Synthese (2018) 195:2001–2019
https://doi.org/10.1007/s11229-017-1312-4

# Guidance control and the anti-akrasia chip

**Chris Ovenden**[1]

Received: 24 February 2016 / Accepted: 4 January 2017 / Published online: 18 January 2017


**Abstract** According to Fischer and Ravizza, an agent has guidance control over some action $A$, whenever $A$ is issued from one of their own moderately reasons-responsive mechanisms. This involves two elements: (i) the process $P$ leading to their action being suitably responsive to reasons-(moderately reasons-responsive); and (ii) their taking an attitude towards processes of kind $P$ such that they see themselves as the agents of the behaviour those processes issue (what they call 'taking responsibility' for a mechanism). For the purposes of this paper, I assume that guidance control amounts to *actually guiding some action*. I present, and defend, a counterexample in which an agent intentionally acts via a suitably reasons-responsive process which they have taken responsibility for and yet, intuitively, does not actually guide their action. On this basis, I argue that taking responsibility for a moderately reasons-responsive mechanism is not sufficient for having guidance control.

**Keywords** Control · Guidance control · Semi-compatibilism · Moral responsibility · John Martin Fischer · Akrasia

## 1 Guidance control and taking responsibility

What is the relationship between control and responsibility?

According to Fischer and Ravizza, we can distinguish between two kinds of control: *guidance* and *regulative*. Guidance control involves freely performing some action $A$, whereas regulative control involves having the dual power to perform one action $A$

✉ Chris Ovenden
Chris.ovenden@manchester.ac.uk

[1] Department of Philosophy, University of Manchester, Manchester M13 9PL, UK

or some other action *B* instead (1998, p. 31). They provide the following example to illustrate this distinction:

> *Driving instructor*
> Sally is driving with her driving instructor in a dual-control car. She comes to a turn that she needs to make and, at the appropriate time, slows the car, signals, turns the steering wheel, and carefully guides the car around the corner. We can image that her driving instructor is quite happy to allow Sally to guide the car on her own, but that if she had shown any inclination to cause the car to go in some other direction he would have intervened and caused the car to go around the corner himself (just as it actually does).[1]

They go on: 'insofar as Sally actually guides the car in a certain way… she has 'guidance control.' [over the car]' (1998, p. 31). Given that she cannot make the car do anything other than it actually does (due to the presence of her driving instructor), Sally lacks regulative control over the car (1998, p. 31).

On this basis, and for the purposes of this paper, let's assume that guidance control amounts *actually guiding* some action through its performance (and doing so 'under one's own steam'). Regulative control, on the other hand, amounts to having the power to determine whether or not *A* is performed.

For Fischer and Ravizza, guidance control (rather than regulative control) is the freedom relevant condition for having moral responsibility. They base this intuition on a familiar kind of case from Harry Frankfurt[2]:

> *Jones and Black*
> Jones intends to kill the mayor and Black, an evil scientist, approves of his doing so. To ensure that Jones carries out his plan, Black implants a device in Jones's brain that, should he show even the slightest inclination of not following through on his intention, will force Jones to act on his original intention to kill the mayor. As it happens, Jones kills the mayor of his own accord (without the intervention of Black or the device).[3]

Intuitively, Jones is morally responsible for killing the mayor; however, he was unable to bring about any other course of action. The parallel to *Driving Instructor* is clear: Jones lacks regulative control over his actions, in that he could not have done otherwise; however, insofar as he, and no one else, actually guides his killing the mayor, Jones has guidance control over his actions. It is therefore guidance control that is 'the freedom-relevant condition necessary and sufficient for moral responsibility' (1998, p. 241, footnote 2).

According to Fischer and Ravizza, an agent has guidance control over an action *A* whenever *A* issues from one of their own moderately reasons-responsive mechanisms (1998, p. 241). This analysis involves several elements:

[1] This example is paraphrased from two cases presented in Fischer and Ravizza (1998, pp. 30–32).

[2] See Frankfurt (1969).

[3] Fischer and Ravizza (2000) present their own paradigm Frankfurt case, 'Assassin', on pp. 29–30.

First, a *mechanism* can be understood as 'the process that leads to [an] action,' or the 'way [an] action comes about' (1998, p. 38).[4]

Second, a mechanism is *moderately reasons-responsive* (henceforth MRR) to the extent that it can issue in different actions given the presence of, and in accordance with, different reasons for acting; what's more it must respond to the presence of reasons in a regular and intelligible pattern.[5]

Put more formally:

**MRR**: An actual sequence mechanism $K$ is MRR if (holding fixed the operation of a $K$-type mechanism) for a set of possible worlds $W$ with the same physical laws as the actual world and in which there exists sufficient reason to do otherwise, (i) the agent's reason-recognition across the members of $W$ gives rise to an understandable pattern (where some of the reasons recognised are moral); and (ii) in at least one of the possible worlds in $W$ in which the agent recognises a sufficient reason to do otherwise, the agent does otherwise *for that reason*.[6]

Finally, an agent can *make* a mechanism their own by *taking responsibility* for it, which involves their forming an attitude towards the behaviour issuing from that mechanism such that they see themselves as the source of that behaviour and as an apt target for the reactive attitudes that it might generate (1998, p. 241). Call this a *responsibility attitude*:

**Responsibility attitude**: $S$ has a responsibility attitude towards a kind of behaviour $B$ if (i) $S$ sees $B$ as an upshot of his agency in the world; and (ii) S sees himself as a fair target for any reactive attitudes that $B$ might generate in others.

An agent has guidance control, then, whenever their actions issue from a MRR mechanism for which they have taken responsibility.

In this paper, I shall be largely concern the ownership element of guidance control, which has received some discussion already: Mele (2000), for instance, has argued that takingresponsibility is not necessary for moral responsibility,[7] whilst Stump (2002) and Long (2004) have challenged the sufficiency of the guidance control conditions, more generally, by proposing cases in which agent's take responsibility for a suitably reasons-responsive mechanisms and yet are not in control of, or morally responsible for, their behaviour.

[4] It is important to note, however, that the mechanism cannot be the entire process, since the entire process would include both the mechanism itself and the inputs to that mechanism. We need to differentiate these two things so that we can make sense of the same mechanism responding to different inputs in alternate sequences. (Thanks to an anonymous referee for this point).

[5] This amounts to the mechanism's being what Fischer and Ravizza term weakly reasons-reactive and regularly reasons receptive (1998, pp. 71–5).

[6] I rely on this formulation of MRR, put in terms of possible worlds, rather than the one given by Fisher and Ravizza Fisher and Ravizza (1998, pp. 243–244) in order to give a clearer statement of the two counterfactual conditions that a mechanism must meet in order to be MRR.

[7] See Fischer and Ravizza (2000) for their reply Mele (2000) for a response.

My argument in this paper follows a similar line to Stump and Long: specifically, I'll argue that taking responsibility for a MRR mechanism is not sufficient for actually guiding the behaviour it issues, and so not sufficient for having guidance control. My central contention, roughly, is that believing yourself to be in control of some action (in some relevant way) cannot be conceptually prior to your actually being in control: that is, I can rightly or wrongly *take responsibility* for an action based upon whether I was in control of it; but I cannot give myself control by *taking responsibility*.

In Sect. 2 I present a counterexample to the guidance control analysis outlined above—-in which an agent takes responsibility for a MRR mechanism and yet fails to actually guide the behaviour it issues. In Sect. 3 I defend this example from several potential objections, and then in Sect. 4 I briefly look at what my example suggests about how agents come to have ownership of their behaviour along with some wider implications for the link between control and moral responsibility.

## 2 The anti-akrasia chip

Suppose that a robotics company design and manufacture an 'anti-akrasia chip' (henceforth AAC), a neural implant designed to help weak willed individuals act on their judgements about what they ought to do. The chip is surgically inserted into the brain and, once activated, begins monitoring the user's mental states. If the user recognises a sufficient reason to $\varphi$ but does not have the strength of will to act in accordance with that reason, the AAC stimulates their brain in such a way as to cause the formation of an effective intention to $\varphi$.

This involves two elements: (i) if the user has not already formed one, the AAC will implant in them an intention to $\varphi$, and (ii) the AAC will ensure that the user maintains and acts upon their intention to $\varphi$ at the appropriate time. We can imagine that the first element is achieved through direct stimulation to the relevant part of the user's brain, whilst the second element is achieved by whatever means Black ensures that Jones maintains and acts upon his intention to kill the mayor.

John is a particularly weak willed individual: regularly acting contrary to his judgements about what he has most reason to do when those judgements are accompanied by conflicting desires or attitudes.[8] As soon as he hears about the AAC he goes to have it installed, and by the mid-afternoon his actions are in perfect accord with his normative judgements; the majority of these actions issuing from his newly active *AAC mechanism*, which operates as follows:

(i) John recognises that there is a sufficient reason to $\varphi$;
(ii) if John does not form an intention to $\varphi$ by the appropriate time, the AAC causes in John the formation of an intention to $\varphi$;

[8] This is not intended to imply that John is incapable of acting in accordance with his judgements: an agent who is always, or almost always, weak willed might simply fail to be the kind of agent that, intuitively, can have control over and be morally responsible for their actions. On the contrary, John *is* able to act in accordance with his judgements—-his akratic actions are not compulsive—he is simply weak: more often than not he lacks the *oomph* to do what he thinks is best and is too easily seduced by other temptations. Thanks to an anonymous referee for pushing me on this point.

(iii) the AAC ensures that John maintains his intention to $\varphi$ and that it is effective in bringing about his $\varphi$-ing; and
(iv) John intentionally $\varphi$'s at the appropriate time.[9]

This mechanism looks MRR: John's AAC actions are issued as a direct response to his reasons-recognition (which we can assume is operating in a regular manner) and he *would* do otherwise if he recognised a sufficient reason to do so.

Suppose, also, that John comes to form a responsibility attitude for his AAC behaviour, thereby *making it his own*: his peers treat him as responsible for his AAC behaviour and he too, aware of the link between his judgements and resulting behaviour, comes to think of himself as the agent of that behaviour. That being the case, his AAC behaviour would be issued from his own MRR mechanism.

Is John *actually guiding* his AAC behaviour? I think not: Suppose that a week before having the AAC installed John had been unfaithful to his wife. After the chip's activation, John begins to feel guilty and recognises that he has sufficient reason to come clean; nonetheless, he does not *want* to confess, it is not in his character to do so, and in the absence of the AAC he would not even entertain doing so; rather, he would commit himself to keeping this infidelity a secret. However, with his AAC mechanism fully operational, the recognition that there is a sufficient reason to come clean activates the chip, which in turn implants in John an intention to confess the whole affair (which he does—in sordid detail).

John's AAC behaviour can still constitute his acting against his will (or at least, against the will that he wants): John is internally divided in a similar way to Frankfurt's unwilling addict, the difference being that whilst the addict acts on a desire he does not identify with, John is caused to act by a judgement that he does not want to have made. In each case, the agents are moved to action by forces that they do not identify with; therefore, just as the unwilling addict is not properly in control of (and not actually guiding) his behaviour, neither is John properly in control of his AAC behaviour.[10]

If John enjoys any kind of control over his behaviour, it certainly is not the 'actually guiding' kind enjoyed by Sally in *Driving Instructor,* and he therefore lacks guidance control (despite meeting its official conditions).

## 3 Objections and replies

Let's formalize my argument from the previous section as follows:

1. John's AAC mechanism is MRR.
2. John takes responsibility for the behaviour issued by his AAC mechanism.

[9] Assuming a broadly causal theory of action according to which one's actions are intentional just in case they are appropriately caused by one's reasons.

[10] John might still be morally responsible for his AAC behaviour: after all, there might be many conditions that are independently sufficient for being morally responsible, besides having guidance control. Perhaps, in this case, John's recognising that he has a reason to *A*combined with his free decision to have the AAC device implanted in himself jointly make him morally responsible for subsequently *A*-ing (especially to anyone who takes morality to reduce, ultimately, to rationality). However, whether or not John is morally responsible for his AAC behaviour, he does not *actually guide* it in the same way that Sally does her actions in *Driving Instructor*. Thanks to an anonymous referee for pushing me on this point.

3. John does not have guidance control over the behaviour issued by his AAC mechanism.
4. Therefore, taking responsibility for a MRR mechanism is not sufficient for having guidance control over the behaviour issued by that mechanism.

In the following four sections I'll consider a number of objections to each of these premises.

### 3.1 Individuation, inputs and fair opportunities

First, I want to look at some of Fischer's responses to critiques by other authors, in particular his (2004) reply to Stump and (2010) reply to Long, and consider how they might be applied to the AAC. Both of these responses suggest a tension between the manner in which Fischer intends the guidance control account to be applied and its official conditions, and further examination of this tension reveals that there may be extra conditions on having guidance control, not explicitly stated in *Responsibility and Control* (1998).

#### *3.1.1 Response to stump*

Stump's (2002) worry is much the same as my own: that the Fischer–Ravizza account allows for agents to have guidance control over mechanisms that clearly involve a high degree of manipulation so long as those mechanisms are suitably responsive to the *agent's* reasons. She imagines a case (based upon Robert Heinlein's *The Puppet-masters*) in which an agent, Sam, has had his mind taken over by an intelligent alien 'master' as part of a wider scheme to conquer Earth. When the master takes over Sam's body, it takes his consciousness 'off-line,' leaving it to run pretty much as it always does but removing from it the ability to affect his behaviour. It is the master's consciousness alone that determines how Sam acts (2002, pp. 47–48). She goes on:

> Since it is crucial to the alien plan that their taking over human beings be undetected in the early stages of the invasion, they are careful to make the behavior of people like Sam correspond to the behavior Sam would normally have engaged in had he not been infected with the alien. So when, under the control of the alien, Sam does A, it is also true that if there had been reason sufficient for Sam in his uninfected state to do not-A, the alien would have brought it about that Sam in his infected state did not-A. In this case, then, Sam acts on a mechanism that meets Fischer and Ravizza's condition for being strongly reasons responsive. (2002, pp. 47–48)

Stump then imagines that the master alien reveals itself to Sam and convinces him to take responsibility for this mechanism, thereby making it his own (2002, pp. 49–50). From then on, whenever he is being controlled by the alien master, Sam is acting on his own MRR mechanism; clearly, though, Sam is not the one guiding his behaviour.

Fischer's response to this is that, of course, if you individuate the mechanism on which Sam acts as broadly as 'manipulation by an external source,' then it will turn out to be moderately reasons-responsive; but the Fischer–Ravizza account's proposed

way of dealing with manipulation cases is to individuate the relevant mechanism in a far narrower manner: 'manipulation of this specific sort,' for instance (Fischer 2004; pp. 152–153). The relevant mechanism to hold fixed in Sam's case, is not 'manipulation of Sam by the alien master,' but something like 'manipulation of Sam's brain such that he is caused to *A*;' and with that narrower mechanism held fixed Sam will *A* irrespective of any reasons to not-*A* in alternate sequences.

Stump's alien master case shares many features with my AAC case. However, an important difference is that in her case, an agent is manipulated *by the actions of another agent* whereas in the AAC case, the process leading to John's actions does not involve any agents other than John himself. Nonetheless, a similar response might be given: of course, the AAC comes out as MRR if you individuate the mechanism issuing in John's behaviour so broadly as: 'manipulation by the AAC in accordance with John's reasons.' However, when the AAC forces John to intentional *A*, we should assess the reasons-responsiveness of the mechanism issuing in his behaviour by holding fixed something like 'the AAC manipulating John's brain in such a way that he intentionally *A*s.' With this narrower mechanism being held fixed, it is clear that John would not *A* in any alternate sequences in which there was sufficient reason to refrain from *A*-ing: the AAC will manipulate his brain in exactly the way it does in the actual sequence: as though there *is* a sufficient reason to *A*. That being the case, the AAC mechanism is not MRR and John does not guide the behaviour it produces.

It is difficult to know exactly how to respond to this kind of objection because, as Fischer himself acknowledges (2004, pp. 166–167), the Fischer–Ravizza account does not contain an explicit account of mechanism-individuation, so there is no principled reason for holding fixed the narrow rather than the broad manipulation mechanism in the AAC case, this is simply left up to intuition.

The obvious response is to simply restate the above admission: there is no principled reason why we *should not* hold the broad rather than the narrow AAC mechanism fixed: it is in no way clear, even if left to intuition, that the narrower mechanism is the relevant one to assessing John's AAC mechanism; after all, remember that we can describe the AAC case such the device be integrated with John's brain to such a degree that it is functionally identical to the deliberative faculties of a 'normal' strong-willed agent (and still, John would not guide his behaviour). Unfortunately, butting heads does not get us very far.

I think, though, that there is some indication of a principle at work in this objection, and by drawing it out we can see, again, that it is the ownership element of the guidance control analysis that is at issue.

First, consider the following question: if we must hold fixed the narrower mechanism in the AAC case, why ought we not hold fixed an analogously narrow mechanism in cases of normal un-manipulated action? That is, when John's AAC causes him to intentionally *A*, we are told that we ought to hold fixed the narrowly individuated mechanism type:

**n-AAC** 'the AAC manipulates John's brain such that he *A*s'

rather than the more broadly individuated mechanism type:

**b-AAC** 'the AAC manipulates John such that he acts in accordance with his reasons-recognition.'

Whereas, when a normal non-manipulated agent intentionally *A*s we ought *not* hold fixed the narrowly individuated mechanism type:

**n-NORM** 'the agent's deliberative faculties operate such that she As'

but instead hold fixed the broadly individuated mechanism type:

**b-NORM** 'the agent's deliberative faculties operate such that she acts in accordance with her reason-recognition.'

If we *were* to hold fixed **n-NORM** in cases of normal un-manipulated action, because the action issued is part of the description of the mechanism's type, no deliberative mechanism would turn out to be MRR: in any alternate sequence, no matter what the strength of reasons for not *A*-ing, **n-NORM** will issue in the agent *A*-ing.

Why should the two cases be treated differently? Potentially, because the AAC case involves manipulation, whereas the 'normal' case does not. However, to draw this distinction, between a manipulation and non-manipulation case, *before* considering whether an agent has taken responsibility for the mechanism issuing in their behaviour suggests that there is already some notion of ownership at play here!

Whether or not a mechanism is MRR is conceptually independent, and perhaps also conceptually prior, to whether an agent has *taken responsibility* for that mechanism and thereby taken ownership over it. If the AAC case involves manipulation this can only be because the AAC mechanism is not properly owned by John: the mechanism, alien to John's own agency, manipulates his response to reasons so as to produce actions that are perfectly in tune with his reasons-recognition. But given that we are considering whether the AAC mechanism is MRR in isolation of whether John has taken responsibility for it, it must be some other kind of ownership that he lacks over the AAC. Call this *ownership**.

Now, this objection might plausibly be applied in Stump's case because in her example the mechanism by which Sam acts involves the actions of *some other agent;* namely, the alien master. In that respect, there is an intuitive way establishing that this is a case of straightforward manipulation: the mechanism leading to Sam's actions contains embedded within it the operation of a sub-mechanism that is quite clearly owned by another agent

In the AAC case, however, the AAC mechanism contains no embedded sub-mechanisms or actions: the AAC is not an agent, and it does not act, it merely plays a functional role in the processing of John's reasons-recognition into action. If all that mattered for ownership was *taking* responsibility, then we would not be in a position to say whether John's case, or the 'normal' case, involved manipulation at the outset. What's more, once we remove all mention of manipulation from the descriptions of **n-AAC** and **b-AAC**'s operation, both the narrowly individuated **n-AAC** and **n-NORM** and the broadly individuated **b-AAC** and **b-NORM** will conform to the same narrow and broad mechanism types:

**n-MECH** *K* operates such that *S A*s
**b-MECH** *K* operates such that *S* acts in accordance with *R*

where *K* is the kind of mechanism operating and *R* is some reasons-responsive faculty.

To distinguish between the two cases then, without first assuming that the 'normal' agent *owns** their deliberative faculties whilst John does not *own** the AAC mechanism, simply looks ad hoc.

If there is a notion of ownership* at play here, it is not clear why one also need to take responsibility of a mechanism in order to *guide* one's behaviour.[11] Surely, if a mechanism is psychologically one's own and it is suitably responsive to reasons, it simply does not matter whether one thinks of oneself as an agent when that mechanism operates. Without a more worked out explanation of ownership* we are not in a position to say, but there is certainly the threat of making *taking responsibility* redundant by introducing some more fundamental notion of ownership.

### *3.1.2 Response to long*

Moving on to Long's (2004) worry, Long considers a case in which a manipulator feeds inputs into an agent's MRR mechanism right before they come to a decision in order to change the output to their liking.

For instance, Block might want Schmidt to vote for Hitler to be given supreme power over Germany. Sensing that Schmidt is about to vote against Hitler having this power, Block adds new inputs to the very same mechanism that operates in the actual sequence (suppose that these come in the form of reasons for voting in favour of Hitler) so that when the time comes Schmidt's normal deliberative faculties issue in his voting for Hitler.

Schmidt, in this case, appears to act on his own MRR mechanism, and so to have guidance control over his behaviour according to the Fischer–Ravizza account, but, clearly, he is being manipulated by Block (and is not really responsible for his actions) (Long 2004, pp. 157–159).

Fischer's responds to this case as follows:

> The suggestion in question is that the implantation or manipulative induction of "reasons" is done immediately prior to the choice and subsequent behavior, and that there is thus no reasonable or fair opportunity for Schmidt to reflect on or critically evaluate the new input in light of his standing dispositions, values, preferences, and so forth.
> I contend that when "inputs" are implanted in a way that does not allow for a reasonable or fair opportunity for the agent to subject those inputs to critical scrutiny in light of his or her normative orientation, then such manipulation does indeed remove moral responsibility. Such manipulation typically "changes the mechanism." (2004, pp. 180–181)

Schmidt is not morally responsible for his voting for Hitler, because the mechanism on which he acts is not the one that he has taken responsibility for: the addition of extra reasons immediately before his decision without the opportunity to filter these

[11] That is not to say (of course!) that taking responsibility might not be required for *moral responsibility,* only that one can guide one's behaviour without having formed an attitude of ownership towards it.

reasons through his normative orientation changes the kind of mechanism issuing in his behaviour.

Perhaps a similar complaint can be lodged against my AAC case: The AAC produces intentions that are appropriately based upon John's reasons to act, but because it also ensures that he acts upon these intentions it does not provide him with a fair and reasonable chance to filter those intentions through his normative orientation.[12]

Again, it is not clear whether the Fischer–Ravizza account is entitled to this kind of response given its official conditions. This requirement that the inputs to one's mechanisms be filtered through one's normative orientation is presumably part of guidance control's ownership element; but it is not at all obvious how this requirement is supposed to fall out of the conditions for appropriately taking responsibility for a mechanism. This appears to be another extra condition on having guidance control not explicitly stated in *Responsibility and Control* (1998). Adding this condition might help to alleviate the worry raised by the AAC case, but only by suggesting that there is more to owning a mechanism than *taking responsibility* for it, and again raising the question of why we need to *take responsibility* at all!.

In any case, it is not clear that the AAC mechanism should be conceived of in the same way as Long's: the AAC example is set up in such a way that the AAC does not simply add an input to John's normal deliberative mechanisms; rather, it replaces those mechanisms and, once he takes responsibility for it, becomes *his* mechanism for translating reasons into action (according to the Fischer–Ravizza account). The intentions it produces do not need to be filtered through John's normative orientation because, once the AAC mechanism has been made his own, its operation just *is* the filtering of reasons through his normative orientation. Of course, the intention could be re-filtered through John's normative orientation and deliberative faculties, but this would simply involve re-running it through the AAC mechanism.

To sum up, if we conceive of the AAC mechanism as adding inputs into John's deliberative mechanisms, this response may well show that John does not have guidance control over his AAC behaviour; it only does so, however, by adding extra conditions onto the account of appropriately taking responsibility and thereby showing that the official account is insufficient. If, on the other hand, we conceive of the AAC as a mechanism of John's into which inputs are fed, the response does not seem to apply at all.

### 3.2 The AAC mechanism is not MRR

Moving on to some more direct responses to the AAC case, an obvious line of objection would be to deny premise (1), that John's AAC mechanism is genuinely MRR. If it is not then, whether he takes responsibility for it or not, John cannot be exercising guidance control through its operation.

Remember that, in order to be MRR, the AAC mechanism needs to involve both (i) John's reason-recognition across a set of possible worlds in which there exists sufficient reason to do otherwise giving rise to an understandable pattern (where some

[12] Thanks to an anonymous referee for this point.

of the reasons recognised are moral); and (ii) John's doing otherwise in at least one of the possible worlds in which he recognises a reason to do so, and doing so for that reason in the appropriate sense. Importantly, John's actions in the actual and alternate sequences must be *intentional.*

We can suppose that the AAC is linked directly into John's normal reason-recognition faculties—when he recognises a reason to $\varphi$, it electronically stimulates his brain so as to produce in him an effective intention to $\varphi$—so, assuming that John can recognise reasons in a sufficiently understandable pattern, and lets just stipulate that he can, the AAC mechanism will be sufficiently receptive to reasons to meet condition (i).

In that case, plausible objections to (1) will have to show that the AAC mechanism fails to meet condition (ii): that it does not involve John's intentionally acting otherwise for the reasons that he recognises in the alternate sequences.

Such an objection could take one of three forms:

(i) John's behaviour is not properly intentional;
(ii) John's behaviour is causally deviant; or
(iii) John does not act for a reason in the relevant sense.

I'll take these points in turn.

### *3.2.1 John's behaviour is not properly intentional*

First, because it is the AAC that causes John's intention to do otherwise in the alternate sequences, a defender of the Fischer–Ravizza account might claim that his actions are not properly intentional: an agent, we might think, cannot by caused to have a genuine intention by a source external to their own psychology. If John's actions are not intentional, in the actual or alternate sequences, then the AAC mechanism is not MRR.

This particular line of objection does not appear to be open to the Fischer–Ravizza account: as noted in my introductory remarks, Fischer and Ravizza rely on Frankfurt cases such as *Jones and Black* to motivate their focus on guidance rather than regulative control. In doing so, they endorse the possibility of agents being caused to act intentionally by entities external to their own agency: in order for *Jones and Black* to involve Jones's being unable to do otherwise, Black needs to cause Jones to *intentionally* kill the mayor in the alternate sequences; otherwise, we could maintain that Jones *can* do otherwise as his killing the mayor intentionally and killing the mayor unintentionally through coercion or stimulation are different act types (if the latter counts as an action at all). That being the case, and given the stipulation that the AAC ensures that John acts on his intention in the same manner that Black ensures Jones kills the mayor, it looks unprincipled for a defender of the view to argue that the AAC mechanism cannot cause John to intentionally $\varphi$.

### *3.2.2 John's behaviour is causally deviant*

Second, it might be argued that, even though his behaviour is caused by his recognition that he has certain reasons to act, John's AAC behaviour turns out to be causally

deviant: whilst his behaviour *is* caused by his reasons, John's AAC behaviour is not caused by his reasons in *the right kind of way*.

Consider the following case from Donald Davidson:

> *Nervous Climber*
> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold. (Davidson 1973, p. 79)

The climber in this case certainly drops the man *because* he wants to rid himself of the weight and danger associated the man he is holding and knows that he can do so by loosening his grip on the rope; however, his action is not guided by those reasons, or by him, in the right sort of way: his action is brought about by those reasons somewhat accidentally and so, we might think, he does not really act at all when he loosens his hold. If something similar is going on in the AAC case then it looks like the AAC mechanism will not meet the requirements for MRR after all: if John's AAC actions are causally deviant, then they are no actions at all (according to the standard casual theory of action, at least).[13]

As above, however, the Fischer–Ravizza account does not appear to be entitled to this objection: causal deviance occurs whenever the relevant behaviour to some rationale is produced by the wrong kind of mechanism; more precisely, by a mechanism that is not expressive of the agent's own agency or guidance. But, according to Fischer and Ravizza's conditions, the AAC mechanism is of precisely the right kind for producing morally responsible action. What's more John has *made* it his own by taking responsibility for it. If it is still the wrong kind of mechanism to produce non-deviant action, we are owed some explanation of casual deviance that does not simply terminate in an agent's ownership of the mechanism issuing in their behaviour.

From another perspective, the problem of causal deviance just *is* a problem of guidance: an agent's behaviour is caused by and in accordance with their reasons, but they do not genuinely *guide* that behaviour. Now, Fischer and Ravizza do not make explicit whether their account is supposed to apply only to non-deviant action, but given this link between guidance and deviance it would be surprising if it was completely silent on such cases. What's more, if it turns out that a mechanism can only be MRR if it issues in non-deviant (guided) actions, there may be some circularity in the account: a mechanism can only be MRR if it issues in non-deviant actions (those guided by the agent), and an action is only guided by the agent (non-deviant) when it issues from one of their own MRR mechanisms.

### *3.2.3 John does not act for a reason in the relevant sense*

Finally, an objection might be made that whilst John's behaviour is performed for reasons in the sense that it is prompted by his recognition of certain reasons to act, John does not really act *for a reason* in the relevant sense for intentional action.

[13] Thanks to an anonymous referee for this point.

Unfortunately, the sense of *acting for a reason* relevant to intentional action is notoriously difficult to pin down. Fischer and Ravizza seem to endorse Robert Audi's account of *acting for a reason,* according to which it is enough for an agent's having $\varphi$'d for a reason $r$ that he would give $r$ as the reason for his action if he were asked for an explanation (Fischer and Ravizza 1998, p. 64)[14]; and, as it is John's recognition of reasons *qua* reasons for action that activates the AAC mechanism, it certainly seems plausible that John would cite the reasons he recognised as the reasons for his AAC behaviour (knowing, as he does, how the AAC mechanism works).[15]

For an objection of kind (iii) to work, it will need to provide a different account of acting for reasons that rules out John's case, and until such an account has been provided it seems unprincipled to insist that John's AAC behaviour is not performed for a reason.

### 3.3 John has not taken responsibility for the AAC mechanism

The second line of objection would be to reject (2), that John has taken responsibility for the AAC mechanism. If he doesn't have the relevant responsibility attitude then he can't have ownership of the mechanism, and therefore will not have guidance control over the behaviour it issues.

Such an objection might plausibly be based upon a constraint that Fischer and Ravizza place on taking responsibility: that a responsibility attitude must be developed, in an appropriate manner, on the agent's evidence (1998, p. 213).

This constraint, they acknowledge, is 'intended (in part) to imply that an individual who has been electronically stimulated to have the relevant view of himself… has not formed this view of himself in the appropriate way' and therefore lacks guidance control despite meeting all of the other requirements (1998, pp. 235–236). Due to the manner of the AAC's operation, perhaps it will be contended that John has not, and perhaps cannot, develop his responsibility attitude, in an appropriate manner, on the evidence that he has.

Fischer and Ravizza do not offer an analysis of this appropriateness condition, so it is hard to assess why John's taking responsibility might be inappropriate. Presumably, it will because either: (a) the process by which he formed the attitude was inappropriate; (b) the attitude he formed was inappropriate given the evidence he had; or (c) the evidence upon which it was based was inappropriate (for forming a responsibility attitude).

Of the three options, (a) and (b) look the most plausible, especially if we consider the type of case the condition was introduced to prevent. In response to (a), John forms his attitude by coming to see himself as the originator of his AAC behaviour (on the basis that it is his judgements that end up causing his actions) and on the basis of

[14] See Audi (1986).

[15] Interestingly, returning to the previous objection on causal deviance, whilst John plausibly would cite the reasons he recognised as the reasons for his actions, the climber in Davidson's case plausibly *would not*, making this look less like a standard case of causal deviance.

the reactive attitudes of his peers. This seems like an appropriate process given how Fischer and Ravizza describe the usual manner for forming responsibility attitudes[16]:

> As a child grows up, he is subject to moral education (imperfect as it may be). The child's parents—and others—react to the child in ways designed (in part) to get the child to take certain attitudes toward himself… [As a result,] the child typically acquires the view of himself as an *agent*… he sees that upshots in the world depend on his choices and bodily movements. Further, the child comes to believe that he is a fair target for certain responses… as a result of the way in which he exercises his agency. (1998, p. 241)

At the very least, it does not seem from this description that there is anything obviously wrong about the way John forms his responsibility attitude.

In response to (b), his responsibility attitude itself looks appropriate given that it was based upon his peers treating him as though he was morally responsible for his AAC behaviour. However, he may also be aware of his strong desires to act contrary to his better judgement (his desire not to tell his wife about his infidelity, for instance), and we might think that this awareness makes it less plausible that his responsibility attitude is appropriately based upon all the evidence he has: perhaps if one strongly desires to not-$A$, then it is inappropriate to see oneself as responsible for any $A$-ing that one might produce.

This objection cannot be right, though: we quite often act responsibly whilst having strong desires to act otherwise and, what's more, in many of those situations it will be entirely appropriate to hold a responsibility attitude towards ourselves. Think of any case in which an agent overcomes a strong fear: for instance, James is terrified of spiders and yet faces his fear and intentionally picks up a tarantula and allows it to walk across his palm; his strong desire not to pick up the spider ought not prevent him from being able to see himself as responsible for this behaviour; after all, he conquers his fear and deserves praise for doing so. Neither, then, should John's strong desires not to do what he judges best make it inappropriate for him to take responsibility for his AAC mechanism.[17]

This leaves (c), that the evidence itself was inappropriate. This looks like the least plausible reading of Fischer and Ravizza's condition: 'in an appropriate manner' suggests that the process by which the attitude was formed must be appropriate, rather than the evidence that it was based upon. However, perhaps it will be maintained that developing a responsibility attitude on evidence in the appropriate way involves developing it on appropriate evidence.

Potentially, the dissonance that John experiences between what he wills and what he actually does might make his taking responsibility in this case seem inappropriate. Intuitively, his experience of his AAC behaviour ought to tell him that he is not actually guiding that behaviour. John's peers are *wrong* to treat him as if he is responsible for his AAC behaviour and John himself is wrong to base a responsibility attitude on his first-hand experience of AAC behaviour: it is not *genuine* evidence of control.

[16] We could even change the example to stipulate that the AAC was implanted in John's brain at birth so that he develops his responsibility attitude as part of his moral upbringing.

[17] Thanks to an anonymous referee for pushing me on this point.

Unfortunately, this response is not available to the Fischer–Ravizza view: *taking responsibility* is conceptually prior to being in control, so being in control cannot be a condition on appropriately taking responsibility. If John does not have guidance control because he cannot appropriately take responsibility, and cannot appropriately take responsibility because he is not actually guiding his behaviour, then the analysis turns out to be circular.

Unless we can give a non-circular reading of Fischer and Ravizza's appropriateness condition (and no such reading is particularly forthcoming), objections to premise (2) look like non-starters.

### 3.4 John does have guidance control over his AAC behaviour

A certain kind of response, which we might call the *hard-headed* option, would be to deny (3) and assert that, in fact, John *does* actually guide, and thereby have guidance control over, his behaviour. As before, there are a number of lines that this kind of objection could take:

(i) the AAC gives John control by allowing him to overcome his akratic tendencies;
(ii) the AAC acts as a prosthetic, allowing John to guide his behaviour more effectively;
(iii) John would have guidance control by the 'tracing method' (to some extent); or
(iv) guidance control was not supposed to be an analysis of any pre-theoretical notion of control or guiding, anyway.

I take these points in turn.

#### *3.4.1 The AAC gives John control*

First, someone might argue that, because the AAC prevents John from acting akratically it actually helps him to remain in control by allowing him to remain continent and exercise self-control over his behaviour (to bring it in line with his judgements).

I think we can dismiss this kind of initial objection quickly: simply having your behaviour conform to your judgements does not guarantee that you actually guided that behaviour. We can imagine, for instance, that in Stump's case the alien master sometimes causes its host to behave in a way that they actually judge best. That in no way entails that the host guides that behaviour, though: in fact, the case makes it clear that they do not!

#### *3.4.2 The AAC is a prosthetic device*

Second, we might think that, due to the way that it integrates with John's brain, the AAC should not be thought of as an external manipulator but, rather, as a prosthetic device for helping John to connect his reasons with the appropriate actions. After all, the chip serves as a conduit between John's judgements and his resulting intentions and actions, and once it is installed and fully integrated with his brain John's neural psychology will be functionally identical to any 'normal' (non-manipulated) strong-willed agent:

he will recognise reasons and translate those into the relevant intentional actions. So, if normal strong-willed agents are capable of actually guiding their behaviour, then John ought to be capable also.[18]

This is an extremely tricky area and there certainly is a case to be made for the AAC being a kind of prosthetic; however, if we are careful to distinguish between (i) *prosthetics that enable an agent to exercise guidance* and (ii) *prosthetics that guide on an agent's behalf*, I think it will become obvious that the AAC really does prevent John from guiding his behaviour.

Consider a variant of the ACC called 'the Deliberator.' Like the AAC, the Deliberator is a neural implant that monitors the reasons that an agent recognises, along with their relative strengths, and, when the time for a decision arrives, feeds those inputs into an algorithm to determine what the agent has most reason to do. Unlike the AAC, however, the Deliberator only produces a judgement about what the agent ought to do, which the agent is then free to act upon or ignore. We can further imagine that the Deliberator takes into account all of the agent's values and characteristics when coming to a decision so that it produces exactly the judgement that they would have come to, were their deliberative faculties functioning at their very best, but in one-hundredth of the time.

As a prosthetic device, the Deliberator allows for an agent to more effectively exercise guidance over their actions: it speeds up the slow organic deliberation of normal agents by replacing it with quick and accurate electronic calculations, and it allows agents who are unable to effectively deliberate on their own (perhaps due to some neurological disorder) to function as a psychologically standard agent. Importantly, the Deliberator does *not* cut the agent out of the picture altogether: the agent receives a judgement that they can then choose to act upon or veto if it is not to their liking.

Contrast this to the AAC case: The AAC not only deliberates for John, it also forms an intention to act and ensures he sticks to it, so he has no chance to veto any of the decisions it makes on his behalf. The AAC insulates John's deliberation and intention formation processes from his conscious will, cutting him out of the picture altogether. It is still his reasons-recognition that starts the process off, but from that point on he is a bystander, and there is a good deal more to guiding one's actions than having them conform to one's reasons in a regular pattern: one needs *get involved* in this process and mediate between the reasons that one has for acting, adding one's own motivational force to whichever motive one favours (where this is not necessarily the motive that one identifies as being best justified).

The AAC, I submit, is a prosthetic that guides *on the agent's behalf* and therefore one that removes, rather than enables or enhances, an agent's ability to guide her behaviour. It does not matter how integrated the prosthetic becomes, if it subverts the agent's will in the process leading to their actions, it can only remove, rather than enable, their ability to actually guide their behaviour.

Suppose that someone says: sure, the AAC subverts John's conscious willing of what he thinks he wants to do, but, by forcing him to intentionally act on his judgements, it brings his actions into line with his most deeply held values (the same values that

[18] Thanks to an anonymous referee for this point.

caused him to judge that so acting was the right, or best, thing to do). By doing so, the AAC grants John a morally significant kind of control and guidance over his behaviour.[19]

My response to this is two-part:

First, I am happy to concede that John may be morally responsible for his AAC behaviour (for the reasons noted in footnote 10, above) but that does not entail that he is *actually guiding* it. Whilst guidance control might be sufficient for responsibility in some circumstances, it need not be necessary (and this case might well demonstrate that point).

Second, if the AAC does allow John to guide or control his behaviour in some sense, it is not the sense Fischer and Ravizza seem to intend by 'guidance control'. Even if we concede that John's behaviour is in some way guided by his deeply held values, it remains the case that he does not *actually guide* his behaviour in the way that Sally does in *Driving Instructor.*

### *3.4.3 John has control via the 'tracing method'*

Third, we might think that the AAC behaviour John exhibits is a good candidate for being treated to what Fischer and Ravizza refer to as the 'tracing method.'

Roughly, whilst a particular action may not issue from a MRR mechanism, Fischer and Ravizza claim that we can nonetheless look for guidance control 'at various places *along the way to the action…* More specifically, we can look for [MRR] in the *formation* of the relevant trait, its *retention,* or its expression' (1998, pp. 87–88).

In John's case he has no control over whether he retains the trait—let us suppose that he maintains his judgement, much to his despair, that he is better off with the chip than without it, and so cannot intentionally have the chip removed due to its continued operation—but he did have control over its formation, and we might think that he has some control over its expression: if he actively does what he judges best himself, before the AAC kicks in, then the mechanism will not operate, so he can prevent himself from losing control by doing what he judges best as soon as he so judges; also, he might be able to avoid doing things he does not want to do just by trying not to think about them and thereby not become aware of what he has sufficient reason to do.

The thing to note in response to this objection is that the tracing method is supposed to establish that an agent is morally responsible for some action which issued from a non-MRR mechanism, not that they have guidance control over that action's performance. Fischer and Ravizza's own examples of the tracing method focus on agents who are not in control of their behaviour and yet *are* morally responsible for what they do due to their having exercised guidance control in getting themselves into their current situation (see 1998, pp. 49–51 and pp. 87–89).

In John's case, his AAC behaviour *does* issue from a MRR mechanism which he takes responsibility for, so Fischer and Ravizza's conditions imply that he ought to have exactly the same kind of control that Sally exercises in driving her car over his

[19] Thanks to an anonymous referee for this point.

AAC behaviour. The conditions entail, that is, that John is *actually guiding* the AAC behaviour, not just that he is accountable for it.

As above, I am happy to grant that John may be morally responsible for his AAC behaviour, given that he got himself into this situation by his own choice. However, this in no way entails that he must be actually guiding it!

#### *3.4.4 Guidance control is not an analysis of pre-theoretical 'guiding'*

A final hard-headed objection to consider might run as follows: John *does* have guidance control, but this is just a placeholder for having his actions issue from a MRR mechanism for which he has taken responsibility. That is, it is irrelevant whether John is actually guiding his AAC behaviour, or has control in a commonplace or folk sense, since guidance control was not supposed to be an analysis of any pre-theoretical conception of control or guidance anyway.

It seems unlikely that this is what Fischer and Ravizza intend by 'guidance control' given their initial gloss; however, if this *is* the case, and guidance control is just a placeholder for the state one is in when one can be held morally responsible then (i) it's not clear why Fischer and Ravizza even bother to talk in terms of control: it would be clearer to simply give all of their claims in terms of moral responsibility (for example: 'an agent is morally responsible when their behaviour issues from one of their own MRR mechanisms'); and (ii) guidance control is really nothing like the control that Sally has over her car: Sally's case displays a kind of control, perhaps more fundamental than guidance control, which amounts to actually guiding some process, that intuitively *is* relevant to moral responsibility. If guidance control does not capture this notion then my example may well fail to show that the conditions Fischer and Ravizza stipulate for having guidance control are insufficient; however, once we drive a wedge between guidance control and actually guiding, it's not clear why we should want an analysis of moral responsibility in terms of guidance control.

## 4 Summary and prospects

I have argued that taking responsibility for a MRR mechanism is not sufficient for an agent's having guidance control over the behaviour issuing from that mechanism. An attitude on the part of the agent, as the AAC case shows, is not able to make the relevant kind of difference to whether an agent is actually guiding a particular action.

Responses to this case tend to either force the Fischer–Ravizza account into circularity or further demonstrate that *taking responsibility,* as it is described in *Responsibility and Control* (1998), is insufficient for mechanism ownership (by introducing further constraints on making a mechanism one's own); indeed, many of the potential objections to the AAC case, as well as some of Fischer's responses to manipulation cases, seem to suggest that there is some other notion of ownership at work in the analysis that is not conditional upon taking responsibility and which prevents the analysis from applying to cases such as the AAC. Until this condition is made more explicit we are not in a position to determine whether John's case is a genuine case of guidance control. In any case, I think that this suggests that the ownership element included in

the official conditions for guidance control is not suitable for an account *of actually guiding some action*.

That is not to say that actually guiding some action does not have an epistemic element: we may very well think that to action an action you have to be aware of it and see yourself, in some way, as its source. Perhaps *taking responsibility* is necessary for you to actually guide your behaviour, but it is not sufficient, as the AAC shows.

We might think, of course, that having a responsibility attitude is necessary for morally responsibility. Indeed, it may be that unless one understands what it means to be a moral agent and sees oneself as such, one is not in a position to be held accountable for one's actions; what's more, control in the form of 'actually guiding some action' *is* plausibly a good indicator that one is morally responsible for one's actions. However, this should not prompt us to build considerations about moral responsibility (such as taking responsibility) into an account of guidance.

**Acknowledgements** I would like to thank several anonymous referees for their helpful comments and suggestions on this paper. Special thanks, also, to Joel Smith and Thomas Smith for countless readings and re-readings of this paper over the years—we did it!



## References

Audi, R. (1986). Acting for reasons. *Philosophical Review*, *95*, 511–546.

Davidson, D. (1973). Freedom to act. In: T. Honderich (Ed.), *Essays on Freedom of Action* (pp. 137–156). London: Routledge and Kegan Paul. (Reprinted in Davidson 1980, pp. 63–81).

Fischer, J. M. (2004). Responsibility and manipulation. *The Journal of Ethics*, *8*(2), 145–177.

Fischer, J. M. (2010). Manipulation and guidance control: A reply to long. In J. K. Campbell, M. O'Rourke, & H. Silverstein (Eds.), *Action, Ethics, and Responsibility* (pp. 175–186). Cambridge: MIT Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

Fischer, J. M., & Ravizza, M. (2000). Replies. *Philosophy and Phenomenological Research*, *61*(2), 467–480.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. In G. Watson (Ed.), *Free will (2003)* (2nd ed., pp. 167–176). Oxford: Oxford University Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*(1), 5–20.

Long, T. R. (2004). Moderate reasons-responsiveness, moral responsibility, and manipulation. In J. K. Campbell, M. O'Rourke, & D. Shier (Eds.), *Freedom and Determinism*. Cambridge, MA: MIT Press.

Mele, A. (2000). Reactive attitudes, reactivity, and omissions. *Philosophy and Phenomenological Research*, *61*(2), 447–452.

Stump, E. (2002). Control and causal determinism. In S. Buss & L. Overton (Eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (pp. 33–60). Cambridge, MA: MIT Press.