent is generally successful at identifying what the right thing to do is in most ordinary contexts, they have the general capacity to do so. If they fail on a particular occasion, we can investigate whether there might be some factors present which degraded their capacity, but ignorance itself is not always not capacity-compromising.

We can still apply the same kind of tests to work out what factors degrade epistemic capacities, including reasons-recognition. For example, people regularly fail to recognise certain features of their situation when they also have to complete some other task simultaneously, when they have limited time, or when they are intoxicated. Furthermore, their performance with these factors present typically wouldn't improve with added incentives, and at a certain level of e.g. intoxication they won't perform at all. This is good evidence that alcohol, additional tasks and reduced time reduce an agent's capacity to recognise features of their situation.

We can now see why the previous attempts to assess capacities are appealing. The reasonable person standard often gives the right result because most agents are generally reasons-responsive, we can usually recognise what kinds of factors don't provide agents reason to change their behaviour, and so the higher the number of people who fail to follow reasons, the more likely it is that their capacities are being affected. Counterfactuals are often useful because they can establish that an agent has a general capacity to Φ. If we then see no reason why the agent shouldn't have the specific capacity to Φ, this suggests that the agent's failure is attributable to them

Footnote 19 continued
a contrary desire of a particular strength' and *then* look at success rates. This is compatible with thinking that some instances of weakness of will are excused, perhaps those involved in severe depression, because depressive episodes regularly decrease agents' performance at various tasks to a large degree.

[20] Settling this debate requires engaging with the further question here about what justifies blame, which I will have to set aside. Levy (2011) appeals to desert considerations, and argues that agents who did not knowingly choose to do wrong do not deserve blame. Capacitarians are generally more inclined to argue that blame is justified on account of our occupying certain roles (Rudy-Hiller, 2017), or having a zone of competence (Amaya & Doris, 2015), or that certain patterns of responses to reasons are the result of cognitive architecture which constitutes our agential control (Murray & Vargas, 2020).

simply failing to exercise their capacities. Base rates, in tandem with plausible explanations, are convincing because *if* that explanation is complete (in that we've ruled out all other possible explanations and contributing factors) and the factor identified works by decreasing capacities, base rates *just do* measure the degree to which agents' capacities are degraded by that factor, as such patterns of responses *just are* the degree to which agents' reasons-responsiveness is deviating from its normal standard. And how understandable a pattern of responses appears will depend upon things like an agent's motivations, our background beliefs about an agents' capacities, what factors we think to hold fixed, and in most cases we can accurately assess these. Each of these approaches often draws the lines between blameworthy and excused in approximately the right place. But as argued earlier, these tests can give the wrong result when there are unrecognised differences between agents, when particular environmental factors have effects in ways that we don't yet recognise, when we haven't demonstrated that our explanation proves reduced capacity, and when our assessments of patterns rely on mistaken assumptions.

Let's return to the role that explanations and base rates play. Rudy-Hiller (2020) started with a proposed explanation—pluralistic ignorance—and, by reflecting on the nature of that phenomenon, argued it reduced capacities. I noted that this requires assuming some things about capacities and what impedes them, that the explanation might be compatible with agents possessing but not exercising capacities, and that even if it does impede capacities, we need to rule out the possibility that it may do so slightly. My approach proceeds in the other direction: examine patterns of responses beyond that of the experiment, observe that, given what we know about agents' motivations and incentives, said patterns cannot plausibly be explained by anything other than a reduction in capacities, and then arrive at pluralistic ignorance, interpreted as a capacity-degrading phenomenon. This approach doesn't have to assume anything contentious about capacities because the agent's ordinary success is evidence that they possess a general capacity to Φ. Since the patterns of behaviour after certain factors are introduced cannot be explained by anything other than a reduction in capacities, this rules out worries about alternative explanations or unexercised capacities. Once this is in hand, we indeed can use base rates to assess the degree to which capacities are compromised.

Moreover, this in turn means our conception of capacities squares with Von Kriegstein's (2019) conception of difficulty. To say that a factor increases the difficulty of a task is to say that it decreases the likelihood of success. Using base rates to assess relative difficulty, understood as probability of success, is not unhelpfully circular as one's capacity to act on moral reasons just is the degree to which one responds to moral reasons in relevant settings while holding certain factors fixed (e.g. the presence of others; the agent's evidence) but not others (e.g. intentions and ignorance). Thus, the extent to which a factor decreases degree of success in relevant situations just is how much that factor increases the difficulty. All-else-being-equal, tasks which require more effort, skill, or sacrifice typically have a lower chance of success, which is why we take such tasks to be more difficult, and correspondingly, for agents to have less capacity to perform them. However, in order for base rates to be useful in this manner, we need to first have

some idea of what factors to hold fixed, and to have clarified the role that explanations play in our assessments.

It may seem somewhat laborious explaining the structure of our reasoning when making capacity assessments and specifying the background knowledge we draw upon, given how easily we make these kinds of inferences in most settings. But it is necessary to do so because, as mentioned earlier, we risk making incorrect assessments if our background knowledge is incorrect, or is based on certain assumptions, or if we're considering a new set of circumstances which impede capacities in ways that we do not yet understand. I believe that some of the situationist experiments are just such examples. Our task now is to investigate what particular factors explain the variation in subject behaviour, and assess whether these factors are the kind of things likely to affect subjects' capacities.

6 The situationist experiments reconsidered

6.1 Dime

The most commonly offered explanation for why things like dimes affect behaviour is that finding a dime is experienced by subjects as a piece of good luck, putting them into a better mood (Isen & Levin, 1972). Although moods are thought by some psychologists to make people more motivated to help, the effect of moods on attention is thought to be particularly significant (Carlson et al., 1988). It was initially thought that positive moods broaden our attention (Conway et al., 2013; Fredrickson & Branigan, 2005; Wadlinger & Isaacowitz, 2006), but more recent research suggests this is mistaken. Instead, many emotions generate an urge to act in some way (e.g. fear generates an impulse to get away from something dangerous), and studies show that the intensity of such urges narrow our attention (see Harmon-Jones et al., 2012 for a summary). Intense emotions that generate urges to act in some way narrow subjects' attention, causing them to notice things in their peripheral vision less, pay less attention to any peripheral items they do notice, focus more on parts and details rather than wholes, and be more concrete/less abstract in their conceptual categorisations. This plausibly explains why Mathews and Canon (1975) observed that subjects were less likely to help strangers when in the presence of a loud lawn mower. Less intense emotions and moods, such as the happiness one might feel after finding a dime, have the opposite effect.

Now that we have an explanation for the variation in subjects' behaviour between conditions, what should we think regarding subjects' culpability? I believe that they remain mostly blameworthy. The confederate dropping his papers may be less noticeable than he would be had subjects found a coin, and subjects may be feeling some inclination to continue walking. This plausibly makes subjects who fail to help in the experimental condition less blameworthy than they would otherwise be. But this difference is minor. Dimes and noises are just part of the ordinary furniture of moral life which agents are familiar with. If somewhat loud noises meaningfully reduced people's capacity to perceive or act on moral reasons, it would be a wonder how pre-school teachers could ever do the right thing. Doris (2002, p. 31) reports

from personal conversation with Isen and Levin that some subjects in *Dime* actually trampled on the papers dropped by the confederate. Given the wrongs in these experiments are all relatively minor, subjects likely just don't care as much as they ought to. We have no evidence that the subjects are not sufficiently aware of the events around them, or the moral nature of their actions.

6.2 Bystander

Evidence that the subjects in these experiments genuinely have a different perception of their environment than we expect comes from passages like this concerning *Fire*:

> Subjects who had not reported the smoke also were unsure about exactly what it was, but they uniformly said that they had rejected the idea that it was a fire. Instead, they hit upon an astonishing variety of alternative explanations, all sharing the common characteristic of interpreting the smoke as a nondangerous event… subjects claimed, not that they were unworried by the fire or that they were unwilling to endure the danger; but rather that they had decided that there was no fire at all and the smoke was caused by something else. They failed to act because they thought there was no reason to act. Their "apathetic" behavior was reasonable given their interpretation of the circumstances.
> Latané and Darley (1969, pp. 252–253).

Interpretations of what the smoke was included smog, air conditioning vapours, and "truth gas".

More direct evidence that subjects really did not interpret their situation as we do comes from studies on what factors exacerbate or mitigate the bystander effect. Studies consistently show that subjects are much more likely to help when the situation is unambiguously dangerous, or when the victim explicitly labels themselves as in need of help (see Fischer et al., 2011 for a meta-analysis). Subjects are also much more likely to help when other bystanders visibly react in surprise to the event (Darley & Batson, 1973), are blind (Ross & Braband, 1973), children (Ross, 1971), or visibly unable to help (Bickman, 1971). Cramer et al. (1988) found that nurses, who notably have experience in both giving assistance and recognising when people need medical attention, did not fall prey to the bystander effect. Beaman et al. (1978) found that people who had been educated on the bystander effect are much less susceptible to it.

These studies cast doubt on the hypothesis that people experiencing the bystander effect see the moral considerations but simply fail to care about them, or have normal use of their capacities but fail to exercise them. Subjects' rate of success tracks the degree to which the victim's need for help is *recognisable* more than any other factors we could attribute as relevant to subjects' decision-making. Indeed, if subjects were being motivated by objectionable attitudes (e.g. lack of concern), we would expect them to be *less* likely to help in unambiguously dangerous situations since intervening in dangerous situations carries more risk to one's self. Learning about the bystander effect is also unlikely to substantially change one's level of

concern for others, or reduce the risk of embarrassment that comes with intervening, and yet it has a significant effect on people's likelihood of helping. The best explanation of why training reduces the bystander effect is that it enables subjects to recognise that the victim needs help. This, in turn, enables them to ignore other people's failure to help when interpreting their situation, and trigger the thought that helping is an action they are capable of performing.

These considerations count in favour of taking the subjects to have a reduced capacity to recognise moral reasons. However, given agents can be blameworthy for at least some unwitting omissions, we need some benchmark by which to assess the degree of mitigation. I think this can be provided by *Fire*. It is important to note that, had the smoke in this experiment been real, all the subjects who did not leave the room would have *died*. Given the lengths most humans will go to for self-preservation, to insist that the subjects retained the capacity to leave the room looks less like a criticisable unexercised capacity and more like pure irrationality. If many agents want to keep on living, regularly behaves in ways consistent with wanting to keep on living, have many strong reasons to keep on living, but in one very isolated setting with no features that are normatively relevant to dying regularly acts in ways that would get them killed, it becomes much more plausible to think that something in that setting is interfering with their usual reasoning capacities. The strength of the bystander effect on perceptions relevant to self-preservation, and the regularity with which it happens, gives us reason to believe that it would also have a similar capacity-inhibiting effect on perceptions of other agents who need aid. Given this, the blameworthiness of most subjects in *Bystander* is (at least) significantly mitigated because they have a significantly reduced capacity to do the right thing.

6.3 Samaritan

The subjects who failed to help in this experiment are a heterogeneous group. Some subjects seemed to not notice the confederate while walking by him:

> When we talked to them later, some of them seemed to be thinking about the victim for the first time, as if saying 'Gee, maybe the guy was sick and needed help'. They really hadn't noticed him.
> Darley, quoted in Shearer (1971, p. 11).

These subjects plausibly suffered from inattentional blindness, a phenomenon where attending to particular features of a situation causes agents to miss other aspects of their situation that normally seem obvious. An example of how strong this can be comes from Chabris et al. (2011), who had subjects follow a confederate around campus for three minutes and count how many times he touched his head. While doing this, subjects passed a staged fight only eight metres away, in which two people appeared to be loudly beating up a third person. Though this fight was in subjects' line of sight for 30 s, 58% of subjects did not notice the fight at all.[21]

[21] Most readers are probably also familiar with the closely related studies on selective attention, where people watching an object being passed around fail to notice the addition of a person in a gorilla suit (Chabris & Simons, 2010).

The subjects Darley is referring to may have 'seen' the confederate, in that, upon reflecting on their memory of the walk, they realised that there had been a person present. But it seems plausible that at the time of walking, they did not notice that a person was lying on the ground. This phenomenon of 'looked but failed to see' is thought to explain why drivers are at a much greater risk of crashing when on their phone despite keeping their eyes on the road (Hyman Jr, 2016).[22] This also plausibly accounts for why so many more subjects helped when told they were not late. Since they didn't have to concentrate on getting to the room quickly and avoiding mistakes, their attention wandered more, enabling them to notice something out the ordinary on their walk, and realise that a person was lying on the ground.

Some subjects may have noticed the confederate in a slightly stronger sense, but still failed to 'see' him in the relevant way, i.e. that he needed help. Consider the following quote:

> Our seminarians in a hurry noticed the victim in that in the post-experiment interview almost all mentioned him as, on reflection, possibly in need of help. But it seems that they often had not worked this out when they were near the victim… it would be inaccurate to say that they realized the victim's possible distress, then chose to ignore it; instead, because of the time pressures, they did not perceive the scene in the alley as an occasion for an ethical decision. Darley and Batson (1973, p. 107).

Based on our earlier treatment of *Bystander*, it seems plausible that some subjects genuinely didn't interpret the confederate as needing help. They likely had a low-to-medium grain representation of a person lying down, but failed to see some finer-grained properties, like that he was unconscious and appeared to have collapsed rather than lain down. If this characterisation is accurate, then these subjects will be only marginally blameworthy for their failure to help, as will the subjects who did not notice the confederate at all.

Not all subjects who failed to give help were like this though. Some seemed to have noticed the confederate and made a decision to continue walking:

> For other subjects it seems more accurate to conclude that they decided not to stop. They appeared aroused and anxious after the encounter in the alley… because the experimenter, whom the subject was helping, was depending on him to get to a particular place quickly.
> Darley and Batson (1973, p. 107).

It seems plausible that the factors which contributed to some subjects not seeing the confederate as needing help, or not seeing him at all, would still have some effect on these anxious subjects' construal of their situation too. But one particular consideration counts against excusing these subjects entirely. If these subjects

[22] Note such drivers are blameworthy precisely because a lot of effort is made to warn drivers of the dangers of driving while using their phones. The same cannot be said for the subjects in *Samaritan*.

were anxious upon arrival, it seems likely that they were consciously considering whether to go back and check on the confederate.[23] This would mean that they had been deliberating about whether to go back for the remainder of their walk to the room. The longer they deliberated, the less likely it is that they were genuinely unable to switch tasks. Subjects would be aware that they could go back, and could have had the sort of thoughts necessary to trigger the relevant plan, and overcome any inclinations against helping. These subjects would merely be suffering from akrasia, which isn't in itself mitigating.

6.4 Obedience

Though we can't be certain that all the obedient subjects experienced their situation in the same way, I believe there are grounds for significantly mitigating the level of blame that they deserve. A number of factors prevented them from recognising that they had a choice, and without realising this, those subjects couldn't form the right kind of intention or resolution to disobey. Subjects who did disobey seemed to do so precisely because they had this thought, plausibly due to previous experiences, exposure to certain prompts, or luck in what they were attending to.

The main piece of evidence I want to draw upon concerns at what points subjects disobeyed, which gives us some insight into *how* they disobeyed, allowing us to assess whether the obedient subjects could have done the same. Psychologists have noted that one of the factors crucial to producing the high obedience rate is the *gradual* increase in voltage of the shocks (Burger, 2009; Gilbert, 1981; Zimbardo, 2007). That this played an important role also seems supported by the fact that in multiple variations of the experiment, most subjects who disobeyed did so when the learner first complained of heart trouble. These factors are able to be explained by studies showing that humans are reliably affected by hysteresis, a phenomenon in which the point at which we change our judgements, perceptions, or attitudes regarding some variable depends on which direction our judgements, perceptions, or attitudes regarding that variable came from. For example, most philosophers know that when constructing a phenomenal sorites series and leading students through the inductive premise, which vague predicate is applied to items in the middle of the series depends on which end of the series was chosen as the starting point. Although real adults presented with heaps of rice or thermostats will change which vague predicate they apply at some point (i.e. they don't call a single grain of rice a 'heap'), the *point* at which they change from 'heap' to 'not heap' itself changes depending on which direction they come from (Raffman, 2014).[24]

Our susceptibility to hysteresis seems to explain both why the gradual increase in shocks is such an important factor, and why most subjects who disobeyed did so when the learner first complained of heart trouble. Complaining of heart trouble

[23] It is still possible that their capacity to switch tasks or form an intention to go back was impeded, but our evidence doesn't yet show this.

[24] Hysteresis has been shown affect perceptions and judgements of motion (Nichols et al., 2005), other people's emotions (Sacharin et al., 2012), speech (Tuller et al., 1994), hearing (Chambers & Pressnitzer, 2014), ambiguous sentences (Rączaszek et al., 1999), and dating behaviour (Tesser & Achee, 1994).

acted as a 'line' for some subjects to pick out, making them interpret any new shocks as being different to previous shocks. It seems likely that had there been larger increases between the shock voltages, and correspondingly larger increases between the learner's cries, more subjects would have disobeyed. The gradual increase in the shocks, which were initially harmless and thus *morally permissible* anchored subjects' construal of their situation. It would have been difficult to see administering shocks at 120 V as impermissible, given everything about this decision looks almost identical to the decision faced at 115 V. The gradually increasing scale acted as a psychological slippery slope for subjects, making it difficult to pick out any particular point at which they should stop.

In addition to Dimow's testimony about how he interpreted his situation, another source of evidence that something changed in the perception of the subjects who disobeyed comes from looking at the effectiveness of the experimenter's prods. It is notable that of the four prods used to get the subjects to continue, the last—'You have no choice, you must go on'—was completely ineffective. Not a single person who heard this prod continued to administer shocks. This is curious given the prod is much *more* direct than the previous three, and so acted as a *stronger* form of social pressure, and yet was less effective. Consider these transcripts from Gibson (2013), where the refusals of Milgram's disobedient subjects show they are aware that they can choose to not continue:

> Experimenter: You have no other choice you must continue.
> Participant 2032: I have another choice. I won't continue
> Experimenter: You have no other choice, sir, you must go on.
> Participant [unknown]: If this were Russia maybe, but not in America.
> Experimenter: You have no other choice you must [continue.]
> Participant 2036: [Oh I] certainly do have, you can have your cheque back sir.
> Experimenter: You have no choice, really.
> Participant 2005: Why?

The reason the fourth prod was ineffective seems to be that it made salient to subjects that they had a choice. By telling subjects they 'must' continue, subjects' attention was drawn towards the supposed force of this 'must', which in turn made it much easier to recognise that no such obligation existed. This plausibly triggered something like a gestalt switch in how they saw their situation, resolving the ambiguity and enabling them to clearly see the moral considerations and their options. This then allowed them to attend to certain features of their situation, ignore others, deliberate, and form an intention to disobey. It seems likely that if the experimenter had explicitly asked the obedient subjects what their final decision was, or if there was a large sign saying 'you have a choice', more subjects would have disobeyed. These prompts would have triggered a certain way of perceiving their situation, which would make disobeying much easier.

Although it seems unambiguous to us what is happening in the experiment, and very clear what courses of action the subjects can take, a number of factors plausibly had a distorting effect on subjects' interpretation of their situation and awareness of their options. First, the experimenter seems unperturbed by the learner's behaviour, which would make most reasonable people question their evaluation of the

situation. The experimenter looks like someone who has a greater understanding of what is going on, and who is much more familiar with the experiment. Subjects didn't have anyone else to look to for cues, which would make it hard for them to reconcile their initial interpretation of the shocks as harmful with what the experimenter's words and manner seem to be suggesting. Subjects would also have the reasonable belief, at least for some of the experiment, that the experimenter would eventually come to share their concern and stop the experiment.

Worth considering also are the many things that subjects in *Obedience* attended to and thought about, which plausibly distracted them from cues which might lead them to realise they could disobey.[25] Subjects had clearly arrived at the conclusion that they ought to register their concern to the experimenter in order to get him to stop the experiment. Having formed this intention, the experimenter's lack of uptake would be very puzzling, and subjects' attention would plausibly remain on trying to figure out why he wasn't responding in the expected manner, thinking about whether there was something else they could say to impress their concern upon him, and weighing up the likelihood that the learner might get the next answer right. These (very understandable) targets of subjects' attention would make it harder to consider other possibilities, such as what would happen if they chose to leave the room, or call for help, or steadfastly refuse to continue.[26]

Also notable is that the obedience rate decreased when subjects were in the same room as the learner, when they had to force the learner's hand onto a shock plate, when orders were given by the experimenter by phone, and when there was another teacher present who voiced some concern. While none of these factors greatly change the reasons that subjects have to disobey, what they have in common is that they all seem to increase the *salience* of either the wrongness of continuing, or the option of disobeying. That salience of reasons can affect quality of opportunity plausibly explains why many of us feel that failing to save a drowning child is more blameworthy than failing to donate enough money to charity to save one life, even if we accept Singer's (1972) argument that the two are equally morally wrong.

One objection to the claim that subjects' construal of their situation was ambiguous or confusing regards their distress: if subjects couldn't perceive their situation accurately, why were they so anxious about what they were doing? I don't think the obedient subjects believed that what they were doing was permissible. As people with extreme phobias can attest, perceptual seemings can be very hard to ignore when we deliberate or act, even if they conflict with our explicit beliefs. Subjects may have had thoughts like 'the learner is in pain' and 'I don't want to do this' or 'I should convince the experimenter to stop', and this would explain their

[25] The option to keep shocking may also have been much more salient than the option to disobey because of the effects of perceptual affordances (Gibson, 1979). Subjects are immediately in front of the electrocution box, and have only the experimenter to talk to, who keeps directing their attention back to the task with prompts to continue. See Ye et al., (2009) on how affordances can make us aware of the possibility of performing certain actions.

[26] "The thought of quitting never occurred to me … just to say: 'You know what? I'm walking out of here'—which I could have done. It was like being in a situation that you never thought you would be in, not really being able to think clearly."—Participant Bill Menold, quoted in Perry (2013b).

distress. But they may have also simultaneously failed to form any firm thoughts like 'I could refuse to keep shocking.'

Even if subjects formed an intention to cease shocking at some points, Cohen and Handfield (2010) point to a capacity which is typically not identified in discussions of reasons-responsiveness: the capacity to suppress or cease one's deliberation.[27] This is needed because our agency itself is temporally extended, many actions take place over an extended period of time, and it is often helpful to make resolutions in order to prevent ourselves from changing our decisions at a later time. Someone who came to a decision and then abandoned it in favour of another decision, before repeating this process over and over, may count as reasons-responsive at each moment that they form an intention. But such behaviour seems pathological, rather than a manifestation of ordinary agency. Like other capacities, the capacity to cease or suppress deliberation is one that comes in degrees. But it is also the kind of thing which can be degraded by certain factors, and in *Obedience*, the obedient subjects' "wills are not merely weak but positively anemic" (Doris, 2002, p. 134). Even if subjects thought to themselves 'he's clearly in pain, now I'll stop', the behaviour and directives from the experimenter caused them to begin deliberating again about whether to continue. Intentions often come with let-out clauses (Holton, 1999), but if one doesn't consider what ought and ought not to count as a legitimate clause, one may find that their intention alone isn't enough to produce action. Those who successfully disobeyed not only formed an intention to disobey, they seemed to frame it in such a way that 'stopping' was taken to be inclusive of 'choosing to disobey any future directives'. The importance of the capacity to suppress deliberation also seems clear if we imagine the effect of forewarning subjects about what was going to happen, and how most people behave in the experiment. 'If that many people obey, I better take care to keep my wits about me' is likely to be the kind of thought which enables a much higher rate of disobedience.

Individually, each component I have pointed to doesn't necessarily show that the subjects had a diminished capacity to avoid wrongdoing. But put together, they act as evidence that many subjects construed their situation differently to how we expect, did not realise that disobeying was something they were capable of choosing, or were unable to maintain their resolve anytime they intended to stop. Given the variety of demographics that these studies have been replicated on (Blass, 1999; Burger, 2009; Doliński et al., 2017; Edwards et al., 1969; Shanab & Yahya, 1978), subjects who disobey don't seem any more disposed to exhibit virtue, care or strength of will than the obedient subjects in other settings. Instead, the main thing that seems to distinguish them from the obedient subjects is the way they happened to interpret their situation. To interpret any given situation, our perceptual capacities depend greatly on cues from our environment, other people, and previous experiences, and we often don't notice what effect these have. When these ordinary sources of support are removed and replaced with factors specifically intended to

[27] The 'capacity to maintain one's resolve' falls into this category, but the broader capacity to cease deliberation is needed to avoid e.g. Buridan's Ass-type situations. Though such situations appear able to be avoided with the capacity to form an intention alone, ceasing deliberation is necessary to exercise this capacity.

mislead our interpretations, it becomes very difficult to know what to attend to in order to resolve ambiguities and deliberate clearly. Given this, I believe that many of the obedient subjects had a significantly reduced capacity to avoid wrongdoing, which significantly mitigates their blameworthiness.

7 Objections

One might worry this argument generalises in such a way that risks excusing too many wrongdoers. After all, there are thousands of situationist experiments, featuring factors which we are in the presence of all the time. In response, it should be emphasised that the majority of situationist experiments are more like *Dime*, and have only minor effects on our overall capacities. Additionally, many factors are likely to cancel out each other's effects unless they happen to be influencing us in the exact same way, and our capacity to mitigate such factors improves the more familiar we become with environments. In contrast, experiments like *Obedience* and *Bystander* are highly unusual, subjects have a reduced capacity partly in virtue of their unfamiliarity with the setting, and we have good reason to suspect that agents improve their capacity to do the right thing in such cases with education or forewarning. This limits the risk that my conclusion generalises in ways that excuse too many agents.

Still, one might worry that we've ended up with a conception of capacities that is too narrow, or maintain that these experiments are too similar to other instances in which we blame agents for failing to notice certain considerations or failing to interpret their surroundings in a particular way. The obedient subjects had the knowledge that shocking strangers is wrong, and they knew that they were shocking a stranger, so isn't that enough to ground culpability given they should have seen that therefore they should stop?

The regularity with which people successfully interpret their situation and use their moral knowledge to arrive at right action is great evidence that these agents had the general capacity to do so. It is also very natural to want to focus on these cases to make inferences about a present failure. But insisting that they therefore had the specific capacity to do the right thing makes both their wrongdoing and the kinds of factors which reduce the chances of wrongdoing all the more inexplicable, and this cannot be ignored in our assessments. When we realise that many factors influence success, and then notice those factors independently tend to inhibit people's capacities and judgments in other settings, it becomes more plausible that agents are having their capacities degraded. It is precisely because people are so susceptible to the influence of others and slippery slopes that we continually caution against them, trying to instil good habits and ways of thinking that increase their capacity to resist or mitigate said influence.

Arguing that the subjects are blameworthy because they 'could have worked it out' requires us to provide more than just a "disengaged comment on possibility" (McGeer & Pettit, 2015, p. 165), while also avoiding ending up with a conception of capacities that is too broad and renders intuitively excused agents blameworthy. Part of the reason we tend to think that children lack the capacities needed for

blameworthiness is that they regularly fail to do the right thing in a variety of settings. While most adults regularly succeed at doing the right thing and regularly manifest the relevant capacities, I've tried to show that they also regularly fail when in the presence of certain environmental factors, and this failure doesn't seem attributable to anything else that could both explain their behaviour and ground culpability.

In many ordinary cases when someone fails to interpret their situation a certain way, their general rate of success shows they could and should have interpreted their situation differently. But in highly unusual, confusing settings that we lack any prior familiarity with, we should naturally expect our performance to be much lower than average. We shouldn't think that our ordinary levels of success are a given and that we can maintain them everywhere for any task. Rather, we should appreciate how our ordinary levels of success and thus the capacities that enable them rely heavily upon our environments and other agents. Such factors are why philosophers have recently given more attention to how our capacities are 'socially scaffolded' (McGeer, 2012) or can rely on a certain moral ecology to be sustained (Vargas, 2018).

8 Conclusion

This paper has examined how we assess when agents lack the capacities necessary for moral responsibility, in order to assess the blameworthiness of subjects who act wrongly in the situationist experiments. I considered some existing proposals, noting problems with each. Appeals to the reasonable person standard risk missing differences between subjects regarding their level of capacities. Appeals to relevantly similar counterfactuals risk misidentifying factors that are capacity-compromising. Appeals to predictive likelihood of wrongdoing require us to first have certain kinds of explanations which demonstrate a factor is capacity-compromising, and that this accounts entirely for the decrease in success rates. And appeals to the understandability of patterns of responses requires us to already know something about agents' motivations and capacities. I instead argued that we can attribute a reduced capacity or opportunity to an agent when this is the best explanation of a range of behaviour. This occurs, for instance, when the agents regularly fail despite being sufficiently motivated to exercise any capacities they do have. From here I argued that most of the situationist experiments involving minor environmental factors such as *Dime* do not significantly reduce subjects' blameworthiness. Many subjects in *Bystander* and *Samaritan* are excused because they are not able to recognise that someone is in need of help. Subjects in *Samaritan* who do realise, and consider going back to help but do not, are blameworthy. Finally, many of the subjects in *Obedience* who obeyed are mostly excused because being able to disobey first required subjects to realise that they had a choice, intend to disobey, and maintain this intention, and these are things that the setting prevented them from being able to do.

Acknowledgements Thanks to Luke Russell, Caroline West, Dana Nelkin, David Brink, Isabelle Wentworth, and audiences at the AAP's annual conference for helpful discussions and comments. Thanks

also to the Latam Free Will, Agency, and Responsibility Project for support. This publication was made possible through the support of the grant #61255 from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

References

Adams, R. M. (2006). *A Theory of Virtue: Excellence in Being for the Good*. Oxford University Press.

Amaya, S., & Doris, J. M. (2015). No excuses: Performance mistakes in morality. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics* (pp. 253–272). Springer.

Austin, J. L. (1956). Ifs and cans. *Proceedings of the British Academy, 42*, 109–132.

Badhwar, N. K. (2009). The Milgram experiments, learned helplessness, and character traits. *The Journal of Ethics, 13*(2–3), 257–289.

Beaman, A. L., Barnes, P. J., Klentz, B., & McQuirk, B. (1978). Increasing helping rates through information dissemination: Teaching pays. *Personality and Social Psychology Bulletin, 4*(3), 406–411.

Bickman, L. (1971). The effect of another bystander's ability to help on bystander intervention in an emergency. *Journal of Experimental Social Psychology, 7*(3), 367–379.

Blass, T. (1999). The Milgram Paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology, 29*(5), 955–978. https://doi.org/10.1111/j.1559-1816.1999.tb00134.x

Bradford, G. (2015). *Achievement*. Oxford University Press.

Brink, D. O. (2013). Situationism, responsibility, and fair opportunity. *Social Philosophy and Policy, 30*(1), 121–149.

Brink, D. O., & Nelkin, D. K. (2013). Fairness and the architecture of responsibility. In *Oxford Studies in Agency and Responsibility* (Vol. 1, pp. 284–313). Oxford University Press.

Browning, C. R. (2001). *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. Penguin.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*(1), 1–11. https://doi.org/10.1037/a0010932

Carlson, M., Charlin, V., & Miller, N. (1988). Positive mood and helping behavior: A test of six hypotheses. *Journal of Personality and Social Psychology, 55*(2), 211.

Chabris, C. F., Weinberger, A., Fontaine, M., & Simons, D. J. (2011). You do not talk about Fight Club if you do not notice Fight Club: Inattentional blindness for a simulated real-world assault. *i-Perception, 2*(2), 150–153.

Chabris, C. F., & Simons, D. J. (2010). *The Invisible Gorilla: And Other Ways Our Intuitions Deceive Us*. Harmony.

Chambers, C., & Pressnitzer, D. (2014). Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics, 76*(5), 1271–1279.

Cohen, D., & Handfield, T. (2010). Rational capacities resolve and weakness of will. *Mind*, *119*(476), 907–932.

Conway, A. M., Tugade, M. M., Catalino, L. I., & Fredrickson, B. L. (2013). The broaden-and-build theory of positive emotions: Form, function and mechanisms. *Oxford handbook of happiness* (pp. 17–34).

Cramer, R. E., McMaster, M. R., Bartell, P. A., & Dragna, M. (1988). Subject competence and minimization of the bystander effect. *Journal of Applied Social Psychology, 18*(13), 1133–1148.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*(1), 100.

De Renzi, E. (2000). Disorders of visual recognition. *Seminars in Neurology, 20*(4), 479–486.

Dimow, J. (2004). Resisting authority: A personal account of the milgram obedience experiments. *Jewish currents*. Retrieved from http://www.jewishcurrents.org/2004-jan-dimow.htm.

Doliński, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycha, K., Martynowska, K., & Trojanowski, J. (2017). Would you deliver an electric shock in 2015? Obedience in the experimental paradigm developed by Stanley Milgram in the 50 years following the original studies. *Social Psychological and Personality Science, 8*(8), 927–933.

Doris, J. M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.
Doris, J. M. (2007). Out of character: On the psychology of excuses in the criminal law. In H. Lafolette (Ed.), *Ethics in Practice* (3rd ed., pp. 519–530). Malden: Blackwell Publishing.
Edwards, D. M., Franks, P., Friedgood, D., Lobban, G., & Mackay, H. (1969). *An experiment on obedience*. Johannesburg, South Africa: Doctoral Thesis, University of the Witwatersrand.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
Fischer, J. M., & Tognazzini, N. A. (2011). The physiognomy of responsibility. *Philosophy and Phenomenological Research*, *82*(2), 381–417.
Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517.
Franklin, C. E. (2013). A theory of the normative force of pleas. *Philosophical Studies, 163*(2), 479–502.
Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion, 19*(3), 313–332.
Gibson, J. J. (1979). *The Ecological Approach to Visual Perceptions*. Routledge.
Gibson, S. (2013). Milgram's obedience experiments: A rhetorical analysis. *British Journal of Social Psychology, 52*(2), 290–309.
Gilbert, S. J. (1981). Another look at the Milgram obedience studies: The role of the gradated series of shocks. *Personality and Social Psychology Bulletin, 7*(4), 690–695. https://doi.org/10.1177/014616728174028
Guerrero, A. A. (2017). Intellectual difficulty and moral responsibility. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition*. Oxford University Press.
Haney, C., Banks, C., & Zimbardo, P. G. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology & Penology, 1*, 69–97.
Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society, 100*(2), 223–226.
Harmon-Jones, E., Price, T. F., & Gable, P. A. (2012). The influence of affective states on cognitive broadening/narrowing: Considering the importance of motivational intensity. *Social and Personality Psychology Compass, 6*(4), 314–327.
Herdova, M., & Kearns, S. (2019). Difficult circumstances: Situationism and ability. *Journal of Ethical Urban Living, 2*(1), 63–91.
Herdova, M., & Kearns, S. (2017). This is a tricky situation: Situationism and reasons-responsiveness. *The Journal of Ethics, 21*, 1–33.
Herdova, M., & Kearns, S. (2015). Get lucky: Situationism and circumstantial moral luck. *Philosophical Explorations, 18*(3), 362–377. https://doi.org/10.1080/13869795.2015.1026923
Hieronymi, P. (2007). Rational capacity as a condition on blame. *Philosophical Books, 48*(2), 109–123.
Holton, R. (1999). Intention and weakness of will. *Journal of Philosophy, 96*(5), 241–262.
Hyman, I. E., Jr. (2016). Unaware observers: The impact of inattentional blindness on walkers, drivers, and eyewitnesses. *Journal of Applied Research in Memory and Cognition, 5*(3), 264–269.
Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology, 21*(3), 384.
Jaster, R. (2020). *Agents' Abilities*. New York: deGruyter.
Jaster, R. (forthcoming). The ability to do otherwise and the new dispositionalism. *Inquiry*. https://doi.org/10.1080/0020174X.2021.1904632.
Kamtekar, R. (2004). Situationism and virtue ethics on the content of our character. *Ethics, 114*(3), 458–491.
Latané, B., & Darley. J. M. (1969). Bystander "Apathy". *American Scientist*, *57*(2), 244–268.
Latané, B., & Darley, J. M. (1970). *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts.
Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology, 5*(2), 189–202. https://doi.org/10.1016/0022-1031(69)90046-8
Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.
Mathews, K. E., & Canon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology, 32*(4), 571.

McGeer, V. (2012). Co-reactive attitudes and the making of moral community. *Emotions, Imagination and Moral Reasoning, 4*, 299–326.
McGeer, V., & Pettit, P. (2015). The hard problem of responsibility. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility* (Vol. 3). Oxford University Press.
McKenna, M., & Warmke, B. (2017). Does situationism threaten free will and moral responsibility? *Journal of Moral Philosophy, 14*(6), 698–733.
Michelle, C. (2013). Situationism moral responsibility and blame. *Philosophia, 41*(1), 179–193.
Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371–378. https://doi.org/10.1037/h0040525
Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies, 177*(3), 825–843.
Nelkin, D. K. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy, 29*(1), 181–206.
Nelkin, D. K. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs, 50*(2), 356–378.
Nichols, D., Huisman, A., Rivera, M., Hock, H., & Bukowski, L. (2005). Dynamical vs. judgmental comparison: Hysteresis effects in motion perception. *Spatial Vision, 18*(3), 317–335. https://doi.org/10.1163/1568568054089393
Perry, G. (2013a). *Behind the Shock Machine: The Untold Story of the Notorious Milgram Psychology Experiments*. Scribe Publications.
Perry, G. (2013b). Taking a closer look at Milgram's shocking obedience study. *NPR*. Retrieved from https://www.npr.org/2013/08/28/209559002/taking-a-closer-look-at-milgrams-shocking-obedience-study.
Quine, W. V. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*(1), 20–43.
Rączaszek, J., Tuller, B., Shapiro, L. P., Case, P., & Kelso, S. (1999). Categorization of ambiguous sentences as a function of a changing prosodic parameter: A dynamical approach. *Journal of Psycholinguistic Research, 28*(4), 367–393.
Raffman, D. (2014). The competent use of vague words. In D. Raffman (Ed.), *Unruly Words: A Study of Vague Language*. Oxford University Press.
Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives, 18*(1), 295–313.
Ross, A. S. (1971). Effect of increased responsibility on bystander intervention: The presence of children. *Journal of Personality and Social Psychology, 19*(3), 306.
Ross, A. S., & Braband, J. (1973). Effect of increased responsibility on bystander intervention: II. The cue value of a blind person. *Journal of Personality and Social Psychology, 25*(2), 254.
Rudy-Hiller, F. (2017). A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly, 98*(S1), 398–426.
Rudy-Hiller, F. (2020). Reasonable expectations moral responsibility and empirical data. *Philosophical Studies*, *177*(10), 2945–2968.
Sabini, J., & Silver, M. (2005). Lack of character? Situationism Critiqued. *Ethics, 115*(3), 535–562.
Sacharin, V., Sander, D., & Scherer, K. R. (2012). The perception of changing emotion expressions. *Cognition & Emotion, 26*(7), 1273–1300.
Shanab, M. E., & Yahya, K. A. (1978). A cross-cultural study of obedience. *Bulletin of the Psychonomic Society, 11*(4), 267–269. https://doi.org/10.3758/BF03336827
Shearer, K. (1971). A Princeton Study Explores Why People Are - And Are Not - Good Samaritans. *Princeton Alumni Weekly*, *72*, 11.
Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs, 1*(3), 229–243.
Small, W. (2017). Agency and practical abilities. *Royal Institute of Philosophy Supplement, 80*, 235–264.
Solomon, R. C. (2003). Victims of circumstances? A defense of virtue ethics in business. *Business Ethics Quarterly, 13*(1), 43–62.
Solomon, R. C. (2005). What's character got to do with it? *Philosophy and Phenomenological Research, 71*(3), 648–655.
Sreenivasan, G. (2002). Errors about errors: Virtue theory and trait attribution. *Mind, 111*(441), 47–68.
Talbert, M. (2009). Situationism, normative competence, and responsibility for wartime behavior. *Journal of Value Inquiry, 43*(3), 415–432.
Tesser, A., & Achee, J. (1994). Aggression, love, conformity, and other social psychological catastrophes. In R. R. Vallacher & A. Nowak (Eds.), *Dynamical systems in social psychology*. Academic Press.
Tuller, B., Case, P., Ding, M., & Kelso, J. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 20*(1), 3.
Vargas, M. R. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford University Press.

Vargas, M. R. (2018). The social constitution of agency and responsibility: Oppression, politics, and moral ecology. In M. Oshana, K. Hutchinson & C. Mackenzie (Eds.), *The Social Dimensions of Responsibility*. Oxford University Press.

von Kriegstein, H. (2019). On being difficult: Towards an account of the nature of difficulty. *Philosophical Studies, 176*(1), 45–64.

Wadlinger, H. A., & Isaacowitz, D. M. (2006). Positive mood broadens visual attention to positive stimuli. *Motivation and Emotion, 30*(1), 87–99.

Wielenberg, E. J. (2006). Saving character. *Ethical Theory and Moral Practice, 9*(4), 461–491.

Ye, L., Cardwell, W., & Mark, L. S. (2009). Perceiving multiple affordances for objects. *Ecological Psychology, 21*(3), 185–217.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics, 107*(3), 410–426.

Zimbardo, P. G. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.