Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Unlearning Agency: Lessons from Cults

JOSEPH METZ
Widener University

This paper investigates a type of agential impairment that has not been fully appreciated in the philosophical literature: unlearning agency, a process which undermines or destroys its victim's sense of self and agential competence. This type of manipulation is prominent in cults and is vividly displayed by the puzzling phenomenon of cult members seeming to harm themselves despite doing what they claim they want to do, as when several men in Heaven's Gate cult volunteered to castrate themselves in accordance with their beliefs. Drawing on the literature on cults and using Heaven's Gate as its test case, this paper distinguishes unlearning agency from other agential impairments in the philosophical literature. It then explores why unlearning agency matters. Understanding this type of agential damage not only highlights a new type of exemption from moral responsibility, but also outlines methods of rehabilitating agency. Furthermore, it provides an explanation for why cults can harm members who are doing what they want to do, without having to rely on an objective theory of well-being. Finally, this paper explores how unlearning agency can arise in other contexts besides cults. Thus, we should take this threat to agency seriously.

1. Introduction

During 1994 and 1995, several men in the Heaven's Gate cult were surgically castrated.[1] This was not part of a mystic ritual or sacrifice, nor was it demanded by their leader. Instead, these men wanted to leave their sexual—and, more broadly, human—desires behind. This was in accordance with their cult's belief that they needed to transcend to the Next Level as beings who could travel by spaceship to what us Earthlings call "Heaven." The castration was not only

1. See, e.g., Lalich (2004), Tweel (2020), and Zeller (2014) for details about how Heaven's Gate operated.

Contact: Joseph Metz <jwmetz@widener.edu>

https://doi.org/10.3998/ergo.8582

apparently voluntary; the men were so eager to undergo it that they flipped a coin for who would have the privilege of going first. When the first in-house surgery was botched, all further castrations were stopped—but only until the group was later able to locate doctors willing to perform the procedure. Then the remaining members who wished to be castrated, including the cult leader, had the procedure performed.

It seems clear that those cult members volunteering for castration were impaired in their choices,[2] but it's not obvious what exactly went wrong agentially. Cults raise interesting challenges about impaired agency, particularly when cult members are ostensibly doing what they want to do. Although cults can involve many kinds of agential impairments and although some of these agential impairments have been discussed by philosophers, cults also often involve a distinctive impairment that has not yet been fully appreciated in the philosophical literature: *unlearning agency*. As we will see, unlearning agency results from a special kind of manipulation akin to brainwashing that does a distinct kind of damage to the victim's agency. Victims of this kind of manipulation undergo a process of what I call a "reverse agential education" in which they are taught not to think of themselves as beings who are qualified to make decisions for themselves, problem solve, or find alternative answers besides those they are directly given. Unlearning agency comes in degrees, but in the extreme, the victims—though still agents in a minimal sense—lack deep features of agency that are important for things like moral responsibility.

Using the Heaven's Gate cult as its test case, this paper investigates this previously underappreciated agential impairment in three ways. First, it argues that this impairment often found in cults is importantly distinct from other noted agential impairments in the philosophical literature. Having drawn this contrast, this paper then clarifies what unlearning agency is, exploring the contours of the associated agential damage to the victim's sense of competence and sense of self. Third, it explores why unlearning agency matters. Unlearning agency can constitute an important potential exemption from moral responsibility, and understanding it suggests methods of rehabilitating agency that have some empirical backing. Additionally, unlearning agency allows us to explain why cult members can be harmed even when they are true believers and are doing what they want to do in accordance with their cult's beliefs and practices, and it can do so without committing us to an objective theory of well-being or relying on criticizing a cult's beliefs for being false. Finally, this paper explores how unlearning agency can even arise in a wide range of abusive, but non-cult-based, settings. This commonly includes abusive romantic relationships, and though

2. It might be possible for some people to legitimately want to castrate themselves in a way that reflects intact agency. As we will see, however, that is not what happened to the cult members in Heaven's Gate, though it may have happened to the leader of Heaven's Gate.

less common, it can also include religious, military, and corporate settings. Thus, it is an agential threat that is worth taking seriously.

2. Clarifying "Cults"

Let's begin by clarifying what is meant by "cult." Following the sociological literature, cults, sometimes also called "high-control groups" or "new religious movements" are, roughly, groups with a closed-off, hierarchical structure; a charismatic leader; a transcendent belief system; a high level of commitment from most members; and some methods of control and enforcement of group rules (Lalich 2004; Stein 2021). These are somewhat technical terms, but the fine details are not necessary for our purposes here. Instead, it's worth flagging that cults' transcendent belief systems are not necessarily religious, and being a religious group is neither necessary nor sufficient for being a cult. For instance, the Democratic Workers Party was a political cult in the 1970s and 1980s that was part of the New Communist Movement and had a transcendent belief system based on Marxist-Leninist ideology rather than religious doctrine.[3] Despite not being religious, it still had all the standard hallmarks of being a cult, including a charismatic leader and intense methods of control. Conversely, although religious groups can be cults and damage the agency of their members, many organized religions lack the relevant features of cult-hood and do not force their members to unlearn their agency, as will be discussed in §5.

Cults' methods of control, commitment, and enforcement can vary. This can include coercion with explicit threats. For instance, the Peoples Temple started in California, established the settlement Jonestown in Guyana, trapped cult members who moved there by confiscating their passports, and later used armed guards to force hundreds of them to drink poisoned Flavor-Aid and administer it to their children in an event now called the Jonestown Massacre (Winfrey 1979). Alternatively, other cults rely on intense social pressure, as with the Heaven's Gate cult using "check-in" partners when members left the group on "missions" as a way of constantly surveilling and socially pressuring those members to ensure compliance. Often there is only minimal control of fringe members, who are allowed to come and go as they please (as in, e.g., the Heaven's Gate and NXIVM[4] cults). However, this is sometimes used as a way to weed out

3. For more information on the Democratic Workers Party, see, e.g., Lalich (2004).

4. NXIVM was primarily a self-help group and multi-level marketing scam for the majority of its peripheral members, who were relatively uninhibited in leaving. It was only a sex-trafficking cult replete with branding, blackmail, and sexual slavery that made it nearly impossible to leave for a much smaller group of women who were ensnared by its inner circle. For more information, see, e.g., Grigoriadis (2018).

less-dedicated members, thereby leaving the remaining members feeling even more committed and even more social pressure to remain.

There are, of course, borderline cases where it is unclear whether a group is a cult. This paper investigates cults qua agency, not cults qua cult-hood. Hence, we need not focus on strict necessary and sufficient conditions for cult-hood, and we can instead focus on groups that are commonly agreed upon as being cults. Our test case fits the bill: Heaven's Gate's leader openly called it the "cult of cults" (Tweel 2020). Later on, we will examine other settings in which the relevant agential damage also arises. Again, what matters for our purposes is not whether these settings or groups count as cults, but whether they involve the relevant kind of agential impairment—unlearning agency.

3. A Distinct Threat to Agency

There are many different and overlapping factors that impair the agency of cult members, and not all cults nor all cult members involve or experience each of these factors. Nevertheless, the agential impairment and harm that this paper investigates is commonly found in cults and is importantly distinct from other impairments already noted in the philosophical literature. Let's examine why.

At first glance, people in cults might appear not to be agents at all, and cult members are often portrayed in popular media as insane, glass-eyed puppets mindlessly doing their leader's bidding. However, the reality of their agency is far more complex.

For starters, impaired agents in cults aren't obviously insane, nor are they obviously behaving compulsively or exhibiting other signs of severe, agency-undermining mental illness. For instance, one castration-seeking member of Heaven's Gate was referred to a psychiatrist, who evaluated him and wrote that the cult member "revealed no extreme or unusual beliefs (outside of his desire for an orchiectomy [i.e., castration]), and there was no evidence of psychotic symptoms. The patient did not appear impulsive…" and the clinical impression was that the patient's "wish for castration was authentic, long-standing, and nonpsychotic in nature … No symptoms of a current, full psychiatric syndrome were observed … [and] it was concluded that no strictly psychiatric contraindications to an orchiectomy were evident in this patient" (though an alternative was recommended) (Roberts et al. 1998: 416).

Relatedly, the agential impairment often found in cults is importantly different from impairments to cognitive abilities. For instance, some philosophers have argued that psychopaths might have severe enough empathetic impairments that they are not morally accountable for their behavior (Shoemaker 2015). In contrast, the cult members we are investigating don't have special problems

with empathy. Nor are they necessarily intellectually disabled or especially gullible—millions of intelligent, educated, and seemingly ordinary people can and do get drawn into and trapped by cults (Lalich 2004; Stein 2021). Some people might be more predisposed towards being in cults than others (Zablocki 2001), but it is far from clear that this predisposition, by itself, is more agentially restrictive than many other "normal" dispositions—such as the disposition to be slightly more trusting of strangers.

The impairment in question is also importantly different from that in Wolf's (1987) classic example of JoJo, a man raised by an evil dictator to become an evil dictator. JoJo fully endorses torturing and otherwise harming his subjects, but despite knowingly and willingly causing such harm, he might not be a competent moral agent because he is morally insane and cannot recognize the difference between right and wrong. Related issues like having skewed formative circumstances can clearly matter for agency, as well, and some unlucky people are born into cults and have bad moral upbringings as a result. Nonetheless, many cult members had ordinary moral upbringings and joined as adults. Hence, identifying the impairments that adult cult members face goes beyond settling whether impaired moral upbringing can excuse or exempt one from blame.

The impairment in question isn't simply coercion either. Many cults clearly involve coercion, as in the aforementioned mass-"suicide" of 909 people in Jonestown at gunpoint. However, as noted earlier, some cults exert much control over members despite using little explicit coercion and instead relying on intense social pressure (Zablocki 2001). Furthermore, not all behaviors in coercive cults involve explicit coercion; not every action of every member of Jonestown was at gunpoint.

Unlearning agency also differs from situationist concerns about our local environments (as raised by, e.g., Vargas 2013); the pressure the men in Heaven's Gate felt to self-castrate is different in kind from situationist findings such as that being late for a meeting tends to make people more reluctant to stop and help strangers. The agential impairment in question is also different from concerns that oppression and society-level structures can sometimes constrain our agency and responsibility (as discussed in Vargas 2018). Many members of Heaven's Gate were not members of oppressed groups, and their agential impairment went beyond oppressive societal structures, biases, prejudices, decision-making shortcuts and heuristics, etc.

In contrast to these other kinds of agential threats and impairments, the threat we are investigating here involves a distinct kind of manipulation and brainwashing—one not yet discussed in the philosophical literature. Manipulation is at the heart of many important debates within the free will and moral responsibility literatures, but the kinds of manipulation discussed there importantly differ from the manipulation often found in cults. For instance, some

philosophers such as Pereboom (2001) have argued certain extreme forms of manipulation—such as controlling someone's brain and reasoning processes via radio signals—ultimately are not relevantly different from determinism, and since this manipulation seems freedom-undermining, determinism is, too. Others, such as McKenna (2008), disagree. Other manipulation-based debates center on whether compatibilist requirements for freedom and responsibility ought to include historical components, such as having certain kinds of causal histories free from manipulation. To illustrate with Mele's prominent Ann and Beth case (e.g., Mele 2019: 40–41): Suppose Beth's balanced preferences, desires, hopes, dreams, etc. are replaced overnight with Ann's workaholic ones. Significant debate has ensued over whether Beth has been too manipulated to be free and responsible for her new behaviors.[5]

In contrast to these kinds of cases, real-world brainwashing of the sort relevant to unlearning agency is messy, resistible, and takes time to implement (Lalich 2004; Stein 2021; Zablocki 2001). This isn't a problem for Pereboom and Mele's cases and arguments about the similarities between manipulation and determinism. However, the differences between the more stylized cases in the philosophical literature and the messier cases of real-world brainwashing also highlights that the slower, imperfect manipulation that impacts cult members is quite different from the instantaneous, continuous, and/or total manipulation found in the extant philosophical literature, as we will see in more detail in the next section.[6]

Thus, this agential impairment often found in cults goes beyond the impairments that have already been discussed in the philosophical literature.

4. Unlearning Agency Via Undergoing a Reverse Agential Education

Having suggested why unlearning agency is different from other agential impairments that have been discussed in the philosophical literature, let's clarify more precisely what this impairment is. Cults exert strong pressure on members to conform to the cult's beliefs and practices. This pressure is so strong that it causes some cult members to undergo what I dub a "reverse agential education," in

5. For more information on the debates between historical and ahistorical compatibilists, see, e.g., Frankfurt (1971), McKenna (2012b), and Mele (2019).

6. Even Pereboom's third case—in which the victim is trained from a young age in his home and community to reason in a certain kind of way—is different from the manipulation in cults. Cult members often join as adults and did not always reason and behave the way they come to do while in the cult, and their behavior is not necessarily determined, unlike the behavior in Pereboom's cases.

which they are trained not to think of themselves as agents anymore. In essence, they are trained to unlearn their agency. As we will see later, unlearning agency actually isn't unique to cults, but cults provide a particularly illuminating source of information about it and are a good place to start exploring it.

As a foundational element of typical moral education, children and young adults are taught to acknowledge and appreciate their agency—to appreciate that their behaviors impact themselves, others, and their surroundings in a way that merits accountability. For, understanding and appreciating what it means to be a full, morally responsible agent requires first recognizing oneself as an agent. When cult members undergo a reverse agential education, they are agentially infantilized; they are trained to take on the preferences and desires of the cult, and they are trained *not to think or act for themselves anymore*. They stop thinking of themselves as the kinds of beings who can or should make decisions for themselves, entrusting those decisions to others, typically the cult leader.

There are different senses of agency that philosophers care about, and victims of reverse agential educations are still agents in a minimal sense. For instance, these victims are importantly different from rocks and other things governed only by natural forces, and they still retain certain basic agential capacities like mean-ends reasoning. However, there are additional, deeper senses of agency that philosophers also care about. These deeper senses ground important claims such as whether someone's behaviors really reflect meaningful exercises of her agency, whether she can fairly be held morally accountable for her behaviors, and whether she is meaningfully the author of her own life.

Unlearning agency undermines two essential aspects of this deeper sense of agency. First, victims' beliefs that they are full-blooded, competent agents are undermined. There is an important conceptual difference between *not being* a competent agent and (possibly mistakenly) *not believing* that you are a competent agent. Nonetheless, it seems highly plausible that a key constitutive element of being a competent agent is (accurately) believing that you are a competent agent (see also, e.g., Velleman 2000). Hence, even if the victims are "only" manipulated into not believing that they are competent agents, their overall agency is still importantly damaged as a result.

Second, victims have their sense of self and individuality undermined or—in the extreme—destroyed. Cult members undergoing a reverse agential education not only take on the preferences of the group (typically, the group leader), but also gain higher-order preferences not to have other distinct individual preferences in the future. Their goal is to assimilate, to functionally *lose themselves as individual agents* in the process. Individuality and independent thoughts, desires, plans, etc. are often punished in these contexts, leading cult members to doubt, fear, and mistrust exercises of their agency. They don't merely question their competence and agential capacities; they also believe they shouldn't think of

themselves as individual agents, so they have their sense of self undermined in the process. I contend that another plausible constitutive element of deep agency is having a sense of yourself as an individual. Thus, unlearning agency thereby threatens a second key aspect of agency, too.

To further clarify this agential damage, let's contrast unlearning agency via reverse agential education with philosophers' extant examples of instantaneous brainwashing or posthypnotic suggestion. Instantaneously brainwashed or hypnotized people can still form future preferences, still believe themselves to be agents, and do act as full agents in the future—albeit as agents with bad causal histories of certain desires, preferences, etc.—so, despite their bad histories, they have not lost their major agential capacities. Hence, even if such agents are not morally responsible for their behaviors at the moment that their preferences are altered, they could still be responsible for their future behaviors as morally responsible agents who have endorsed their new preferences.[7] A reverse agential education is harder to overcome. "Fixing" a hypnotized agent involves fixing her beliefs, preferences, etc. back to the way they originally were, whereas fixing an agent who has undergone a reverse agential education involves fixing her preferences *and retraining her to understand that she is an individual who is capable of making meaningful decisions for herself*.

Let's now turn to the evidence that reverse agential educations and this associated agential damage occurs in cults. Again, we will focus primarily on the Heaven's Gate cult as our test case, but these phenomena exist in many other cults and even arise in some non-cult settings.

To start, consider how brainwashing actually works as a real-world phenomenon. "Brainwashing" is a controversial term (Zablocki & Robbins 2001), and some sociologists and religious studies scholars dislike it and regard it as debunked since attempts at real-life total mind control have all failed (see, e.g., Zeller 2014). However, there is significant psychological and sociological research showing that real-life brainwashing and indoctrination exist as powerful controlling forces that can be wielded as *retention* tools to keep longstanding members, even though they don't seem to work as *recruitment* tools to instantaneously create new converts via robotic-like mind control (Zablocki 2001). As Stein (2021: 22) details in her recent book:

> There are several alternative terms scholars have used to name this process [i.e., brainwashing]: coercive persuasion (Schein), thought reform (Lifton), resocialization (Berger and Luckmann), total conversion (Lofland), mind control (Singer, Hassan), coercive control (Stark), or,

7. See, e.g., McKenna (2012b). With his Suzy Instant case, McKenna (2012b) also points out that it is not obvious that hypnotized or instantaneously manipulated agents are not responsible for their current behaviors.

> most recently by Lalich, bounded choice. All these thinkers describe variants of the same essential process: the alternation of love and fear within an isolating environment resulting in a dissociated, loyal and deployable follower who can now be instructed to act in the interests of the leader rather than in his or her own survival interests.

Despite being in a large group, people who become ensnared in cults are socially isolated even from each other while having a highly variable relationship with the charismatic leader. On Stein's (2021) model, this relationship, along with other coercive and fear-driven pressures, creates a traumatic "anxious dependency" on the group and the leader. Some members try to seek comfort and resolve inner psychic contradictions or tensions by giving up their sense of individual agency and committing wholesale to the group. This, argues Stein, is when the leader swoops in and manipulates the beliefs of these members.

Again, this process is far from absolute and is not applied to all members, and if the pressure is ratcheted up too quickly, people leave. But for those who remain, the exit costs—the psychic, spiritual, emotional, economic, etc. costs of leaving the group—become much higher, and follower retention rates notably increase as time passes (Zablocki 2001). The effect is that the remaining members are largely stripped of their old self in a process Zablocki (2001) characterizes as a death and rebirth. For more evidence that this process shatters its victims' sense of self and belief in their agential competence, consider the testimony of two former cult members from different cults:

> They ask you to betray yourself so gradually that you never notice you're giving up everything that makes you who you are and letting them fill you up with something they think is better and that they've taught you to believe is something better.

> In the frame of mind I was in [at the time], I welcomed the brainwashing. I thought of it like a purge. I needed to purge my old ways, my old self. I hated it and I felt really violent toward it … I wanted to wash it all away and make myself an empty vehicle for [the guru's] divine plan. Our ideal was to be unthinking obedient foot soldiers in God's holy army.[8]

Members of Heaven's Gate faced these same sorts of brainwashing pressures. They were socially isolated but never given full privacy, even when venturing on "missions" away from the cult, because of the constant presence of

8. Both passages are quoted in Zablocki (2001: 200). Zablocki also discusses similar testimony from different cult members from various other cults, and he notes that none of these cult members were ever affiliated with anticult groups.

check-in partners who supervised every phone call and social interaction.[9] This limited individuality-affirming connections with the outside world. Followers were also given the same gender-neutral clothing and new, diminutive names—usually three consonants followed by "ody" (e.g., "Jstody"). Additionally, they were heavily pressured to conform: They were given a notebook that detailed how exactly they were supposed to do every task. This included the minute details of daily tasks—not only to make pancakes for breakfast and spaghetti for dinner; but also, exactly how large to make the pancakes, how to cut those pancakes, and how to twist their spaghetti against their forks. Hence, these agents were trained not to think of themselves as competent decision-makers for even the most trivial situations.

As part of their pressure to conform, members of Heaven's Gate were also intensely pressured to leave their individuality behind. Members were told that they could—literally—undergo a metamorphosis and become their true alien selves and travel via spaceship to the Kingdom of God, but to do this, they had to leave their humanity and individuality behind. Prospective members were told, "You would have to literally overcome every human indulgence and human need … it is the most difficult task that there is … you have to lose everything. You will sever every attachment that you have."[10] They were taught the tenets of the cult in sessions that were called "the classroom." They were told that human urges had to be left behind to achieve salvation, especially sexual urges: Members were forbidden from having sex, and the men had to sign a prominently displayed sheet of paper whenever they had a nocturnal emission. The desire to remove these sexual urges led the members mentioned in the introduction to seek out castration. Heaven's Gate members were also taught to suppress expressions of emotion and to speak with an even-toned, gender-neutral voice—to all sound alike. Their desire to conform and their fear of failure was intense: One man spoke too excitedly and with too deep and masculine of a voice, and he was mocked for doing so by the charismatic leader. In an interview decades later (Tweel 2020), the former member still struggled to speak due to the trauma he suffered in that moment. At the cult's conclusion, 39 cult members ritually committed suicide. The night before, they went out for a final dinner. They all not only ordered the same meal; they also ordered and ate it in the same precise way.

Many members left the cult during these brainwashing and indoctrination processes—some were even given money to go home. However, this functionally ratcheted up the pressure on those who remained, and they became infantilized conformers to the extreme, complete with extremely high commitments to the group and severed attachments to the rest of the world.

9. See, e.g., Lalich (2004), Tweel (2020), and Zeller (2014) for further details of the practices of members of Heaven's Gate.

10. Quoted in Zeller (2014: 38).

It is not particularly controversial to hold that these remaining cult members experienced intense coercive pressures, but my contention is that what is distinctive about their particular agential impairment is that they had functionally been trained not to view themselves as competent individual agents anymore.[11] They were afraid of failing to conform to the desires and goals of the group, and they had been pressured to believe that expressing individuality, distinct planning, or independent thought would put them at risk for missing out on an eternity in Heaven. Acknowledging their agency meant risking losing all that they valued, so they abandoned it. In short, they had undergone a reverse agential education and had unlearned their agency.

5. The Significance of Unlearning Agency

Now that we have clarified what unlearning agency is, let's explore why it matters.

5.1 The Impact on Responsibility and Agency

The first reason unlearning agency is important is that it has significant impacts on understanding responsibility and agency—both for understanding the precise kinds of damage suffered by people ensnared in cults and for understanding what successful exercises of agency look like. As I have argued, unlearning agency is a distinct kind of threat to agency that differs from the forms of manipulation that have already been considered by philosophers. Destroying someone's sense of being a competent individual agent is a particularly vicious kind of agential damage, so it's worth paying attention to.

Additionally, understanding exactly how agency goes awry in cases of reverse agential educations sheds light on the contours of successful agency. Since believing you are a competent agent is plausibly an important part of successfully being an agent in a deep sense, this suggests a path for rehabilitating cult members: re-teaching them to believe that they are individual agents with meaningful capacities for self-determination and the ability to exercise control over their lives. And, there is at least some empirical evidence that accords with this. As Stein (2021) discusses, one successful strategy for helping cult members escape and regain control over their lives is to help them gain a connection with

11. Perhaps the cult members were being trained to be part of a collective agent—and the Heaven's Gate cult leaders told their followers to idolize the hivemind Borg from Star Trek. If there are genuine collective agents, perhaps these cult members could have been a member of such a collective agent, but their individual agency was still destroyed.

someone or something that reaffirms their individuality and that breaks their anxious dependence on the cult and its leader. Sometimes this is a secret phone-based relationship to a friend outside the cult. Other times it is secretly listening to banned music. Still other times it is finding a safe venue to express doubts about the cult with another cult member. The key idea is building a sense of self identity that exists independently of this cult. As further evidence of the importance of developing an independent sense of self, consider that when cult members temporarily escape and have a hard time reintegrating into society and redeveloping their sense of individuality, they often fall back into the cult, believing it is the only place where they belong.[12] Hence, understanding exactly how the victims' agency is damaged in cases of reverse agential educations provides some clues on how to rebuild it.

Another reason unlearning agency is significant is that, as a form of agential damage, it thereby also constitutes a threat to moral responsibility, given the reasonable assumption that moral responsibility is grounded in the deep sense of agency that we have been considering.[13] Moral responsibility—being an appropriate target of blame and praise (as in the tradition of Strawson 1962) and/or being morally accountable for one's behaviors (as discussed in Shoemaker 2011)—is commonly thought to require sufficient control and awareness of one's behaviors (see, e.g., Fischer & Ravizza 1998; Robichaud & Wieland 2017). In some cases, though, people are excused or exempted from responsibility. Following the standard dichotomy,[14] excuses function by showing that although someone is still a full-blown moral agent, she—contrary to appearances—did not fail to meet a key basic moral demand. For example, she might have rear-ended your car, but it turns out that this only happened because someone rear-ended her car and she blamelessly had no way of avoiding subsequently hitting yours. In contrast, exemptions arise when someone is not (currently) the right kind of being that can appropriately be held responsible or have basic moral demands placed upon her since her agential capacities and/or moral understanding are significantly compromised. Common exemptions include young children, psychopaths, and people who have been hypnotized.

Because unlearning agency is a type of agential damage, it threatens the agential competence of its victims and can constitute a previously unappreciated moral excuse. Although the people who have unlearned their agency differ from children, psychopaths, and people who have been hypnotized, they have damaged senses of self and they doubt that they are agentially competent. When they

12. See, e.g., Stein (2021) and Tweel (2020) for more examples.

13. See, e.g., McKenna (2012a: 6–30) for a discussion of how responsibility is grounded in agency.

14. See, e.g., Strawson's (1962) two types of pleas and Watson's (1987: 118–119) discussion of exemptions.

commit harm as a result of their unlearned agency, they are not failing to meet basic moral demands that they understand. Instead, because of their agential damage, they don't belong to the right category of being that can appropriately be said to have these moral demands in the first place. Consider a victim of a reverse agential education who decides to harm someone else because it is what her cult leader demands of her. It's not that—contrary to appearances—she wasn't failing to meet a basic moral demand in making that bad decision. Rather, it seems unfair to hold her accountable for this "decision" because her agency is so compromised that she didn't think she was qualified to make it for herself or question her leader's desires. Unlearning agency does not necessarily give rise to a permanent exemption, for if these victims are lucky enough to escape their destructive situations and relearn their agency, they can thereby regain their status as morally responsible agents. Nonetheless, unlearning agency shows that the category of exemptions from responsibility is broader than commonly recognized.

A few clarifications are in order about exemptions from responsibility. First, exempting people who have unlearned their agency from responsibility might appear to reinforce their agential impairment. The worry is that these victims would be treated as incompetent, which might further entrench their beliefs that they are not competent agents, and, as we have seen, believing you are a competent agent is a plausible requirement for being a competent agent. In response, it is possible to exempt someone from responsibility while still treating them as if they are ultimately capable of becoming responsible for their behaviors. Young children are exempted from responsibility, but in teaching their children how to eventually become responsible adults, parents often still treat their children as if they are capable, morally responsible agents—without actually holding them morally accountable as they learn. This ultimately bolsters children's agency, and this approach can plausibly be extended to rehabilitating cult members. Hence, temporarily exempting the cult members from responsibility as they regain their agency might actually be a key part of their rehabilitation.

A second clarification: Agential impairments come in degrees, and it seems possible for people to unlearn their agency to varying extents. For instance, someone might believe themselves not to be an agent with respect to financial behaviors while still recognizing their agency with respect to medical ones, or they might view themselves as capable of making minor decisions towards achieving some goal their cult leader set, but not capable of deciding whether they ought to pursue that goal in the first place. In such cases of partial impairment, the person might still be entirely or partially responsible for many of their behaviors. In the extreme, though, when people forcibly undergo a thorough reverse agential education and no longer believe they are competent individual agents in a meaningful sense, they arguably are not morally responsible for their behaviors—on the grounds that they are not agents of the relevant kind anymore.

Third, nothing in this paper denies that many cult leaders and lieutenants in cults are morally responsible for their morally atrocious acts, for these leaders and higher-ups usually do *not* undergo reverse agential educations and still retain their agential faculties. In contrast to their victims, cult leaders often don't practice what they preach about the cult's restrictive tenets and don't suffer the associated isolation and agential harm. Hence, the leaders lack this exemption from responsibility, and we can still straightforwardly maintain that cult leaders are blameworthy for their behavior. The case of Marshall Applewhite—the leader of Heaven's Gate—raises interesting questions on this front, since he, too, volunteered for castration. One reading of the situation is that his agency was intact: He despised his sexuality and sexual urges, and he freely volunteered for castration in a way that was agentially unimpaired. Another reading is that the carefully orchestrated manipulative structure he established in Heaven's Gate sucked him in and manipulated him, too. In this case, his agency was impaired, but he might still be indirectly free and responsible for this decision because it was negligent not to be more cautious around a dangerous system which he culpably played an instrumental role in establishing. Though we may never know the precise contours of his agency, we can still explain his responsibility either way.

As a final clarification, the exemptions from responsibility defended in this paper would not license people to absolve themselves of moral responsibility by willingly unlearning their agency. If it turns out to be possible for someone to teach herself to unlearn her agency or if someone willingly volunteered for this process, she would still be morally responsible for that choice since she was still an agent when she freely made that choice. In contrast, the victims of involuntarily unlearned agency are not responsible for their own agential damage because it was forced on them and because they had little reason to believe it would happen to them. It was not reasonably foreseeable, for example, that attending a public meeting hosted by people who were interested in UFOs would lead to castration and ritual suicide in matching shoes.

5.2 Explaining Why Cults Are Harmful to Their Believers

A second set of reasons to care about unlearning agency stems from how it allows us to account for why cults are harmful to their members—even while taking seriously the idea that the members believe the cult's teachings and are doing what they "really" want to do in acting in accordance with the cult's practices. Continuing with our test case of Heaven's Gate, even if the castration-seeking cult members really believed that they were aliens trapped in human bodies and really believed that having sexual urges might prevent them from

metamorphosizing into aliens and reaching Heaven, it is still important to be able to explain why they were harmed.

One such benefit of the proposed account of unlearning agency is that it allows us to explain how cults are harmful to their members without having to adopt an objective theory of well-being. For, the issue is that the cult members' agency is damaged during the acquisition of their beliefs and desires—not whether those desires or the satisfaction of them are objectively valuable.

In the literature on well-being, there are three central families of views. One family—objective list theories—is objective about well-being. The other two—hedonistic theories and desire-satisfaction theories—are subjective theories of well-being.[15] Objective list theories hold that well-being and what is good for someone consist in a list of things with objective value, such as knowledge and friendship, even if those things are not desired by someone. Objective views of well-being such as objective list theories can straightforwardly explain why cults are harmful to their members: Unlearning one's agency surely violates important things on the objective list and is therefore harmful for someone even if it is what someone wants. Hence, even if a cult member wants to be castrated or if he wants the cult leader to make decisions for him, the objective list theorist can still hold that these are harmful for him because they conflict with things on the list.

In contrast, some subjective theories of well-being like hedonism seem to struggle to explain why cults are harmful to their members. Hedonistic theories hold that something contributes to someone's well-being only to the extent that it contributes to his pleasure or reduces his pain. Hedonists have a difficult task in explaining why the castration was harmful to the members of Heaven's Gate, since they seemed to experience pleasure in it. However, hedonistic theories of well-being are highly controversial,[16] and there are other subjective theories that can draw on the harms of unlearning agency to explain why cults are harmful to their members.

In particular, consider desire-satisfaction theories, which hold that well-being consists in getting what you want. Desire-satisfaction views often build in constraints on these desires, requiring that they are informed or rational, and it seems plausible that one cannot rationally or in an informed way desire that one's agency be undermined. Hence, even though some cult members do satisfy some of their desires—including the desire to castrate themselves—desire-satisfaction views still have the resources to hold that these cult members are harmed because they cannot rationally or in an informed way desire that they undergo a reverse

15. For an overview of these central views as well as discussion of other prominent views on well-being, see Tiberius & Haybron (2022).

16. For a prominent argument against hedonism and that other things besides pleasure could be intrinsically valuable, see Nozick's (1974) experience machine thought experiment. For recent commentary on the literature about the experience machine, see, e.g., Lin (2016).

agential education and unlearn their agency while acquiring those desires.[17] In short, an advantage of identifying unlearning agency as a type of agential damage is that we don't need an objective theory of well-being to explain why cults are harmful to their members.

We also don't need to criticize cults' belief systems to hold that they can harm their members. This is a further major benefit since we can thereby sidestep a concern raised by some sociologists and religious studies scholars. Some of these scholars argue that many cults really are best understood as new religious movements and that the cult members' behaviors reflect sincerely-held beliefs, so many criticisms of cults are unfair attacks on non-standard—but genuine—worldviews.[18] It is beyond the scope of this paper to assess whether or not these sociologists and religious studies scholars are right about how best to understand the genuineness of the cult members' beliefs (and it's worth noting that many former cult members disagree with these scholars[19]).

Nonetheless, the view defended by this paper doesn't require us to criticize a cult's beliefs for being false or unusual in order to hold that the manipulated acquisition of those beliefs is harmful to cult members. The issue with unlearning agency is not what the victim's beliefs are, but how she came to acquire them. A group could have perfectly ordinary, mundane, and true beliefs about how the world works and still damage the agency of its members by forcing them to acquire those beliefs as part of a reverse agential education.[20] Thus, the beliefs may be sincere, but it is their acquisition—not their content—that is agency-undermining in cases of reverse agential educations. There are important further questions about whether and how these beliefs can also harm cult members, but that is beyond the scope of the issues covered in this paper.

17. For similar reasons, other subjectivist theories like value fulfillment theories can also hold that cults are harmful to cult members because you cannot desire for your broad values to be undermined by manipulation of the sort that arises in cases of unlearning agency.

18. See, e.g., Zeller's (2014) argument that the behaviors of the Heaven's Gate cult members are best understood as reflecting genuine religious beliefs and that Heaven's Gate's belief structure is best understood as a fusion of New Age, Evangelical Christian, and other beliefs.

19. See the testimony of former cult members in, e.g., Lalich (2004) and Stein (2021). See also the point emphasized in Stein (2021) and noted earlier in this paper that many cults are not religious.

20. For this reason, cults are importantly different from conspiracy theories, even though the two are often grouped together in popular discourse. People who believe conspiracy theories have both strange beliefs and compromised ways of obtaining evidence—to the extent that many of them think they are on equal epistemic footing with actual experts. Still, conspiracy theorists seem to retain their agency, especially since many conspiracy theorists pride themselves on their ability to reason "better" than the "sheep" who don't believe in the conspiracy, and many conspiracy theories lack a single clear leader who controls the behavior of the lower-level conspiracy theorists. Some cults involve conspiratorial beliefs, but the relevant issue for this paper is the agential damage that some cult members experience and conspiracy theorists don't, and a cult with an ordinary, epistemically justified belief system could still damage the agency of its members. Clarifying the precise issues with conspiracy theories goes beyond the scope of this paper, but it is worth highlighting this important way in which cults and conspiracy theories differ.

5.3 Where Else Unlearning Agency Arises

Although unlearning agency is made particularly vivid in cults, it can arise in other kinds of settings, which is yet another reason it is worth our attention. One prominent way reverse agential educations arise is in abusive romantic relationships. To illustrate, consider Stein's (2021) observation that the model for isolation and control found in cults is similar to that found in abusive romantic relationships, though the abuser often acts alone in isolating, surveilling, and controlling their victim. Consider also that a common tactic by abusers in abusive romantic relationships is gaslighting, in which an abuser undermines their victim's sense of basic competence as a way of preventing disagreement and further cementing the abuser's control.[21] For instance, gaslighters charge their victims not just with being mistaken but also with being crazy. On Abramson's (2014: 8) account of gaslighting, abusers make accusations "about the target's basic rational competence—her ability to get facts right, to deliberate, her basic evaluative competencies and ability to react appropriately: her independent standing as deliberator and moral agent."

We can use unlearning agency as a way of further understanding gaslighting: Undermining the victim's confidence and trust in her own judgments via gaslighting is one brutal method of implementing a reverse agential education. The precise method of manipulation and outcomes of unlearning agency in one-on-one settings like gaslighting is different than those in collective settings like cults. For instance, gaslighting doesn't typically involve a complete destruction of the victim's sense of self in the way that happens in cults, but gaslighting does also often involve isolating the victim from friends and family, which does still harm the victim's sense of self. The end result of unlearning agency, both in cults and in these abusive romantic relationships, is relevantly parallel: The victims no longer believe themselves to be competent agents and have damaged senses of self. Of course, unlearning agency is not part of all abusive romantic relationships, but the concept of reverse agential educations can help further elucidate the precise nature of the harms in some of these cases. Similar methods of forcing victims to unlearn their agency plausibly arise in other types of abusive, non-romantic interpersonal relationships, too, such as abusive parent-child relationships.

Another place unlearning agency can arise is in the military. Settings in which people are trained to obey orders without question, conform, forego their individuality in the name of the greater collective good, and follow a charismatic leader are ripe for manipulation of the sort this paper has explored. However, many modern militaries importantly do not force their soldiers to unlearn their

21. See, e.g., Abramson (2014) for further analysis of gaslighting and its harms.

agency or undergo its associated agential damage. While soldiers are taught to follow orders, they are expected to push back if an order breaks the law. They swear allegiance to their country, not their military's leader. Crucially, they are taught to problem-solve and find answers to challenges they face so that, for instance, they can overcome obstacles when their commanding officers are not present. These types of measures preserve key aspects of their agency. Recall, members of Heaven's Gate were not allowed to problem-solve what size to make their pancakes, let alone what to make for breakfast. Thus, although it is possible for soldiers to be manipulated into unlearning their agency, this does not seem to be the norm for most modern militaries.

Religion is another setting where people can be manipulated into unlearning their agency, but as with military settings, unlearning agency is not the norm here either. Some cults are explicitly religious and use their religion to indoctrinate and manipulate their victims. However, although religious groups can, for example, demand strict adherence to rules and practices, be led by charismatic leaders, preach the unimportance of the individual and that all must give up everything for God, these features can be present without damaging their practitioners' agency. The issue for unlearning agency is whether the participant is forced to give up her reasoning and decision-making. For instance, a Buddhist's deep metaphysical commitment to relinquishing the ego does not undermine her sense of herself as a competent individual agent who is capable of making major decisions about her life (including whether to continue being Buddhist). Similarly, a nun can decide to take vows of poverty, chastity, and obedience while still retaining the agential capacities to independently assess her convent's plans. In short, some religious groups do undermine the agency of their followers, but many people can still be deeply committed to their religion without losing their beliefs that they are competent individual agents.

One final place unlearning agency can occur is in corporate environments. As evidence, consider the following allegations from an ongoing lawsuit (as of this writing) filed by a Panda Express employee (Complaint, *Spargifiore v. Panda Restaurant Group, Inc.*). According to her testimony, employees at Panda Express who wanted to be promoted allegedly had to attend work retreat seminars. At one such seminar, she and her fellow victims had their cell phones taken away and were placed in a room with blacked-out windows. They were isolated and told to remain completely silent, and then waited for an hour. Suddenly, a person from the seminar burst into the room and screamed at the Panda Express employees in various languages and berated them for doing what they had been told to do, such as sitting quietly. The situation further deteriorated from there. The Panda Express employees were not allowed to be alone, even when running from the room to go vomit, and the experience culminated with several of them

being forced to strip to their underwear while being ogled and filmed while crying and shouting their shortcomings to their fellow victims.

Given that she filed a lawsuit against her employer, this particular employee does not seem to have unlearned her agency, and of course, most workplace environments are not cults and do not involve serious agential damage. Nonetheless, this case involves many of the key ingredients for unlearning agency: psychological trauma with total compliance to the group or leader offered as an escape as part of a calculated effort to get the victim to conform. Recall that in cults, some members leave when the pressure to conform is ratcheted up too quickly but that for those who remain, the exit costs significantly increase, as does the pressure to unlearn their agency. In the Panda Express case, although this pressure was extreme and was ratcheted up quickly, note that it still was highly effective in getting this employee and her peers to conform until the conference ended. This kind of workplace environment and the broader corporate culture that allegedly condoned it seem like fertile grounds for unlearning agency. Other toxic workplace environments that abuse workers to convince them that they are worthless and lack the ability to be employed elsewhere are good candidates for the proposed analysis of agential damage, too.

This is not an exhaustive list of all the situations in which reverse agential educations could occur. Still, it highlights that this kind of agential damage is not unique to cults, making it all the more worthwhile to care about.

6. Conclusion

In summary, I have outlined unlearning agency as a distinct kind of manipulation and agential damage that differs from the other kinds of agential impairments already discussed in the philosophical literature. Understanding reverse agential educations and unlearning agency allows us to explain what exactly went wrong with the castrated men's agency in Heaven's Gate. More broadly, understanding this threat to agency provides the tools to explain why cults harm their victims, even when those victims are true believers, and it accounts for why some victims are exempted from moral responsibility for their behaviors. Understanding this agential harm also sheds light on what successful agency involves, and it suggests a way of reestablishing agency via reestablishing individual-affirming connections and relearning to think of oneself as a competent individual agent. Finally, this paper highlights that although cults provide particularly vivid examples of unlearning agency, this kind of agential harm is not unique to cults and can arise in other kinds of ordinary contexts. Thus, unlearning agency is a serious agential threat that merits our attention.

Acknowledgements

I would like to thank Carolina Sartorio, Kay Chronister, Aaron Eli Segal, Karolina Wisniewska, and two anonymous referees and an anonymous area editor at *Ergo* for their detailed feedback on this paper. Thanks also to the audiences at the Western Michigan University Graduate Student Philosophy Conference; Austin Graduate Ethics and Normativity Talks; and the Florida State University Free Will, Moral Responsibility, and Agency Conference for helpful discussions of an earlier version of this project. Finally, thanks to Catherine Fakler, Lauren Jacobson, and Tim Salzer for introducing me to Heaven's Gate.

References

Abramson, Kate (2014). Turning Up the Lights on Gaslighting. *Philosophical Perspectives 28*(1), 1–30. https://doi.org/10.1111/phpe.12046

Complaint at 1, *Spargifiore v. Panda Restaurant Group, Inc., Alive Seminars and Coaching Academy, Does*. Case 21STCV07909 (C.A. Sup. Ct. 2021). Retrieved from https://mms.businesswire.com/media/20210304005286/en/863082/1/Complaint.pdf?download=1

Fischer, John M. and Mark Ravizza (1998). *Responsibility and Control*. Cambridge University Press. https://doi.org/10.1017/cbo9780511814594

Frankfurt, Harry (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Grigoriadis, Vanessa (2018, May 30). Inside NXIVM, the 'Sex Cult' That Preached Empowerment. *New York Times Magazine*. https://www.nytimes.com/2018/05/30/magazine/sex-cult-empowerment-nxivm-keith-raniere.html

Lalich, Janja (2004). *Bounded Choice: True Believers and Charismatic Cults*. University of California Press. https://doi.org/10.1525/california/9780520231948.001.0001

Lin, Eden (2016). How to Use the Experience Machine. *Utilitas*, *28*(3), 314–332. https://doi.org/10.1017/s0953820815000424

McKenna, Michael (2008). A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument. *Philosophy and Phenomenological Research*, *77*(1), 142–159. https://doi.org/10.1111/j.1933-1592.2008.00179.x

McKenna, Michael (2012a). *Conversation and Responsibility*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199740031.001.0001

McKenna, Michael (2012b). Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism. *Journal of Ethics*, *16*(2), 145–174. https://doi.org/10.1007/s10892-012-9125-7

Mele, Alfred (2019). *Manipulated Agents: A Window to Moral Responsibility*. Oxford University Press. https://doi.org/10.1093/oso/9780190927967.001.0001

Nozick, Robert (1974). *Anarchy, State, and Utopia*. Basic Books.

Pereboom, Derk (2001). *Living Without Free Will*. Cambridge University Press. https://doi.org/10.1017/cbo9780511498824

Roberts, Laura, Michael Hollifield, and Teresita McCarty (1998). Psychiatric Evaluation of a "Monk" Requesting Castration: A Patient's Fable, With Morals. *American Journal of Psychiatry*, *155*(3), 415–420. https://doi.org/10.1176/ajp.155.3.415

Robichaud, Philip and Jan W. Wieland (Eds.) (2017). *Responsibility: The Epistemic Condition*. Oxford University Press. https://doi.org/10.1093/oso/9780198779667.001.0001

Shoemaker, David (2011). Attributability, Answerability, Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, *121*(3), 602–632. https://doi.org/10.1086/659003

Shoemaker, David (2015). *Responsibility from the Margins*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198715672.001.0001

Stein, Alexandra (2021). *Terror, Love and Brainwashing: Attachment in Cults and Totalitarian Systems* (2nd ed.). Routledge. https://doi.org/10.4324/9781003030959

Strawson, Peter F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187–211.

Tiberius, Valerie and Daniel Haybron (2022). Prudential Psychology: Theory, Method, and Measurement. In Manuel Vargas and John Doris (Eds.), *The Oxford Handbook of Moral Psychology* (600–628). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198871712.013.31

Tweel, J. Clay (Director) (2020). *Heaven's Gate: The Cult of Cults*. HBO.

Vargas, Manuel (2013). Situationism and Responsibility: Free Will in Fragments. In Till Vierkant, Julian Kiverstein, and Andy Clark (Eds.), *Decomposing the Will* (400–416). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199746996.003.0017

Vargas, Manuel (2018). The Social Constitution of Responsible Agency: Oppression, Politics, and Moral Ecology. In Marina Oshana, Katrina Hutchinson, and Catriona Mackenzie (Eds.), *The Social Dimensions of Responsibility* (110–136). Oxford University Press. https://doi.org/10.1093/oso/9780190609610.003.0005

Velleman, J. David (2000). From Self Psychology to Moral Philosophy. *Philosophical Perspectives*, *34*(14), 349–377. https://doi.org/10.1111/0029-4624.34.s14.18

Watson, Gary (1987). Responsibility and the Limits of Evil. In Ferdinand Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (256–286). Cambridge University Press. https://doi.org/10.1017/cbo9780511625411.011

Winfrey, Carey (1979, Feb 25). Why 900 Died in Guyana. *New York Times*. https://www.nytimes.com/1979/02/25/archives/why-900-died-in-guyana.html

Wolf, Susan (1987). Sanity and the Metaphysics of Responsibility. In Ferdinand Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (46–62). Cambridge University Press. https://doi.org/10.1017/cbo9780511625411.003

Zablocki, Benjamin (2001). Towards a Demystified and Disinterested Scientific Theory of Brainwashing. In Benjamin Zablocki and Thomas Robbins (Eds.), *Misunderstanding Cults: Searching for Objectivity in a Controversial Field* (159–214). University of Toronto Press. https://doi.org/10.3138/9781442677302-008

Zablocki, Benjamin and Thomas Robbins (Eds.) (2001). *Misunderstanding Cults: Searching for Objectivity in a Controversial Field*. University of Toronto Press. https://doi.org/10.3138/9781442677302

Zeller, Benjamin (2014). *Heaven's Gate: America's UFO Religion*. New York University Press. https://doi.org/10.18574/nyu/9781479825394.001.0001