Answering for the past: Exploring the conditions of answerability over time

NICOLE RAMSOOMAIR *McGill University*

ABSTRACT: Whether seen after a vehement denunciation of one's former values or a subtle change of heart, it is often thought that significant change to one's evaluative character could undermine responsibility for past wrongdoing. In this article, I explore this intuition by analyzing Angela Smith's concept of "responsibility as answerability." I introduce an alteration/replacement distinction to define the limits of answerability over time. These limits are then further qualified by drawing on Delia Graff's work on Sorites type cases to argue that persons are answerable for past wrongdoing if they remain "saliently similar" in some relevant respects

RÉSUMÉ : On pense souvent qu'un changement significatif de notre caractère évaluatif pourrait saper notre responsabilité par rapport à nos méfaits passés. Dans cet article, j'explore cette intuition en analysant les concepts de responsabilité et de l'obligation de rendre des comptes, tels que présentés par Angela Smith. J'introduis une distinction entre « modification » et « remplacement » pour définir les limites de la responsabilité au fil du temps. Ces limites sont ensuite précisées en s'inspirant des travaux de Delia Graff sur les arguments Sorites pour soutenir qu'une personne est responsable d'actes répréhensibles passés si elle demeure « manifestement similaire » à certains égards.

Keywords: moral responsibility, answerability, social philosophy, character change, Sorites paradox, salient similarity

***Dialogue* 60 (2021), 359–377**

doi:10.1017/S0012217321000081

1. Introduction

On 1 December 2005, one of the founding members of the notorious street gang, "the Crips," Stanley "Tookie" Williams, was put to death. Governor Arnold Schwarzenegger rejected an executive appeal for clemency to commute his death penalty sentence to life in prison. There was no new evidence to back any suggestion that anyone other than Williams had committed the violent murders. The governor detailed Williams' crimes and included a "strong and compelling" list of evidence that left "no reason to second guess the jury's decision of guilt or raise significant doubts or serious reservations …" (Schwarzenegger, 2005, p. 3). However, when Williams' supporters yelled, "The state of California just killed an innocent man!" in the witness media room after execution, they did not refer to a mistake within the judicial process itself (Warren & Dolan, 2005, para. 5). Williams' lawyers and supporters alike suggested that Williams was worthy of clemency due to the changes in his evaluative character while he was in prison. The outrage that followed his death sprang from the notion that the man facing the sentence was recognizably different from the one who committed murder in 1979 in ways that seemed to mitigate continued responsibility attribution. While on death row, Williams renounced his former gang affiliations and actively spoke against them. His work against gang violence through ongoing community projects and authoring children's literary works eventually led to a Nobel Peace Prize nomination.

Williams' execution was undeniably controversial. Schwarzenegger aimed to answer whether Williams' redemption was "complete and sincere" or "just a hollow promise" (Schwarzenegger, 2005, p. 5). However, my question in this article is not about the details of this case or how to determine the veracity of Williams' testimony. Instead, I want to explore the claim of innocence made by Williams' supporters and ask whether shifts in one's evaluative character could ever warrant mitigation of continued responsibility attribution. These individuals seem to no longer be able to inhabit the lives they once lived, psychologically speaking. Their attitudes and judgements have fundamentally changed. I will explore this claim by exploring Angela Smith's (2005, 2008, 2012, 2015a, and 2015b) concept of "responsibility as answerability" and I will analyze the conditions of answerability over time (Smith, 2015b, p. 99). I will argue that, if sufficiently qualified, there is some truth to the claim made by Williams' supporters. The conditions for continued answerability diminish when there are significant changes to one's evaluative character. Moreover, I also suggest that factors such as profound repudiation or a lack of evaluative coherence — usually cited in such cases — do not excuse but merely provide evidence of the kind of relevant similarity needed for temporal comparison of the person's evaluative beliefs. When this similarity holds, so do the conditions for answerability.

I will unpack this claim as follows: in Section One, I outline what it means to be answerable as a foundation for my analysis. Section Two will see examples of

various kinds of change drawn from the novel and movie, *A clockwork orange,* by Anthony Burgess (1962/1995) and Stanley Kubrick (1971). I suggest that, despite what might amount to a loss of responsibility, each scenario introduced provides a clear example of continued answerability. In Section Three, I explain each example by introducing an alteration/replacement distinction to determine the parameters for when answerability no longer holds. Section Four will then qualify this claim by drawing on arguments from Sorites-type cases and the work of Delia Graff. Changes to one's moral character only mitigate responsibility when specific aspects of one's moral constitution have been replaced. Thus, answerability holds over time if persons remain "saliently similar" in some crucial respects (Graff, 2000, 68). Section Five explores some implications within the context of the earlier examples. Section Six further refines the account by distinguishing it from attribution theories that focus on identification and integration. Finally, I apply the full analysis of answerability over time to Williams' case. I argue that, although he may indeed remain answerable, this fact does not necessarily justify his punishment.

2. Responsibility as answerability

With a few qualifications, Smith's rational relations view might support the claim made by Williams' supporters due to its deep connection between responsibility attribution and an assessment of character. Like other attributionist theories of responsibility, Smith's view not only attends to the exercise of the will in the agent's actions or choices themselves. Instead, attributions of moral responsibility must reflect one's moral character. She argues, "moral praise and blame, unlike assessments of beauty or native intelligence, seem to go beyond mere unwelcome description, and involve something that might be called a reactive entitlement" (Smith, 2008, p. 380). (Smith, 2008, p. 380). That is, her account does not merely provide a means to assess persons according to some proposed standard as we might with traits such as beauty, intelligence, or even height. Persons can be described as more or less beautiful or intelligent, but this assessment would not warrant reactive attitudes because these attributes do not "reflect [their] own judgment or assessment of reasons" (Smith, 2008, p. 380). If an act or state reflects one's reason when it is "rationally connected" to actions, attitudes, or attributes, then it "makes sense to ask [them] to defend or justify it" (Smith, 2012, p. 579). Otherwise, persons may be excused. For instance, if asked why a person is tall, beyond citing biological facts, no other justificatory answer could be provided because being tall "bears no relation to [one's] own judgmental activity" (Smith, 2008, p. 380). "Rational sensitivity" might then be understood as a conditional of the form: "… if one sincerely holds a particular evaluative judgment, then the mental state in question should (or should not) occur" (Smith, 2005, p. 257, 253). Thus, to hold answerable is a direct challenge to agents *qua* moral agents because it targets only those actions, attitudes, or attributes contingent on their evaluative beliefs.

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

Responsibility as answerability allows persons to be responsible for whims, emotional reactions, and even telling omissions. For instance, if I forgot my partner's birthday, responses, such as 'I did not choose to forget' and 'forgetting was out of my control' would not alleviate criticism because my partner is not asking about the mechanics of forgetting. My partner is asking why I give the birthday a low valuation. Forgetting is counterfactually expressed by the idea that I would have remembered the birthday if I had cared enough. Indeed, the close relationship we share and the care I have for them should have sufficiently raised the chances I would be reminded about the upcoming birthday. Even if it turns out that the forgetfulness was not revealing my valuations and a legitimate excuse can be provided, it remains intelligible to initiate a call to answer as long as the act is rationally connected to evaluative beliefs and attitudes. Persons are answerable if it is at least in principle legitimate to request a response, whereas "rational sensitivity" marks the class of actions open to this demand (Smith, 2005, p. 257).

Note that despite Smith's use of the words "judgment" and "rationally" when speaking of attributability in terms of what is "rationally sensitive to our evaluative judgments," this usage should be taken more as a catch-all and much less rationalistic than it initially appears (Smith, 2005, p. 257). She states, "I want to make clear that the judgments I am concerned with are not necessarily consciously held propositional beliefs, but rather tendencies to regard certain things as having evaluative significance" (Smith, 2005, p. 251). These judgements may involve a wide array of mental activity, including cares and emotional reactions. She uses the word "judgment" despite the potential confusion to denote the stability of the kinds of dispositions that concern her. They are "standing commitments" and not "merely one time assessments" (Smith, 2005, p. 251, n. 27). For the sake of clarity, I would like to rename Smith's judgement sensitivity as 'evaluative sensitivity.'

If certain cares, judgements, whims, or any variety of mental activity depend on one's specific evaluative constitution, persons may be considered answerable for those states and actions or attitudes connected to them. I will call a person's evaluative constitution or collection of evaluative judgements as a whole their 'evaluative profile,' with the beliefs, cares, and judgements constituting that whole, the 'evaluative aspects.' If we ever find ourselves unsure if an aspect is evaluatively sensitive, we can ask: 'If the agent's evaluative activity about the issue at hand changed, would the behaviour, action, or state continue?' Through such a counterfactual question, I propose that we determine whether evaluative sensitivity holds by seeing whether the action or activity would occur regardless of the agent's evaluative profile. If so, it would make as much sense to request a response for such aspects as it would ask why a person is tall. Regardless of one's evaluative profile, the evaluative aspect would occur. If not, it seems to say something about who the person is, which that makes answerability demands appropriate.

Although there are sure to be several objections to thinking of responsibility in terms of answerability, I see Smith's view as an appropriate foundation for its close connection between responsibility attribution and one's moral character. In what follows, rather than exploring these criticisms, I will concentrate on the less-explored answerability conditions over time. In the following section, I will consider several scenarios of character change and what each implies about answerability.

3. The changing "raskazz"[1]

As the question of guilt surrounding Williams' case is controversial, I would like to pull an example from literature and film to provide a clear-cut example of guilt that also involves differing depictions of transformation. I will draw from *A clockwork orange* (Burgess, 1962/1995), and I will focus on the violent protagonist, Alex. A teenager engaged in a criminal life, Alex and his comrades make a habit of committing and revelling in violence. After the murder of an older woman, Alex is sent to prison. While incarcerated, the prison officials offer an experimental procedure, "Ludovico's technique," as a way of minimizing his sentence (Burgess, 1962/1995, p. 30). The technique is a form of associative learning said to rehabilitate Alex in a matter of weeks. As part of the procedure, Alex receives a drug that causes debilitating nausea and is subjected to movies depicting excessive violence. Afterward, an intense sickness follows his murderous impulses and the once aggressive criminal is now subdued. The retention of his evaluative aspects is apparent when Alex is humiliated by being forced to grovel at an attendant's feet:

> Now I knew that I'd have to … get my cut-throat brivita out before this horrible sickness whooshed up …. But, O'brothers, as my rooker reached for the brivita in my inside carman I got this like picture in my mind's glazzy of this insulting chelloveck howling for mercy with red red krovy all streaming out of his rot … and I viddied that I'd have to change the way I thought about this rotten veck. (Burgess, 1962/1995, p. 228)

While Ludovico's technique is operative, at most, we see a criminal with his hands metaphorically tied, whereas who he is and whom he conceives himself to be remain the same. Kubrick's film adaptation vividly depicts this lack of change. Following a reversal of the technique, the closing scene depicts Alex resuming his former fantasies and menacingly telling the audience, "I was cured alright" (Kubrick, 1971). Let us refer to this version of Alex as 'Movie Alex.'

At the end of the novel, in the original United Kingdom edition, Alex seems to experience a second, internally motivated change following a reversal of the

[1] This term can be found in the fictional language, "Nadsat," used by Alex and his fellow gang members. The word "raskazz" may be translated to mean "story" (Burgess, 1962/1995, p. 351).

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

procedure. Alex notices that not only is there a change in his aesthetic tastes, but he also finds himself yearning for domestic life. He is bored with his violent, criminal life and finds himself rather envying an old friend who has made a new life with a wife and child. Alex no longer appreciates the booming concertos that were once a soundtrack to his violent acts, develops a taste for ordinary beer over "moloko," and finds himself strangely annoyed at the lack of sentimentality of his former companions (Burgess, 1962/1995, p. 101). He thinks to himself, "Perhaps I was getting too old for the sort of jeezny I had been leading, brothers. … I felt this bolshy big hollow inside my plott, feeling very surprised too at myself. I knew what was happening, O my brothers I was like growing up" (Burgess, 1962/1995, p. 340). Drawing on his internal dialogue, 'who he is' seems to be changing. If we imagine this initial portrait's conclusion, perhaps we might even see an eventual reversal in Alex's desires. I will call this Alex, at the end of the novel, 'Novel Alex.'

Given the changes Alex undergoes, would he still be answerable? I suggest that, despite the differences between Movie and Novel Alex, both remain answerable. At least, if we think he is answerable for crimes committed in one scenario, arguably, we should also think him answerable in the other. To see this, consider what would happen if Ludovico's technique remained operative for Movie Alex. Let us call this version of Alex 'Cured Alex' to avoid confusion. As a successful application of Ludovico's technique, the process may externally limit his sadistic desires and restrict him from acting on the basis of his evaluative profile. However, debilitating nausea might encourage him to form something like an adaptive preference for domesticity. Eventually, Cured Alex's change might resemble the kind of change Novel Alex undergoes. He may alter his preferences so that he would no longer gravitate toward violence once it is no longer a possibility.

Adapting one's preferences to the situation does not require anything as extreme as Ludovico's technique. The same could be said of a maturing partygoer declining a drink. They may decide to forgo a night out not only due to maturing tastes, but those tastes may be partially formed due to an inability to recover from a hangover like they once did. This physical and evaluatively insensitive aspect of themselves can change their preference to drink, but not in a way such that we would consider them no longer responsibility-apt for that decision. Indeed, the reality of our situation and the world around us may impress on agency not unlike Ludovico's technique if we are given time for it to work. If one is open to a responsibility attribution, then we should think the other is too. Thus, despite invoking different intuitions, the difference between Alex's three versions may only be of degree and not kind. All would remain answerable.

If continued answerability is compatible with a wide variety of changes a person can undergo, when might it no longer hold? It might be illustrative to imagine a fourth way Alex could change. Rather than being subject to forced associative learning, this Alex has his preferences, desires, and evaluative

judgements wiped clean. His evaluative profile is then replaced with a new set through extensive brainwashing. It is not hypnosis that would act on him like some alien force disconnected from his will. It represents the implantation of an entirely new evaluative profile — a comprehensive attitudinal overhaul. Let us call this Alex 'Brainwashed Alex,' to separate him from his counterparts. I argue that Brainwashed Alex would no longer be answerable for actions before the brainwashing because his case constitutes a replacement of one's evaluative profile rather than just an alteration of it. I call this the alteration/replacement distinction, to be explored in the next section.

3. The alteration/replacement distinction

Alterations to one's evaluative profile, as I understand it, might be likened to a change in one's subjective values, as seen in J. S. Mill's work. Mill argued that after having experienced higher and lower pleasure, agents would strongly prefer the former due to an expansion of subjective values (Mill, 1871/2016, pp. 32–34). It is not a loss of one's evaluative aspects but an addition that may cause some rearrangement on what the agent considers most worthwhile. Perhaps, like Alex, after listening to Beethoven, Mill's sophisticated hedonists prefer his "Symphony No. 9" as a soundtrack to their exploits. They now prefer Beethoven, but this does not entirely rid them of their love for burgers and other lower pleasures. They have added to their evaluative profiles without experiencing a loss within them.

Likewise, Movie, Novel, and Cured Alex all remain open to a responsibility attribution because at no point does it remain unintelligible to consider them answerable. I would also argue that part of the intelligibility of requesting a response necessitates an answer to be personal because it involves reasons that the agent currently holds. Past actions may remain connected to the agent in the present due to being part of their history. However, the answerability test requires more than an explanation of events that occurred if it speaks to 'who' the agent is and not only 'who' they used to be. For instance, if responsibility attribution relied on choice somewhere in the lineage of the action or attitude, the assessment would be detached from what is being assessed. As Smith argues, this sort of strategy "forces us to view our own attitudes as the mere products of our own actions" (Smith, 2015a, p. 127). One does not usually praise another's act of goodwill because, 20 years ago, that individual adopted a mantra of repeatedly 'paying it forward' that only now has become second nature. There is a tenuous connection between the act and the individual's current state that "fails to capture the special fact about attitudes, which is that they are judgment-sensitive responses to the world around us" (Smith, 2015a, p. 127). The response would be far removed from the agent's evaluative character in the present, and it would serve as a mere explanation of why these events occurred. Thus, when answerability includes a personal connection, it tells a more connective story for responsibility attribution by joining wrongdoing

with how agents interact within the world, as they currently stand, not who they once were.

The alteration/replacement distinction might be compared to a recipe.[2] When there is additive change, we are still working with the same ingredients as before, even if the amounts are altered, and new ingredients are added. Just as tomato soup with more basil and onion is still a tomato soup, an altered constitution still shares many original ingredients. Consider Movie Alex, who looks fondly at past violent crimes with a sense of yearning and anger at his inability to continue his criminal life. He remains the same because his criminal desires are outweighed but not necessarily replaced. A change in circumstance — including the reversal of the Ludovico technique — could reignite an appetite for violence because his circumstances, rather than his constitution, stop him. The reason he desisted has no connection to his evaluative constitution. The recipe remained the same.

Similarly, Novel Alex would also be open to responsibility attribution because the subtle alteration he undergoes can be likened to normal maturation as the gradual ebb and flow of altering and rearranging one's values. He would also remain answerable as the changes are additive to the original recipe, at least at this point. This slowly maturing Alex may still remember the excitement he once "viddied" when engaging in violence, and he may even feel a certain sense of nostalgia when passing the "Korova Milkbar" (Burgess 1962/1995, p. 11). Novel Alex may not be entirely sympathetic to his past but, at this point, I doubt that Alex has undergone a loss of evaluative aspects even if such loss may be pending.

Like making soup from scratch, varying the proportions amounts to additive change, but it is the same type of soup, generally speaking. If we start replacing the ingredients, we might begin to question whether it is the same type of soup. Contrast these instances of alteration with Brainwashed Alex. He is different from the previous three, not merely because he was brainwashed, but because of what brainwashing does to his evaluative profile. He may have been answerable at the time, and the action itself was motivated by evaluatively sensitive influences and judgements. However, just as it would be inapt to call a bland soup spicy, once those former evaluative aspects are lost and no longer influence who he is, any demand for an answer will be met with an explanation without reference to the evaluative aspects he currently holds. Any response given by Brainwashed Alex is far removed from his current constitution and provides only an explanation of how he came to be this way.

[2] This analogy is derived from Marya Schechtman's work. She compares the composition of one's perspective or self to a "stew or soup" as "each ingredient contributes to the flavor of the whole and is itself altered by being simmered together with others" (Schechtman, 1996, p. 149).

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

A couple of questions remain that reveals a weakness of this distinction: What if the brainwashing was not complete? What if it only changed some aspects of his evaluative profile? Would he still be responsible if evaluative aspects similar to those he had at the time of the crime were added? These questions show that, while the alteration/replacement distinction provides an essential foundation, it remains too general to inform more ambiguous cases. In the following section, I will address this problem and, in the process, define the threshold necessary to maintain answerability over time.

4. Potential problems

Suppose that a soup recipe is continually altered over time. If I gradually used less basil or swapped it out for another herb, would we still have the same tomato soup? Does it count if there is only a pinch of basil left? What about a smidgen? If not, when exactly did the soup change? The problem is that I have not made it clear whether answerability could fail to persist even in the absence of wholesale loss or replacement of evaluative aspects. The alteration/replacement distinction neglects the much more common gradual day-to-day change typically experienced in a lifetime.

How then might we understand subtle, responsibility-undermining change over time? Perhaps we could take a page from Derek Parfit's book and argue that responsibility overtime should be understood by degree. He states:

> When some convict is now less closely connected to himself at the time of his crime, he deserves less punishment. If the connections are very weak, he may deserve none. This claim seems plausible. It may give one reason why we have Statutes of Limitations, fixing periods of time after which we cannot be punished for our crimes. (Suppose that a man aged ninety, one of the few rightful holders of the Nobel Peace Prize, confesses that it was he who, at the age of twenty, injured a policeman in a drunken brawl. Though this was a serious crime, this man may not now deserve to be punished.) (Parfit, 1986, p. 326)

Parfit's convict closely resembles the example that began my inquiry. Further, his solution seems right, at least to an extent. Responsibility seems to admit of degrees. However, while the notion of survival could very well be best served by something like Parfit's account of psychological continuity and connectedness (I remain agnostic about this), his concerns are much broader than mine. Parfit is primarily concerned with the different questions of re-identification, identity, and survival. As a result, the quantitative solution he suggests is insufficiently fine-grained to capture the persistence of answerability/responsibility taken in isolation. Thus, to draw on some of Parfit's insights and shift the focus to the persistence of one's evaluative aspects, the emphasis needs to be placed on whether the evaluative aspects can be retained by degree. The quantitative solution translated to a question of responsibility would ask whether one's evaluative constitution could be understood as having a quantitative

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

threshold beneath which the convict would no longer be open to responsibility attributions.

Compare the shared experience of gradual but steady change to a Sorites paradox concerning predicates. Words like 'bald' are thought to be vague predicates that carry no clear and universal conditions for satisfaction. When a single hair is lost every day until a person is bald, it is not clear when the transition occurred. Like evaluating the number of individual strands required to apply the predicate 'bald,' it is likewise challenging to define a threshold point in which responsibility attribution no longer holds. If Alex lost an evaluative aspect here and there until his evaluative profile was entirely replaced, when did he become no longer answerable? The complications here involve two related problems: First, the problem of triviality asks whether the retention of certain evaluative aspects rather than others is more important in determining responsibility. Second, the problem of degree asks about possible limitations of continued responsibility attributions over time due to the loss of several evaluative aspects. I suggest that we employ an analogous strategy to Graff's proposed solution to the Sorites paradox to resolve both issues.

4.1 Problem of triviality

Andrew Koury and Benjamin Matheson argue that what matters for blame is not the extent of change, but what changes. Although persons readily change throughout life, when one's "distinctive psychological features," as those features essential to performing a specific act, are replaced, the grounds for blameworthiness are undermined (Koury & Matheson, 2018, p. 214). All other changes may be trivial. For instance, concerns about predictability may inform criminal treatment because knowing whether one's previous evaluative profile has undergone a replacement provides a reasonable, pragmatic ground to determine the likelihood of recidivism. If a crime involved a release of animals for medical testing, then the belief in animal rights would be relevant in ways it would not be if the agent had embezzled funds. The evaluative aspects in question must be relevant to the line of inquiry.

The focus on relevance might also inform continued answerability. If there is "no lasting trace of the 'springs of action' that gave rise to" the act in question, then any answer given would be impersonal and detached from who the person is now (Koury & Matheson, 2018, p. 214). To no longer be answerable, Alex may not need to go through any radical or complete change. A loss of answerability is not responsive to all kinds of character change, but only those changes that relate to the reasons for interrogating the agent.

4.2. The problem of degree

Focusing on relevance helps to answer the problem of degree by setting the parameters for when a threshold of evaluative aspects is met. My argument will be like Graff's in that continued answerability "rests on the idea that two things that are qualitatively different in some respect, even when they are

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

known to be different, can nonetheless be the same for present purposes" (Graff, 2000, p. 67). The reasons for defining a boundary determine whether an individual is the same for present purposes. When an agent is similar enough for a particular purpose or "*saliently similar*," this marks the parameters needed to satisfy a claim to answerability based on the constitution of one's evaluative profile (Graff, 2000, p. 68). Thus, if we asked whether agent A was answerable for action B, we would ask of the significant evaluative aspects X, Y, and Z and whether they remain to an agreed-upon threshold (where it might be enough to retain X and Z if not Y, or X and Y and not Z). If so, A displays enough salient similarity for present purposes, and in being similar enough, A remains answerable.

As similar to the head of a balding man, there is indeterminacy between two clear-cut cases. At the one end, we have Alex as a violent teenager, while at the other, a potentially reformed man with a whole set of new beliefs and attitudes. In between, we have someone who is neither entirely the same nor entirely different. Alex would only remain answerable if his current outlook is similar enough to his criminal self for our purposes, but similar enough need not be incredibly precise. Importantly, no exact amount of evaluative aspects determines whether answerability holds; it just needs to be reasonable to think the agent will approach the world in the same manner as before. Here, I find Graff's analogy to coffee making instructive. She argues that we are unconcerned with the exact proportion of grains within two scoops because "my coffee-making purpose permits me to behave as if the two amounts were the same since the purpose is in no way thwarted by my behaving as if they were the same … " (Graff, 2000, p. 67). Coffee making allows for a fair amount of variation without being "boundaryless" (Graff, 2000, p. 48). It is tolerant of a few spilled grains here and there before becoming too weak, strong, satisfying, or unsatisfying. As far as coffee making is concerned, the scoop with the few grains spilled and the one without is both what she calls "live options" to achieve my purposes (Graff, 2000, p. 68). Likewise, there may be some interest-relative threshold to be met when determining answerability, but one that is generally tolerant of variations in satisfaction conditions. Thus, continued answerability does not require the exact recreation of evaluative aspects, but a general cluster of them relevant enough to the act in question that it would satisfy a claim of salient similarity.

Determining which aspects should be under consideration and determining the threshold to be met complicates responsibility attribution. How is it even possible to know which evaluative aspects motivate an agent? Would A be saliently similar if evaluative beliefs X and Y were replaced, but not Z? Does X or Y even influence the agent in ways that should concern us? These epistemological concerns are representative of some larger methodological questions that I can acknowledge but not fully address here. Part of the problem is that these questions might never have clear answers because whether one is answerable is primarily informed by the circumstances and the nature of the offender's

wrongdoing. There may be no threshold point or degree of similarity that would hold for all cases, but this is not necessarily disadvantageous. Rather, it frames such attributions as a much more nuanced pursuit than traditionally and historically thought. For instance, if the crime is serious, greater care may be taken to determine what is meant by 'similar enough' than if one were answerable for a forgotten birthday. Vague or general resemblance might matter in one instance, where it might be dismissed in another. Persons may be more or less responsible in a way that requires careful consideration to better represent the complexities involved in cases of radical character change, while also drawing our attention to the role that interpretation plays in responsibility attribution.

5. Implications

The focus on salient similarity thus produces some interesting results. Just as much as having a slightly bigger or smaller scoop of coffee will not undermine one's particular interest in making coffee, it is not the exact recreation of Alex's current beliefs that determines responsibility attribution. Instead, responsibility requires a general cluster of similar evaluative aspects concerning the relevant criminal preferences that would satisfy a salient similarity claim. Alex might be answerable despite radical change. It is also possible that answerability might similarly diminish due to smaller-scale changes.

Indeed, responsibility attribution does not require an extensive change to one's evaluative character. One may be generally unrecognizable and have otherwise undergone a radical change but may still be considered responsible if salient similarity holds. For instance, I doubt that the prison administration would be too pleased if Movie Alex were radically changed in several respects but still maintained sadistic impulses. His love of violence is significant in a way that other aspects are not, and if that persists, he will remain open to continued responsibility attribution.

A similar result holds in the other direction as well. The loss of continued responsibility attribution might occur without profound change. If the hints at reform prove lasting, Novel Alex may continue to mature into domestic life and away from the violent desires that once characterized his outlook. He may nevertheless retain many of his old preferences and preserve his character more generally.

We might also describe Cured Alex as no longer answerable. After forming adaptive preferences that push him away from his violent desires and guide a reluctant reformation, we could potentially call the procedure a success even though he underwent little change. We might be hesitant to describe him in such terms, but I do not think this is because responsibility attribution still holds. Instead, intuitive doubt might be caused by the precariousness of the change. Unlike Novel Alex, there was no effort to change on his part and the more similar he is to his old self, the less reason we have to think this change is lasting. He may not revert to his former violent self as quickly as his Movie counterpart if the physical restraints were removed, but the possibility is there in a way it would not be if he had changed due to his own volition.

By contrast, Novel Alex inspires confidence because, as he changes, so do whole swaths of his evaluative character that could support this change. This confidence might even be heightened if were to engage in activities that speak to a possible redemption. While he remains responsible, these acts signal potential change. We might even be confident enough to justify more lenient treatment even when he remains similar enough to remain answerable to some extent. Equally, this confidence would be undermined if Novel Alex's former evaluative profile simply atrophied and further change was only the result of mere happenstance, rather than any effort to change. Any changes he undergoes would be described as precarious at best.

Brainwashed Alex's situation is more complicated than the others. If having received a complete mental wipe of his evaluative aspects, he would not be saliently similar. However, if he was implanted with the same evaluative beliefs or those closely resembling his former beliefs, we might consider him open to responsibility attribution despite their usual formation. It may not be his fault that he is the way he is, and the brainwashing may mitigate applying punishment if that punishment were simply a means of deterrence. However, the fact that brainwashing occurred does not necessarily undermine salient similarity as I have framed it. There would still be good reasons to treat this newly implanted person as the same, given that he would exhibit the same criminal tendencies requiring the same sort of rehabilitation as if this mental tampering never occurred. He remains answerable, despite the lack of choice in the formation of those evaluative aspects. After all, it is the same soup with the same recipe but made in a different pot. This response may raise further questions concerning personal identity, but insofar as attributability is concerned with warranted requests for a response, my inclination would be that answerability holds unless we can give a good reason to think that bodily continuity is a necessary condition for this attribution. If so, then the usual means of formation might act as an excusing condition. Likewise, if Alex should accidentally suffer a loss in his evaluative profile, let us say to head trauma or a similar accident, it is possible that he would no longer be answerable. We may doubt the permanence of the change due to a lack of evaluative scaffolding that might otherwise be present in a gradual change. Even so, it may be the same pot, but with a different recipe altogether.[3]

6. Complexity of persons

Notice in the above examples that determining salient similarity involves a direct comparison and does not require repudiation or alienation to say whether answerability has been retained or lost. Indeed, I see it as support for this account that it recognizes the complexity and capriciousness of one's evaluative

[3] I would like to thank one of the anonymous referees for pointing out the issues I address in this section.

https://doi.org/10.1017/S0012217321000081 Published online by Cambridge University Press

constitution. Before reviewing Williams' case, I will now briefly explore these implications to argue that persons may repudiate the past and be deeply conflicted psychologically speaking without undermining an attribution of responsibility. As I suggested when considering the reasons for change, continued answerability provides the base for an interpretation, whereas factors such as integration, and repudiation are more prognostic.

Marya Schechtman, who once argued for a loss of self by employing affective discontinuity, has more recently argued (Schechtman, 2016) that states of deep shame and repudiation cannot represent a loss because they require conflict. She argues that experiences such as religious conversion do not constitute a loss of sinful impulses but a "rejection of them" (Schechtman, 1996, p. 26). The sinner does not feel the "need to give weight to former impulses" and is estranged from those experiences' past affects (Schechtman, 1996, p. 26). Despite affective dissociation, the desire for sin creates the internal conflict that could arguably form the basis for the feeling of shame. To feel these extremes of shame and repudiation, I would even go as far as to argue that it requires the person who "inhabits the point of view of the convert still has the point of view of the sinner in her experiential repertoire" (Schechtman, 1996, p. 26). It is possible that Novel Alex's desire for violence still occupies a space in his evaluative profile, even if it is unlikely that he would act on such preferences due to potential feelings of shame. If it occupies space, then we can say that he has undergone an alteration, not a replacement. The difference between one who matures and one who grows to repudiate the past is relatively thin. Both are additive changes, with the primary difference being with the sort of attitude expressed afterwards. Be it pride, nostalgia, or shame, simply because one causes negative affect does not make the past any less attributable. There is no loss as the "… convert does not claim to no longer be a sinner, but only to repudiate his sin" (Schechtman, 1996, p. 26). As shown with Brainwashed Alex, it might only make a difference in the level of confidence one has with those changes.

Responsibility attribution might also hold when the agent is acting on apparently lesser evaluative aspects. For instance, there may be evaluative aspects that form part of one's profile as a whole but are not readily attributable to the agent. I may inadvertently laugh at an off-colour joke for which, as Smith argues, "it seems not only morally objectionable to *express* amusement; it seems morally objectionable, and blameworthy, to *be* amused" (Smith, 2015a, p. 118). I may argue that it is generally not like me to laugh at such jokes, and I usually do not. The inadvertent laugh does not cohere with my more settled and integrated disposition, and, admittedly, I would not want to be identified with a lesser part of myself.

As Neil Levy and others have argued, the problem is that, for an action to express one's evaluative profile, it must express their "global perspective on what matters morally, and not a single attitude or a set of attitudes that falls short of constituting the agent's evaluative stance" (Levy, 2011, p. 256). Levy argues that a single controversial attitude, evaluatively sensitive whim,

or accidental lapse inadvertently manifesting itself in one's actions does not express the agent's evaluative profile "in the right way" because these do not constitute attitudes adequately filtered through the agent's broader moral beliefs (Levy, 2011, p. 251). The attitude is "*too alien* to the self to ground moral responsibility" (Levy, 2017, p. 14). If he were right, then this would mean that evaluative sensitivity alone (as I have suggested) does not determine answerability because the attitude may be "crucially at odds with the states with which we can most securely identify the agent" (Levy, 2017, p. 14).

Likewise, Nomy Arpaly and Timothy Schroder defend what they call the "Whole Self" view of attributability. Their view amounts to the claim that "[O]ther things being equal, agents are praiseworthy (or blameworthy) for the good (or ill) they do to the extent that the morally relevant beliefs and desires which led them to act were well-integrated (assuming that the act met some very minimal standard of integration)." (Arpaly and Schroder, 1999, p. 175).

If Levy and Arpaly and Schroder are correct, then evaluative sensitivity and alteration may be too thin for answerability. However, as profoundly connected to character, answerability is concerned with 'who the person is,' not who that person *mostly* is, and internal conflict is quite common. A begrudging misanthrope, not unlike the one seen in Kant's *Groundwork for the metaphysics of morals* (1785/2005), may even deserve acclaim for acting on a morally praiseworthy whim despite not having clear motivations to do so. More importantly, it still seems intelligible to request a response as to why he did what he did. As David Shoemaker writes, "integration, which is just about the relationship between psychic elements, has no obvious connection to mattering" (Shoemaker, 2005, p. 136). Again, as with brainwashing and repudiation, I suspect a lack of integration is often considered to mitigate responsibility due to this predictive function. For example, being a product of deeply integrated identification may make it the case that the belief or attitude is featured more often in decision-making. A more isolated attitude may even be more likely to fade into indifference and strain attribution, but not so much as to rend apart the connective thread to the current evaluative profile. It matters, but not in the way it is usually framed.

Furthermore, I also find that the importance of integration is overstated: we tend to think of ourselves as more unified — not just in our aims, but also in our psychological constitution — than we actually are. Whole-self views and an emphasis on integration assume that normal functioning corresponds to an ideally rational, coherent, smoothly functioning agent. However, the reality is far messier, and internal conflict is quite rampant. The focus on relevance shows that we need not consider the entire evaluative profile because the lesser aspects could potentially be what is most relevant in the circumstances. Some evaluative aspects may represent a less integrated aspect of our character, but being a lesser aspect would not necessarily undermine salient similarity, even if it gives reason to think such similarity is likely to fade in the future.

7. A claim of innocence

If continued responsibility attribution is not undermined by repudiation and internal conflict, does this deny the claims made by Williams' supporters? Did answerability, and thereby responsibility attribution, hold? To this question I would answer, quite unsatisfyingly (at least initially), 'it depends.' Responsibility and continued responsibility attribution are complicated and they depend on several facts concerning the kind of change that Williams underwent, and the facts of this case are not entirely clear. Due to his sentence's finality, like the ending of *A clockwork orange*, we will not see what could have eventually happened to Williams. All we have is Williams and his supporter's claims of transformation and the state's apparent denial of it.

On the one hand, there is an argument that Williams was quite similar to the criminal who first entered the prison. Law enforcement officials and victims' rights activists argued that Williams' changes were overstated. After all, he had numerous violations in San Quentin and this behaviour seemingly implicated him as the same violent criminal he once was. One persistent reason that was given to deny him clemency was that his behaviour in prison was riddled with citations and aggression, but also that, until the end, he refused to inform on his former gang associates. Schwarzenegger questioned his change as "hollow" because Williams had remained "loyal to the gang member street code of ethics" and refused to be debriefed by the prison authorities and provide information on how the gang operated (Schwarzenegger, 2005, p. 3). In a "60 Minutes" interview, Williams stated, "I have to say that the word 'debriefing' is a euphemistic term for snitching. And my convictions won't allow that" (Leung, para. 33, 2004). Williams' actions and responses to investigators seemingly show him to be the same self, constituted by the same sort of motivations and values.

On the other hand, it is not entirely clear that Williams is *saliently similar* enough to render him answerable. Despite remaining loyal to some of the values from his life before prison, Williams also underwent much personal change. His work with similarly situated youth and how he considered himself to be transformed raised questions of his continued guilt. However, as Koury and Matheson argue, while redemptive acts can serve as evidence for profound psychological change, "what matters is a relevant change to the psychology itself" (Koury & Matheson, 2018, p. 219). To this extent, Williams seemed conflicted about his past values and exhibited optimistic narratives to reframe his future. Despite retaining some loyalty to his former criminal life, Williams also considered himself to be "a student of sociology and psychology" and memorized words in the dictionary to improve his vocabulary (Williams, 2004, p. 224). He would draw portraits and sketch family and famous figures in a process that he claimed to have a "halcyon affect" as a means of "calming the beast within" (Williams, 2004, p. 224). This process and a newfound inner calmness moved him to write books that targeted at-risk youth. Those arguing

on Williams' behalf might also point to his repudiation of former gang associations as evidence for a change in the relevant evaluative aspects. Williams wrote, "In a cold sweat I shook myself out of this awful reverie, consumed by sadness — not for Crippen, but for the lives of all the Crips who had died, for the innocent black lives hurt in the crossfire, for the decades of young lives ruined for a causeless cause" (Williams, 2004, p. 279). The change seen here may represent an alteration that occurred with "day-to-day improvement" (Williams, 2004, p. 295). Many of his evaluative beliefs shifted priority or started to wither away. Nevertheless, an analysis of continued responsibility would not reduce to a single continued trait or evaluative aspect, but a determination including whether the myriad of evaluative aspects he retained is sufficient for determining similarity.

Williams felt shame about his criminal life and experienced inner conflict; he was loyal to his former gang associates but repudiated this loyalty at the same time. Consequently, the changes Williams underwent should be described as alteration rather than replacement. It may be the case that his newfound passion for writing, art, and the spoken word is telling of the potential and eventual loss of answerability, but until that happens, it seems arguable that there is sufficient salient similarity to his former self.

Thus, Williams may be genuinely conflicted and experience a deep affective break with his past, but remain open to responsibility attribution for his past crimes. Neither change of perspective, internal conflict, nor even fierce repudiation guarantees a loss of answerability. I argued that these factors confuse the question of responsibility attribution with potential recidivism and are based on an ideal of human psychology. Williams was nevertheless changed, and even if he may be said to be answerable, remaining answerable would not necessarily justify his punishment. If his acts of service, repudiation of the past, and internal conflict are indeed heuristics that help to predict future absolution, then there is good reason to think that a loss of salient similarity is foreseeable if not likely. The change he underwent could have, quite possibly, continued until there was a complete replacement if his appeal for clemency had been granted.

8. Conclusion

As I have argued, salient similarity is concerned with whether one has retained relevant aspects of character as to remain answerable and open to continued responsibility attributions. Determining the extent of such similarity will never be exact and will always be open to error. Its virtue, however, is the clarity it brings to an often-fraught subject of inquiry. It serves as a reminder of what is most important about character change, and it allows us to shift the significance of other aspects of responsibility attribution without losing sight of them altogether. Repositioning the importance of repudiation and integration as predictive alters their importance, but it does not render these factors unimportant or morally inert. They are indications of what could be. Perhaps Williams was not rehabilitated — at least, not yet. The execution pre-empted the possibility

of the rehabilitation that could have likely occurred, given the kinds of changes he already had undergone. California may not have executed an innocent man who was no longer answerable to his past crimes, but they likely executed someone we have good reason to think would eventually be dissimilar enough as to no longer be answerable.

Acknowledgements

As this article was derived from my Ph.D. dissertation conducted at McGill University, I would like to thank my advisor, Natalie Stoljar, for her unwavering support. I would also like to thank my M.A. supervisor at Wilfrid Laurier University, Kathy Behrendt, for helping me grapple with some of these ideas in a nascent form and sparking my interest in questions of identity. Many thanks as well to Hugo Cossette-Lefebvre, who so graciously helped translate my abstract, and Marya Schechtman, whose work and commentary were key in reframing aspects of my thesis for journal publication. Finally, thank you as well to all the manuscript workshop participants for their invaluable insights.

References

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies*, *93*(2), 161–188. https://doi.org/10.1023/A:1004222928272

Burgess, A., (1995). *A clockwork orange*. W. W. Norton and Co. (Original work published 1962).

Graff, D. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, *28*(1), 45–81. https://doi.org/10.5840/philtopics20002816

Kant, I. (2005). *Groundwork for the metaphysics of morals*. (A. Kingsmill, & Denis, L. Trans.) Peterborough: Broadview Press. (Original work published 1785).

Khoury, A., & Matheson, B. (2018). Is blameworthiness forever? *Journal of the American Philosophical Association*, *4*(2), 204–224. doi:10.1017/apa.2018.17

Kubrick, S. (director) (1971). *A clockwork orange* [Film]. Warner Bros.

Leung, R. (2004, May 21). *Rewriting the past: Former Crip leader teaches children how to avoid gangs*. CBS news. https://www.cbsnews.com/news/rewriting-the-past/

Levy, N. (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy*, *52*(4), 243–261. https://doi.org/10.1111/j.2153-960X.2011.00543.x

Levy, N. (2017). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research*, *94*(1), 3–26. https://doi.org/10.1111/phpr.12352

Mill, J. S. (2016). *Utilitarianism*. (A. Bailey, Ed.). Broadview Press. (Original work published 1871)

Parfit, D. (1986). *Reasons and persons*. Claredon Press. https://doi.10.1093/019824908X.001.0001

Schechtman, M. (1996). *The constitution of selves*. Cornell University Press.

Schechtman, M. (2016). "A mess indeed: Empathic access, narrative, and identity." In Julian Dodd (Ed.), *Art, Mind and Narrative: Themes from the Work of Peter Goldie* (pp. 17–34). Oxford University Press.

Schwarzenegger, A. (2005, 12 December). *Statement of decision: Request for clemency by Stanley Williams* [pdf file]. Retrieved from https://graphics8.nytimes.com/packages/pdf/national/Williams_Clemency_Decision.pdf

Shoemaker, D. (2005). Ecumenical attributability. In R. Clarke, M. McKenna, & A. Smith (Eds.), *The nature of moral responsibility: New essays* (pp. 115–140). Oxford University Press.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life*. *Ethics*, *115*(2), 236–271. https://doi.org/10.1086/426957

Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, *138*(3), 367–392. https://doi.org/10.1007/s11098-006-9048-x

Smith, A. M. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, *122*(3), 575–589. doi:10.1086/664752

Smith, A. M. (2015a). Attitudes, tracing, and control: Attitudes, tracing, and control. *Journal of Applied Philosophy*, *32*(2), 115–132. https://doi.org/10.1111/japp.12107

Smith, A. M. (2015b). Responsibility as answerability. *Inquiry*, *58*(2), 99–126. https://doi.org/10.1080/0020174X.2015.986851

Warren, J., & Dolan, M. (2005, 13 December). *Tookie Williams is executed.* LAtimes.com http://www.latimes.com/local/la-me-execution13dec13-story.html

Williams, S. T. (2004). *Blue rage, black redemption: A memoir*. Damamli Pub. Co.