2 Responsibility beyond Control

Ibo van de Poel and Martin Sand

2.1 Introduction

Our modern highly technological society seems to be confronted by a paradox of control. On the one hand, we can – at least collectively – increasingly control our environment, nature, society, and ourselves through new technological means. On the other hand, this very development seems to have led to an increase in humanly-induced technological and natural risks that we cannot, or hardly, control; think of climate change and COVID-19, or of the potential perils of new technologies like climate engineering and synthetic biology. In particular, in relation to autonomous technologies, Ezio di Nucci has recently argued that these technologies are employed in order to gain more control over traffic safety or military operation. As a result, he argues, we have to cede control: "In order to increase (or improve) control, we must cede it, and this is what I argue is paradoxical. […] The reason for this is simple enough: software – whether it is installed on a car or, as we will see shortly, on many other things – is better than we are at controlling, so that if we really care about control, we must let software take care of it for us – and not just for software or cars" (Di Nucci 2021: xiv).

These new risks emerging from increased control raise profound questions about responsibility. Who, if anyone, is responsible for them? In some cases, like climate change, it seems obvious that we are at least collectively responsible while it remains unclear how such collective responsibilities translate into individual responsibilities. For sure, individuals have some responsibilities and obligations with respect to climate change, like seeing to it that collective agreements to abate it are reached (van de Poel et al. 2012) or making reasonable individual contributions (Björnsson 2021). However, it is eccentric to assume that everyone is *individually* responsible for the whole of excessive climate change, as that is obviously beyond individual control.

In the case of risks of new technologies, like climate engineering, synthetic biology, and artificial intelligence, we may also lack knowledge

DOI: 10.4324/9781003276029-3

about the exact risks. That is to say, the risks may be uncertain or even unknown; we may not only not know the risks beforehand, but even be unable to come to know them before they have occurred (van de Poel 2017). This lack of control, again, raises the question to what extent we can individually and collectively be responsible for such risks.

The above considerations raise some serious questions about the relation between control and responsibility. It is commonly assumed – and echoed by many philosophers[1] – that we cannot be responsible for things beyond our control. Of course, we can inquire what type of control is exactly required for responsibility, or, more precisely, for what type of responsibility like blameworthiness, accountability, or forward-looking responsibility. However, it seems unfair to attribute responsibility to some agent i for some φ if i had no control over φ.

Still, there also appear to be cases in which people take responsibility for something that is at least initially beyond their control. One may think of Greta Thunberg, Nelson Mandela, or Martin Luther King who committed themselves to a greater cause. Typically, these people take a forward-looking responsibility to correct some evil in the world (like abating hunger or injustices), or to see to it that some future risk or hazard (like a climate disaster) is forestalled. They are typically unconcerned about their range of control when taking responsibility; they somehow feel they should take responsibility. This does not mean that they ought to take such responsibility from a moral point of view. Often, this is not the case, and we are usually inclined to judge their responsibility-taking as morally supererogatory (i.e., as morally praiseworthy but not required).

In line with such cases, we will suggest that at least under some conditions it can be permissible and reasonable to take on new forward-looking responsibilities, even if the object of such responsibilities is initially beyond our control. This is not to say that control is irrelevant in these situations, quite the contrary. We will suggest that in these cases, the typical relation between control and responsibility is reversed. Rather than being responsible for what is already under our control (and, perhaps, because it is under our control), we are sometimes moved by the call of responsibility, and as a result of taking responsibility, we aim to increase control.

Our aim in this contribution is to further tease out this idea. In order to do so, we first inquire what type of control is required for responsibility. Since most of the philosophical literature on control and responsibility has focused on backward-looking responsibility, and more specifically on blameworthiness, we start out with a delineation of the control condition for blameworthiness by building strongly on Fischer's and Ravizza's (1998) theory of moral responsibility. Next, we show how these insights translate to the case of forward-looking responsibility and spell out what

control would be required for forward-looking responsibility. We then argue that we need to distinguish between being forward-looking responsible for some φ (or others attributing such responsibility to us) and cases in which an agent *takes* forward-looking responsibility. We argue that the latter type of case allows room for taking forward-looking responsibility for things that are still beyond the agent's control. Next, we discuss whether taking on such responsibilities is merely morally supererogatory or not, and whether we can also assume "too much" responsibility. Lastly, we argue that one might understand the reciprocal relation between control and responsibility by zooming in on the underlying notion of moral agency.

2.2 What Control Is Needed for Responsibility?

Control has mainly been discussed as a condition for backward-looking responsibility and more specifically for blameworthiness in the philosophical literature. We take blameworthiness here to mean that it is appropriate to blame an agent i for some action or state-of-affairs φ. So understood, blameworthiness means that i is a proper target for blame (with regard to φ), but it does not necessarily mean that i is also actually blamed, or that it would necessarily be obligatory, or even desirable, to blame i for φ (Sand and Klenk 2021).[2] We also leave open the possibility that if i is not blameworthy in the responsibility-sense here intended, it might nevertheless be possible to appropriately blame her on other grounds.[3]

Blameworthiness is not the only sense of backward-looking responsibility. We may, for example, also distinguish accountability and liability (e.g., van de Poel, Royakkers, and Zwart 2015). Here, we will focus on blameworthiness as the main type of backward-looking responsibility, which is at the fore of much of the recent philosophical literature. Roughly, the idea is that for an agent i to be blame-responsible for some φ, there needs to be a certain connection between i and φ. The question is what minimal conditions need to be met to make it appropriate (or fair) to blame i for φ.

There are several conditions that the connection (between i and φ) may have to meet (e.g., foreseeability), but a necessary condition in any theory of responsibility seems to be *control*.[4] The basic idea is that without some control over φ by i, it would be inappropriate to blame i for φ. Originating from Thomas Nagel's work on moral luck, this intuitive idea has been put into a standard formulation by Dana Nelkin and is known as the control principle (CP): "We are morally assessable only to the extent that what we are assessed for depends on factors under our control" (Nelkin 2013). In a recent publication, one of us (Sand 2020) argued that this formulation of CP is too broad; there are types of moral

assessment where control plays no or only a subordinate role. Hence, in the following, we will endorse the alternative and more specific formulation of CP focusing not on moral assessment but responsibility as blameworthiness: "People are blameworthy only for things within their control" (Sand 2020).[5]

But what type of control is exactly required for blame-responsibility? Fischer and Ravizza (1998) argue that what is required is not regulative but guidance control. Regulative control involves the possibility to act otherwise, or to bring about other consequences. Guidance control does not require that; it only requires that the action (or consequence) is in a more limited sense under the control of the agent. For actions, guidance control requires that the action results from a reason-responsive mechanism that is the agent's own. Fischer and Ravizza (1998) argue that the condition of guidance control better meets our intuitions about blame-responsibility in a number of cases than regulative control.

For consequences (or states-of-affairs), the conditions for guidance control are somewhat more complicated. Fischer and Ravizza (1998) distinguish here between what they call consequence-particulars and consequence-universals. Consequence-particulars refer not just to a state-of-affairs but also to the (specific) way in which it was brought about (e.g., "the mayor was killed by me"). For consequence-particulars, they propose the following condition for guidance control: "An agent S has guidance control over a consequence-particular C just in case S has guidance control over some act A, […] and it is reasonable to expect S to believe that C will (or may) result from A" (Fischer and Ravizza 1998, 121).

For consequence-universals, this condition does not hold as Fischer and Ravizza (1998) point out. The reason is that consequence-universals can also be realized by what they call triggering events, like actions by others, or external events. For example, the consequence-universal "the mayor was killed" might be caused by me killing her but also by somebody else doing so or by a natural event like a lightning stroke. In such cases, they argue, guidance control needs to be split between the internal process leading to the action (bodily movement) and the external process from the agent's action to the outcome. For the former, the guidance control conditions for action apply (reason-responsive own mechanism). For the latter, they argue that the agent's action (bodily movement) "must be sensitive … in roughly the following sense: if the actual type of process were to occur and all triggering events that do not actually occur were *not* to occur, then a different bodily movement would result in a different upshot (i.e., … a different consequence-universal)" (Fischer and Ravizza 1998, 112). This condition implies that the outcome of the external process (i.e., the consequence-universal in which we are interested) needs to be *action-responsive* in the right way to the action of the agent.

Later authors have pointed out that this criterion (which is more formally formulated in Fischer and Ravizza [1998]) does not work in some cases in which the actions of different agents *jointly* determine an outcome without having engaged in a joint action (i.e., the actors act independently and unaware of each other) (e.g., Björnsson 2011; Brown 2011). The following is an example of this (the case is called *The Lake* and introduced in Björnsson [2011]): suppose three individuals pour an amount of a substance into a lake, unaware of each other. Two amounts of the substance are enough to poison the lake. Who is responsible for the consequence-universal "the lake is poisoned"?

If we apply Fischer's and Ravizza's criterion, it would seem none of them. The actual type of (external) process here is that all three pour an amount of substance and, therefore, if one of them would have acted differently the same outcome would still apply (as two amounts are enough to poison the lake). Such responsibility attribution, however, seems wrong. We are inclined to say that all three are equally responsible.[6] To deal with this type of case, we might want to weaken the action-responsive condition.[7] We may, for example, formulate a weaker action-responsiveness condition along the following lines: "There is at least one scenario (possible world) in which whether agent i doing or omitting her action makes a difference for the outcome."[8] Such a scenario factually exists for each of them. Consider, for example, the scenario that agent A and agent B, but not agent C pour their amount, then *in this scenario* the outcome is action-responsive to the actions of both agent A and agent B; and we can formulate similar scenarios for agents A and C, and for B and C.

This new action-responsiveness condition is quite weak, and it might well be possible to imagine other cases that show that it is too weak.[9] The point, however, is that there is a reasonable condition for action-responsiveness between the apparently too strong version of the action-responsiveness condition proposed by Fischer and Ravizza (1998) and this rather weak version of the condition. If this is indeed on the right track, it seems to show something important, namely that, as Fischer and Ravizza (1998) suggest, action-responsiveness is the right kind of control condition for the external process, even if we might not yet be exactly sure how to spell it out.[10]

We conclude then that blame-responsibility minimally requires guidance control and that in the case of consequence-universals this guidance control has two components, namely an internal reason-responsive mechanism that is the agent's own and which results in the action of the agent, and an external process that is action-responsive (i.e., the consequence-universal needs to be action-responsive in the right sense to the agent's action).

2.3 Control and Other Types of Backward-Looking Responsibility

Our aim now is to investigate whether the established control condition also applies to forward-looking responsibility. Before we do so, it is worthwhile to briefly consider the question whether the previously discussed condition of guidance control also applies to other kinds of backward-looking responsibility besides blameworthiness. We will briefly consider two other main types here, namely accountability and liability.

We take it that if we hold an agent i accountable-responsible for some φ, in which φ again is an action or outcome, we ascribe an obligation to i to account for the occurrence of φ (or at least i's role in the occurrence of φ). It seems that for such ascription to be appropriate, it would not be required that (we know for sure that) i had control over φ but only that we have a reasonable suspicion (expectation) that i had control over φ.

Take the following simple case: you are having a conversation with someone else and suddenly you slap that person in the face. It would seem completely appropriate for her to ask: why did you slap me in the face? And by asking this, the person demands you to account for what you did. Now, perhaps, you are able to provide an explanation of your action that shows that it was not under your control. Maybe you have a condition that sometimes, unexpectedly, causes seizures of sorts, like slapping others, that is not under your control (because it is not reason-responsive). While this may be a perfectly acceptable explanation, which also shows that it would be inappropriate to blame you, it does not mean that the initial ascription of accountability was inappropriate. On the contrary, by holding you accountable, your counterpart confirms that you are a moral agent, who under normal conditions is able to control herself and hence is responsible for her actions, albeit not for this specific action (cf. Watson 2004, 8).

Something similar may well apply to liability-responsibility, which we take to be the obligation to rectify some φ (for example, by compensation or repair). Some authors hold that you can only be *morally* liable for some φ if you are also blame-responsible for that φ (e.g., Hart 1968). If that were the case, it would follow that you can only be morally liable for things under your (guidance) control. Others hold that sometimes causal responsibility, rather than blame-responsibility, may be enough to be morally liable (e.g., Honoré 1999). Consider again the case of you slapping someone in the face. This time the other person is seriously hurt and in pain. It would seem appropriate to say that you are morally liable in this situation (assuming you have regained control over your actions) to help that person and to call a doctor, for example. Such cases still require some control (i.e., same basic control over one's actions and control over some action that rectify φ), but they do not require past control over the occurrence of φ.

2.4 Forward-Looking Responsibility

Let us now focus on the control condition for forward-looking responsibility (other authors have used somewhat different terms here like "prospective responsibility" or "active responsibility") (e.g., Bovens 1998; Cane 2002). We take forward-looking responsibility to mean that the agent i has an obligation to see to it that φ (with φ a state-of-affairs) (cf. Goodin 1995).[11] While we can talk about both forward-looking and backward-looking responsibility in the past, present, or future tense, what distinguishes the two is that when we ascribe backward-looking responsibility, we do so *from the viewpoint* that φ has already occurred (even when this φ is in the future); we may, for example, ask whether agent i would be backward-looking responsible, if φ were to happen in the future. But in answering this question, we take an imaginary viewpoint at some future moment in time in which φ has already occurred and can no longer be changed. Conversely, if we ascribe forward-looking responsibility, we do so from the point of view that φ has not yet occurred. Of course, we can ask whether an agent i was forward-looking responsible for some φ that happened (or did not happen) in the past, but we should judge the responsibility ascription from the viewpoint that φ has not yet occurred. These distinctions will turn out to be important when it comes to the question what type of control is needed for forward-looking responsibility. They also underline, as did the previous section on accountability and liability, that we cannot simply assume that the control condition applies equally to different kinds of responsibility.

Therefore, to tease out the control condition for forward-looking responsibility, we will start with a rather general characterization of forward-looking responsibility and the type of control that seems required. We have seen that forward-looking responsibility can be understood as the obligation to see to it that φ, from the viewpoint that φ has not yet occurred. In terms of control, this seems to require that the responsible agent has some forward-looking control, or what we may call *causal efficacy*, with respect to φ.

One way in which we may understand such causal efficacy is as the capacity to *ensure* φ. This is, however, a quite strong condition because in order for i to have the capacity to ensure φ, i must be able to realize φ *under all possible external conditions*. Effectively, this means that i should have regulative control over φ.[12] But perhaps there is another plausible way for understanding causal efficacy that does not require regulative control but only guidance control. To see whether that is indeed possible, let's look at what is typically expected from an agent who has forward-looking responsibility for φ.

Goodin (1995, 83) suggests that forward-looking responsibility "require[s] certain activities of a self-supervisory nature from [agent] A. The standard form of responsibility is that A sees to it that X. It is not enough that X occurs. A must also have 'seen to it' that X occurs. 'Seeing to it that X' requires, minimally: that A satisfy himself that there is some process (mechanism or activity) at work whereby X will be brought about; that A check from time to time to make sure that that process is still at work, and is performing as expected; and that A take steps as necessary to alter or replace processes that no longer seem likely to bring about X."

A few things are important here. First, the most important criterion in fulfilling one's forward-looking responsibility is *not* that φ (or X in Goodin's terminology) occurs, but rather that agent i (A in Goodin's terminology) *has seen to* it that φ occurs. Second, it is not required that i brings about φ by an action of her own, it is enough that there is a process P that results in φ and that i has certain abilities with respect to that process P (monitoring it, intervening in it or switching to a process P*). This gives i some discretionary room in deciding how φ is to attain.[13] It is for this reason that it seems proper to conceive of the obligation to see to it as a responsibility rather than as a duty, as duties typically refer to (specific) actions that an agent should do or refrain from (van de Poel 2011).

One consequence of the above is that it seems possible that i has fulfilled her obligation to see to it that φ without φ actually attaining. This also seems in line with intuitions about when it is appropriate to attribute forward-looking responsibility. Consider the following example: it seems appropriate to ascribe the responsibility to see to it that passengers are *safely* transported from A to B to a public transport company, or its director(s). Now, this responsibility, among others, implies that we expect the company director to see to it that qualified drivers are hired, that they are instructed to drive safely, that the company buys safe vehicles, that these vehicles are inspected and maintained regularly, and so forth. In other words, we expect the director to see to it that certain processes are in place that, at least in normal circumstances, would guarantee the safety of the passengers. We typically do not expect, however, the company director to be able to prevent all possible accidents, as there can still be cases like, for example, a storm or a terrorist attack that the company director cannot prevent. We accept, thus, that there are scenarios in which the passengers turn out not to be safe, despite the fact that the director has fully discharged her forward-looking responsibility. And the fact that these cases are beyond the company director's control does not invalidate the ascription of forward-looking responsibility beforehand; it is still perfectly appropriate to say that the company director has a forward-looking responsibility for the safety of the passengers of the company when traveling in the companies' vehicles.

This suggests that in order to appropriately attribute forward-looking responsibility for φ to an agent i, i need *not* be able to ensure φ *in all circumstances*. Rather, we would require that i must be able to ensure φ *under normal circumstances* (van de Poel, Royakkers, and Zwart 2015). We propose the following set of conditions as a first approximation to express this:

1. Agent i knows at least one feasible process P that results in φ
2. i can undertake a set of supervisory activities that allow i to monitor P and to intervene in P (if necessary) so that i can ensure that P occurs and results in φ under normal circumstances

It should be noted that these conditions do not require that there is an alternative process P* that achieves φ and to which agent i can switch if P gets blocked. This is not required as a minimal condition for appropriately attributing forward-looking responsibility. This possibility has emerged since we no longer require that i can ensure φ *in all circumstances*.

The proposed set of conditions, thus, does not require i to have regulative control over φ, but only some form of guidance control. Similarly, to the case of blame-responsibility for consequence-universals, this guidance control has an internal and external component. The internal component is that i should have guidance control over the mentioned set of supervisory actions; this means that these supervisory actions should result from a reason-responsive process that is the agent's own. The external component is that the occurrence of φ should be action-responsive to the exercising of these supervisory actions. In this case, this action-responsiveness is cashed out in terms of the outcome φ being responsive to the monitoring of, and potential intervention in a process P by i. Although this is a somewhat different condition for action-responsiveness than in the case of blame-responsibility, it still is an action-responsiveness condition. Whereas in the case of blame-responsibility, action-responsiveness would minimally require that there is a set of (perhaps counterfactual) circumstances (i.e., in a possible world) in which i can prevent the consequence-universal φ from occurring, in the case of forward-looking responsibility it requires minimally that there is a set of (perhaps counterfactual) circumstances in which i can make φ occur.[14]

2.5 Taking Responsibility

Now that the control condition for forward-looking responsibility has been clarified, we will look at cases in which agents actively take or assume responsibility, rather than being held or ascribed a responsibility by others.

We take the following to be the basic form of any responsibility ascription:

Agent j attributes to agent i the responsibility for φ

Using this scheme, we can understand taking responsibility as a special case of responsibility ascription, namely as the case in which j = i.

However, the conditions under which agents can meaningfully take responsibility for φ are somewhat different, and less strict it would seem, than the conditions under which responsibility can be attributed by other agents.

Another way of phrasing this might be to say that an agent i can attribute responsibility to herself from two different perspectives. The first perspective is the third-person perspective in which i asks what responsibility can reasonably be attributed to her (by others, but perhaps also from a general, moral point of view), and a first-person perspective, from which she asks the question: "what do I feel responsible for?" or "what do I aspire or want to take responsibility for?" While the third-person perspective may set limits on what she should take responsibility for, the first-person perspective creates room for taking more responsibility than what one is strictly required to do.[15]

This seems particularly the case for forward-looking responsibility and control, on which we will focus here. We suggest that we can reasonably take forward-looking responsibility for things not yet under our control, but over which we can reasonably expect to gain (some)[16] control, if we seriously try. This possibility can be illuminated by briefly comparing backward-looking responsibility (and in particular blameworthiness) and forward-looking responsibility again and emphasize the differences to which we alluded earlier. The difference is this: when we ascribe backward-looking blame-responsibility (either to ourselves or to others), we do it from the viewpoint that φ has already occurred. In other words, we do it from the viewpoint that we can no longer execute control over φ (as we cannot change the past). However, this is different in the case of forward-looking responsibility, which we ascribe from the viewpoint that φ has not yet occurred. Things that haven't occurred yet do not automatically fall within the range of anyone's control (e.g., volcanic eruptions). While it may be inappropriate for others to ascribe forward-looking responsibility for things currently beyond our control, it seems that we can reasonably assume such responsibility provided that it is reasonable to assume that we can acquire the required control at some not-too-distant point in the future.[17] This ascription merely presumes that one is in control of being able to obtain the required control over φ in a not too-distant future.

The following example illustrates this idea: suppose someone is worried about traffic safety in her neighborhood. She is aware of a number of possible measures that can improve the situation, like the placement of traffic

lights, the lowering of speed limits, speed bumps, or other road reconstruction measures. However, she lacks the control over the introduction of such measures that are required to see to it that the traffic situation is reasonably safe in the neighborhood. In this situation, it would clearly seem unreasonable to ascribe a forward-looking responsibility to her to see to it that the traffic situation is reasonably safe in her neighborhood. This responsibility, so it seems, should be attributed to the relevant civil servants or perhaps to the city council. Nevertheless, it is conceivable that they fail to act and that she is so (morally) upset by the situation that she decides to take responsibility for seeing to it that the traffic situation becomes reasonably safe. Despite initially lacking the control to exercise that responsibility, she may look for ways to acquire such control, e.g., placing warning signals, organizing the neighborhood, or running to be elected into the city council.

As this example suggests, taking responsibility may be considered rational and reasonable under a set of conditions like the following:

- i reasonably believes that she has, or can acquire knowledge of, at least one feasible process P (mechanism, causal pathway) resulting in φ
- There is a set of supervisory actions A through which i can monitor and intervene in P so that
  - The occurrence of φ (through P) is action-responsive to A
  - i reasonably believes that she has or can acquire guidance control over A

While the conditions are somewhat similar to the case in which forward-looking responsibility is ascribed from a third-person perspective, there are two important differences. The first and most important difference is that in taking responsibility the agent does not already need to have the required control but only needs to reasonably believe that she can acquire the required control. Secondly, and related to this, it seems that in the case of ascribing responsibility to others, we typically attune the responsibility that we can reasonably ascribe to an agent i to the control i already has, while in the case of taking responsibility, the responsibility seems to come first, and we then attune the required control in order to be able to fulfill that responsibility.

2.6 When Should People Take Forward-Looking Responsibility for Things beyond Their Control?

Taking forward-looking responsibility for something that is beyond one's control may be seen as a voluntary commitment. This suggests that taking such responsibilities is, at least usually, not morally required. Moreover,

in many cases, it would seem morally praiseworthy to take on new responsibilities. This suggests that assuming such responsibilities is morally supererogatory (van de Poel and Sand 2021). This, however, needs to be qualified: situations are conceivable in which it is morally undesirable to take on new responsibilities, as well as situations in which it may be morally required to take on new responsibilities.

In so far as taking responsibility equals a voluntary commitment, its moral status is somewhat similar to that of promising. Promising is in itself not morally good or bad; it very much depends on what is promised. For example, one must not promise to do morally bad things (e.g., to kill somebody for money), nor should one make promises that cannot be kept. But even if certain promises are neither immoral nor unfeasible, there may be reasons why it is (morally) undesirable to make them.

One concern is that promises introduce new obligations, the fulfillment of which may conflict with the fulfillment of other (moral) obligations the agent already has. So even if the new obligations can be fulfilled, the fact that their fulfillment comes at the expense of fulfilling other moral obligations may, at least in some cases, be a reason why one should not make the promises in the first place.

Something similar applies to taking forward-looking responsibility for things beyond our control. Assuming such responsibilities introduces a range of new (moral) obligations for the agents, not just the obligation to see to it that φ, but also an obligation to increase one's span of control so that one can see to it that φ. Acquiring such control may, depending on the case, require quite some efforts on behalf of the agent and therefore conflict with other obligations.

Moreover, increased control – as a result of taking responsibility – may itself introduce new responsibilities, even beyond the responsibilities that were initially taken by the agent. Take the earlier example of the local resident who takes responsibility for traffic safety in her neighborhood. Assume she decides to try to get elected in the city council, and she succeeds; this obviously leads to many new responsibilities beyond the responsibility for traffic safety in her neighborhood, for which she took responsibility.

The more general point is that responsibility and control may mutually reinforce each other. An example of global scale is the attempt to develop geoengineering as a way to mitigate climate change. While such attempts have been criticized in the philosophical literature as a technological fix that undermines the motivation to solve the "real" problem (i.e., too high emission levels) (cf. Gardiner 2010), it may also be interpreted in a more positive light as an attempt to increase humanity's control so that we can collectively better take forward-looking responsibility for mitigating climate change. The worry that this reply to climate change nevertheless raises is that by trying to increase control over the

climate, we may well introduce *new uncontrollable* risks. Although it is conceivable that these can eventually also be brought under human control, one might not only worry that this is an endless process but also that somewhere along the road, new risks are introduced that are (clearly) unacceptable.

Another worry that may be raised by the potentially mutually reinforcing dynamics of responsibility and control is that there are things one should accept to be beyond control. We are not thinking here about collective or global issues such as climate change, poverty, and environmental degradation. Rather, on the individual level, there are some things which one should not aspire to control (at least directly) and, hence, should not take responsibility for. There are, for example, limits to the extent to which one not only can, but also should, take responsibility for one's own happiness.[18]

The above should not be interpreted as a plea against taking responsibility. As the examples in the introduction show, there are many situations in which taking responsibility is morally praiseworthy. Nevertheless, there are also situations in which it is not praiseworthy and perhaps even morally undesirable to take on certain new responsibilities.

On the other side of the spectrum, one may wonder whether there are situations in which it is morally obligatory to take on new responsibilities. We suggested earlier that one should at least assume responsibilities that others can reasonably attribute to us. So, even if others do not actually, overtly attribute such responsibilities, we should probably assume them ourselves. Moral responsibility does not need a spokesperson.

However, we have also suggested that such attributable responsibilities are typically limited to what is currently within our control. Still, one might wonder whether others can also not sometimes reasonably or appropriately attribute responsibilities to us for things beyond our control. Alfano and Robichaud (2018) briefly mention an example in which someone (a diplomat or politician) is tasked with the responsibility to solve the Middle-East conflict. Such a political position requires the agent to acquire control in a sense that she usually doesn't have at the moment when she is accepting the task. As suggested by the example, it seems true that others can attribute (or delegate) forward-looking responsibilities to us for things that are beyond our control, but it would also seem that such attributions are only *appropriate* attributions of *moral* forward-looking responsibility, if they are *voluntarily accepted* by the responsible agent. That is to say, the attribution may sometimes be inappropriate because the agent to whom responsibility is attributed (or delegated) lacks the capability to exercise the responsibility (as is obviously the case with regard to Jared Kushner as Alfano and Robichaud [2018] correctly point out). But even when the attribution is not inappropriate, it would only seem to be an attribution

of (political, legal) *task responsibility*, not of moral responsibility. It only becomes a moral forward-looking responsibility, if and once the agent to whom this responsibility is attributed voluntarily accepts the responsibility or at least accepts the task that accompanies the responsibility. This type of cases differs from the prototypical case of an agent voluntarily taking forward-looking responsibility for things beyond her control that we discussed before; it is in any cases crucial that the agent *voluntarily accepts* the forward-looking responsibility attributed to her by others for things beyond her control.

Still, there may also be situations in which it is not just praiseworthy but even morally obligatory to take on new responsibilities. Three types of considerations seem to be relevant here (cf. Miller 2001). First, the seriousness and urgency of a certain moral situation. The more serious or urgent the situation, the greater the moral demand for someone to take responsibility for it. Second, the degree to which an agent has or can acquire unique capabilities to address the problem.[19] A third consideration seems to be the agent's current connection with the problem ("connection" is here understood broadly). There may, for example, be cases in which one is (partially) morally blameworthy or morally liable for the problem, which may introduce an obligation to take responsibility for it. While a causal connection alone (without blame or liability) is probably not enough to introduce an obligation to take a responsibility, it may be a factor among the other mentioned considerations. Yet, another way that one may be connected to the problem is that it is in one's realm of authority (e.g., as politician) without necessarily already possessing the required control to solve it.

2.7 Moral Agency

We have suggested that responsibility and control have a reciprocal relation. While in many cases control precedes responsibility and it may be unfair or inappropriate to hold someone responsible for actions or consequences beyond that person's control, in other cases taking (forward-looking) responsibility may precede control and may motivate expanding one's scope of control. Still, there seems to be an important way in which these two types of situations are similar despite their apparent difference. We suggest that in both cases, the relation between responsibility and control suggests a particular notion of moral agency.

As Fischer and Ravizza (1998) point out, attributions of moral responsibility to an agent are historically preceded by that agent having taken responsibility for her actions in a more general sense. With taking responsibility, they do not mean that an agent takes a specific responsibility, as we have used the phrase above. Instead, they mean that humans at some

point in their upbringing begin to see their actions as their *own*. At some point in their upbringing, humans take or accept moral authorship for their actions, and the consequences of these actions. This acceptance of moral authorship by an agent is in their view a (historical) precondition for guidance control.[20]

By accepting moral authorship for one's action and their consequences, one typically also starts to conceive of oneself as a proper target of praise and blame, or sanction and reward. In other words, one starts to think of oneself as a being that can properly be held responsible by others, or oneself. A third aspect of moral agency (in addition to accepting moral authorship for one's actions, and conceiving of oneself as a proper target of reactive attitudes) is to start seeing oneself, and being recognized by others, as part of a larger moral community, a community that to some extent shares certain moral norms and values, where it is considered appropriate to hold another accountable for living by these moral norms and values (cf. Kutz 2000).

While becoming a full-blown moral agent may, as Fischer and Ravizza (1998) suggest, historically precede the attribution of specific moral responsibilities, we would like to suggest that the scope of our moral agency, and hence the scope of our moral responsibility, is not given but may change over time. And it may do so in two ways, namely (1) by extending (or reducing) our span of control in the world, we increase (or decrease) the scope of our moral agency in the world and hence the scope of our moral responsibilities, and (2) by (voluntarily) taking on new (forward-looking) responsibilities, we extend our moral agency, and to effectuate that extended moral agency, we may need to increase our scope of control.

From our point of view, the traditional discussion about responsibility has focused only on the first route. It was assumed that control is a precondition for responsibility and that the only way in which our moral agency and responsibility can increase is through a preceding increase in control. However, there is also a second possibility, where we start with (voluntarily) extending our moral agency and hence our responsibility, and as a result of such (voluntary) commitment need to try to extend our scope of control. The existence of such a route is indeed suggested by the fact – laid bare by Fischer and Ravizza (1998) – that all responsibility attributions are grounded in an agent having taken responsibility in a more fundamental and basic sense.

2.8 Conclusion

Traditionally, control is seen as a precondition for responsibility. We have sketched an alternative view. On this view, there is still a strong (conceptual) connection between control and responsibility, but control does

not always precede responsibility. Rather, the relation may be reversed. Responsibility might sometimes precede control. The main reason is that we can reasonably take responsibility also for things that are not yet under our control.

Taking responsibility is not only important as a way to acquire specific forward-looking responsibilities, including for things not yet under our control. It is also a more fundamental phenomenon that precedes any appropriate responsibility attribution in a more fundamental sense as Fischer and Ravizza (1998) already suggested. In order for certain actions to be the agent's own and to be under her control, she first needs to accept moral authorship or agency over her actions.

On the picture that arises, moral agency is not something given but something that has been acquired and assumed (typically during upbringing). Moreover, moral agency comes in degrees, and human agents can assume less or more moral agency, with more moral agency not necessarily being better because – as we have seen – taking on new responsibilities is not always desirable or morally permissible.

The sketched view has a number of implications regarding responsibility for the risks of new technologies. It suggests that we can sometimes take responsibility for technological (or other) risks that are still beyond our control. At the same time, it suggests that taking such responsibilities will typically also require the agent to increase her span of control, and that may not necessarily always be good or desirable. Hence, there is a limit to the extent that agents not only can but also should take on new responsibilities.

Acknowledgments

This publication is part of the project ValueChange that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321. This publication also contributes to the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

Notes

1 Thomas Nagel considers the control principle not as a philosophical artefact, but as being deeply rooted in common sense morality: "Prior to reflection it is intuitively plausible that people cannot be morally assessed for what is not their fault, or for what is due to factors beyond their control" (Nagel 1979: 25).

2 There might be all kinds of reasons, including pragmatic ones, why it may not be obligatory or desirable to blame an agent that is blameworthy. We take the attribution of blame-responsibility, thus, to be an attribution of blameworthiness, not an attribution of blame.
3 For example, some form of legal blame (and penalty) may be appropriate also in cases an agent is not morally blameworthy in a responsibility-sense.
4 Not all authors consider control explicitly as a responsibility-condition, but as far as we see most, if not all, assume it implicitly in one way or the other. As Sand (2020) points out, those who reject the control principle (e.g., Hanna 2014) have to develop a theory of blameworthiness that explains why blaming people for random harms or the wrongs of other people is unacceptable, something to which CP has a clear answer.
5 Remarkably few defenses of CP have been developed in the philosophical literature. One of the authors of the present paper defended CP in another publication with an appeal to simplicity (Sand 2020).
6 One might debate whether they are all three (equally) responsible for the "consequence-universal" (that the lake is poisoned) or perhaps for something else (like contributing to the poisoning). It is, in any case, clearly wrong to say that they are not responsible.
7 Björnsson (2011) himself proposes another solution that focuses on whether the actions might *explain* the outcome.
8 This is our proposal, not that of Björnsson (2011) or Fischer and Ravizza (1998), although it is intended to be in line with Fischer's and Ravizza's proposal.
9 A main worry about the weak criterion seems to be that the actual process by which φ is achieved is irrelevant to it, while it intuitively would seem to matter what the actual process was that led to φ. Another worry might be that the criterion cannot distinguish between (relatively) more substantial contributions (like in *The Lake*) and small contributions. Consider, for example, the case of climate change. Here, also a certain threshold of individual contributions needs to be passed in order for the collective (undesirable) effect to occur (although this partly depends on how one exactly understands the relevant physical mechanisms). But, contrary to *The Lake*, much more than two individual contributions are required for the collective effect to occur. If we apply the weak action-responsiveness criterion, climate change – maybe somewhat surprisingly – seems to be under individual control, as for each individual there is at least one scenario in which the contribution of that individual is decisive for whether the threshold is passed or not (depending on how exactly the physical mechanism at play are understood). While in the case of climate change, there may be some individual blameworthiness, it would seem excessive to say that each individual is blameworthy for the total effect (as we are inclined to do in the case of *The Lake*). Perhaps, this needs to be explained by the fact that the cases are different in terms of other responsibility conditions, like wrong-doing.
10 This is not meant to suggest that action-responsiveness exhausts the control condition. Perhaps more is required for control than action-responsiveness (and reason-responsiveness as earlier discussed), like knowledge of the consequences or at least the ability to know the consequences, or – alternatively – one might understand 'knowledge' as an additional condition for proper attribution of moral responsibility, in addition to control.
11 It is not meaningful to talk about forward-looking responsibility for actions, at least for the agent's own actions. Those are better called duties or obligations.

12 This can be seen as follows. Suppose that there is some process P that leads to φ under normal circumstances. Now also suppose that i has guidance control over P. Now, by having guidance control over P, i also has guidance control over φ (because P would under normal circumstances result in φ). However, such guidance control is not enough to have the capacity to ensure φ because due to external events or actions (i.e., what Fischer and Ravizza call triggering events) something may happen that blocks P or the path from P to φ. Now in order to ensure φ, i should be able to switch to another process P* that also results in φ. This means that agent i should have guidance control over both process P and P* (and perhaps in real-world scenarios over even more processes). Such dual guidance control is effectively a form of regulative control as Fischer and Ravizza (1998) point out.
13 Interestingly, φ doesn't even have to be brought about (by anyone). It could be a state that is the result of a natural process and forward-looking responsibility ought to ensure that no one is interfering with it.
14 With 'minimally' we do not mean that these are the conditions under which we can appropriately attribute backward-looking or forward-looking responsibility, but rather that any further specification of the action-responsiveness condition (for appropriate responsibility contributions) should at least be as strong as this minimal criterion. There might in fact be other reasons why ascriptions of forward-looking responsibility are inappropriate. For example, Alfano and Robichaud (2018) suggest that ascriptions of forward-looking responsibility are inappropriate if the standing of the attributer doesn't permit the attribution (e.g., due to lack of authority) or if it overburdens the agent. Overburdening might mean that fulfilling the forward-looking responsibility requires too big of a sacrifice (cf. Fischer and Tognazzini 2011).
15 Since both perspectives are eventually hers, there can be a misalignment between what she believes her moral obligations to be and what her moral obligations really are. At the same time, she might mistakenly judge her aspirations to be beyond what she is morally obliged to do, while both coincide. In some sense, she can then count herself lucky for doing the right thing (though most likely for the wrong reasons).
16 How much control is required strongly depends on the context of action and the exact forward-looking responsibility assumed.
17 A related, yet distinct, idea is that it is sometimes desirable that we take – or at least try to take – responsibility for things that might remain beyond our control. A weaker version of this view is clearly defensible. Whether we get climate change "under control" is currently not predictable, but our chances certainly increase if people give a wholehearted try (an effort that oftentimes motivates others to join). The stronger version is less defensible: in medical situations, when there is literally no way of saving a patient, it is unreasonable to continue with the effort. We thank Adriana Placani for making us aware of this and Sven Ove Hansson for suggesting the formulation "responsibility to try".
18 As Aristotle already suggested happiness may well be a by-product (i.e., something that is attained in aiming for other things rather than something that can be aimed at or deliberately achieved).
19 If a choice between agents can be made, it is most reasonable to choose someone who has the relevant control to handle the situation (a doctor to help an accident survivor rather than asking someone, figuring out how to handle a patient).
20 Remember that guidance control requires that actions originate from a reason-responsive mechanism that is (recognized as) the agent's *own*.

References

Alfano, Mark, and Philip Robichaud. 2018. "Nudges and Other Moral Technologies in the Context of Power: Assigning and Accepting Responsibility." In *The Palgrave Handbook of Philosophy and Public Policy*, edited by David Boonin, 235–48. London: Palgrave MacMillan.

Björnsson, Gunnar. 2011. "Joint Responsibility Without Individual Control: Applying the Explanation Hypothesis." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Jeroen van den Hoven, Ibo van de Poel and Nicole Vincent, 181–200. Dordrecht: Springer.

Björnsson, Gunnar. 2021. "On Individual and Shared Obligations: In Defense of the Activist's Perspective." In *Philosophy and Climate Change*, edited by Mark Budolfson, Tristram McPherson and David Plunkett, 252–80. Oxford: Oxford University Press.

Bovens, Mark. 1998. *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Brown, Alexander. 2011. "Moral Responsibility and Jointly Determined Consequences." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole A. Vincent, Ibo van de Poel and Jeroen van den Hoven, 161–79. Dordrecht: Springer Netherlands.

Cane, Peter. 2002. *Responsibility in Law and Morality*. Oxford: Hart Publishing.

Di Nucci, Ezio. 2021. *The Control Paradox: From AI to Populism*. Lanham: Rowman & Littlefield.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*, *Cambridge Studies in Philosophy and Law*. Cambridge: Cambridge University Press.

Fischer, John Martin, and Neal A. Tognazzini. 2011. "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82 (2): 381–417.

Gardiner, Stephen M. 2010. "Is "Arming the Future" With Geoengineering Really the Lesser Evil?: Some Doubts About the Ethics of Intentionally Manipulating the Climate System." In *Climate Ethics: Essential Readings*, edited by Stephen M. Gardiner, Simon Caney, Dale Jamieson and Henry Shue, 284–312. Oxford: Oxford University Press.

Goodin, Robert E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Hanna, Nathan. 2014. "Moral Luck Defended." *Noûs* 48 (4):683–98. https://doi.org/10.1111/j.1468-0068.2012.00869.x

Hart, Herbert L. A. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press.

Honoré, Tony. 1999. *Responsibility and Fault*. Oxford: Hart.

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

Miller, David. 2001. "Distributing Responsibilities." *The Journal of Political Philosophy* 9 (4): 453–71.

Nagel, Thomas. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.

Nelkin, Dana K. 2013. "Moral Luck." In *The Stanford Encyclopedia of Philosophy (Winter 2013 ed.)*, edited by Edward N. Zalta. Stanford: Stanford University, Metaphysics Research Lab. https://plato.stanford.edu/entries/moral-luck/

van de Poel, Ibo. 2011. "The Relation between Forward-Looking and Backward-Looking Responsibility." In *Moral Responsibility. Beyond Free Will and Determinism*, edited by Nicole Vincent, Ibo van de Poel and Jeroen Van den Hoven, 37–52. Dordrecht: Springer.

van de Poel, Ibo. 2017. "Society as a Laboratory to Experiment with New Technologies." In *Embedding New Technologies into Society: A Regulatory, Ethical and Societal Perspective*, edited by Diana M. Bowman, Elen Stokes and Arie Rip, 61–87. Singapore: Pan Stanford Publishing.

van de Poel, Ibo, Jessica Nihlén Fahlquist, Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2012. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18 (1):49–68. https://doi.org/10.1007/s11948-011-9276-0

van de Poel, Ibo, Lamber Royakkers, and Sjoerd D. Zwart. 2015. *Moral Responsibility and the Problem of Many Hands*. London: Routledge.

van de Poel, Ibo, and Martin Sand. 2021. "Varieties of Responsibility: Two Problems of Responsible Innovation." *Synthese* 198: 4769–87. https://doi.org/10.1007/s11229-018-01951-7

Sand, Martin, and Michael Klenk. 2021. "Moral Luck and Unfair Blame." *The Journal of Value Inquiry*. https://doi.org/10.1007/s10790-021-09856-4

Sand, Martin. 2020. "A Defence of the Control Principle." *Philosophia*. https://doi.org/10.1007/s11406-020-00242-1

Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.