Self-induced Moral Incapacity, Collective Responsibility, and Attributability

Abstract:

Niels de Haan (2023) defends the possibility of holding collectives morally responsible against a challenge posed by *the problem of self-induced moral incapacity.* Self-induced moral incapacity seems to introduce a responsibility gap that corporate agents might exploit to avoid responsibility. De Haan argues that the problem does not introduce responsibility gaps because collective moral agents become responsible for actions they committed while they were incapacitated once they *reacquire* moral capacity. I argue that de Haan's argument is incomplete. Simply because a group *reacquires moral capacity* does not also necessarily mean that the group *retains attributability* for the wrongs in question. This means we must supplement de Haan's argument with an *Attributability Proviso*: To overcome the problem of self-induced moral incapacity it must be the case that the group reacquires moral capacity *and* that we can still attribute the earlier actions to the group despite the changes that it has gone through.

Felix Lambrecht

Keywords:

collective moral agency; responsibility gaps; identity; attributability

In a recent article published in this journal, Niels de Haan (2023) defends the possibility of holding collectives morally responsible against a challenge posed by *the problem of self-induced moral incapacity*. Self-induced moral incapacity seems to introduce a *responsibility gap* that collective agents might exploit to avoid responsibility. De Haan argues that the problem does not introduce responsibility gaps because collective moral agents become responsible for actions they committed while incapacitated once they *reacquire* moral capacity.

In this paper, I argue that de Haan's argument is incomplete. Section 1 reconstructs de Haan's argument. Section 2 presents a problem for this argument. De Haan says that a corporate moral agent bears responsibility for the actions it committed while incapacitated once it reacquires moral capacity. However, I argue that the changes the collective agent undergoes in the process of losing and reacquiring moral capacity may also change the features that make it possible to *attribute* the earlier actions to the later collective agent. This means that for de Haan's argument to succeed we must supplement it with an *Attributability Proviso*. Section 3 sketches the *Attributability Proviso* and concludes.

1.0 de Haan's Argument

De Haan attempts to vindicate the *collective moral agency thesis* (CMA). This thesis says that it is possible to hold a collective agent (CA) responsible for an action even if we cannot hold any individual fully responsible for that action. Holding CAs responsible fills a *responsibility gap* (de Haan 1-4). We often cannot attribute full responsibility to the individuals who are part of a CA when a CA commits a wrongful action. For instance, we may not be able to hold any one individual responsible for a massive oil spill caused by a corporation's negligence. If we could not hold CAs responsible for their actions then there would be a gap in responsibility for such wrongs, where such wrongs would not go accounted for (2-4). CMA allows us to attribute responsibility to the *collective itself* and ensure that some agent is responsible for such actions.

CMA seems threatened by the *problem of self-induced moral incapacity*. The problem is as follows. Only *moral agents* can be held morally responsible for wrongs. We can only hold agents responsible when they have the capacity to respond to moral reasons and act on moral evidence (2-5). A CA, such as a corporation or state, has moral capacity and is a collective *moral* agent when it has an *organizational structure* that can respond to moral reasons and process moral evidence in a way that gives the CA moral capacity (10-13). However, a CA's organizational structure may change such that the CA can no longer process or respond to moral reasons. When this happens, the CA ceases to be a moral agent. In some cases, a collective moral agent may change *its own* organizational structure to one that is unable to process moral reasons or evidence (11-13). These are cases of *self-induced moral incapacity*. Self-induced moral incapacity poses a problem because it reintroduces a *new* responsibility gap in two ways. First, it introduces moral loophole for the actions committed while the group was incapacitated. If CAs can self-induce capacity before committing a wrong and thereby avoid responsibility, then this seems like a loophole in moral responsibility. CAs could simply self-induce moral incapacity to avoid responsibility for wrongs they want to commit. We need some explanation for why CAs are not permitted to do this. Second, the problem of self-induced moral capacity might seem to challenge CMA *entirely*. De Haan suggests that a moral system which allows this sort of loophole should be called into question (2-3). If CMA comes at the cost of allowing the loophole created by the problem of self-induced moral incapacity, then this might give us reason to doubt CMA altogether (3-4). For both reasons, we need some solution to the problem: We need to explain how there is some agent we can hold responsible for actions the CA committed when incapacitated even though these actions were committed by an incapacitated agent.

De Haan introduces the following argument to solve the problem. When a CA undergoes organizational structure change that is *self-induced* such that it loses its moral capacity and then undergoes a *second* organizational structure change such that it *reacquires* moral capacity, this CA acquires *diachronic moral responsibility* for the actions it committed while it lacked moral capacity (14-18). At time t_1, CA g with an organizational structure adequate for moral capacity commits action Φ which self-induces organizational structure change, making g lose its moral capacity. Φ is a wrong because a CA has an obligation to maintain its moral capacities (15-16). At t_2 g commits wrong Ψ while incapacitated. Then, at t_3 g changes its organizational structure to one that is adequate for moral capacity. At t_3 g acquires moral responsibility for Φ and Ψ. Ψ was committed by a CA that lacked moral capacity. However, because Ψ resulted from the CA having self-induced moral incapacity (action Φ), and because Φ was committed by a CA which had moral capacity, when the CA reacquires moral capacity, it also becomes responsible for Ψ.[1] This argument prevents the responsibility gaps seemingly created by the problem of self-induced moral incapacity. A CA cannot simply self-induce incapacity to avoid responsibility for a wrong since it will be responsible for the wrongs it committed while incapacitated once it reacquires capacity.

[1] De Haan's argument relies on the claim that when an agent self-induces incapacity, that agent can be held responsible for any of the wrongs that the resulting incapacitated agent commits, even if those wrongs are not foreseeable by the agent at the time of self-inducing incapacity (15-16). This means de Haan must only show that the agent at t_3 is responsible for Φ, since that entails the agent is also responsible for Ψ. I will accept this claim for the purposes of my argument.

2.0 A Problem

De Haan's argument faces the following problem. De Haan's argument relies on saying that a later CA which has reacquired capacity has *diachronic responsibility* for the earlier CA's action. However, diachronic responsibility often requires that an earlier agent's action is *attributable* to the later agent. And, as I'll now argue, when a CA reacquires capacity, it does not *also necessarily* reacquire the conditions required for attributability that make it diachronically responsible for the past action.

Attributability is the idea that an action is *one's own* (Shoemaker 2012, Khoury 2013). The general idea is that an action is attributable to an agent when the agent did that action in a way that coheres with that agent's true self or endorsed central features (Shoemaker 2012, Schroeder 2022). In contrast, when an agent performs an action because of manipulation, coercion, or even an unendorsed impulse, the action is not attributable to the agent (Khoury 2013, Schroeder 2022).

Many philosophers think that responsibility requires attributability.[2] This means that for any agent *A* to bear responsibility for any action Φ, *A* must stand in the right kind of *Attributability Relation* with Φ. Some philosophers argue that the Attributability Relation requires *numerical identity*: Φ is attributable to *A* only if *A* is numerically identical to the agent which committed Φ (Glannon 1998). Others argue that the Attributability Relation does not require identity but still requires some form of psychological connectedness: Φ is attributable to *A* only if *A* is

[2] Glannon 1998, Shoemaker 2012, especially fn. 1, Khoury 2013. I consider the objection that responsibility does not require attributability in Section 2.1.

psychologically connected to the agent which committed Φ (Shoemaker 2012; Khoury 2013). For now (though I'll consider an objection to this in a moment), I'll assume these philosophers are correct: Responsibility requires some form of attributability. This also means that *diachronic responsibility* requires attributability. A later agent is responsible for an earlier action only if the earlier action is attributable to the later agent. Importantly, if the agent has undergone changes since the time it committed the action in question, these changes must have preserved the Attributability Relation. Whatever features make it possible for us to attribute the earlier action to the earlier agent must be preserved through the changes the agent has undergone such that we can attribute the earlier action to the later, changed agent.

Attributability creates a problem for de Haan's argument. CAs undergo significant changes to their organizational structures through the process of losing and reacquiring moral capacity. At t_1, CA with an organizational structure adequate for moral capacity undergoes self-induced organizational structure change such that it loses moral capacity (action Φ). Call this CA that had moral capacity and committed Φ to lose capacity "g_1". At t_2 a CA commits wrong Ψ while lacking moral capacity. Call this CA with an organizational structure inadequate for moral capacity "g_2". At t_3 g_2 changes its organizational structure to one that is adequate for moral capacity. Call this new CA with moral capacity "g_3". De Haan argues that g_3 is responsible for Φ and Ψ because g_3 has *reacquired* moral capacity. But for g_3 to have responsibility for Φ and Ψ, Φ must be attributable to g_3. This means that the Attributability Relation must have preserved through the changes that the CA underwent from g_1 to g_3. Just because a CA reacquires moral capacity does not mean that the Attributability Relation required for responsibility for earlier actions is preserved. The changes the CA undergoes may also have changed the features that allow us to attribute the actions to *g_3*. We can see this by looking at numerical identity and psychological connectedness.

First, consider numerical identity. There is no developed literature on CA identity. However, we still may have reason to doubt that CA identity necessarily preserves through the kinds of changes that de Haan's argument concerns. These sorts of changes have the potential to disrupt any of the features that could be plausible candidates for a CA's criterion of identity. For instance, you might think that a CA's identity corresponds to some *essential features* (Diamantis 2018). The changes sufficient to change the organizational structure that g_1 and g_2 undergo might also trigger changes to the essential features of the group. For instance, changes to the decision-making structure that eliminate moral capacity might remove certain members from the group (de Haan 10-13). This might mean that the essential features of the group change. And, when the group reacquires moral capacity when becomes g_3, it might not reacquire *the same* essential features even if it *does* manage to reacquire moral capacity. To be clear, my argument here is not that g_1 is necessarily *not* identical to g_3 following these changes. Nor is my point that the organizational structure is the group's criterion of identity. Rather, my point is that changes to organizational structure that are sufficient to make the group lose moral capacity might also result in or accompany changes to the features that make the group's identity. This means that simply reacquiring an organizational structure sufficient for moral capacity does not guarantee that the later group is identical to the earlier group. If numerical identity is necessary for attributability, then the actions committed by the earlier groups cannot be attributed to the later group even though it reacquires capacity.

Second, let's consider psychological connectedness. Many philosophers argue that attribution does not require identity but, rather, only requires psychological connectedness (Shoemaker 2012, Khoury 2013). This might take the form of having the right kind psychological states that make it possible to hold the agent responsible for the wrong (Khoury 2013). Or attribution might require some sort of core central features of the self that multiple numerically distinct agents might share (Shoemaker 2012). In either case, the view says that agent *A* bears diachronic responsibility for a past action Φ committed by agent *B* when *A* has the same psychological features that made it possible to hold *B* responsible for Φ.

These arguments about psychological connectedness are about *individual* responsibility. A full account of collective diachronic responsibility would have to explore what (if anything) counts as psychological connectedness for collectives. However, for my purposes here, I do not need to establish what collective psychological connectedness would amount to, or whether it is possible. Rather, all I want to suggest is the following. Whatever collective attributability requires, it will likely need to be analogous to individual psychological connectedness to have the same role in attributability. That is, there will need to be some feature analogous to psychological connectedness that makes a later group responsible for the action of an earlier group.[3] Importantly, as with the identity point above, it does not necessarily follow from the fact that the group reacquires moral capacity that the group also retains this feature. Any changes the group's organizational structure undergo from t_1 to t_3 may have removed these features. When this occurs,

[3] Note: Even if you deny that collectives must be psychologically connected you still need to establish that both the earlier and later group are relevantly related. And this alone is enough for the worry I raise here: Simply because a group reacquires capacity does not mean that the features that make the group relevantly related have preserved through the changes it has undergone.

even if the group reacquires moral capacity, it does not follow that the group necessarily reacquires the features that made it possible to hold it responsible for the past actions.

We can now see the problem for de Haan's argument. The changes that a CA undergoes between t_1 and t_3 are sufficient to change the CA's organizational structure such that it loses and then reacquires moral capacity. These changes might also be sufficient to change the features that preserve the Attributability Relation. My claim here is not that the changes the CA undergoes necessarily challenge the Attributability Relation or the possibility of responsibility. Rather my point is that simply because the CA reacquires capacity does not mean that the Attributability Relation is *also* preserved. The CA has undergone significant changes that are sufficient to change its moral capacity, and these changes might also have changed the features required for us to attribute Φ to g_3 and, thus, to hold g_3 responsible for Φ. This re-introduces the possibility of responsibility gaps. The later group might not be responsible for the actions of the earlier group because those actions might not be attributable to the later group. Just as a group could self-induce moral incapacity to avoid responsibility in the way de Haan worries, a group could self-induce moral incapacity *and* undergo change that is significant enough to change its identity or psychological connectedness to avoid responsibility. This means that to avoid the responsibility gap created by the problem of self-induced moral incapacity, we must supplement de Haan's argument in a way that insulates it from the responsibility gap created by attributability. CAs must reacquire capacity *and* do so in a way that preserves attributability. This means that we must supplement de Haan's account with an *Attributability Proviso*.

Before sketching the *Attributability Proviso* I will consider two objections to my argument.

2.1 Objection: Responsibility without Attributability

One might object to my argument by suggesting that responsibility does not require attributability.[4] For instance, Olle Blomberg (2022) has recently argued that an agent may be responsible for the intentional action of another agent. If this kind of argument is correct, then it might seem my argument here is not. We would not need to ensure that the earlier action is attributable to the later agent for the later agent to be responsible for it.

For the purposes of my paper, I remain neutral about whether responsibility generally requires attributability. What matters for my argument here is that the move to deny the attributability requirement does not salvage de Haan's argument. Here's why. To deny the attributability requirement, one must point to some other feature about an agent that allows us to hold that agent responsible for the action in question. In most cases, this is some sort of *foreseeability* criterion: An agent *A* is responsible for action Φ when *A* foresees Φ or a morally significant outcome of Φ *even if* Φ is not attributable to *A* (see Blomberg 2022, 556-558). However, the cases we are concerned with here involve *diachronic responsibility*. This foreseeability is not relevant for such cases. The later group in the cases that concern de Haan's argument cannot foresee the past actions since they are in the past. So, even if one denies that attributability is necessary for responsibility or diachronic responsibility (a point I'll remain neutral on here), one still needs to establish that a later agent stands in the right sort of relation with the earlier agent or action. And, once we require this sort of relation, then we encounter the problem I have raised: It

[4] Thanks to an anonymous reviewer for raising this objection.

is not simply enough to reacquire capacity; the changes later CA goes through must also preserve the feature that allowed us to say the later CA is appropriately related to the past CA and action.

2.2 Objection: Responsibility without a *Present* Agent

One might object in a second way. Perhaps we do not need to identify a *present* agent responsible for the wrong to avoid the responsibility gap de Haan worries about. That is, perhaps we can simply say that the earlier CA which self-induced incapacity (g_1) is responsible for Ψ committed by g_2. Even if g_1 no longer exists and there is no *present agent* responsible for the past actions, this does not mean there is no agent responsible for the past action at all. We can still hold the past (no longer extant) agent responsible for the past action. And, thus, there is no responsibility gap since we can still point to an agent which is responsible.[5]

This approach would not avoid the responsibility gap in the way de Haan attempts. De Haan attempts to avoid a responsibility gap in which an agent could self-induce moral incapacity and thereby avoid being held responsible. A moral system which allows this, he thinks, should not be accepted. This means that a solution to the problem of self-induced moral incapacity must locate responsibility for the past action in a later, *extant* CA. If identifying *some, possibly no-longer-extant* CA that is responsible for the action were enough, then CAs would be able to self-induce capacity, change their features such that the past action is not attributable to them, and thereby avoid being held responsible in precisely the same way that de Haan worries. Put another way, it seems to be a responsibility gap and serious moral loophole if a CA could self-induce capacity,

[5] Thanks to an anonymous reviewer for raising this objection.

commit wrongs, reacquire capacity and then, when asked to repair those earlier wrongs, say "oh, those wrongs were committed by a different agent which no longer exists!". This means that to avoid the responsibility gap and the moral loophole de Haan worries about, the CA we identify as responsible for the past actions must exist in the present.

3.0 The Attributability Proviso

I have argued that the fact that Cas can reacquire moral capacity is not enough to overcome the problem of self-induced moral incapacity. It must also be the case that the Attributability Relation between the later CA and the past action has preserved through changes to the organizational structure. This means that for de Haan's argument to solve the problem of self-induced moral incapacity we must supplement it as follows:

> *Attributability Proviso:* g_3 at t_3 is responsible for the wrong Ψ committed by g_2 when g_2 lacked moral incapacity only if (i) g_3 reacquires moral capacity at t_3 and (ii) the Attributability Relation between g_3 and the action g_1 committed to self-induce moral incapacity (Φ) preserves through changes that the CA undergoes from t_1 to t_3.

Clause (i) is de Haan's argument. However, as I have argued, this is not enough for us to say that g_3 is responsible for Φ and Ψ. This is why we must add clause (ii). g_3 can be responsible for Φ and Ψ only if we can attribute Φ and Ψ to it, that is, if the Attributability Relation preserves despite changes from g_1 to g_3.

A full defense of CMA from the problem of self-induced moral incapacity will need to explore important questions about what attributability requires in the context of CAs. This might include an exploration of collective identity criterion or an equivalent to psychological connectedness for groups. I leave this for future scholarship. All I have aimed to show here is that a defense of CMA from the problem of self-induced moral incapacity requires that the later CA be appropriately related to the past CA.

References

Blomberg, Olle. 2023. “How to Be Morally Responsible for Another’s Free Intentional Action.” *Journal of Ethics and Social Philosophy* 25 (3): 545–79.

Diamantis, Mihailis E. 2018. “Corporate Essence and Identity in Criminal Law.” *Journal of Business Ethics* 154 (4): 955–66.

De Haan, Niels. 2023. “Collective Moral Agency and Self-Induced Moral Incapacity.” *Philosophical Explorations* 26 (1): 1–22.

Glannon, Walter. 1998. “Moral Responsibility and Personal Identity.” *American Philosophical Quarterly* 35 (3): 231–49.

Khoury, Andrew C. 2013. “Synchronic and Diachronic Responsibility.” *Philosophical Studies* 165 (3): 735–52.

Schroeder, Mark. 2022. “Attributive Silencing.” In *Oxford Studies in Normative Ethics Volume 12,* edited by Mark Timmons, 1st ed., 170–92. Oxford University Press.

Shoemaker, David. 2012. “Responsibility Without Identity.” *The Harvard Review of Philosophy* 18 (1): 109–32.