Ethic Theory Moral Prac (2018) 21:537–551
https://doi.org/10.1007/s10677-018-9897-3

Irrational Option Exclusion

Sofia Jeppsson[1]

Accepted: 16 May 2018 / Published online: 5 June 2018

Abstract In this paper, I describe a hitherto overlooked kind of practical irrationality, which I call *irrational option exclusion.* An agent who suffers from this problem does not merely fail to *act* on her best judgement – she fails to realize that the superior action is even an *option* for her. I furthermore argue that this kind of irrationality is serious enough to undermine moral responsibility. I show that an agent suffering from this problem has compromised reasons-responsiveness, does not really express her will through action, and has a hard time doing otherwise; thus, from the standpoint of several popular moral responsibility theories, we ought to conclude that her responsibility is at the very least diminished.

Keywords Moral responsibility · Reasons-responsiveness · Ability to do otherwise · Quality of will · Practical irrationality

In this paper, I will describe a hitherto overlooked kind of practical irrationality, which I call *irrational option exclusion.* An agent is practically irrational when she fails to do what she judges best. This is not always responsibility undermining; it is widely agreed that agents can be responsible for *akratic* acts. But an agent who suffers from irrational option exclusion does not merely fail to *act* on her best judgement – she fails to realize that the superior action is even an *option* for her. I will argue that when an agent's rational thinking is compromised in this way, her moral responsibility is compromised as well – and this is so, regardless of which moral responsibility theory we adopt.

1 The Milgram Obedience Experiments: The Case of Mr. Braverman

For a first illustration of the phenomenon, let us take a look at the much discussed Milgram obedience experiments (Milgram 1974/2010). Prima facie, experiments such as Milgram's seem

✉ Sofia Jeppsson
Sofia.jeppsson@gu.se

[1] Gothenburg Responsibility Project, Department of philosophy, linguistics and theory of science, University of Gothenburg/Göteborgs Universitet, 405 30 Göteborg, Sweden

to threaten the thesis that most people are morally responsible for what they do most of the time. The Milgram experiments and similar ones seem to show that our actions depend more on external factors – in the case of the Milgram one, the presence of an authority figure – than on our characters, values and reasons, which prompts the question of whether we can really be morally responsible for what we do (Nelkin 2005). Some philosophers claim, however, that the Milgram test subjects were likely *akratic*; they failed to act on their best judgement (Mele 2014: 69–70; Doris 2002: 134–136). Since *akrasia* is normally not considered an excuse, those philosophers conclude that the Milgram test subjects were responsible after all; the prima facie threat to responsibility evaporates at closer scrutiny. Things are not, however, quite this simple. *Akrasia* was not the only kind of practical irrationality at play in these experiments.

In Milgram's experiment, test subjects were told that they were to participate in a study of the effects of punishment on learning. The test subjects were given the impression that they were randomly assigned either a 'teacher' or a 'student' role. In reality, all the actual test subjects were assigned 'teachers', whereas the 'student' was a confederate of Milgram. The student was asked to memorize a list of words. When providing the wrong words in response, the teacher was to give him an electric shock as punishment. The shocks were mild at first, but the test leader continuously told the teacher to turn up the voltage until he reached levels labelled dangerous on the electro shock machine and the student seemingly passed out from the shocks. In reality, of course, no shocks were actually administered. Most of the test subjects obeyed the test leader all the way. A few thought the shocks justified (for instance, because they assumed that the test leader, being a scientist and all, must know what he was doing, and would not perform these tests unless it was for the best, Milgram 1974/2010: 89–90), but most seemed to think that they did something very bad indeed. So why did they not refuse to obey the test leader's order?

One test subject, a Mr. Braverman, is quoted as saying that he had the impulse to refuse, but he went on with the experiment anyway (ibid: 54–55). This sounds like a clear-cut case of *akrasia*. Interestingly enough, however, when Braverman talks about 'refusing', he does not refer to a refusal to continue with the experiment at all. What he, somewhat confusedly, means by 'refuse', is to feed the student clues and help him with the words on the list in various ways, ensuring that he gets the words right and thus does need not be shocked. Regarding *this,* Braverman was *akratic*; he 'felt the impulse' to try and help the student – presumably this is what he deemed *best* – and yet he failed to do so. However, I want to focus on the fact that Braverman never even mentions the possibility of an actual refusal to continue with the experiment. It does not seem to have struck him at all that this was a possibility – not even afterwards, when interviewed about his thoughts during the experiment. This is a serious and puzzling failure of rationality on Braverman's part – a failure not just to *do* what best accords with his values, but to even *think* about what would be the right thing to do. (In the following, I will continue to focus on Braverman, but I assume that he was not unique among the test subjects.)

First of all, it should have been evident to Braverman that there was nothing actually stopping him from refusing (henceforth, when I talk about 'refusal', I mean a complete refusal to continue with the experiment, not merely, e.g., trying to help the student out with the words on the list). He knew that he was there voluntarily. He knew that the test leader did not have the authority to punish him if he were to refuse. There is no law according to which you must perform well in psychology experiments or you go to jail. All this is common knowledge; surely Braverman, as an ordinary adult, knew this as well. One does not need to be particularly well versed in the ways of logic in order to draw the conclusion, from these premises, that refusal is an option. So why did the thought of refusal never even strike him?

Of course, there are always countless actions that are possible for us to perform, and yet the thought of doing them never crosses our minds. When I sit in my office philosophizing, nothing stops me from instead counting all the tiles in my floor, jump up and down on one leg, sing the national anthem and so on. In a simple sense of 'can' (disregarding worries about determinism and the ability to do otherwise), I *can* do all these things, and yet they never strike me as possibilities. But this is obviously because I lack good *reasons* to do them. Writing my paper is, right now, my best option, and so I need not consider anything else; or, even if it is not the absolutely best option (perhaps something more far-fetched which I have not considered would be better after all) it is at least a *good enough* option. That I do not constantly consider and deliberate about all the actions that I can do is not in itself problematic; it is, on the contrary, both necessary and beneficial, since I would never get anything done otherwise.

What makes Braverman's failure to even think about refusal so odd is that refusal is obviously his best option, at least subjectively. In reality, of course, no shocks were given, so one might argue that the Milgram test subjects lacked objective reasons to refuse. From Braverman's point of view, however, he was shocking an innocent person. Clearly, he had a strong moral reason not to do this, and a self-interested one as well. Braverman felt *terrible* about delivering the electric shocks. He ought to have refused for both the student's and his *own* sake. Given all this, it is highly surprising that the thought of refusing never even struck him.

I call the kind of practical irrationality at play here *irrational option exclusion,* or IROE. An agent suffers from IROE when she fails to even consider φ-ing (an action, sequence of actions or way of life other than the one she performs or lives) as an option despite the fact that

a) φ is far better, according to her own values, than the action or actions she ends up performing or the kind of life she ends up living. (Values are here construed rather loosely – we *value* something when it *matters* to us in a positive way. See Wolf 1990: 31)
b) she is dissatisfied with the action or sequences of actions that she ends up performing or the kind of life she currently lives
c) the possibility of φ could be easily and quickly deduced, by a person of average logical competence, from facts that the agent knows.

This definition of IROE employs several vague concepts. Firstly, there is no sharp distinction between cases where φ is *far* better than what the agent actually does, and cases where φ is only slightly better. If φ is only slightly better, the agent is not obviously irrational – at least, her irrationality might not seem severe enough to threaten responsibility.[1] Secondly, it is not possible to draw a sharp distinction either between cases where the possibility of φ can be quickly and easily deduced, and cases where it could only be deduced with difficulty. For Braverman, however, the deduction required to reach the conclusion that a refusal was possible would clearly fall on the "quick and easy" side.[2]

An agent who merely satisfies a) and c), but feels satisfied with her choice anyway, still seems irrational, at least to some extent. Still, it would be harder to argue that this agent had

[1] I will not, in this paper, delve into the debate between maximizers and satisficers about rationality. See Pettit 1984 and Slote 1984 for a discussion about this issue.

[2] If an agent fails to conclude that φ is an option where said conclusion would have required highly sophisticated deductive skills, she need not be irrational for failing to do so. One need not be a super logician to be rational. She might still be excused for failing to φ, but for the familiar reason that she was non-culplably ignorant of the possibility.

diminished moral responsibility for failing to φ, at least on some moral responsibility theories. Her satisfaction with her choice seems to reveal something about who she is – possibly this fact can ground moral responsibility for not φ-ing. Since b) at least arguably plays an important part when excusing agents, I include it in my definition of IROE.

The fact that I usually write my papers without even considering whether I would rather count all the tiles in my floor does not show that my rational thinking is compromised in any way. But for Braverman, a-c holds, making his failure to even think about refusal quite seriously irrational – serious enough to compromise his moral responsibility. Or so I will argue.

2 IROE and Reasons-Responsiveness

It is generally assumed that a certain level of rationality is required for moral responsibility. I have suggested that Braverman's irrationality in failing to even consider refusal as an option, despite feeling terrible about hurting the student, is of a fairly serious kind. However, the details of the argument for IROE being responsibility undermining will depend on which theory of moral responsibility we assume. I will therefore investigate the implications that IROE has for moral responsibility one theory at a time.

The two main families of compatibilist moral responsibility theories are reasons-responsiveness theories and quality of will theories (Vargas 2013: 135–136). Libertarian theories tend to focus on the element that sets them apart from their compatibilist competitors, i.e., where to locate the indeterminism in the decision-making process and, for some theories, how to explain agent causation. Libertarians want to *add* to compatibilist theories, not deny the compatibilist conditions. Plausibly, most libertarians will therefore also accept something like reasons-responsiveness or the expression of the agent's quality of will in action as necessary conditions for moral responsibility, even though libertarians deem them far from sufficient. Besides reasons-responsiveness and quality of will theories, and despite the huge impact that Frankfurt scenarios (which allegedly show that an agent can be morally responsible for his action despite an inability to do otherwise) have had on the moral responsibility debate, there are also philosophers who defend a principle of alternative possibilities; they argue that an agent cannot be morally responsible for what she did, or at least not blameworthy for a wrongful action, unless she had the ability to do otherwise (Vihvelin 2013 Ch. 4; Nelkin 2011 Ch. 5; Wolf 1990 Ch. 4).

I will begin by discussing Fischer and Ravizza's reasons-responsiveness theory about moral responsibility, since it is probably the most influential theory of its kind, and show that if we accept this theory, we ought to conclude that IROE at the very least diminishes moral responsibility. Later in this paper, I will show that IROE furthermore makes it very difficult to do otherwise, and finally, discuss whether an agent can fully express his quality of will when suffering from IROE.

Reasons-responsiveness has two parts, according to Fischer and Ravizza: Reasons-receptivity and reasons-reactivity (Fischer and Ravizza 1998: 41–46 and 69). Although they never say so explicitly, it seems clear that Fischer and Ravizza have in mind objective, mind-independent reasons when they discuss reasons-responsiveness; an agent must be receptive and reactive to *those* (including moral reasons) in order to be morally responsible for what she does (ibid: 42 and 76). However, most of what they have to say about reasons-responsiveness does not really hinge on the nature of reasons. Their theory works just as well if we assume that what an agent has reason to do depends on, e.g., what she values, as long as our theory of

reasons as mind-dependent in this way is sophisticated enough to allow for the possibility that agents sometimes fail to see what they have reason to do.[3] My arguments in this paper do not depend on whether reasons are mind dependent or not.

Now, Fischer and Ravizza consistently discuss reasons-responsiveness as a property of the agent's *mechanism* rather than the agent herself (Fischer and Ravizza 1998: 38–39) (The mechanism is not, of course, supposed to be an actual mechanism, but is to be understood roughly as the faculty of practical reason, broadly conceived.). I will, however, consistently write about *agents* being or not being reasons-responsive. Fischer and Ravizza introduce the mechanism talk in order to be able to handle Frankfurt cases, but this is a debate that I will not delve into here. Readers convinced that moral responsibility really depends on qualities of the mechanism rather than the entire agent can take my agent talk as short for mechanism-of-the-agent-talk.

If we first look at the receptivity part of reasons-responsiveness, it is possible, according to Fischer and Ravizza's theory, that an agent has an all-things-considered reason to φ and fails to realize this, despite the fact that she is sufficiently receptive to reasons to fulfil the conditions for moral responsibility. Suzy might think that what she has most reason to do is to go shopping, and proceed to do so, when in reality she has most reason to give her shopping money to charity instead. If Suzy is reasons-receptive, however, she would realize that she had most reason to give her money to charity in an alternative scenario where she had more and/or stronger reasons to do so. Suppose, for instance, that Suzy found herself the only person in the world who could save the life of another by giving him money for life-saving medication. In this scenario, Suzy would be receptive to her moral reason to give her money away to someone who needs it more. Furthermore, in order for Suzy to be morally responsible for what she does, her receptivity to reasons must form an understandable pattern (ibid: 65–68). To illustrate and motivate this requirement, Fischer and Ravizza give us the case of Brown, who is a habitual user of the drug Plezu. The drug causes him to laze about doing nothing, thus interfering with both his other activities and his self-respect. Brown therefore has good reasons to stop taking Plezu, but fails to realize this. He still counts as sufficiently reasons-receptive, because he would recognize that he ought to stop taking Plezu if one dose cost a thousand dollars, and also if one dose cost two thousand dollars and so on – in short, he would recognize that there are good reasons against taking Plezu if the price was too high. Imagine, however, that Brown would recognize that a thousand dollar price was a reason against taking Plezu, while failing to realize that a two thousand dollar price gives him an even stronger reason to stop. In this scenario, Brown is not sufficiently reasons-receptive to be morally responsible for his taking of the drug, because his receptivity fails to show the requisite understandable pattern. Fischer and Ravizza write that when an agent fails to be reasons-receptive, this is typically a sign of psychosis (ibid: 41–42). I will argue that crucial failures of reasons-receptivity can occur in non-psychotic agents too.

When it comes to reasons-*reactivity*, on the other hand, a pretty weak one is sufficient for moral responsibility, according to Fischer and Ravizza (ibid p. 70). Weakness of will, they argue, is intuitively no excuse for bad behaviour, and thus, weak reactivity must suffice. Braverman 'had the impulse' to help the student out – presumably he meant that he thought that he ought to do so. Yet he did not act on this judgement. We might, however, easily imagine scenarios in which

[3] If reasons are mind-dependent, we cannot argue, like Fischer and Ravizza do, that psychopaths lack moral responsibility in virtue of failing to see that they actually have moral reasons not to hurt others, because psychopaths (at least not the possibly fictional psychopaths discussed in philosophy) would not, in fact, have such reasons. A defender of a reasons-responsiveness theory who also embraces the mind-dependence of reasons could either simply accept this, or tack on an independent moral competence condition for moral responsibility.

Braverman does. Suppose for the sake of argument, and in order to focus on *akrasia* rather than IROE for the time being, that Braverman's only two options were to try and help the student to avoid wrong answers, and to not try to help and continue administering the shocks. Suppose, furthermore, that if the student had looked more frail and delicate, Braverman would have helped him out by feeding him subtle cues. If so, Braverman has some reasons-reactivity with regard to helping the student with the answers, and he is therefore morally responsible for not doing so in the actual scenario. Reactivity need only be weak.

Now, let us return to reasons-*receptivity*, and how it is compromised by IROE.

As previously stated, in every moment countless possible actions remain unconsidered. This is normally no problem, nor does it normally undermine our responsibility for what we actually do. I am no less responsible for writing my papers because I did it without asking myself whether I should count all the tiles in my office floor instead. I need not think about counting the tiles, because I obviously have no reason to do so. We assume that I, a regularly reasons-responsive agent, *would* consider counting the tiles in a counterfactual scenario where I *do* have a good reason to do so (if, say, the office floors are to be renovated, counting the tiles is important for this purpose, and every employee has been charged with counting those of her own office). Braverman, however, already has good reason to refuse, and yet the thought of doing so does not even strike him. Since he never thinks about refusal, he does not think about the reasons he has for refusal either. It seems spurious to simply posit that Braverman was, somehow, aware of the reasons he had to refuse on a subconscious level, when this supposed awareness has no effect whatsoever on either his actions or his thoughts – we should rather conclude that Braverman was unaware of the fact that he had reason to refuse. We might try to save the thesis that Braverman is reasons-receptive with regard to refusal by saying that he *would have* realized that he had a reason to refuse, *if* his reasons had been different and/or stronger – just as Brown counts as reasons-receptive despite failing to see that he has a reason to stop taking Plezu, because he *would* realize this *if* Plezu was more expensive. But there are problems with this suggestion.

Let us assume that Braverman, like many other test subjects in Milgram's experiments, would have refused, and thus also have realized that he had reason to refuse, if the scenario had been altered in certain ways. Milgram conducted a number of experiments while alternating the settings. In some versions, two other phony test subjects who were actually Milgram confederates were assigned 'teacher' roles, and then protested against the test leader. Witnessing this, a large percentage of the real test subjects refused to obey (Milgram 1974/2010: 117–122). Likewise, more test subjects refused when they had to force the student's hand onto an electro-shock plate rather than just push a button from a different room (ibid: 35–37), and more also refused when the test leader issued his orders by phone from a different location (ibid: 62). It follows from the fact that the test subjects did refuse in those situations that they also realized that refusal was an option, and we can safely assume that they also realized that there were strong reasons to do so. If Braverman too would have realized he had reason to refuse in those scenarios, it follows that he has *some* reasons-receptivity. But whereas weak reactivity suffices for moral responsibility, Fischer and Ravizza demand more when it comes to receptivity. In order to be sufficiently reasons-receptive for moral responsibility, an agent must have an understandable pattern when it comes to the reasons they recognize. Does Braverman exhibit such a pattern? I think not.

Fischer and Ravizza ask us to imagine an interview with the agent, in order to see whether he can provide a comprehensible explanation as to why he thought he had a reason to φ in one scenario but not in another (Fischer and Ravizza 1998: 71–72). Now, Milgram conducted a

real interview with Braverman, not just an imagined one, in which Braverman utterly failed to see refusal as an option; he even claimed he was "totally helpless" in the situation in which he found himself (Milgram 2010/Milgram 1974: 54–55). But if we imagine that Braverman would have refused, like so many others, in the three alternate scenarios described above, that an interviewer and Braverman himself somehow gained access to these counterfactual scenarios and that the interviewer asks him about them, what would Braverman say by way of explanation? Given sufficient self-knowledge, he could very well come up with causal psychological explanations. He could say, for instance, that seeing two other people refuse made vivid that option, that when he had to force the student's hand onto an electro shock plate he was *so* overcome with feelings of horror that he just recoiled, that when the test leader merely talked to him by phone his authoritative presence was less keenly felt and therefore had less of an influence on his thoughts and which options he perceived and so on. All this is understandable in the way causal explanations are understandable. However, explanations that are understandable in this causal-psychological sense can be given of even the most insane behaviour, and yet Fischer and Ravizza do not want to argue that completely insane people are morally responsible for what they do. It is clear that what they are asking for are understandable *reason explanations*; an explanation as to why the agent, according to his own values and preferences, had *reason* to φ in situation A but not in situation B. If Braverman had thought that he had reasons of strength 0.8 to refuse in the actual scenario, and would have thought that the strength of the reasons rose to 1.0 (or the strength of his countervailing reasons to obey the test leader sank) in the alternate scenarios, there might be an understandable explanation in terms of reasons and values to be had, and Braverman would come out as sufficiently reasons-receptive. But Braverman did not just perceive his reasons as weaker in the actual scenario; he failed to see that disobedience was at all an option, he failed to even think about it, and therefore also failed to see that he had reason to do so. Unless Braverman has some very peculiar values and preferences indeed, there can be no comprehensible explanation as to why he saw *no* reason to refuse in the actual scenario whilst seeing strong reasons to do so in the alternate ones.

We should conclude that Braverman's reasons-receptivity is compromised. Even if we think he still possesses sufficient receptivity to be somewhere in the ballpark of moral responsibility, we should conclude that his responsibility is, at the very least, *diminished* due to his deficient reasons-receptivity.

3 Life Choices, IROE and Reasons-Responsiveness

Braverman found himself in a highly stressful situation. We might think that this explains why he suffered from IROE with regard to a complete refusal. There are empirically based reasons to believe, however, that it is possible to suffer from IROE when it comes to big, long-term choices of how to life one's life as well. I will argue that such agents, too, have their responsibility diminished according to the same reasons-responsiveness theory.

There are therapy techniques directed at repeat offenders whose purpose is to make the offender see a law-abiding lifestyle *as an option*, thus presupposing that at least one important reason why the offender is as yet a habitual criminal is his failure to realize that he has any better alternative open to him. Langlands et al. (2009) explain the so-called *Good Lives Model* (GLM) for rehabilitation of violent offenders through a case study constructed from their experience of working with a number of domestically violent offenders. Mr. X, as they name

the offender, is a man of Maori descent, with past convictions for sexual as well as non-sexual violence against both his female partner and strangers. He was raised in a family where he was exposed to both violence and sexual abuse, and often stayed with his grandfather in order to escape the violence of his parents. As Gary Watson famously points out, we often have mixed intuitions about cases like this (Watson 1987/2013). On the one hand, Mr. X commits really bad acts, but on the other hand, he is a victim as well. Now, I will not say that we *must* appeal to his failure to see a law-abiding lifestyle as an option in order to argue that Mr. X has diminished moral responsibility for his life of crime. Mr. X:s childhood might, for instance, have left him with bad impulse control, making restraining violent impulses very difficult for him (Levy 2011: 194–199). Temptations to, e.g., steal things, might be much more difficult to resist for Mr. X than for us because of the sad circumstances in which he lives (von Hirsch 1996: 106–109). Nothing I say precludes that there might be several more or less independent factors that each diminishes Mr. X:s moral responsibility. However, among other things, he seems to have suffered from IROE. As in the case of Braverman, I will argue that this fact at least diminishes his responsibility.

Initially, Mr. X complied with the rehabilitation program he was assigned to merely because he hoped to gain a parole. Once in GLM therapy, however, the clients are encouraged to tell the therapist what they have always *dreamt* about. Although Mr. X had sought to get his kicks from drugs, alcohol, multiple partners, quick money, violence and gang membership, he came up with very different goals when encouraged to envision what kind of future he really wanted for himself. What he most wanted was to go to university and study Maori history, and have more intimate and fulfilling relationships with friends as well as with a romantic partner. After Mr. X had managed to formulate these goals for himself, he could develop a plan of how to get there, which included steps such as reducing his drug use, distancing himself from anti-social acquaintances, building new relationships with more pro-social peers, taking a driver's license and so on. Eventually Mr. X made a number of profound life changes; he now goes to university, has taken up diving as a new hobby, has made new friends and remains in a committed relationship with no violent episodes.

Langlands et al. do not explain what GLM therapy is supposed to do in terms of increasing practical rationality or combatting IROE. Still, the success of the therapy method depends on the client having *dreamt* about a different kind of life, which indicates that the client is dissatisfied with his current situation, and that a different kind of life would be better according to his own values.[4] And at least in the supposedly typical case of Mr. X, it is also true that the possibility of a different and law-abiding lifestyle could be deduced by a person of average logical competence from known facts. None of the steps that Mr. X took towards a different kind of life seems to have required esoteric knowledge unavailable to him prior to therapy or complicated logical inferences. Thus, from the description that Langlands et al. give of Mr. X, it seems like the main problem he needed to overcome in order to become law-abiding was IROE; the irrational exclusion of a law-abiding life with close relationships and university studies from his set of deliberative options, despite him dreaming about it.

Jeanette Kennett and Doug McConnell likewise discuss how our self-conception, not just perceived rewards, might influence what we do (Kennett and McConnell 2013). Their discussion is centred, not on habitual criminals but on addicts. The alcoholic Crispin Sartwell, frequently discussed in their paper, have gone sober for extended periods of time, which might

[4] We might, of course, daydream about lifestyles we do not really want to pursue, but that is clearly not what dreaming about means in the context of GLM therapy.

lead one to think that he certainly realizes that going sober is an option. In this he is different from Mr. X, who had never tried a different kind of life. Still, Sartwell, from Kennett and McConnell's description, does not seem to think that being sober *for the rest of his life* is a real option; he thinks that being an alcoholic is his "destiny", since most of his family are alcoholics (ibid: 485). Sartwell considers life as an alcoholic completely meaningless and horrible (ibid: 470, 476, 477), he should be able to conclude from his extended periods of sobriety that quitting is possible (ibid: 484), and yet, simply because he has trouble regarding life-long sobriety as a real option, he constantly goes back to drinking.

Criminals from harsh backgrounds and addicts are the kind of cases where we might feel spontaneously inclined to make judgements of diminished moral responsibility. However, being barred from seeing φ as a real option because one just does not see onese`lf as the kind of person who φ's, can presumably happen in less dramatic cases too. There is, for instance, some research done on why working-class youths rarely go to university that indicate that besides very real obstacles, self-conception could also play a part – youths from the working class think that "people like us don't go to university" (e.g., Pearce et al. 2008).

Mr. X, Sartwell and others who are to make decisions concerning their entire way of life are not quite analogous to Braverman. The difference is not just that Braverman was to make a decision about what to do right now in a stressful situation whereas Mr. X and Sartwell were to choose a way of life, but also that Braverman had not even *thought* about refusal, whereas Mr. X had *dreamt* about a law-abiding life, and Sartwell spent much time pondering the sober life. It might seem as if Mr. X and Sartwell, therefore, *were* receptive to reasons, although they did not react to them. One might argue that in thinking of a different kind of life as desirable, they implicitly saw that they had a reason to pursue it. Nevertheless, another argument can be made that they did not. It has been argued, to my mind convincingly, that 'reason' (like 'ought') implies 'can' (Haji 2012; Streumer 2007; Jeppsson 2016). If a child is drowning on the other side of the world, I do not have a reason to try and save her. Were I to have magical powers and knew about her plight I might have such a reason, but since I am an ordinary mortal, I do not. If someone were to advise me that I *do* have a strong reason to save the drowning child on the other side of the world, that person would seem delusional, or perhaps simply mistaken about what 'reason' means. The thesis that 'reason' implies 'can' (just as its more famous sister principle according to which 'ought' implies 'can') is controversial (see, for instance, Zimmerman 2007 pp. 329–330), and I cannot, within the confines of this paper, resolve the issue. But if we agree that 'reason' implies 'can', and if this furthermore is something that laypeople intuitively appreciate, it is plausible that we need to perceive φ as a real possibility in order to be receptive to the reasons we might have to φ. If Mr. X, prior to therapy, has a hard time understanding that it *is* possible for him to break with his old lifestyle and create a new life for himself, and Sartwell never grasps life-long sobriety as an option, it is plausibly the case that although they think of a different kind of life as *desirable*, they fail to see this desirability as a *reason* to pursue it.

Now, on Fischer and Ravizza's account, an agent can be reasons-receptive despite not recognizing reasons for action in the actual world. If we go back to Mr. X, he does not merely recognize that he has reason to change his life in some other possible world, but he even does so in this very world after he has been through GLM therapy. Thus, he clearly has *some* reasons-receptivity. Possibly, there might have been ways to make him see that he had reason to change his life without therapy as well – perhaps he would have done so if, for instance, someone had presented him with a large cash reward for changing his life, thus vividly presenting a change as a real option while simultaneously providing him with extra motivation.

I am not arguing that Mr. X is completely out of the responsibility ballpark. However, as already noted in the discussion about Braverman, Fischer and Ravizza's reasons-responsiveness theory also requires that there is an understandable pattern to the agent's reasons-receptivity. The agent must be able to explain in an imagined interview why he believed that he had reason to φ in this scenario but not in that. Mr. X does not quite live up to that demand. Mr. X had precisely the same reasons to change his life, study and build healthier relationships *prior to* therapy as *after* therapy. He dreamt about doing so both before and after. Given enough self-knowledge, Mr. X (or perhaps a psychologist working with Mr. X) could probably provide a causal explanation as to why he did not see the reasons he had prior to therapy, only after, but no reasons explanation can be given. We should therefore judge that his moral responsibility for living a life of crime prior to therapy was at least diminished.[5]

Thus, it is plausible that agents can suffer from responsibility undermining IROE both when it comes to stressful, proximate decisions and when it comes to choosing an entire way of life. When an agent suffers from IROE, she fails to realize that φ is a possible option, and therefore fails to realize that she has reason to φ, despite φ being much better for her than what she actually does. In a scenario in which the possibility of φ was made more vivid to her (variations on the experiment for Braverman, post therapy for Mr. X), she would recognize that she has strong reasons to φ, despite the fact that nothing about her reasons has actually changed. For most IROE agents, there will plausibly be causal-psychological explanations for this, but no rationally understandable explanation in terms of reasons, values and preferences can be provided for the agent's failure to appreciate her reasons to φ in the actual scenario. Her reasons-receptivity is compromised, and therefore, I argue, she is less than fully morally responsible for her failure to φ. Responsibility undermining IROE might even be fairly common, providing at least partial excuses to agents for whom none of the traditionally recognized excuses apply. At least I have shown that this is so if we assume Fischer and Ravizza's reasons-responsiveness theory of moral responsibility. I will now show that the same holds true if we assume that moral responsibility, or at least blameworthiness, requires the ability to do otherwise.

4 IROE and the Ability to Do Otherwise

Fischer and Ravizza's theory is supposed to explain what kind of control is required in order to be morally responsible for what one does. They explicitly deny *the principle of alternative possibilities,* since it is allegedly proven false by Frankfurt examples (Fischer and Ravizza 1998: .29–.41 and 55–61). Some philosophers, however, still argue that moral responsibility, or at least blameworthiness, requires an ability to do otherwise (Vihvelin 2013 Ch. 4; Nelkin 2011 Ch. 5; Wolf 1990 Ch. 4). When the agent is capable of doing the right thing, but doing so is very *difficult*, blameworthiness still attaches, but is diminished (Wolf 1990 86–87; Nelkin 2016). If we accept these theses, we can build another argument for agents like Braverman and Mr. X having at least diminished moral responsibility.

Overcoming weakness of will might be difficult in the sense that it requires a huge mental effort on part of the agent. Thus, on Nelkin's account, it seems that weakness of will can diminish blameworthiness, but this is perhaps as it should be; we do intuitively judge a

[5] How this diminished responsibility for his entire way of life bears on his responsibility for individual crimes is an interesting and complicated question, which I will nevertheless not try to resolve here.

wrongdoer less blameworthy if she struggled hard and finally fell for temptation despite making a sound moral judgement than if she gleefully and wholeheartedly did wrong. Still, an *akratic* agent, we might think, knows how to overcome her problem: She must grit her teeth together and muster up some more willpower. There are problems with this suggestion, in particular since there is some (controversial) psychological research indicating that willpower is a limited resource (e.g., Baumeister and Tierney 2011 for the 'ego depletion' theory; Hagger et al. 2016 for critique). Still, our common-sense judgement according to which *akratic* agents are blameworthy probably depends on the belief that they *could have* done better if only they had made a bit more of an *effort*.

But what about the agent suffering from IROE? What is she supposed to do in order to think about the option she is currently not thinking about, or begin to see a mere daydream as a real option to deliberate about? It seems that in this situation, the very problem at hand might prevent its own solution. We might suggest that people like Braverman or Mr. X engage in a brainstorming session, and make a conscious effort to think outside the box. But in order to see the need for such strategies, they would first have to realize that they suffer from IROE; that there is a highly desirable option that they have overlooked. We do not, after all, go around brainstorming and outside-the-box-thinking all the time; we do this only when we have a special *reason* to think that we might have missed something and need to generate more options for deliberation. If we did it all the time, we would get nothing done. But precisely because Braverman suffers from IROE, he fails to realize that there are any more options than 'help the student' and 'obey the test leader without helping the student' on the table. Thus, he does not suspect that there is an option that he has overlooked – apparently, he did not even realize this much in the far less stressful situation he found himself in when interviewed about the experiment after the fact. When it comes to Mr. X, it is unclear whether a brainstorming session would even be of much help. He has already *thought* about a different lifestyle; the problem is that this did not seem like an *actual option* to him. He was helped by therapy, but it is unclear – to say the least – whether there was anything that he could have done on his own to turn a mere daydream into an option for serious deliberation.

When agents are *akratic,* they know that they are, and they know (at least arguably) what to do to fix their problem, namely, muster up more willpower. When agents suffer from IROE, on the other hand, it is hard to say what they ought to do in order to spot their problem in the first place. Overcoming IROE might not be difficult in the sense that it requires a huge effort, but it *is* difficult in the sense that it requires an unusual imaginative leap to conclude that this is a problem that one suffers from, and furthermore that one ought to, e.g., brainstorm in order to try to overcome it. If you cannot spot that you suffer from IROE in the first place, you cannot fix it.

Thus, if we assume that blameworthiness for a wrongful action requires an ability to do the right thing, we have a strong case for Braverman and Mr. X not being blameworthy, since there does not seem to be anything that they could have done in order to first spot their problem and then fix it. At the very least, we should conclude that fixing their problems would have been *very difficult* for them, since spotting that they suffered from the problem in the first place (and that there thus was something to be fixed) would have required such unusual imaginative leaps – if so, their responsibility is at least diminished.

5 IROE and Quality of will

Finally, we come to quality of will theories, the main alternative to reasons-responsiveness theories. Quality of will theories take their inspiration from Peter Strawson's classic paper

'Freedom and Resentment' (1962/2013). According to Strawson's theory, an agent is morally responsible and blameworthy for what she did insofar as her action expresses an ill will, or at least indifference, towards others. Excuses (that is, excuses proper, excluding exemptions) function by showing that the agent's behaviour was actually compatible with the demands of good will being met (Strawson 1962/2013: 68). Strawson lists a number of excuses, among them 'he couldn't help it', 'he had to do it', and 'it was the only way'. Those excuses once again bring to mind the principle of alternative possibilities, but maybe a better interpretation is that an agent is not blameworthy if he did what he did *only because* he could not do otherwise[6] – or because he mistakenly *thought* that he could not do otherwise. When this is the case, the agent does not really express his quality of will in his action. The fact that Braverman failed to help the student despite judging it best to do so might say something about his will. But the fact that he did not refuse to continue with the test, when he never realized that doing so was even an option, does not really reveal anything about his regard for the student or how much he cared about the suffering he (seemingly) caused.

Of course, an agent might fail to see φ as an option because she does not really care about φ or anything that could be accomplished through φ-ing – as in the previously discussed example where I did not see counting the tiles of my office floor as an option. Previously, I explained this by referring to my lack of reasons to count the tiles. One might just as well say, however, that I fail to see this as an option because I do not care if the tiles get counted or not. If a Milgram test subject failed to see refusal as an option because it was obvious to him that obeying the test leader was the right thing to do, his obedience *does* reveal something about himself; an indifference towards the suffering of others. But when an agent, like Braverman, suffers from IROE, and fails to see φ as an option despite the fact that φ is far better, according to his own values, than the action that he ends up performing, and he is dissatisfied (to say the least) with doing what he does, *then* we have a situation where the agent's failure to φ does not really express his will. Likewise, an agent who *chooses* a life of crime over a law-abiding one thereby expresses an ill or at least indifferent will towards his fellow citizens, and the same can be said about an agent who never considered anything but a life of crime due to him strongly valuing quick money over other people. But an agent who, like Mr. X, would prefer a law-abiding life but goes along with the criminal one since he fails to see that there is any other real option, does not similarly express his will through his way of life.

The QoW theorist who wants to defend the blameworthiness of agents like Braverman might agree that he cared to some extent about the student's perceived plight, but still argue that his QoW was substandard; if he had cared *sufficiently*, she might insist, he *would* have realized that refusal really was an option. Angela Smith, for instance, argues that not just the actions we chose to do, but also non-chosen phenomena such as our emotions, beliefs and omissions, can reflect our judgments and normative commitments (Smith 2008). One might argue that insofar as we can be morally responsible for our omissions, we can be morally responsible for failing to think of the right action as an option. However, I believe this line of argument will not take us all the way towards full-blown moral responsibility. It might be true that Braverman's QoW is short of optimal; it might be true that had he cared *even more* about the student's plight, he would have realized that refusal was an option and also have acted on it. If so, he might be a little bit blameworthy. But a QoW theorist should still distinguish between Braverman and someone who never considers refusal because he does not care much about others at all; whereas the latter person is fully blameworthy for not even thinking about refusing, Braverman should be considered less so.

[6] Frankfurt also lists this as a valid excuse in his original paper (Frankfurt 1969: 838).

Braverman is likewise not analogous to an ordinary distracted agent who, say, forgets her friend's birthday, where it might be plausibly argued that her so forgetting shows that she did not care much, and therefore is blameworthy. Braverman would be more analogous to an imagined agent who thinks about and feels strongly about the fact that it is her friend's birthday today, but, because of some strange psychological blockage, cannot conceive of any way to wish her a happy birthday. The simple act of picking up the phone and giving her friend a call does not occur to her; she feels helpless in this regard. The latter kind of agent would undoubtedly be a strange person with a strange and unusual problem, but I believe that a QoW theorist should agree that if such an agent existed, she would be far less blameworthy than a regular agent for not calling her friend to wish her a happy birthday.

Thus, we arrive at the same conclusion if we start with a quality of will theory: Agents who suffer from IROE thereby have at least diminished moral responsibility. At least this is so if we focus on the direct variety. But what about indirect moral responsibility?

6 Indirect Responsibility

Some readers might object that I have only shown that IROE diminishes or undermines *direct* moral responsibility, while leaving open the possibility that Braverman, Mr. X and other IROE agents have *indirect* moral responsibility for what they do. It is, after all, popular in the moral responsibility literature to distinguish between direct and indirect moral responsibility. The drunk driver is the stock example here; if the driver was drunk enough, he might not satisfy the necessary moral responsibility conditions when running over a poor pedestrian. He lacked reasons-responsiveness, he did not express hatred or indifference towards the pedestrian through doing what he did, and he was not able to stop or swerve in time. However, if he satisfied the relevant conditions when deciding to drink all that alcohol, he can still be *indirectly responsible* for killing the pedestrian. There is some debate over whether a satisfying moral responsibility theory really need to include this kind of tracing (King 2011; Khoury 2012), but I will assume for the sake of argument that indirect responsibility of this kind exists.

Might Braverman and Mr. X have indirect moral responsibility for obeying the test leader and living a life of crime respectively? That is, was there some earlier point in time at which Braverman, while satisfying relevant moral responsibility conditions, created for himself a character so impressed by authority figures that he failed to even think about a refusal to obey when finding himself in Milgram's experiment? Was there some earlier point in time in which Mr. X, while satisfying those conditions, made himself so narrow-minded that he failed to see a different way of life than his own as a real option? This seems unlikely. Most people were very surprised by the results of Milgram's obedience experiments when they were first published; no one suspected that ordinary people were so susceptible to these kinds of pressures. There was thus no reason for Braverman, before the experiment, to suspect that he suffered from some kind of character failure that he ought to work on. And when we look at Mr. X's tragic background, it seems highly improbable that there ever was a time when a morally responsible Mr. X made himself irrational regarding his future. Thus, it is unlikely that either Braverman or Mr. X have *indirect* moral responsibility for failing to refuse to continue with the experiment and failing to abandon a life of crime respectively. We ought to conclude that their responsibility really was undermined or at least diminished.

It is, of course, possible to flip my line of argument on its head. It might be objected that it is intuitively obvious that people like Braverman and Mr. X are fully morally responsible for

what they do, and that there is a lot of responsibility to go around. If our best theories about moral responsibility imply that this is not the case, so much the worse for our theories. Fischer and Ravizza's thesis that moral responsibility requires only weak reasons-*reactivity* seems to be motivated at least in part by their intuition that *akratic* agents are obviously responsible for what they do. Likewise, someone might want to tweak the reasons-receptivity conditions, or conditions related to the ability to do otherwise or quality of will, in order to ensure that agents suffering from IROE still come out as fully responsible on their preferred theory. One person's modus ponens is another person's modus tollens. I cannot, within the confines of this paper, lay out in detail what such a strategy would look like and then refute it. I will merely conclude by saying that I strongly believe that we must keep an open mind and be willing to accept that more agents than we previously thought ought to be fully or partially excused, if this is something that can be inferred from independently plausible moral responsibility theories.

7 Conclusion

I have, in this paper, identified a hitherto unexamined kind of practical irrationality: Irrational option exclusion, or IROE for short. I have defined IROE in the following way: An agent suffers from IROE when she fails to even consider φ-ing (an action, sequence of actions or way of life other than the one she performs or lives) as an option despite the fact that

a) φ is far better, according to her own values, than the action or actions she ends up performing or the kind of life she ends up living.
b) she is dissatisfied with the action or sequences of actions that she ends up performing or the kind of life she currently lives
c) the possibility of φ could be easily and quickly deduced, by a person of average logical competence, from facts that the agent knows.

I have argued that regardless of whether we subscribe to the thesis that moral responsibility requires reasons-responsiveness, the ability to do the right thing or that the agent expresses her quality of will through the action she performs, there is a strong case to be made for IROE (at least insofar as it is not the agent's fault that she suffers from it) at least diminishing moral responsibility. Furthermore, we do not know how common IROE is. For all we know, it is possible that it is a fairly common phenomenon, and that there is therefore less moral responsibility to go around than compatibilists and libertarians tend to assume.

Acknowledgements I want to thank the participants in the CELAM seminar (the Center for Ethics, Law and Mental Health at the University of Gothenburg) and participants at the Moral responsibility and consciousness workshop 2015 at the VU University of Amsterdam for helpful comments.

References

Baumeister RF, Tierney J (2011) Willpower: rediscovering the greatest human strength. Penguin Press, New York

Doris J (2002) Lack of character. Cambridge University Press, Cambridge
Fischer JM, Ravizza M (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press, New York
Frankfurt HG (1969) Alternate possibilities and moral responsibility. J Philos 66:829–839. https://doi.org/10.2307/2023833
Hagger MS, Chatzisarantis NLD, Alberts H, Anggono CO, Batailler C, Birt AR, Brand R, Brandt MJ, Brewer G, Bruyneel S, Calvillo DP, Campbell WK, Cannon PR, Carlucci M, Carruth NP, Cheung T, Crowell A, de Ridder DTD, Dewitte S, Elson M, Evans JR, Fay BA, Fennis BM, Finley A, Francis Z, Heise E, Hoemann H, Inzlicht M, Koole SL, Koppel L, Kroese F, Lange F, Lau K, Lynch BP, Martijn C, Merckelbach H, Mills NV, Michirev A, Miyake A, Mosser AE, Muise M, Muller D, Muzi M, Nalis D, Nurwanti R, Otgaar H, Philipp MC, Primoceri P, Rentzsch K, Ringos L, Schlinkert C, Schmeichel BJ, Schoch SF, Schrama M, Schütz A, Stamos A, Tinghög G, Ullrich J, vanDellen M, Wimbarti S, Wolff W, Yusainy C, Zerhouni O, Zwienenberg M (2016) A multilab preregistered replication of the Ego-depletion effect. Perspect Psychol Sci 11:546–573. https://doi.org/10.1177/1745691616652873
Haji I (2012) Reasons debt to freedom. Cambridge University Press, Cambridge
Jeppsson S (2016) Reasons, determinism and the ability to do otherwise. Ethical Theory Moral Pract 19:1225–1240. https://doi.org/10.1007/s10677-016-9721-x
Kennett J, McConnell D (2013) Explaining addiction: how far does the reward account of motivation take us? Inquiry 56:470–489. https://doi.org/10.1080/0020174X.2013.806133
Khoury AC (2012) Responsibility, tracing and consequences. Can J Philos 42:187–207. https://doi.org/10.1353/cjp.2012.0017
King M (2011) Traction without tracing: a (partial) solution for control-based accounts of moral responsibility. Eur J Philos 22:463–482. https://doi.org/10.1111/j.1468-0378.2011.00495.x
Langlands R et al (2009) Applying the good lives model to male perpetrators of domestic violence. In: Lehmann P, Simmons C (eds) Strengths-Based Batterer Intervention. Springer, New York, pp 217–235
Levy N (2011) Hard Luck. Oxford University Press, New York
Mele A (2014) Free. Why science hasn't disproved free will. Oxford University Press, New York
Milgram, Stanley (First edition 1974. This edition 2010) *Obedience to Authority. An Experimental View.* New York: HarperCollins Publishers, Inc.
Nelkin D (2005) Freedom, responsibility and the challenge of Situationism. Midwest Stud Philos 29:181–206. https://doi.org/10.1111/j.1475-4975.2005.00112.x
Nelkin D (2011) Making Sense of Freedom & Responsibility. Oxford University Press, New York
Nelkin D (2016) Difficulty and degrees of moral praiseworthiness and blameworthiness. Noûs 50:356–378. https://doi.org/10.1111/nous.12079
Pearce J, Down B, Moore E (2008) Social class, identity and the 'good' student: negotiating university culture. Aust J Educ 52:257–271. https://doi.org/10.1177/000494410805200304
Pettit P (1984) Satisficing consequentialism. P Aristotelian Soc 58:165–176. https://doi.org/10.1093/aristoteliansupp/58.1.139
Slote M (1984) Satisficing consequentialism. P Aristotelian Soc 58:139–164. https://doi.org/10.1093/aristoteliansupp/58.1.139
Smith A (2008) Control, responsibility and moral assessment. Philos Stud 138:367–392. https://doi.org/10.1007/s11098-006-9048-x
Strawson, Peter. (1962) Freedom and resentment. *Proceedings of the British Academy*. 48:1–25. Reprinted in P. Russell O. Deery, eds. 2013. *The Philosophy of Free Will. Essential Readings from the Contemporary Debate:* 63–83. New York: Oxford University Press
Streumer B (2007) Reasons and impossibility. Philos Stud 136:351–384. https://doi.org/10.1007/sl 1098-005-4282-1
Vargas M (2013) Building better beings. A theory of moral responsibility. Oxford University Press, Oxford
Vihvelin K (2013) Causes, Laws & Free Will. Why Determinism Doesn't Matter. Oxford University Press, New York
Von Hirsch A (1996) Censure and sanctions. Oxford University Press, New York
Watson, Gary (1987/2013) "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme". Reprinted in Russell, Paul and Deery, Oisín, eds. 2013. *The Philosophy of Free Will. Essential Readings from the Contemporary Debate*: 84–113. New York: Oxford University Press
Wolf S (1990) Freedom within reason. Oxford University Press, New York
Zimmerman M (2007) The good and the right. Utilitas 19:326–353. https://doi.org/10.1017/S0953820807002622