I, Volkswagen

Stephanie Collins

This is the final draft of a paper whose definitive version will be published in

Philosophical Quarterly

Abstract

Philosophers increasingly argue that collective agents can be blameworthy for wrongdoing. Advocates tend to endorse functionalism, on which collectives are analogous to complicated robots. This is puzzling: we don't hold robots blameworthy. I argue we don't hold robots blameworthy because blameworthiness presupposes the capacity for a mental state I call 'moral self-awareness.' This raises a new problem for collective blameworthiness: collectives seem to lack the capacity for moral self-awareness. I solve the problem by giving an account of how collectives have this capacity. The trick is to take seriously individuals' status as flesh-and-blood material constituents of collectives. The idea will be: under certain conditions that I specify, an individual can be the locus of a collective's moral self-awareness. The account provides general insights concerning collectives' dependence on members, the boundaries of membership, and the locus of collectives' phenomenology.

Key words

Collective agency, collective responsibility, blameworthiness, self-awareness, social ontology, organisations

1. Introduction

In September 2015, the United States Environmental Protection Agency (EPA) issued a 'notice of violation' to car manufacturer Volkswagen. The EPA found Volkswagen had programmed its vehicles to activate emissions controls only during lab testing—not during real-world driving. Consequently, around eleven million vehicles emitted up to 40 times more air pollutants during driving than when they were being tested against environmental regulations (Ewing 2015; EPA 2015). This was not a matter of a few bad apples, but company culture and policy (Leggett 2017). In 2017, Volkswagen pleaded guilty to criminal charges.

How should we react to Volkswagen? Natural reactions include anger, resentment, and indignation. But it's unclear where to target these attitudes. Should we target employees, who may just have been doing their jobs to make ends meet? Should we target directors, who may just have been lowering costs to make shareholders happy, on pain of shareholders replacing them with new directors who would authorise the same (or worse) wrongs? Should we target shareholders, who may have had no way of knowing about the fiddling?

Recent philosophy claims a different answer: collective agents can themselves be blameworthy. Arguments for this derive from theories of rationality (Rovane 1998), judgment aggregation results in social choice theory (List and Pettit 2011), and interpretivism in philosophy of mind (Tollefsen 2015). Yet, obviously, collective agents differ from human agents. Collectives' existence, persistence, and distinctness depends largely on social facts (norms, laws, etc), on which human agents aren't as dependent. And it's relatively easy to demarcate the physical boundaries of a human, while collectives' boundaries are often murkier (consider Uber, whose operatives are 'partners,' not 'employees'). And humans have phenomenology—there's 'something it's like' to be them—while there's apparently nothing it's like to be Volkswagen (List 2018; cf. Schwitzgebel 2015). Generalising: collective agents can seem like "mysterious," "incorporeal," "ghostly" entities, "hovering" over and above their

human constituents. (The quoted words come respectively from Rönnegard 2015, Rönnegard and Velasquez 2017, Velasquez 2003, and Ludwig 2017b; all quoted in Hess 2018a, 36.)

In the spirit of the latter concerns, this paper poses a new problem for collectives' blameworthiness. I then solve that problem, in a way that sheds new light on collectives' agency and metaphysics. The result is not a full vindication of collectives' blameworthiness: other problems may remain. But my proposed solution utilises a general framework for understanding collectives' physical dependence on members, collectives' boundaries, and collectives' phenomenology. This framework may help to solve other problems for collectives' blameworthiness, if those problems arise from the above-mentioned disanalogies between collectives and humans.

Section 2 begins by sketching an ecumenical functionalist theory of collective agency. Section 3 paves the way for the problem: the functionalist theory extends to present-day machines, yet machines aren't blameworthy. Section 4 presents the crux of the problem: I argue that blameworthiness presupposes the capacity for *moral self-awareness*—contra many defenders of collective blameworthiness (Rovane 1998; Gilbert 2002; Pettit 2007; French 2008; List and Pettit 2011; Tollefsen 2015; Björnsson and Hess 2016; Hess 2018b). The problem is that collectives are seemingly incapable of moral self-awareness. Section 5 considers two existing theories that contain potential routes to collective moral self-awareness. I argue neither of these work. Section 6 fills the gap and solves the problem, by providing an account of collectives' capacity for moral self-awareness. The trick is to take seriously individuals' status as flesh-and-blood material constituents of collectives. The basic idea will be: under certain conditions that I specify, an individual member can be the material locus of a collective's moral self-awareness. Section 7 defends this proposal from three objections.

2. Collectives' Agency

We want a permissive and schematic account of collectives' agency: an account that's neutral on, and extendable to, others' more restrictive and concrete accounts. We also need an account with clear criteria for membership. This is because (Section 6 will argue) collectives' moral self-awareness is housed in individual members. We'll need to know which individuals can do the housing.

Using these desiderata, I conceptualise collective agents as follows. A collective agent has a 'rational point-of-view': a bundle of interlocking beliefs, preferences, hopes, fears, and so on (Rovane 1998). It builds its point-of-view with a decision-making procedure, which enables the collective's point-of-view to become wide-ranging, largely internally consistent, temporally extended, and made of mental states that depart from those of members (List and Pettit 2011). Voting is one decision-making mechanism, but the most important mechanisms are often informal, implicit, and ad-hoc (Hess 2018a, 37-38). Regardless of the mechanisms used, a collective's mental states depend (at least in part) on members.

Who, then, are 'members'? I assume members meet three conditions (Collins 2019, 12-14). First, each member is (pro tanto) committed to abide by the procedure's results. Second, each member is permitted to have some input into the procedure. Third, the enactment of some possible decision of the collective requires behaviour from the member—where that behaviour is also attributable to the collective. In French's (1984) words, these behaviours are 'incorporated by' the collective. Such incorporation happens when a member behaves within, because of, and while performing their role. (Section 6 will add to this picture.)

Roles are necessary for the second and third membership conditions: roles are the vehicles through which members have inputs and through which their behaviours are incorporated by the collective. A role is defined by its relations to other roles, such as the relations 'commands,' 'obeys,' and 'collaborates with.' The roles can be represented via a

structure (a collection of roles and relations). The instantiation of the structure can be identified with the collective itself (Ritchie 2013). I interpret 'role' to include not just explicit 'tasks,' but also, for example, the inputs a member is permitted to place into the procedure, the extent to which a member is expected to be committed to the procedure, and the extent to which the procedure facilitates the member's imaginatively adopting the group's perspective.

When enough individuals meet the three membership conditions ('enough' as specified by the decision-making procedure and role structure), and together use the procedure to produce a rational point-of-view including intentions, then the collective is an agent. The collective's intentions can lead to action, via members acting within, because of, and while performing their roles. Some collectives have the procedural and material resources to take moral considerations as inputs to decision-making. The question is: can such collectives be blameworthy?

3. Machine Blameworthiness

The above story is functionalist. Mental states are characterised by their role in a point-of-view, which is characterised by its relation to a procedure, which is characterised by its relations to members, who have inputs, commitments, and behaviours associated with the procedure. Likewise, collectives' actions are characterised functionally by their relation to the point-of-view, procedure, and members. It's all about relations internal to the agential system.

On this picture, a robot can be an agent. List and Pettit embrace this, describing a small robot moving around a table-top, on which there are cylinders (2011, 19-20). Some cylinders are upright, others are on their sides. The robot places the sideways cylinders upright. According to List and Pettit, the robot is an agent in virtue of having representational states (that depict the environment) and motivational states (that require the environment to be some way), where the robot relies on the representational states to make the world conform with the

motivational states. If this makes for agency, then collective agents (as I characterised them) more than qualify.

Now consider blameworthiness, conceptualised as 'fittingness for negative reactive attitudes'—paradigmatically: anger, resentment, and indignation. Assume blame*worthiness* can hold regardless of whether, all-things-considered, we should blame, and that blameworthiness is consistent with determinism (Strawson 1962). Using this conception, suppose some calamity will ensue unless the robot sets the cylinders upright. The robot is programmed to predict and avoid the calamity. It fails to set the cylinders upright. Is the robot blameworthy? No. We would target our blame at the robot's engineers, not the robot itself—or perhaps we would label the event a tragedy. This point extends to complex machines, such as self-driving cars or self-taught algorithms. So: why don't we blame such machines? In this section, I consider three potential answers. I argue none are correct. The next section advocates a different answer.

Pettit implies the first potential answer: the cylinder-straightening robot can't reason. It's unable to ask itself "questions about connections between propositions, say about whether they are consistent or inconsistent, and then do something – pay attention to the inter-propositional relations – out of a desire to have a belief form one way or the other." (2007, 499) This process requires 'meta-propositional beliefs,' that is, beliefs about relations between propositions—including relations of support, consistency, entailment, and so on. When we reason, we use meta-propositional beliefs to revise our first-order attitudes towards propositions, to enhance the relations of support, consistency, and so on, between our first-order attitudes. Pettit suggests reasoning separates robots from humans, and from collectives (2007, 499-501; similarly List and Pettit 2011, 177-178).

However, robots—like humans, like collectives—can be (and often are) reasoners. (List and Pettit deny this of their simple robot (2011, 188), but that's a stipulation that they could

have made false.) Machines can apply modus ponens to the propositional objects of their first-order representational or motivational states; they can revisit their evidence and use this to question their beliefs; and so on. This is how machine learning works: the machine is given a data set, which it analyses to reach conclusions. When given new data, it updates the conclusions as need be. The system decides whether to reject data (old or new) as anomalous or adjust the conclusion to accommodate that data: the propositions in the data and conclusion are entertained meta-propositionally. In that sense, machines reason. But they're not blameworthy. So the capacity for reasoning cannot be the answer.

A second potential answer says: not reasoning, but reasons. Humans can give (moral, justifying) reasons for their behaviour. Present-day machines cannot. Of course, machines can be programmed to do the right thing in morally uncontested situations. And if we provide machines enough cases of agents acting morally, then machines can predict what those agents will do in a new situation—thus making (what we might recognise as) a novel moral judgment. But present-day machines are not good at giving (what we recognise as) moral *justifications* for their predictions or performances (Krishnan 2020). Perhaps the provision of reasons is what separates non-blameable machines on one side, from blameable humans and collectives on the other.

There are three problems with this suggestion. First, if a reason is a proposition that supports an action, then present-day machines sometimes can give reasons. My laptop regularly tells me it will shut down 'so that updates can be installed.' Yet my laptop is not blameworthy. Second, humans—even ethicists—often can't provide inter-personally communicable reasons. We decide based on our intuitions. And when we do provide inter-personally communicable reasons, those reasons are liable to evolutionary debunking, social conformity explanations, and so on. This is particularly true of reasons we might try to provide for blameworthy

decisions. So, machines sometimes give reasons, and humans sometimes can't give (adequate) reasons.

Third, we're ultimately concerned with collectives. Collectives often cannot provide reasons for their decisions, because members endorse the group's decision (if they do) for different reasons. That is, the group's decision is often a 'modus vivendi,' an 'overlapping consensus,' or an 'incompletely theorised agreement' (borrowing terminology from Rawls (1993) and Sunstein (1995)). When we ask a collective the reason for its decision, we'll often get an answer that appeals to the decision-making procedure—for example, 'that's what most shareholders voted for.' After all, if the shareholders voted for mutually contradictory reasons, the collective cannot coherently provide all those reasons. But 'that's what shareholders voted for' doesn't make a decision *justified*; it makes a decision *legitimate* (on this distinction, see, e.g., Rawls 1993, 225). It's the equivalent of asking me why I did something, and receiving the answer 'because that's what my balance of reasons supported.' The substantive justifying reasons for a collective's decision are the reasons why shareholders voted for the decision, not the fact that shareholders voted for it. But the collective (usually) cannot provide those substantive reasons without self-contradiction. So collectives often cannot give justifying reasons for their decisions. Yet we blame them. One might say: so much the worse for collectives' blameworthiness. But, as I mentioned, many *humans* also often cannot give reasons for their immoral actions, yet we blame them. This suggests reason-giving is not the heart of the matter.

Third and finally, perhaps an entity is blameworthy only if it has the capacity to *respond* to reasons (rather than *give* reasons). On one popular view, an entity lacks the capacity to respond to reasons just if "his behaviour would be the same, no matter what reasons there were" (Fischer and Ravizza 1998, 37). That is, responsiveness-to-reasons requires that one's behaviour is counterfactually sensitive to reasons. It also requires that "the mechanism that

actually issues in the action is [one's] own, reasons-responsive mechanism" (ibid., 39). Collectives are responsive to reasons (Björnsson and Hess 2016). The problem is: so are machines (ibid., 294). Perhaps machines do not conceptualise reasons *as* reasons, but their counterfactual behaviour varies with reasons, via a mechanism that is their own.[1] So responsiveness-to-reasons also doesn't appropriately divide entities along blameworthiness lines.

Perhaps these three views can be refined to exclude blameworthy robots, while retaining blameworthy humans and blameworthy collectives. However, given the foregoing, it's worth considering other possibilities.

4. Blameworthiness Presupposes the Capacity for Moral Self-Awareness

Why don't we blame robots? I'll argue it's because blameworthiness presupposes the capacity for 'moral self-awareness.' An entity has moral self-awareness when it has a phenomenal belief-like attitude ('awareness') to the proposition 'I will do wrong,' 'I have done wrong,' or 'I am doing wrong.' There are three components to moral self-awareness: a wrongness component (the 'moral' part), a first-personal self-identifying component (the 'self' part), and a phenomenal belief-like component that is best captured by the notions of awareness, grasping, feeling, or acquaintance (the 'awareness' part). Most adult humans have the capacity for all

[1] Coeckelbergh (2009) agrees, arguing that robots' "performance" renders them morally responsible. But Coeckelbergh is concerned with the "appearance" of responsibility, not with whether robots are "really" responsible (2009, 181). Conversely, Hakli and Mäkelä (2019) argue that robots lack the autonomy required for responsiveness-to-reasons. Regardless, in the next section I will argue that there is another, more problematic, prerequisite on blameworthiness.

three components. Machines have the capacity for the first two, but not the last. This explains why machines cannot be blameworthy. Collectives might appear to be like machines in this regard. Section 6 will argue that this appearance is misleading. In this section, I'll explain moral self-awareness and motivate its importance for blameworthiness.[2]

First, notice that moral self-awareness is not a challenge to functionalism per se. It's a challenge to functionalism as applied to collectives. After all, contemporary functionalists provide theories that account for humans' phenomenal experiences, such as moral self-awareness. These theories rely on human neurology (Wu 2018), or human brains' functional complexity, the likes of which is not exhibited by any present-day collectives (List 2018). So functionalist theories of mind can (arguably) give us *humans'* moral self-awareness. But they do so in ways that are a poor fit for collectives. Section 6 will give an account collectives' moral self-awareness. Importantly, that account can be endorsed whether or not one is a functionalist about humans' mentality.

To understand the need for such an account, we need to know what moral self-awareness is and why the capacity for it is necessary for blameworthiness. The 'self'

[2] Relatedly, Torrance (2008, 510) briefly presents the view that responsibility requires "empathic rationality." However, his argument focuses on arguing that empathic rationality is restricted to biological organisms. Similarly, Baddorf (2017) argues that the capacity for phenomenology is necessary for blameworthiness, by relying on the controversial 'phenomenal intentionality research program.' My argument has no such reliance. Moreover, Baddorf argues (briefly) that collectives *lack* the capacity for phenomenology. I'll argue collectives *have* the capacity for moral self-awareness. Hess (2010) argues organisations can satisfy the 'self' component of moral self-awareness. As we'll see, it's the 'awareness' component that really causes trouble.

component is familiar from discussions of 'de se' beliefs (Perry 1979). To represent content that's de se—'about the self'—List and Pettit's robot must be capable of representing not just 'whichever robot is on the table is about to do wrong' (a 'de dicto' belief), nor even 'this particular robot, which happens to be on the table, is about to do wrong' (a 'de re' belief). Rather, the robot must be capable of representing 'whichever robot is on the table is about to do wrong *and I am the robot on the table*' or 'this particular robot, which happens to be on the table, is about to do wrong *and I am that robot*.' The robot can have such representations in a functionalist sense: it can change course when it's about to do wrong, it can say 'I'm sorry' after doing wrong, and it can have the functionally-characterised mental states entailed by 'I am a wrongdoer,' such as 'I am less good than I could have been' (List and Pettit 2011, 187-188).

The 'moral' component of moral self-awareness requires that the entity has the concept of 'wrongness.' Again, a simple robot can satisfy this. Such a robot can abide by hard constraints. There is no roadblock to them labelling those constraints 'what morality requires' or 'constraints the violation of which would be wrong.' In this functionalist sense, robots can possess the concept of wrongness. Satisfying the moral component doesn't require a complete theory of morality, or even true moral beliefs: a blameworthy entity might have systematically incorrect moral beliefs, as callous humans do. Indeed, some callous humans never bother to operationalize a flawed concept of 'wrongness.' Thankfully for their blameworthiness, my claim is that blameworthiness presupposes the *capacity* to operationalize a concept of wrongness. So the moral component is relatively easy to satisfy.

The 'awareness' component requires the entity to have a belief-like grasp, feel, or acquaintance towards the proposition 'I have done wrong' (or 'I am doing wrong' or 'I will do wrong'). The awareness, grasping, feeling, or acquaintance is 'belief-like' in that it has a mind-to-world direction of fit. But unlike the notion of 'belief' employed so far, this 'belief-like'

state has a particular phenomenology. The phenomenology is perhaps impossible to analyse precisely, but there are evocative examples. Consider David Bourget on the size of the Sun:

> I can draw numerous inferences from the proposition <the Sun has a volume of 1.412×10^{18} km^3>: I can infer that the Sun has a volume greater than 1.3×10^{18} that the Sun has a radius of more than 100,000 km, and so on. I can also draw non-deductive inferences. For example, I can infer that the Sun is larger than my house, that I could not eat the Sun, and that the Sun would make a bad tennis ball. The relevant deductive and non-deductive inferences that I am in a position to make can be multiplied *ad nauseum*. I am also able to behave as required by such reasoning when relevant. My behavior and inferences with regard to the proposition that the volume of the Sun is 1.412×10^{18} km^3 are about as rational as can be expected of anyone towards any empirical proposition they believe, yet, intuitively, I do not really grasp how big the Sun is. (2017, 300)

Bourget is adept with the size of the Sun, just as a robot can be adept with its wrongdoing. A robot can draw all sorts of inferences, and perform all sorts of behaviours, based the belief 'I have done wrong.' So it can believe 'I have done wrong,' in a non-phenomenal sense of 'believe.' This gives an entity a lot, in terms of its relation to that proposition.

But there is something missing. The Sun's size is beyond Bourget's awareness, grasp, feel, or acquaintance. Likewise, a present-day robot's wrongful agency is beyond the robot's awareness, grasp, feel or acquaintance. This is because present-day robots cannot be aware of, grasp, feel, or be acquainted with *any* proposition. Here, I'm using 'acquaintance' to evoke the relation Bertrand Russell (1912) thought connected us to qualia or sense data. That is: there is something distinctively phenomenological or consciousness-based about awareness. At its most intuitive, moral self-awareness can be thought of as a component (though probably not the entirety) of self-directed reactive emotions like remorse, guilt, or regret. If one conjures up

the *feeling* of these emotions, one will conjure up awareness, grasping, feeling, or acquaintance towards 'I have done wrong.'

Three clarifications are in order. First, my suggestion is not that an entity must have the capacity for moral self-awareness in order to master moral concepts. Perhaps present-day robots can master moral concepts, because perhaps mastery is functionally characterizable. Instead, my suggestion is that an entity must have the capacity for moral self-awareness in order to be apt for blame. There may be entities who master moral concepts but are not apt for blame (such as machines). Second, my suggestion is that the capacity for moral self-awareness is necessary (not sufficient) for blameworthiness. Perhaps blame-apt creatures need to grasp not just their own wrongdoing, but also moral reasons or wrongness. If so, I suggest collectives' grasping of moral reasons or wrongness will work in the way I outline for collectives' moral self-awareness in Section 6. I focus on moral self-awareness rather than awareness of moral reasons or wrongness because self-awareness requires reflexivity (the 'self' component): as Section 6 will explain, this reflexivity requires members to be aware *as* the collective, so collectives' moral self-awareness is more challenging than their awareness of moral reasons or wrongness. Third, an entity can be blameworthy in a context without being morally self-aware in that context. Instead, the *capacity* for moral self-awareness is presupposed by blameworthiness, because that capacity is necessary for being the kind of entity that can be blameworthy.

Having clarified moral self-awareness, I will now provide four independent reasons for believing the thesis that blameworthiness presupposes the capacity for moral self-awareness. Together, these four reasons add up to provide sufficient reason to believe the thesis.

First, suppose the table-top robot acquired the capacity for moral self-awareness. Suppose moral calamity would ensue unless the robot set the cylinders upright, where the robot had been programmed to avoid the calamity. Then, I contend, we'd judge it apt to react angrily

towards the robot if it failed to set up the cylinders. In this way, the thesis ('blameworthiness presupposes the capacity for moral self-awareness') doesn't rely on moral self-awareness being impossible in robots. It merely relies on the idea that *if* robots became morally self-aware, *then* they would meet a presupposition of blameworthiness. In this way, the thesis explains our hesitancy around robots.

Second, if blameworthiness presupposed the capacity for moral self-awareness, that would explain why some philosophers find collective agency "mysterious," "incorporeal," "ghostly," or "hovering." The explanation would be: these philosophers infer blameworthiness from agency, then infer moral self-awareness from blameworthiness—and it's collectives' *moral self-awareness* that truly looks mysterious, incorporeal, ghostly, and hovering. This is perhaps what the First Baron Thurlow was alluding to when he famously said that groups cannot be punished, because they have "no soul to be damned and no body to be kicked" (quoted in Coffee 1981). A more contemporary version of Thurlow's point is: collectives lack the neural and physiological systems that provide naturalistic bases for humans' moral self-awareness (whether those explanations are functionalist or non-functionalist); so, they cannot be appropriately blamed or punished. This Thurlow-style inference would make sense if blameworthiness presupposed the capacity for moral self-awareness. [3]

[3] Other philosophers reject collective agents and responsibility for different reasons. For example, Seumas Miller provides an account of individuals' blameworthiness in collective contexts, which he suggests renders collectives' blameworthiness redundant (Miller 2001, ch. 8; Miller 2006). Such arguments are left untouched by the possibility of collective moral self-awareness. My aim in this paper is to develop and solve an overlooked problem for collectives' blameworthiness; I do not claim to respond to all arguments against collective agents and blameworthiness, such as those provided by Miller.

Third, as mentioned above, moral self-awareness can be conceptualised as a component of attitudes like remorse, guilt, and regret. Thus, if blameworthiness presupposed the capacity for moral self-awareness, that would gel with views on which the capacity for remorse, guilt, or regret is necessary for blameworthiness. Famously, Peter Strawson (1962) placed such attitudes at the centre of our responsibility practices. More recently, David Shoemaker has argued that psychopaths are not blameworthy because of their incapacity for these attitudes (Shoemaker 2011) and that sincerely felt remorse is necessary for blameworthiness (Shoemaker 2019). Of course, some philosophers give non-phenomenal accounts of remorse, guilt, or regret—allowing collectives to have these attitudes (Gilbert 2002; Huebner 2011; Björnsson and Hess 2016). But these accounts are highly revisionary. The standard picture is that moral emotions involve phenomenology (Scarantino and de Sousa 2018). The point here is that, if moral emotions involve phenomenology, then this section's thesis dovetails with independently-motivated philosophical views on which moral emotions are central to blameworthiness.

Fourth, consider the function of blame. According to popular accounts (e.g. Macnamara 2011, esp. 90-91; Shoemaker 2015, 110, 171; Fricker 2016), one crucial function of blame is to induce remorse in the target. (This is independent from the previous point: X can have the function of producing Y, even if the capacity for Y is not necessary for X.) The idea here is that 'inducing felt remorse' is to 'blame' as 'cutting' is to 'knives.' If the function of knives is to cut, then an apt target of a knife is something that has the capacity to be cut (air is not an apt target for a knife). Likewise, if the function of blame is to induce felt remorse, then an apt target of blame is something that has the capacity to feel remorse. This is consistent with blame (or knives) sometimes failing in its function, without failing to be blame (or knives). The point concerns what kinds of things it is appropriate for blame to target. If blame's point is inducing remorse, then an appropriate target is one with the capacity for moral self-awareness.

At this point, one might object: it is apt for us to blame unstructured groups, such as lynch mobs or misogynists, without the shadow of a presupposition that the group has the capacity for moral self-awareness. So how can that capacity be presupposed by blameworthiness? To this, I have two responses. First, blame might be a mistake when it targets unstructured groups as such, rather than each member individually (for discussion, see Feinberg 1968; Held 1970). Second, the upcoming proposals in Section 5 might capture the capacity for moral self-awareness in unstructured groups—but, I will suggest, they fail for collectives. For collectives' moral self-awareness, we will need an enriched metaphysical picture, which I will present in Section 6.

I don't claim that these four reasons prove beyond doubt that blameworthiness presupposes the capacity for moral self-awareness. But they place that thesis on a firmer footing—vis-à-vis robots and humans—than any of the three proposals considered in Section 3 (the reasoning, reason-providing, and reason-responsiveness proposals). If blameworthiness presupposes the capacity for moral self-awareness, then the question arises: do collectives have the capacity for moral self-awareness?

5. Plural or We-mode Self-Awareness

In a series of papers, Hans Bernhard Schmid has developed the concept of *plural pre-reflective self-awareness* (esp. Schmid 2014; Pettit (2018) uses Schmid's account to argue for corporations' indexicality). This self-awareness is plural: it's not held by any individual, nor by the group as a singular entity, but by individuals as a plurality. The subject and object of awareness is 'we' (plural), not 'I' (singular), where that 'I' might be individual or collective. The awareness is *pre-reflective*: non-inferential, non-observational, direct, and immediate. And the awareness is *self*-awareness: awareness by us, of us; the subject and object are identical. If the members of Volkswagen have the capacity for plural pre-reflective self-awareness,

combined with some shared conception of wrongness derived from Volkswagen's point-of-view, then Volkswagen might have the capacity for moral self-awareness.

Thus, Schmid (2018, 101) imagines a mugging on a train in which passengers gaze out the window. The passengers see the mugging reflected in the window. Each gazer falsely believes the mugging is happening in a train on the parallel track. Each thinks 'those people should act to prevent that mugging.' Suddenly, each grasps that they are looking at a reflection. It dawns on each that *those people are us*. Schmid contends that this is different from each passenger grasping '*I* am in the mugging train.' It is different because each lacks the power to stop the mugging. Only the plurality can do so—so, it is only when each realises that those people are *us* that each can reasonably feel the moral imperative to play a part in stopping the mugging. Because each feels this imperative, the imperative must derive from the plural perspective, which means 'that's us!' must be felt from the plural perspective. If the mugging continued, perhaps this would transmute into plural moral self-awareness, in which the plurality grasps 'we did wrong by inaction.'

Somewhat differently, Raimo Tuomela has long-defended the idea that individuals are sometimes self-aware in the we-mode (e.g., Tuomela 2006). Tuomela distinguishes the subject, mode, and content of a mental state. Suppose I believe I did wrong. The subject of this belief is me, the belief is in the I-mode, and belief's content refers to me. Suppose I believe we did wrong. Then the subject remains me, and the mode remains I-mode, but the belief has we-content: it's a belief about *us*. Suppose *we* believe that *I* did wrong: here, the subject is us, the content contains me, but the mode becomes ambiguous. We each might *I-mode believe* that I did wrong—each believing from our respective individual perspectives. By contrast, we each might *we-mode believe* that I did wrong—each believing from the group's perspective. Perhaps we-mode beliefs can produce collectives' moral self-awareness: although 'we did wrong' is

grasped phenomenally by each individual (the individual is the subject), this content is grasped by each individual *from the group's perspective*.

Unfortunately, plural pre-reflective self-awareness and we-mode beliefs share a problem: the subject of awareness is either the plurality (in plural pre-reflective self-awareness) or the individual (in we-mode beliefs). But the subject of a collective's moral self-awareness needs to be *the collective that outlives these specific members*—not *me*, not *us*, and not *me in the group mode*. To see this, consider large and hierarchical collectives—like Volkswagen, Oxfam, and Sweden. Call these 'organisations.' Organisations have two features that demonstrate why the subject must be the organisation itself. (These features generalise, less starkly, to smaller and less internally segregated collectives.)

First, the organisation's identity stays constant when membership changes. But when membership changes, there will be changes in whether it's appropriate for members to have plural self-awareness of *us* as wrongdoers, or be the subject of a we-mode belief that 'we' did wrong. New members (who were not around when the wrong was done) were not part of the plurality that did wrong. It would be irrational for them to have the kind of realisation the train passengers have (plural self-awareness), or be the subject of a group-mode belief where the group did not include them (we-mode beliefs). Yet we want to hold the collective responsible for what it did before the change in membership. It's true that new members *could* irrationally have plural self-awareness or we-mode beliefs. But our practices of holding organisations blameworthy shouldn't depend on members' capacities to be irrational.

Second, consider public discourse around wrongdoing organisations. The public doesn't want *these people* to see *their* wrong. The public wants *Volkswagen* to see *its* wrong. Merely changing 'the people at the top'—or even changing all the executive team—does not satisfy us that the corporation *itself* has acknowledged its wrong. That acknowledgement requires a wholesale change in the ethos, culture, or values of the organisation. Often, it

requires a distinctive kind of apology—one done by the right member in the right way. Such apologies come from the collective itself, not from one individual or from a plurality that's non-identical to the collective. Section 6's account will capture such collective apologies.

That said, Schmid's and Tuomela's accounts might be important for blaming *pluralities* (Schmid), or for blaming *individuals in a certain mode* (Tuomela). They just don't help with blaming *collectives*. They give us moral self-awareness with the wrong subject.

6. Human Loci of Collectives' Moral Self-Awareness

Above, I said present-day robots lack awareness (grasping, feeling, acquaintance). Collectives may seem in the same boat—after all, awareness is phenomenal. Even those who advocate strongly for collective agency have denied collective phenomenology (Gilbert 2002; French 2008; Björnsson and Hess 2016; List 2018).[4] If blameworthiness presupposes the capacity for moral self-awareness, this may look like the death knell for collectives' responsibility. In this

[4] Tollefsen (2006, 235, fn. 12) is an almost-exception, saying collective guilt is "felt 'through' … members. But insofar as it is produced by collective acts, results in assessment of the group, and it is the group which has the ability to get rid of such guilt (via reparations, collective apology, and so on), the guilt 'belongs' to the group itself." Tollefsen says this briefly in a footnote, and her latter three conditions don't make the *feeling* of guilt group-level rather than individual-level. Also, the three properties that Tollefsen says make the guilt 'belong' to the collective are non-phenomenal properties. Indeed, Tollefsen suggests a non-phenomenal view of guilt, when she says "[c]ollective guilt is functionally similar to individual guilt and thus is a genuine form of guilt" (2006, 234). Elsewhere, she considers (without defending) a similar approach, which again "does not have the group *itself* feeling the emotion" (2015, 132). So it's doubtful that Tollefsen attributes phenomenology to the group, as I will do.

section, I argue that collectives can have moral self-awareness. Collectives can therefore rise to the challenge that, I have argued, moral self-awareness poses for their blameworthiness. Of course, other challenges to collectives' blameworthiness might remain. But collectives will have cleared an important hurdle to inclusion in our responsibility practices.

To begin, we need to take seriously the fact that collectives (unlike robots) are *constituted by humans*, in the sense of material composition. Humans constitute collectives in the way clay constitutes a statue (Hindriks 2012; Hindriks 2013; Hess 2018a). Collectives are different agents from the humans that constitute them—but this is just as a solar system is a different system from the system of each planet that makes up that solar system, or a forest is a different system from each of the trees that compose it. Humans are to collectives as planets are to solar systems and as trees are to forests. (Hess (2018a, 41) proposes this analogy, for different purposes.)

This picture allows us to view members as the locus of the collective's self-awareness, where that awareness is properly attributed to the collective. This is an intuitive idea: if a tree contains green, the forest contains green; if a planet supports life, the solar system supports life. The tree is the *locus* of the forest's greenness; the planet is the *locus* of the solar system's life-supporting propensities. This is just to say that the tree (or planet) is a *material constituent* of the forest (or solar system), where the constituent bears some property (greenness or life-support) that (due to the nature of the property, the constituent, and the whole) is also a property of constituted object. In what follows, I will say the locus 'houses' the whole's property, in cases where a property transfers from the constituent to the whole. (Being a locus is orthogonal to being a 'proxy,' as theorised by Ludwig (2014).)

Of course, not just any property 'transfers' from the constituent to the whole: a tree might be one metre tall, while the forest is more than one metre tall (because the tallest tree is nine metres tall); a planet might have a mean temperature of 15 degrees Celsius, while its solar

system has a mean temperature much lower than that. In these cases, the tree (or planet) is not the locus of the forest's (or solar system's) height (or mean temperature). For any property, we need a story about how and why it transfers from the constituent to the whole, such that the constituent is the locus of the whole's property. What's the story when the whole is a collective?

To see that some attributes transfer from members to collectives, consider a job search committee. When the committee's chair—a human—makes a job offer that she is authorised to make, the university makes a job offer. The chair acts as a limb or organ of the university; her action is 'incorporated,' in French's (1984) phrase. The same is true not just of some *actions*, but also of some *mental states*, such as beliefs. The search committee has been authorised by the university to find the best candidate. So when the committee believes Tama is the best candidate, the university believes this. If the committee has just one member, then this belief transfers from a singular human to a collective; the hirer is the locus of the collective's belief.

But the university doesn't believe everything the hirer believes. We must place conditions on when the mental states of an individual are mental states of their collective, just as we place conditions on when the properties of trees are properties of their forest. These conditions might be different for different mental states. We are concerned with moral self-awareness. And we are looking for a possibility proof of collectives' capacity for moral self-awareness, because we are looking for a possibility proof of their blameworthiness. So we are looking for *merely jointly sufficient* conditions for a member housing a collective's self-awareness. That said, we want our jointly-sufficient conditions to capture a good number of cases, so it would be good if something in the vicinity of our jointly-sufficient conditions were necessary. And it would be good if our conditions were somewhat explanatory of *why* the awareness is attributed to the collective.

I propose the following. A collective is morally self-aware if, and because, a member is aware that 'this first-personal agent did (or is doing, or will do) wrong' (1) within the remit of their role, (2) because of their role, (3) while performing their role, and (4) while adopting the point-of-view of the collective. I'll now elaborate on this proposal, before Section 7 defends it by responding to objections.

An initial clarification concerns 'this first-personal agent.' Is the first-personal agent the member, in which case, how can this amount to *the collective's* being aware of *itself*? Or is the first-personal agent the *collective*, in which case, how is the locus *the member's* moral self-awareness? The second option is correct: the member is aware of the collective, and so the locus of the collective's moral self-awareness cannot be *the member's individual* moral self-awareness. Instead, the member is morally self-aware *as* the collective. Returning to the three parts of moral self-awareness: the 'moral' part is the collective's functionally-characterised concept of wrongness, as determined by its rational point-of-view; the 'self' is the collective; and only the grasped, felt, or acquainted 'awareness' is the member's.

More concretely, when a member grasps 'we did wrong' while conditions (1) to (4) are met, then the 'we' refers to the collective as a singular entity. In English, this grasping is expressed using 'we' rather than 'I'—but, in this instance, 'we' is the first-person singular. Thus, the proposal is this: if a member of Volkswagen is aware of (feels, grasps, is acquainted with) 'we, Volkswagen, did wrong in fiddling our emissions'—and if this awareness is had within, because of, and while performing the member's role, and while adopting Volkswagen's rational point-of-view, then *Volkswagen grasps* 'I, Volkswagen, did wrong in fiddling my emissions.' Importantly, this mental state has a phenomenal component. The mental state's locus is the human member, who provides a physiological and neural basis for this phenomenology. Functionalists about human phenomenology can breathe easy, while non-functionalists can account for the human member's phenomenology in their usual way.

To see how this works in practice, consider collectives' apologies. Consider, for example, CEO Matthias Mueller's apology for Volkswagen's emissions scandal.[5] Or consider Australian Prime Minister Kevin Rudd's apology for Australia's treatment of Indigenous people.[6] Assuming these apologies were sincere, here we see an individual housing a collective's awareness that 'I, Volkswagen, did wrong' or 'I, Australia, did wrong.' (Again, 'we' is the word used; my suggestion is that 'we' is sometimes singular. This makes sense of Rudd's use of 'We, the Parliament of Australia': Parliament is a singular entity, if any collective is.) During their apologies, Mueller and Rudd were—we hope—grasping, feeling, or being aware of their collective's wrongdoing. What's more, each was aware of this *as* the collective, as evidenced by their use of the *first-person* pronoun 'we.' These apologies were given within the remit of their roles, because of their roles, while performing their roles, and while adopting the collective's rational point-of-view. Taking these apologies at face value, both Volkswagen and Australia have been morally self-aware.

Conditions (1)-(4) are designed to ensure that not just any old member's grasping of Volkswagen's wrong is a grasping *by* Volkswagen. Condition (1) says the member's grasping is *within the remit* of their role—that is, the member's role must *permit* that the member grasps Volkswagen's wrong from Volkswagen's point of view. Only then is the members' grasping *authorised* by the collective—just as a search committee's actions and beliefs are authorised by a university. Of course, this permission (to grasp Volkswagen's wrong from Volkswagen's point of view) is unlikely to be written into any member's contract. More often, the permission will be inferable from the norms, ethos, or culture of the collective. These 'informal' role elements are as real as formal ones (Miller 2001, 163; Hess 2018a). And we should require

[5] https://www.youtube.com/watch?v=t7ne3cwqI4A

[6] https://www.youtube.com/watch?v=b3TZOGpG6cM

only that collective moral self-awareness is *permitted*—rather than *mandated*—by a member's role, because otherwise it would be too easy for a collective to avoid responsibility by failing to mandate any member's moral self-awareness. The permission is much harder for a collective to avoid, consistent with the freedom that is necessary for high-ranking members (at least) to do their job.

Condition (2) says a member has moral self-awareness *because of their role*. This condition can be broken down into two components: the member's role is part of the correct psychological explanation of the member's mental state *and* the member's role is part of the justification the member would give for their mental state if asked. These two components can come apart, since sometimes the true psychological explanations of our action are hidden from us, even after reflection.

Condition (3) says the moral self-awareness occurs while the member is *performing their role*. This condition is tricky to delineate: not all roles have uniforms, circumscribed working hours, or clear performance locations. Generally, an individual's role performance can be identified by asking about the normative expectations surrounding the individual. When an individual performs her role, there is a shift in what others take her to owe them and what they taken themselves to owe to her. For example, when performing her role, a car salesperson suddenly owes customers information about cars, and is suddenly owed the driving of cars by the dealership's valet. Of course, customers and valets can get their expectations wrong. A member's role performance can usually be discerned by a *critical mass* in changed expectations, including a change in the role performer's expectations of themselves. Still, there will inevitably be grey cases of role performance.

Condition (4) says the member *adopts the collective's point-of-view*. This occurs when the collective's bundle of beliefs, preferences, and other mental episodes is the bundle from which the member reasons. To adopt a collective's point-of-view, the member needn't have

access to *all* the collective's mental episodes. This would be too difficult in large complex groups. Instead, the member need only be able to adopt *enough* of the collective's point-of-view for her to reason as that point-of-view requires. Specifically, a member who houses the collective's moral self-awareness must have access to the collective's beliefs about the wrongful action in question, including the collective's belief that the action is wrong.

Each condition admits of a 'subjective' and 'objective' reading: an individual can *believe* she acts within, because of, and while performing her role, and while adopting the group's point-of-view; and an individual can act in a way that accords with the objective facts about what is within, because of, and while performing her role, and about the group's point-of-view. The member's moral self-awareness is the collective's moral self-awareness when *both* the subjective and objective readings of the four conditions are true. This makes the conditions somewhat difficult to satisfy—but, again, the conditions are intended as jointly sufficient. I include subjective as well as objective conditions to accommodate the possibility that one can be morally self-aware only if one is aware of one's moral self-awareness.

As emphasised in Section 4, blameworthiness presupposes that its target has the *capacity* for moral self-awareness—not that the target *is* morally self-aware in every context where it does wrong. I've outlined when a collective *is* morally self-aware. When does a collective have the *capacity*? It has the capacity just if it bears (in a functionalist way) the concepts 'wrongness' and 'self,' and at least one member has the capacity to house the collective's moral self-awareness.

Of course, in large collectives, it's unlikely that all members will house the collective's moral self-awareness on any occasion, or even that all members have the capacity to do so. This might seem contradictory: the collective is both morally self-aware and not morally self-aware; or capable and incapable. But there is no contradiction. We can usefully view the members of a collective at one time as analogous to the temporal parts of an individual across

time (Dietz 2020). There are times when I am morally self-aware and times when I am not. This is a 'contradiction' only if we insist on asking whether I am morally self-aware in some time-neutral sense. Likewise, there is a contradiction within collectives only if we insist that all co-temporal parts of a collective are perfectly in-sync. But the parts of any agent are rarely in-sync. One member housing the collective's moral self-awareness is enough for the collective's moral self-awareness and one member's capacity is enough for the capacity.

This point unlocks two virtues of my proposal. First, it yields a scalar account of collective moral self-awareness. A collective is *more morally self-aware* on a given occasion if a higher proportion of its members house its moral self-awareness on that occasion. This is practically significant: we might reasonably demand that a collective is *more morally self-aware* when it commits worse wrongs, compared to when it commits less worse wrongs. My proposal translates this as: for more severe collective wrongs, we might reasonably demand that a higher proportion of the collective's members house the collective's moral self-awareness.

The second virtue concerns the differential capacities of differently-ranked members. When we blame collectives, we often want *highly-ranked* members to atone. My account explains this. Recall that the 'moral' and 'self' components of moral self-awareness are given a functionalist reading: these components are characterised by their role in the collective agential system, including their role of causing the collective to be disposed to certain behaviours (such as apologies, reparations, and preventions of future wrongdoings). In many collectives, only highly-ranked members have the capacity to induce this causal role, so only highly-ranked members are capable of producing the 'moral' and 'self' components of moral self-awareness. When we demand moral self-awareness from an entity like Volkswagen, we don't just want the phenomenal awareness of 'I, Volkswagen, have done wrong.' We *also* want

the functional profile that comes with moral self-awareness. So we go after the CEO, not the production line worker.

7. Objections

There's a seeming tension in the account: if collectives are materially constituted by members, and if their moral self-awareness (and other exercises of agency) depend on members, then isn't the collective at the mercy of members? And if so, then isn't it unfair to hold the collective blameworthy?[7] In response, my account does not place collectives at the mercy of members. Collectives hold various kinds of control over their members: editorial control (Pettit 2004), structural control (Strand 2012), guidance control (Fischer and Ravizza 1998), and programme control (List and Pettit 2011). I'm neutral between these accounts of control, which differ slightly in the details. The common core to each is that the collective ensures a particular (type of) action will be performed, even though members are the ones who perform the action and who determine in what precise way it will be performed (and, in particular, a given member might determine whether they are the one to perform it). Issues of collective control are beyond the scope of this article; suffice to say that moral self-awareness does not produce any new control-related problems that collectives' actions did not already face (and that have been discussed at length by others).

Next, one might object that I have set the bar too high, if my aim is to provide sufficient conditions that are explanatory and that approximate necessity. Specifically, one might worry that large powerful collectives will easily be able to set things up such that *no* member has the capacity to house the collective's moral self-awareness, so that the collective lacks the capacity for moral self-awareness, so that the collective cannot be blameworthy. Most straightforwardly,

[7] I thank an anonymous reviewer for raising this.

perhaps a collective can do this by never introducing the concept of wrongness into its point-of-view. Perhaps criminal organisations do this, or cut-throat for-profits. Yet surely such collectives are blameworthy!

My answer is two-fold. First, such collectives are extremely rare in the real world. The Mafia has definite views about moral right and wrong, so it has the capacity for moral self-awareness (even if it is systematically mistaken, as many callous individuals are). Almost all for-profit corporations pay lip service to values such as human rights, seemingly committing to a functionally-characterised belief that human rights should be upheld. As thin as such commitments are, they let the concept of wrongness into the collective's point-of-view. So almost all real-world collectives have the concept of wrongness. Second, for the few collectives that genuinely lack the capacity for moral self-awareness, I bite the bullet: they are apt targets of containment, deterrence, and restraint—but not of blame. Crucially, though, there will almost always be blame in their vicinity. It's plausible that agents have obligations to teach collectives the capacity for moral self-awareness. These agents include founders, managers, regulators, donors, shareholders, and stakeholders. If those agents have reneged on those obligations, then those agents are blameworthy. This is analogous to parents being blameworthy for not teaching their children the capacity for moral self-awareness.

Finally, conversely, one might think I have set the bar too low. Why think my conditions are sufficient for Volkswagen being morally self-aware *in its own right*?[8] My answer is: for the same reasons we think members' actions under certain conditions are sufficient for Volkswagen acting in its own right. (Arguments that members' actions sometimes transfer to collectives are given by, e.g., French 1984, Hindriks 2013, Rovane 2014.) Thus, my argument is one of parity: if members are sometimes the loci of collectives' actions, then members are

[8] I thank an anonymous reviewer for raising this.

sometimes the loci of collectives' moral self-awareness. I believe the onus is on those who want to say that contract-signings, or job-offerings, are *dis*analogous to moral self-awareness. It is a curious fact about the literature that so many philosophers accept that actions (like contract-signings) can transfer from members to collectives, yet none so far has accepted that phenomenal states (like moral self-awareness) can transfer from members to collectives.

Naturally, however, not all philosophers accept collectives' actions (Miller 2001; Ludwig 2017a). These philosophers will remain unmoved by my point that collectives' phenomenal states can be derived in much the same way as collectives' actions. My aim in this paper has been to pose—and solve—just one new problem for collectives' blameworthiness, namely, the problem that blameworthiness presupposes the capacity for moral self-awareness. I have assumed that collectives can perform actions, under the account of collective agency given in Section 2. I have not addressed or refuted problems for collective blameworthiness that arise from arguments that deny collectives' actions. For those swayed by such arguments, work remains to be done by advocates of collectives' blameworthiness.

8. Conclusion

Plausibly, a blameable entity has numerous features. I have focused on one: the capacity for moral self-awareness. I have argued—contrary to others who advocate collective moral responsibility—that moral self-awareness is inadequately addressed by pointing out that collectives satisfy a functionalist picture of moral agency, which robots also satisfy. Instead, I argued that blameworthiness presupposes the capacity to grasp one's wrongful agency. Fortunately, collectives can grasp their own wrongful agency, via members that grasp content within, because of, and while performing their roles, and while adopting the group's perspective. This picture provides us with a general way of understanding collectives' physical dependence on members (it's the dependence of physical constitution), collectives' boundaries

(which are determined by the limits of members' roles), and collectives' phenomenology (which is sometimes housed within members). There may be sound objections to collective blameworthiness, but moral self-awareness is not one of them. Our blame of Volkswagen lives to die another day.[9]

References

Björnsson, Gunnar and Kendy Hess. 2016. "Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents." *Philosophy and Phenomenological Research* 94(2), 273-298.

Bourget, David. 2017. "The Role of Consciousness in Grasping and Understanding." *Philosophy and Phenomenological Research* 95(2), 285-318.

[9] I wrote the first draft of this article while a Visiting Research Professor at the University of Vienna. I particularly thank Herlinde Pauer-Studer for hosting me in Vienna and for numerous helpful conversations about these arguments while I was there. Research for this article was funded by the European Union's Horizon 2020 Research and Innovation programme under grant agreement No. 740922, ERC Advanced Grant 'The Normative and Moral Foundations of Group Agency'. I subsequently worked on this article while receiving financial support under the Australian Research Council's DECRA scheme (project number DE200101413). For comments on written drafts, I thank Nevin Climenhaga, Luara Ferracioli, Simon Goldstein, John Hawthorne, Ole Koksvik, Richard Rowland, Hannah Tierney, and participants at a workshop on 'Collective and Shared Responsibility' at the 2019 Mancept Workshops in Political Theory at the University of Manchester. For helpful feedback on presented versions, I thank audiences at Monash University, St Andrews University, the University of Sydney, and the University of Wollongong.

Coeckelbergh, Mark. 2009. “Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents.” *AI & Society* 24, 181-189.

Collins, Stephanie. 2019. *Group Duties: Their Existence and Their Implications for Individuals*. Oxford: Oxford University Press.

Dietz, Alexander. 2020. “Are My Temporal Parts Agents?” *Philosophy and Phenomenological Research* 100(2), 362-379.

Environmental Protection Agency (EPA). 2015. “EPA, California Notify Volkswagen of Clean Air Act Violations / Carmaker allegedly used software that circumvents emissions testing for certain air pollutants.” Available at https://web.archive.org/web/20170302172909/https://yosemite.epa.gov/opa/admpress.nsf/a883dc3da7094f97852572a00065d7d8/dfc8e33b5ab162b985257ec40057813b

Ewing, Jack. 2015. “Volkswagen Says 11 Million Cars Worldwide Are Affected in Diesel Deception.” *The New York Times.* 22 September.

Feinberg, Joel. 1968. “Collective Responsibility.” *Journal of Philosophy* 65(21), 674-688.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control*. Cambridge University Press.

French, Peter. 1984. *Collective and Corporate Responsibility*. Columbia University Press.

French, Peter. 2008. “Responsibility with No Alternatives, in Loss of Innocence, and Collective Affectivity: Some Thoughts on the Papers by Haji, McKenna, and Tollefsen.” *APA Newsletter on Philosophy and Law* 7(2), 13-18.

Fricker, Miranda. 2016. “What’s the Point of Blame? A Paradigm Based Explanation.” *Nous* 50(1), 165-183.

Gilbert, Margaret. 2002. “Collective Guilt and Collective Guilt Feelings.” *Journal of Ethics* 6(2), 115-143.

Hakli, Raul and Pekka Mäkelä. 2019. “Moral Responsibility of Robots and Hybrid Agents.” *The Monist* 102, 259-275.

Held, Virginia. 1970. “Can a Random Collection of Individuals be Morally Responsible?” *Journal of Philosophy* 67(14), 471-481.

Hess, Kendy M. 2010. “The Modern Corporation as Moral Agent: The Capacity for ‘Thought’ and a ‘First-Person Perspective.’” *Southwest Philosophy Review* 26(1), 51-69.

Hess, Kendy M. 2018a. ‘The Peculiar Unity of Corporate Agents.’ Pp. 35–60 in Kendy M. Hess, Violetta Igneski, and Tracy Isaacs (eds), *Collectivity: Ontology, Ethics, and Social Justice*. Rowman and Littlefield.

Hess, Kendy M. 2018b. “Does The Machine Need a Ghost? Corporate Agents as Nonconscious Kantian Moral Agents.” *Journal of the American Philosophical Association* 4(1), 67-86.

Hindriks, Frank. 2012. ‘But Where Is the University?’ *Dialectica* 66(1), 93–113.

Hindriks, Frank. 2013. ‘The Location Problem in Social Ontology.’ *Synthese* 190(3), 413–37.

Huebner, Bryce. 2011. “Genuinely Collective Emotions.” *European Journal for Philosophy of Science* 1, 89-118.

Krishnan, Maya. 2020. “Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning.” *Philosophy & Technology* 33, 487-502.

Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge University Press.

Leggett, Theo. 2017. “VW Papers Shed Light on Emissions Scandal.” BBC News. 12 January.

List, Christian. 2018. “What Is It Like To Be A Group Agent?” *Nous* 52(2), 295-319.

List, Christian and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Ludwig, Kirk. 2014. “Proxy Agency in Collective Action.” *Nous* 48(1), 75-105.

Ludwig, Kirk. 2017a. *From Plural to Institutional Agency: Collective Action II*. Oxford University Press.

Ludwig, Kirk. 2017b. "Do Corporations Have Minds of Their Own?" *Philosophical Psychology* 30(3), 265-297.

Macnamara, Coleen. 2011. "Holding Others Responsible." *Philosophical Studies* 152(1), 81-102.

Miller, Seumas. 2001. *Social Action: A Teleological Account*. Cambridge University Press.

Miller, Seumas. 2006. "Collective Moral Responsibility: An Individualist Account." *Midwest Studies in Philosophy* XXX, 176-193.

Perry, John. 1979. "The Problem of the Essential Indexical." *Nous* 13(1), 3-21.

Pettit, Philip. 2007. "Responsibility Incorporated." *Ethics* 117, 141-201.

Pettit, Philip. 2018. "Consciousness Incorporated." *Journal of Social Philosophy* 49(1), 12-37.

Rawls, John. 1993. *Political Liberalism*. Columbia University Press.

Ritchie, Katherine. 2013. "What Are Groups?" *Philosophical Studies* 166(2), 157-72.

Rönnegard, David. 2015. *The Fallacy of Corporate Moral Agency*. Springer.

Rönnegard, David and Manuel Velasquez. 2017. "On (Not) Attributing Moral Responsibility to Organizations." Pp. 123-142 in Eric Orts and N. Craig Smith (eds), *The Moral Responsibility of Firms*. Oxford University Press.

Rovane, Carol. 1998. *Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton University Press.

Russell, Bertrand. 1912. *The Problems of Philosophy*. Oxford University Press.

Scarantino, Andrea and de Sousa, Ronald. 2018. "Emotion." *Stanford Encyclopedia of Philosophy*.

Schmid, Hans Bernhard. 2014. "Plural Self-Awareness." *Phenomenology and the Cognitive Sciences* 13(1), 7-24.

Schmid, Hans Bernhard. 2018. “Collective Responsibilities of Random Collections: Plural Self-Awareness among Strangers.” *Journal of Social Philosophy* 49(1), 91-105.

Schwitzgebel, Eric. 2015. “If Materialism is True, the United States is Probably Conscious.” *Philosophical Studies* 172, 1697-1721.

Shoemaker, David. 2011. “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics* 121(3), 602-632.

Shoemaker, David. 2015. *Responsibility from the Margins*. Oxford University Press.

Shoemaker, David. 2019. “Blameworthy but Unblameable: A Paradox of Corporate Responsibility.” *Georgetown Journal of Law & Public Policy* 17, 897-917,

Strand, Anders. 2012. “Group Agency, Responsibility, and Control.” *Philosophy of the Social Sciences* 43(2), 201-224.

Strawson, Peter. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48, 1-25.

Sunstein, Cass. 1995. “Incompletely Theorised Agreements.” *Harvard Law Review* 108(7), 1733.

Tollefsen, Deborah. 2006. “The Rationality of Collective Guilt.” *Midwest Studies in Philosophy* 30(1), 222-239.

Tollefsen, Deborah. 2015. *Groups as Agents*. Polity.

Torrance, Steve. 2008. “Ethics and Consciousness in Artificial Agents.” *AI & Society* 22, 495-521.

Tuomela, Raimo. 2006. “Joint Intention, We-Mode and I-Mode.” *Midwest Studies in Philosophy* 30(1), 35-58.

Velasquez, Manuel. 2003. “Why Corporations Are Not Morally Responsible for Anything They Do.” *Business Ethics Quarterly* 13(4), 531-562.

Wu, Wayne. 2018. “The Neuroscience of Consciousness.” *Stanford Encyclopedia of Philosophy*.