LOYOLA UNIVERSITY NEW ORLEANS | UNIVERSITY HONORS PROGRAM

Value Drift in the Effective Altruism Movement: A Qualitative Analysis

An Honors Thesis

Presented to

The University Honors Program of

Loyola University New Orleans

In Fulfillment

Of the Requirements for the Degree of

Bachelor of Arts, with University Honors

By

Marisa Jurczyk

December 12, 2019

Leonard Kahn, D.Phil
Associate Professor of Philosophy

Carol Ann MacGregor, Ph.D.
Associate Professor of Sociology

Joseph Berendzen, Ph.D.
Director, University Honors Program

Table of Contents

Acknowledgments

First and foremost, I owe my many thanks to Dr. Leonard Kahn and Dr. Carol Ann MacGregor for being model thesis advisors throughout this project. Dr. Kahn was enormously helpful in guiding me through all of the philosophical nuances of the project, and Dr. MacGregor provided invaluable advice for refining my research methods and analyses. Both made this thesis a genuinely enjoyable learning experience, and I am immensely grateful for their guidance and support over the past year.

Additionally, I was fortunate to have the support of many people in the effective altruism community throughout all phases of this project. Firstly, I thank David Moss and David Janku for helping me select this topic. I would also like to thank David Moss, as well as Richenda Herzig and Catherine Low, for their feedback on and assistance with my interview questions and methodology. Additionally, I am grateful for Darius Meissner and Daniel Gambacorta for discussing their knowledge on the topic with me.

Finally, many of the participants of this study also offered valuable resources, connections, and feedback that I believe helped improve the overall quality of this research. Though I cannot thank them by name, I appreciate the many ways in which they contributed to the success of this project.

Introduction

Five years ago, while browsing his local bookstore, Albert stumbled across Peter Singer's book, *The Life You Can Save*. Intrigued by the title, Albert flipped through the first few pages of the book, decided to take it home with him and read it in its entirety over the next week. Albert was so motivated by the ideas of Singer's work that he decided he wanted to devote his life to doing good per the principles of "effective altruism", the theme of Singer's book.

Albert decided to check out his local effective altruism meetup groups. Within a couple of months, he became a regular attendee, and eventually, he began to help organize these meetups himself. Albert soon quit his job in sales to chase a six-figure salary in managerial consulting and signed the Giving What We Can pledge[1], committing himself to donating fifty percent of his newly-acquired income to one of GiveWell's top recommended charities, the Against Malaria Foundation. Albert had set off on what he thought was the beginning of his journey to making the world a better place inspired by the philosophy of effective altruism.

Now, despite Albert's abundance of wealth, he donates just one-percent of his salary to his child's school. Albert stepped down from leading his local effective altruism group after meeting his now-fiancé and relocating to an expensive condo in New York City. Albert's wardrobe is full of high-end, designer clothes; he proudly shows off his shiny new Ferrari on his commute to and from work each day; and he takes biannual cruise ship vacations with his family. When an article about effective altruism appears on his Facebook feed, Albert feels a bit nostalgic, and even a tad guilty, as he reminisces on the days when he identified as part of the movement.

---

[1] The Giving What We Can pledge is a commitment to donate ten percent of one's income to high-impact charities. More information can be found at givingwhatwecan.org.

Albert's experience is one that people in the effective altruism movement refer to as *value drift*: broadly, a shift away from one's previously held values. (I will explore this definition more thoroughly in the section to follow.) Changes in people's values are far from unheard of. Still, little social research has explored the topic of value drift. Our values guide how we interact with the world, in good ways and in bad. From a moral perspective, then, understanding value drift may be useful in understanding how individuals, organizations, and social movements can do as much good as possible in the world. This paper explores who may be most at risk of value drift, along with factors that may increase or decrease the risk of value drift. I find that people in the effective altruism movement do not report experiencing value drift, nor do they tend to expect to experience value drift in the future. I also find that social networks seem to play a particularly significant role in influencing values, and that competing values, realistic expectations, moral identity, and personality traits also affect people's tendency to experience value drift.

I begin this paper by defining "value drift", along with other changes in values and behaviors that affect our moral decisions. I explain what effective altruism is, and why value drift is particularly important to understand in the context of the effective altruism movement. I also explore academic literature regarding the prevalence of value drift and its possible causes. Then, I use data analyzed from eighteen interviews with people in the effective altruism movement to further understand people's perceptions of value drift and identify factors that are associated with value drift. Finally, I use these results to try to answer the question of whether people should try to avoid value drift and how one might do so.

What is Value Drift?

The academic literature has yet to adopt the term "value drift," though the term has been used informally within the rationalist community[2], widely considered a parent to the effective altruism community, since as early as 2010 (see: Nesov 2010). Value drift is generally described as a change in one's values. In short, a value is a belief about what is important in life. Values are generally discussed in abstract terms and include concepts like happiness, beauty, diversity, equality, and fairness. However, we might have concrete values that follow these abstract values. For example, if one values generosity (an abstract value), that might lead them to value donating to charity (a concrete value). Or, if one values happiness in others, they might value volunteering or other actions that promote others' happiness.

To illustrate how values work in our lives: say "Sarah" values fairness very strongly and considers it one of her most strongly-held values. Sarah's actions are all guided by her desire to be fair and often come at the cost of her own pleasure and happiness. Sarah takes her turn doing the dishes even if she does not feel like it; she leaves enough food for everyone even if she wants more; and she credits the creators of the content she shares on social media even if it takes attention away from herself. However, over the course of several years, months, or, more rarely, a few days, Sarah might begin to value her own happiness over fairness. As a result, she might refuse to do dishes even when it is her turn, take a second course of food even when not everyone has had their first, or post content on social media from other sources without crediting the source as to not avert attention from herself. A change like this is an example of value drift.

---

[2] Rationalists seek to use evidence and reason to make the best decisions possible. The rationalist community is often credited for playing some role in the development of the effective altruism movement, as effective altruism essentially applies rationalism to doing good.

Why did Sarah begin to value happiness over fairness? Sarah might have always valued happiness, but over time, she may have begun to feel that happiness was more important than fairness, or that fairness was less important than happiness. Such a change I refer to as *internal* value drift. On the other hand, Sarah could not have valued her own happiness originally, but something happened in her life that made her realize that happiness was important, and even more important than fairness. I refer to this type of change as *external* value drift.

Some have argued that our values are not motivated by reason and morality, but rather, are motivated by incentives; hence, changes in values are a result of changes in incentives to follow certain values (Alfrink 2019). For example, Sarah may have become part of a community where being fair no longer led to people perceiving her as a good person as a result, or where being fair caused her to lose more money, time, or other resources than she was willing to give. However, each of these changes in incentives is rooted in a value – in this case, social belonging or whatever comfort having more money and time provided her with. This suggests that our motivation is guided by our values, rather than the reverse. Further, reason and motivation are not mutually exclusive, and Sarah's values could have been influenced by a combination of both.

In applying value drift to the context of the effective altruism movement, three possible directions for value change exist: 1) becoming *more* aligned with the ideas of effective altruism; 2) being *just as aligned* with the ideas of effective altruism, but valuing different aspects of the philosophy of effective altruism; or, 3) becoming *less* aligned with the ideas of effective altruism. Unsurprisingly, most discussion about value drift in the effective altruism community concerns the third, as this type of value drift is a threat to, from an EA's perspective, one's personal impact and the effective altruism movement at large. As such, in the context of this

paper, when I refer to value drift, I use it in the context of the third direction unless I specify otherwise.

Our values and the behaviors that follow can change in many ways. I distinguish four types of moral value and behavior changes as *value drift*, *moral drift, lifestyle drift,* and *ethical drift*. Research indicates that our values can, and usually do, change over the course of our lives, typically in accordance with our psychosocial development; for example, people tend to value excitement less and interpersonal relationships as they get older, and their values may be altered as they tend to their families (Gouveia et. al. 2015). In some circumstances, value changes are seen as morally positive, especially when in response to careful reasoning and high-quality evidence. For example, one might encounter evidence about climate change, in response, decide to value helping the environment more than they did prior to knowing of such evidence. However, value changes can also have morally negative consequences. One might go from strongly holding values that motivate one to help others, to strongly holding values that cater to one's self-interest, though often still fall in line with the expectations of mainstream society. I refer to these changes as *value drift.*

However, the possibility of shifting away from moral values towards actively malicious values also exists. Nazism is a (rather extreme) example of this phenomenon; many people who turned to Nazism likely held beliefs that were widely considered normal before becoming Nazis. I refer to these changes as *moral drift.*

While some value changes seem more obviously negative than others, the line distinguishing these can be blurry and can heavily depend on one's personal philosophies. For example, one might perceive a shift away from vegan values as unethical and hence, might classify such a case as moral drift; others may perceive this failing commitment as falling short

of an ideal, though still acceptable, classifying this change as one related to value drift; and still others may not see vegetarianism or veganism as morally good at all. Creating an objective definition of morality, from a social scientist's perspective, is difficult and likely to spark controversy. Hence, I use value drift to refer to a shift away from a supererogatory value, or a value that surpasses the moral expectations of the average person, (for example, going vegetarian or vegan, or donating ten percent of one's income to charity) and towards a value that conforms with society's expectations, whereas I see moral drift as a shift away from an action that meets society's standards and towards a less socially acceptable action.

A distinction between changes in values and changes in the behaviors that follow can also be made. For example, one can value *personally* donating to charity, while still failing to do so (though one could value *others* donating to charity without donating to charity themselves, in which case, their behaviors might still be consistent with their values). This phenomenon has been referred to as *lifestyle drift* – a change in behaviour in which one fails to act upon their values, usually as a result of their circumstances (Meissner 2018). Moving from socially acceptable behavior to deviant behavior is also possible. This phenomenon has been called *ethical drift* (Kleinman 2006). Ethical drift is often a result of the "slippery slope" effect, in which someone commits one small unethical act with the intention of not committing the act again; however, after committing the unethical act once, it becomes easier for them to justify committing it again, and they find themselves slowly making a habit of this behavior. Fraud in business settings, for example, is often a result of the slippery slope effect. Someone might commit fraud "just once" to cover up a bad quarter, but find that they are expected to do so again and again after, leading them to become trapped by social pressure.

Figure 1 below shows the relationships between value drift, lifestyle drift, ethical drift, and moral drift.

	Changes in Values	Changes in Behaviors
Supererogatory to Conforming	**Value Drift:** Changes in *values* from what is considered altruistic to what is considered normal.	**Lifestyle Drift:** Changes in *behavior* from what is considered altruistic to what is considered normal.
Conforming to Morally Deviant	**Moral Drift[3]:** Changes in *values* from what is considered normal to what is considered deviant and unethical.	**Ethical Drift:** Changes in *behavior* from what is considered normal to what is considered deviant and unethical.

Figure 1. Distinguishing value drift, lifestyle drift, moral drift, and ethical drift.

What is Effective Altruism?

According to The Centre for Effective Altruism (2016), effective altruism is "a research field which uses high-quality evidence and careful reasoning to work out how to help others as much as possible. It is also a community of people taking these answers seriously, by focusing their efforts on the most promising solutions to the world's most pressing problems." Those who identify with the effective altruism movement (referred hereafter as EAs[4]) often decide on what problems to work on using three main factors: scale (how many people are affected),

[3] Not currently used in academic literature,

[4] Some people involved in the effective altruism movement have encouraged moving away from the term "effective altruists" (see Toner 2018), as many believe that the name implies that people in the movement are completely effective and altruistic, which is essentially impossible. While I generally agree with this argument, I use the acronym "EAs" for brevity and to prevent confusion.

neglectedness (how many resources are being directed to the problem), and tractability (how likely it is that we can make relatively significant contributions toward alleviating or eliminating this problem) (Wiblin 2017). The most common problems, or "cause areas", that people in the effective altruism community prioritize, at the time of this writing, are health in developing countries; animal welfare, particularly among farmed animals, and more recently, wild animals; and existential risks (that is, problems that have the potential to end the human race), particularly those related to the rise of artificial intelligence. However, a wide variety of other causes have gained substantial support within the community, and new causes are regularly being researched and evaluated.

The effective altruism movement does not call for people to partake in any one specific action or set of actions; as such, how EAs try to maximize the good they do can vary significantly. Though many EAs use similar strategies, no two people are required, nor, necessarily, encouraged, to take the same route in pursuing good. Rather, effective altruism encourages using one's strengths, skills, reasoning, and resources to make the biggest possible impact on the world. Often, EAs refer to one's relative strengths as "comparative advantage" (Todd 2018). These skills and reasoning will vary from person to person, yet EAs remain united by this one, albeit broad, common goal and interest in many similar causes.

Most people get involved in effective altruism in one of two ways: 1) donating to effective charities, or 2) pursuing careers in effective cause areas. Less commonly, people may get involved by volunteering for effective causes or promoting effective altruism to others. People also often involved in local effective altruism meetup groups or online groups, such as the EA Forum, in which they connect and exchange ideas with others in the community.

Who Makes Up the Effective Altruism Community?

The Centre for Effective Altruism (n.d.) uses the "funnel model" to visualize the different ways people can be involved in the effective altruism movement. The idea behind the funnel model, inspired the concept of the sales funnel used in business, is that there are varying degrees to which one can be involved in effective altruism. A large number of people are likely sympathetic to the ideas of effective altruism, but only a small number of people are willing to commit all or almost all of their resources to the cause, and other degrees of commitment exist in between.

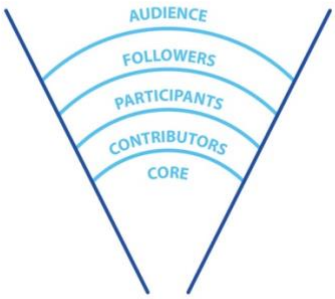

Figure 2. The Funnel Model. *The Centre for Effective Altruism.* Retrieved from https://www.centreforeffectivealtruism.org/the-funnel-model/.

More recently, the effective altruism community has moved away from the metaphor of a funnel and towards that of "filters" (Agarwalla 2019). This metaphor acknowledges that not everyone will – or, perhaps, should – move to the bottom of the funnel.

The Centre for Effective Altruism defines the characteristics of the above tiers as follows:

- The *audience* is not engaged with effective altruism, but may be sympathetic to it. Those in the audience are not yet considered EAs, but could potentially become EAs.
- *Followers* have engaged with the ideas of effective altruism through reading EA content (e.g. books, newsletters, etc.) or perhaps attending a few EA events.

- *Participants* have begun to act on the ideas of effective altruism, perhaps by donating to effective charities and/or exploring effective career routes.
- *Contributors* have committed to EA ideas by taking the Giving What We Can pledge, volunteering for EA organizations, and/or leading local groups.
- The *core* has dedicated most of their resources to the ideas of EA, perhaps by working for an EA organization or by earning-to-give while contributing to the growth of the effective altruism movement. Some people identify *leadership* as a sub-group of the core. Leadership includes people who influence the ideas of the movement as a whole.

Most people pass through each of these stages as they become more involved in the effective altruism movement, though with additional resources or support, some may skip stages. For example, people with direct connections to the core, or people who receive career coaching, may move to the core without becoming contributors or even participants (Agarwalla 2019). The current model, however, does not track how people get less involved in the movement or move out of it altogether.

Value Drift in Effective Altruism and Why It Matters

Values related to effective altruism generally fall into two categories: effectiveness and altruism. As such, value drift in effective altruism might involve a change in the weighting of effectiveness, or altruism, in one's value hierarchy. In the case of the former, one might still care about helping others, but shift towards doing so in a way that follows one's emotions, rather than using rationality and evidence. For example, someone in the effective altruism community might have a family member affected by a particular illness and, as a result, they might devote their life to finding a cure for this illness, even if working on finding a cure for the illness is not

considered to be neglected, large in scale, or solvable. In this case, the person is still trying to do good in the world, but perhaps not in the most effective way.

In the case of the latter, one might stop caring about doing good in the world more generally. For example, someone in the effective altruism movement might donate ten-percent of their salary to charity for several years, but eventually value their own pleasure more than giving to others and stop donating as a result. Such would be an example of a shift away from altruistic values.

EAs might experience value drift in the sense of changing cause areas within effective altruism (for example, shifting from prioritizing global poverty to prioritizing risks from artificial intelligence), though doing so is not typically seen in a negative light by others in the community. Changes in cause areas are not uncommon among EAs, especially for people who have been involved in the movement for long periods of time. In fact, it may be indicative of deeper involvement in effective altruism. As people get more involved in effective altruism, they may be increasingly influenced by the effective altruism community and less attached to mainstream ideas of what causes are most important. The 2018 EA General Survey[5], which surveyed 3,537 people who identified as effective altruists or subscribed to the ideas of effective altruism, found that people who are more involved in effective altruism groups, both online and in-person, are more likely to prioritize artificial intelligence and long-term future as cause areas, and less likely to prioritize global poverty and climate change. (Both of the latter are more generally seen as more conventional causes than the former.)

[5] More information about the 2018 EA General Survey can be found at https://www.rethinkpriorities.org/blog/2018/10/17/ea-survey-2018-series-distribution-and-analysis-methodology.

Preventing value drift away from moral values is important for anyone wishing to be a moral person, but it has especially important implications in the context of the effective altruism movement. EAs sometimes place emphasis on maximizing the amount of good that people can do throughout the course of one's entire career, which often means taking high-earning jobs and finding ways to build career capital early on for the purpose of increasing one's capacity to do more good in the long-term, by having more money to donate and gaining skills that can be applied to effective causes, later. However, in these situations, value drift may cause people to do less good than they otherwise might have if they had taken low-paying jobs with charities, if they stop valuing effectiveness or altruism before they have the opportunity to put these skills to use. Hence, if one is prone to value drift, they should prioritize having an immediate, short-term impact rather than preparing to have an impact in the long-term. Understanding value drift is also important because it affects whether EAs should try to lock themselves into altruistic behaviors or allow themselves flexibility so that they can learn about the best possible ways to do good. If one is prone to value drift and they want to donate money to charity, perhaps they should sign up for automatic monthly donations or put the money they plan to donate into a donor-advised fund, where it must be donated to a nonprofit. On the other hand, if one is not prone to value drift, then they might want to invest that money and take the time to research in-depth what charities are most effective. Additionally, EAs who pursue indirect ways of doing good, such as through the government or academia, may do less good than intended, or to ultimately do more harm than good, if they stop valuing effectiveness or altruism, as these career routes do not directly lead to good outcomes if one's motivation to do good is not sustained.

Understanding value drift is also important to the life and growth of the effective altruism movement because the effective altruism movement depends on moral commitment. If value

drift happens frequently and causes people to become inactive in or to leave the movement, value drift could put the effective altruism movement at-risk. Some have suggested that EAs may be especially vulnerable to value drift because of its younger demographic. About half of EAs are between the ages of twenty and twenty-nine, while another quarter is between the ages of thirty and thirty-nine (Whetstone 2018). Since people's values tend to stabilize as they grow older (Gambacorta 2019), this might mean that leaders of the effective altruism movement should be even more concerned about value drift than leaders of other social movements.

The exact rate at which people leave the effective altruism movement is difficult to calculate, as keeping track of people over several years can be difficult. However, some EAs have begun to research the topic. A longitudinal analysis released as part of the EA Survey 2018 Series suggested that about sixty percent of EAs stay after four to five years (Hurford 2019). Another analysis in the same series found that about sixty percent of people who take the Giving What We Can pledge do not fulfill their pledge in any given year. Further, a study by Joey Savoie (2018) that surveyed thirty-eight donors found that roughly fifty percent continued to donate that same amount over five years, with those who had committed to donating fifty percent sustaining their commitment at a higher rate than those who had committed to donating just ten percent. If being an EA leads someone to do more good in the world than not being an EA (as the philosophy of effective altruism suggests), then, this data suggests that, over the course of five years, forty to fifty percent of people who become involved in the movement at any point do not do as much good in the world as they could be doing with the knowledge they have, which likely has negative consequences on the world at large (though the extent would depend on what people do after they leave the effective altruism movement). Hence, if we care about making the world a better place, then we should care about people leaving the effective altruism movement.

Value drift is likely to be at least one cause of people leaving the effective altruism movement, so if we care about people leaving the effective altruism movement, then we should care about value drift as well.

What Might Cause Value Drift Among EAs?

Social Networks

Social behavior is contagious. One person's behavior influences not only that person's friends' behavior, but also their friends' friends' behavior, and their friends' friends' friends' behavior. For example, in one study, when one person became obese, their friends were then fifty-seven percent more likely to become obese, their friends' friends were twenty percent more likely to become obese, and their friends' friends' friends were ten-percent more likely to become obese (Christaki 2011). The study found similar results among smokers, and the authors argued that this could be applied to other personal characteristics such as political views and creativity. These results remained consistent even when controlling for variables such as homophily, the tendency of people to be drawn to people similar to them, and confounding, the tendency of friends and family to share experiences and be influenced by them in similar ways. Importantly, this study also found that people are only influenced by those we consider part of our in-group. Behaviors spread by influencing what we perceive as normal, and what we perceive as normal is heavily influenced by our in-group. For example, if one person in a group is eating large portions, those who observe this will feel less out-of-place by eating big portions themselves, if they perceive the person to be similar to them. For this reason, in general, people who we believe we are similar to and that we associate with have the most influence on us.

Further, having a strong social network is positively associated with volunteering more and a desire to help others more generally (Tong, Hung, and Yuen 2010). Though this

relationship has been documented as an association rather than a causal relationship, one could imagine that having a strong social network might encourage people to want to give back to their community, might lead people to presume the good in others and be more inclined to help them as a result, and might cause more happiness and life satisfaction, which might encourage them to help others. Hence, EAs may be less likely to experience value drift, particularly away from altruistic values, if they feel strongly connected with any community.

Non-Value-Aligned Behavior

As explained previously, changes in values can result in changes in our behavior. However, the reverse may also be true. People want their behaviors to align with their identities, as conflicts between our behaviors and identities cause cognitive dissonance (Cooney 2011). When people notice conflicts between their actions and values, their values change to alleviate this cognitive dissonance, particularly when changing their behaviors is difficult or impossible. Research has supported this, showing that people who put their time into a cause without a prior interest in the cause, later see themselves as people who support the respective cause and might come to value that cause as a result (Chaiken and Baldwin 1981; Reinders and Youniss 2006). As such, value drift, contrary to what its name may suggest, may not always, or even typically, be caused by a simple erosion of one's values caused solely by the passage of time. Rather, value drift could be a result of circumstances that make it difficult for a person to reach their moral goals, causing them to spend gradually less effort on achieving these goals.

Burnout

The World Health Organization (2019) defines burnout as "a syndrome conceptualized as resulting from chronic workplace stress that has not been successfully managed." Burnout can lead to fatigue and withdrawal from one's job, which can lead someone to perform less than their

best or avoid their job entirely. People working in high-stress jobs are particularly at risk for this. For EAs, these jobs might include politics, academia, and charity entrepreneurship. While not all EAs work high-stress jobs, those who do may be more prone to burnout.

The World Health Organization specifically says that burnout only applies to the workplace and should not be applied to other aspects of life. However, the possibility of tiring out of caring for others and "burning out" of altruistic values still exists. This phenomenon is called "compassion fatigue" and has often been studied among people working with victims of trauma and illness, such as people in the healthcare industry. Most EA-recommended careers do not involve working directly with people in distress, however, so perhaps EAs are less prone to experiencing value drift that is a result of compassion fatigue.

Openness to Change

Developmental psychologists Anne Colby and William Damon (1992) interviewed twenty-three people who exemplified extraordinary moral commitment to look for common characteristics that these people had in common. One recurrent theme from their interviews was a coexisting stable sense of moral commitment accompanied by an openness to change. Colby and Damon found that the people they interviewed held their commitment to morality as an abstract value; however, they remained open-minded to others' beliefs and allowed these beliefs to shape their own, which often led to changes in their concrete values. Though their commitment to morality remained constant over time, their perceptions of morality and how to be moral were constantly refined by their mentors.

In the context of effective altruism, one way we might see this phenomenon evidenced is by a change of cause areas. Perhaps people who change cause areas within effective altruism experience less value drift because they can refine their effective altruist goals and identity,

rather than abandon it altogether, when they encounter criticisms of their prioritized cause. Additionally, staying in one cause area may lead to boredom or even something similar to burnout, which may result in value drift.

Wealth

Much of effective altruism is centered around earning and donating money, but what effect does this have on our ability to keep our moral commitments? Research suggests that having excess wealth might make people more vulnerable to unethical behavior (Gino and Pierce 2009). Based on this research, we would expect that people who pursue high-earning career paths, such as earning-to-give and politics, might be more vulnerable to value drift than people who work in lower-paying career paths, such as nonprofit work and academia.

Moral Identity

One of the most prevalent characteristics found in Colby and Damon's interviews was a connection between the personal and the moral self – that is, their moral goals overlapped with their personal goals. Hence, holding an identity that relies heavily on one's morality may be correlated with experiencing less value drift.

Knowing this, we would expect that effective altruists whose life goals are primarily related to effective altruism will experience less value drift. This characteristic is likely most present in people in the core, whose lives are often centered in effective altruism. People in the middle aspiring to be part of the core may also exemplify this characteristic and, hence, be less vulnerable to value drift.

Is Value Drift a Bad Thing?

The implicit assumption much current research about value drift holds is that it ought to be avoided. However, some critics say that assuming value drift leads us to be less moral than we

could be assumes that our current values are the most moral values we can possibly hold, and further changes to our values can only be bad. Many people believe we likely do not hold the most moral values that we can possibly hold, and there may be ways in which we are neglecting to do good or are even harming others without realizing it. For example, for many years, many people saw no issue with slavery, but most of the world now sees slavery as morally wrong. In this case, the changes of values that mainstream society saw were, by most people's standards, morally positive. For this reason, many believe that we ought not to actively prevent value drift, as value drift may help us gain more moral values in the long-run.

Perhaps this argument holds some validity. However, the assumption that value drift is ultimately good also presumes that our values cannot possibly get worse. Yet, most people can think of examples of people whose values, by mainstream society's standards, have gotten worse, such as in the case of serial killers or perpetrators of genocide. Even if our values do not change quite this drastically, the risk exists of becoming simply apathetic and losing a value than leads one to pursue doing good. The challenge, then, for someone who is trying to do as much good as possible, is how to prevent their values from getting worse while allowing themselves to be open to new ideas that could improve their pre-existing values.

How should one decide whether to privilege pre-existing values or new ideas? No easy answer to this question exists, but nonetheless, the question is still an important one to ask. Throughout this paper, I try to focus on forms of value drift that most people would want to prevent. However, moral uncertainty should be considered, as the possibility exists that the types of value drift most people would want to prevent may not be moral to prevent. If people are skilled at determining whether or not their values are moral, preventing value drift might be advantageous to sustaining moral behavior, if we can set moral values in the first place.

However, if people often miscalculate the morality of their values, then value drift might be beneficial. Determining whether our values are actually as moral as we think they are is difficult, if not entirely impossible, but the extent to which the evaluation of our moral values is accurate is an important factor in determining whether value drift is helpful or harmful.

Ultimately, this paper is a sociological analysis and not a philosophical one; hence, the focus will be on people's experiences with value changes and what is commonly perceived as value drift. The question of whether we ought to prevent value drift will be explored on a surface level using data from this study, but generally, I will leave the philosophy to the philosophers.

Research Methods

Sample

Eighteen people were interviewed in this study. Sixteen participants were found through effective altruism Facebook groups; two were found through referrals from the initial participants. Participants were required to confirm that they had identified as an effective altruist, currently or formerly, for at least six months, and that they had taken some sort of action related to effective altruism, such as attending an effective altruism local group meetup, volunteering for an effective altruism organization, or taking the Giving What We Can pledge.

Design and Procedure

I used a semi-structured interview format with questions about effective altruism involvement, personal experiences with value drift, and perceptions of value drift within the community. Interviews lasted from about half an hour to an hour in length, with the median length of recorded interviews being approximately fifty minutes. Most interviews were recorded and transcribed, except two participants who requested not to be recorded. All interviews were

evaluated qualitatively using a grounded theory approach[6] to identify common and relevant themes among participants and tally how frequently each theme was discussed.

Results

Description of Participants

Participants in this study seemed to have committed more time and resources to effective altruism than the average EA, as demonstrated in Table 1. The difference is likely partially because of selection bias (that is, those who are willing to commit an hour to an interview are probably more likely to commit to other aspects of EA), partially because of the study's qualification criteria, and partially because of the recruitment method. Even though I received referrals of people who were thought to have withdrawn from or left the effective altruism movement, people in the social circles of more involved EA's are likely also very involved, and those who are less involved may still be less likely to respond. Hence, the sample is skewed towards more involved EAs, and hence, the results may reflect beliefs that are more prominent among more involved EAs. However, because the people in this sample are more involved, they may know more people in the effective altruism movement and people who have left the effective altruism movement, and, as a result, they may have more robust observations on why others might engage in value drift. Still, because they are so strongly attached to effective altruism, they may have a more difficult time seeing value drift in themselves. Of course, neither observations of ourselves nor observations of others are fully accurate to begin with, so the above would likely be a limitation even with a more diverse sample. Interviewing more people who have withdrawn from or left the effective altruism movement might have provided different

[6] Grounded theory is a social research methodology that uses allows the researcher to discover themes without testing a specific hypothesis. See Tie, Birks, and Francis 2019 for more information.

results; however, the data still provides insight that may be at least partially accurate and can inform further research on people who have left the movement.

Table 1. Participants' roles in EA[7].

	Sample	**Community-wide based on the 2018 EA Survey**
Full-Time Student (not including recent graduates[8])	16.67%	27%
EA-aligned career (not including earning-to-give)	55.56%	---
Member of Local or Uni Group	77.78%	39.1%
Leader of Local or Uni Group	42.11%	---
Current Giving What We Can Member	61.11%	36.15%
Has donated to an EA endorsed cause	88.89%	---
Mean years involved in EA	4.41	3.96*
Median years involved in EA	5	3*
Mode years involved in EA	5	2*

* Calculated from raw data

[7] I did not collect demographics such as age, gender, and race. Based on personal observation and assumptions, I suspect that the demographic distribution of race is skewed white, of age is skewed younger (mostly 20-29-year olds), and of gender is skewed female (with the possibility of non-binary identifying people), relative to the community-wide rates.

[8] Two people who had graduated about a month before being interviewed were included in the sample. Including recent graduates, the sample was 27% full-time students.

Perceptions of Value Drift

Fifteen of the eighteen participants felt that they had not experienced significant value drift, in the sense of becoming less aligned with effective altruism, in their time as an EA. Of the three who reported becoming less aligned with effective altruism values, all three worked in careers not directly related to effective altruism. One did not entirely identify as someone who was "earning-to-give" because of their relatively low pay, though they did donate some of their income to charity. One participant expressed that they had experienced some value drift after temporarily moving away from a city with a large EA population. "I was with people who were different, and different parts of me took over," the participant said. But, when returning to the city, they heard what others in the community were working on and began to "feel I'm not doing anything good with my life," and returned to EA as a result.

Another participant said that they had experienced value drift after overcommitting to EA for some time:

> "I have [experienced value drift], but I think it's a good thing because I don't think that I necessarily value EA less so much as I value my health and my sanity a lot more. And part of it is that I've come to the extremely obvious but difficult conclusion that it's really hard to get anything done if you don't take care of yourself. But it's also, I don't want to feel terrible all the time."

Some other participants expressed smaller instances of value drift and lifestyle drift, often related to conflict between frugality for the sake of giving more to charity and other values such as personal comfort and social expectations. These participants were convinced by Peter Singer's

drowning child argument[9] – that, if one has the opportunity to save a life at a relatively low cost, they have an obligation to do so – and wanted to spend as little money as possible so that they could give more to charity and potentially save lives. However, when applying this to their lives, participants found this to be difficult, unsustainable, and in extreme cases, mentally unhealthy. Hence, they usually decided that some amount of personal comfort, for the sake of preserving their mental health, was necessary to maintain to be sustainable. One participant (who has donated ten percent of their income over the last ten years), expressed this well:

> "I could spend more of my time beating myself up that I don't go for higher money and promotions, and that I waste money going to the cinema and going on holiday rather than giving it all to effective causes, but you have to set realistic benchmarks if you want to not burnout of your values. I feel like the times when serious value drift happens is when you've burnt out of a value. You've gone, 'no, I can't do this; therefore, in order to continue believing that I'm a good person, I need to reject this value, because otherwise I will just continue feeling like I'm a failure.'"

Most participants expressed experiencing value drift, in the broader sense of value changes more generally, but considered these changes to be morally positive (for example, deciding that animals are morally relevant). Many participants also expressed getting better at implementing their pre-existing values since joining the effective altruism movement:

---

[9] Peter Singer's "drowning child" argument is more thoroughly explored in his book, *The Life You Can Save.* Singer notes that the cost of buying a malaria net, which has the potential to save a child from dying from malaria, costs about $5 in USD. Charity evaluator GiveWell estimates that the Against Malaria Foundation, which distributes anti-malaria nets, can save one life for every 460 nets distributed, making the cost of saving a life by donating to the charity about $2,300 in USD.

> "I've been looking for this movement since I was [a teen]. And when I found it, it's like, oh, this is all of my values, except people have thought about it more strongly than I have, and I know how to implement it better now."

Another participant expressed similar feelings:

> "I think that the motives of making the world a better place and working as efficient, or effective as I can has been always with me. I would say effective altruism even gave me the tools to fulfill the values that I already had. Maybe it gave me the perspective of looking at something from a moral sort of view, maybe it gave me some broad view about the global issues… I would say, I used to have this values, but effective altruism somehow was a way for me, a tool for me to pull these values into practice."

One participant reported having more difficulty executing EA values over time because of their circumstances, but these people said that this did not affect their values themselves. The participant said:

> "Operationally, [having a family] makes things a lot more complicated. I view it as a constraint to the optimization rather than a change to what it is I'm optimizing for… but, if you care enough about the optimization, then you minimize constraints in ways that would change things, and I'm definitely not attempting to do that."

Some participants also noted being more worried about circumstances interfering with their ability to act on their EA values in the future:

> "If I did drift away from EA, I would assume that circumstances are pretty bad, causing me to, having lost my job or I'm just super depressed or something, so I wouldn't want that to happen… I'm more worried about my ability to execute EA rather than not caring about EA."

All participants, except the three who reported experiencing significant value drift, said that they have either stayed equally aligned or have become more aligned with EA values over time, and no participant could identify a moment since first discovering EA that their alignment with EA values decreased. However, one participant discussed the possibility of being perceived by others as becoming less aligned with effective altruism if the ideas of the movement as a whole changed:

> "The EA Handbook, when that was released, there was so much splashback about how they just really focused on AI safety, and some people said, 'that might be the inner circle's main cause area, but that's not everyone's main cause area', and it was written in a really biased manner. I think the only way [value drift I experience] could be viewed negatively would be if I stay focused on, maybe I drift to treatment, as malaria treatment becomes more and more available right now… and other EAs start focusing on MIRI [the Machine Intelligence Research Institute], and I could see that as a bit looking like we're different. But I don't think there's a change in my core values. There might be a change in the community's values that I don't progress to."

Another participant observed similar behavior in others:

> "In EA, I don't know if it's value drift exactly, but there have been changes in the values of the movement overall, or at least, in what org[anizational] leadership is pushing, relative to what the average person who comes to a meet-up twice a year thinks. I think there's a lot of change in emphasis on, definitely a movement towards far future, a movement towards elitism even more unfortunately. I don't necessarily think, at least, [that] the far future stuff is bad necessarily, but I can definitely see someone leaving the movement because they're staying in place."

Only one participant reported that they felt they were likely to leave EA in the future. This person felt that their priorities were likely to change because "regression to the mean" – that is, that people in general are likely to move closer to average in their beliefs over time. They also felt like they did not have as much reason to stay involved in the community, as they had no plans to do direct charity work and experienced many conflicts with the EA community.

For other participants, particularly those who had been involved for long periods of time, their confidence in their ability to retain their altruistic values was primarily because their values had changed little, if at all, over recent years. Most participants seemed to think they were unlikely, and often less likely than the average EA, to experience value drift. This belief was partially based on their involvement and sustained interest in the effective altruism movement, and partially because they found that being involved in effective altruism was beneficial to them. As one participant said:

> "[Being aware of the risk of value drift] hasn't actually affected decisions that I've made. People say that if you're worried about value drift, be involved in the community in a social way and it's like, well, I already do that, and I have lots of reasons to do that because I just enjoy it."

Participants generally had complex feelings about whether they felt they should actively avoid value drift. Two people seemed adamant that changes in values are not usually bad, as these changes are the results of learning and new experiences, and both said that they deliberately try to embrace value drift because of these beliefs. One participant emphasized that "it's okay not to be an EA" and that EAs should stop shaming people for leaving the effective altruism movement. Another participant said:

> "I used to be [concerned about value drift], now it's kind of not that big of a deal if I would. I just do whatever I want, and what I want will be what I want. I don't necessarily want to change that."

One participant noted how the value drift they have observed in EAs seems to be morally positive:

> "I have seen value drift, but in the direction that effective altruists will consider as good. I've seen people who are more willing to identify as altruists and to care about the world. And I've seen fewer people gone in the [other] direction."

No participant was entirely anti-value drift. Others either did not comment on the nature of value drift or had mixed feelings about it. Most participants do not actively prevent value drift, but they reported that many of the activities they do for other reasons help prevent value drift: for example, reading the EA Forum, attending local group meetups, and living in or visiting a city with a large EA population.

Many participants felt that the average EA should be concerned about value drift, given the base rate for drop off and, more generally, the essence of human nature. However, most people seemed to feel that EAs are less prone to value drift than the average person because EAs have goals that they are working towards, whereas many people outside of EA do not. As one participant said,

> "I think EAs are probably on average less likely to be susceptible to value drift than other people because… an effective altruist has already jumped several hoops in having fairly strong, stable values, and having sat down and thought about your values and considered what your values are, which is a thing that many people just haven't done at all. I think

> people who haven't explicitly sat down and considered what they value are more likely to value drift than people who haven't considered what they value and why."

This relationship seemed causal; a couple of people reported that before being involved in EA, they experienced a significant amount of value drift because they had no values to anchor to, but after getting involved in EA, they experienced significantly less value drift. As one participant said: "I experienced [value drift] most radically before I was involved in EA. I was still an altruist but I didn't had an outlet, and I kind of even forget about that for years sometimes."

One participant said that they thought people in general should be concerned about value drift and that EAs should not necessarily be exempt from this:

> "In other people, I worry about [value drift] somewhat, but I'm heartened by the fact that people are thinking about it and talking about it… I think I've seen too many people just following the wind without any principle at all. I've arrived at this belief that the average person could probably do with half an hour of thinking about this."

The same person said that people's tendency to experience value drift might be rooted in Western cultural norms:

> "I do feel like our culture promotes a sort of throw away your past self, like this is a way to not feel embarrassed about that stupid thing you said at the high school dance, and this is a way to not feel bad about the years you wasted on your first major, like our culture sort of endorses, like it's ok, you're allowed to shed your skin and be like the new version of you, that's great."

Several other reasons why people might experience value drift were also reported, which will be explored in the next section.

Factors Affecting Experiences with Value Drift

Participants reported many ways in which they have observed themselves and others moving away from EA values. Participants were also asked to think of reasons for why they, and others, might hypothetically leave the effective altruism movement in the future. Below are some of the most common themes discussed.

1. Connection to the EA Community

Community was the most prominent theme in the interviews, with fifteen participants reporting losing connection with the EA community as a reason that they, or others, might leave the effective altruism movement, with some attributing their involvement in effective altruism to their connection with the effective altruism community. As one participant said:

> "If I grew up in whatever country lacks any EAs, and all it is is this online community -- I don't know anyone in real life, I've never met anyone else who's an EA -- it probably would be a lot harder for me to become as interested in effective altruism. So I think personally, also, the fact that I can go speak with other people who are interested in the same ideas causes me to become more active in the community."

As mentioned, most of the participants reported not experiencing significant value drift. Fourteen of the eighteen participants interviewed were involved in a local effective altruism group, and those who were not had other EA-aligned communities they could connect with online or through work, implying that perhaps being involved in a local group or other EA-aligned communities might be correlated with less value drift because people in this sample generally did not experience a significant amount of value drift. One of the people who reported experiencing value drift reported doing so because of temporarily leaving an EA community, and another mentioned having conflicts with the EA community. In addition, those who felt most connected

to and had positive feelings towards the community generally seemed more involved in the movement, and those who felt less positively towards the community tended to limit their involvement in the community more (though this did not seem to affect private actions, such as how much they donated).

Some also mentioned that connection with people who share the same values also creates pressure to sustain those values, as drifting from those values may cause conflict or a loss in status or respect from others. One participant mentioned:

> "I sometimes felt that reasons why I do this work… have little to do with animals anymore. It's all for some sort of social pressure, some kind of wanting to show somebody something."

Seven people specifically mentioned having a conflict with the community as a reason for leaving the movement. Two of the participants reported mostly negative experiences with the effective altruism community: one of these participants remained marginally involved with the community itself and pursued most of their impact through donating to charity. The other participant expressed a strong likelihood of moving away from the community in the near future. Some participants also mentioned knowing of other people who have left the community due to negative perceptions of or conflicts with the community. One person told the story of someone they knew who had left the movement, saying: "He basically just decided that effective altruism was just a bunch of elitist assholes who didn't care about normal people, so he dropped out of the movement." Many participants acknowledged the possibility of negative perceptions with the EA community leading to negative perceptions of EA values.

2. Competing Values

Twelve participants cited competing values as a possible reason for experiencing value drift in the future, or as a reason that people they know have experienced value drift. Family was an especially significant value that many felt might affect their current value hierarchy; ten people expressed that starting a family may cause, or has caused, them or others to become less interested in effective altruism in the future. One participant said:

> "I think if there was a reason why I wasn't involved, it could be that I have a family at that point… It could just be that I'm so overwhelmed with being a parent and running a household that I don't have the extra cash and I don't have the extra time to be involved with the community at that point."

Another participant mentioned:

> "I would be happy giving more than [ten percent of my income], but I would not be ok with giving more money for that if I was sacrificing things for my kids. I buy the utilitarian ethics piece, and I'm also selfish. I try to be selfish in ways that are non-harmful, but that's kind of the balance that I have."

A third participant expressed similar feelings:

> "There might be a lot of reasons [for leaving the EA movement in the future] I can imagine. For example, having a child, or twins or triples, and focusing all my income on them. It's not effective according to effective altruism, but when I have my own children, I would spend and do my best to provide resources for them."

Other competing values include personal comfort, social expectations, or other non-EA causes and social movements. Participants did not typically perceive these value conflicts negatively. For example, one participant said:

> "I value social comfort and my community I think more than EA, being that the two really work closely for me. But when it comes to it, there's meals that my mom has made since I was a child that it means a lot to her that I eat. So if she's cooking a turkey, I'm going to go, and I'm going to eat the turkey. And no matter how I feel about animals, that's always been the case."

Another participant noted a similar experience:

> "Socially, there's an expectation that you support some of the community organizations that you're involved with. And I'm fine with that, and I think that it makes sense to be kind of paying your fair share for public goods, local public goods, so I'm happy to kind of contribute in that way."

One participant noted a conflict between frugality and gaining social status, though both of these conflicts were rooted in the desire to do good:

> "So, I've probably become less frugal in some ways than I was a couple years ago, and I think that's been to my advantage, I think it's helped… If there's a chance to vocalize with interesting people, I shouldn't be like heavily optimizing on exactly how much food I order, and as a heuristic, to be more willing to get to do what's socially appropriate in a situation, in order to increase the chance of making positive connections, it's still with the ultimate EA goals in mind. But I do it more from a heuristic lens, and if I decide this is something would be useful to have to improve conveniences in my life, I just get it instead of trying to spend a lot of time thinking about whether the time gains or stress reduction gains would be worth having a little less money."

Another participant noted that often, when people move out of alignment with the movement, it seems to be in alignment with other values they hold:

> "I suppose that I see some people whose circumstances have changed, and they got a new job and now they dedicate more time to their job than to effective altruism. But in all those cases, it makes - it seems consistent with the values of the person. It's not like the values have changed. It's more like the circumstances."

3. Involvement in EA

As alluded to in the community section, regularly engaging with effective altruism seems to be an effective way to retain EA values. Ten people reported that they engage with EA content, attending local group events, and being involved in the EA community in other ways and noted that this seems to affect their willingness to retain their values, regardless of whether they engage with these activities specifically to prevent value drift.

Taken a step further, having responsibility within and putting resources into the movement also seems to be associated with less value drift. As one participant noted:

> "I think that taking more responsibility is sort of -- it's not so much going to activities but actually having to organize them or something like that. It makes you more aligned or more concerned with EA."

4. Open-Minded Thinking

Counterintuitively, people who did not experience external value drift, in the sense of becoming less aligned with effective altruism, often reported readily accepting new ideas and experiencing concrete value drift quite a bit. Nine participants mentioned changing cause areas at least once during their time in effective altruism or being open to new and unusual ideas more generally. One participant reported:

> "I was gonna say I haven't experienced value drift, because originally I was thinking about it as, value drift is a bad thing where you leave EA or something, but then I was

like, well, I've changed where I donate to, and I've changed my values, like becoming vegan and becoming more concerned about the long-term future."

Another participant said:

"I'm very open to new ideas and weird ideas. I don't have a lot of commitment to any strong set of beliefs or something that I feel like can't be challenged."

Participants did not, however, specifically attribute their openness to new ideas as a reason for not experiencing value drift, or for others experiencing value drift.

5. Sustainability and Realistic Expectations

Five participants talked about setting boundaries with EA to be more sustainable -- for example, donating less than what they could feasibly do. As one participant mentioned:

"I'm committed to Giving What We Can, and I think that that's really important. I'm also really comfortable with the idea that I'm not currently in a position where I want to start trying to give a lot more than that because I want it to be sustainable."

Another participant talked about setting boundaries as it relates to being involved with EA activities:

"I'm actually worried about the extent to which [effective altruism]'s taking over, not just my time, but like, what I'm thinking about in my spare time. So I'm actually, if I can get two weeks where no one needs me, I'm actually thinking I'm going to do a brief EA detox at some point, just to be in a healthier place with regards to having other things in my life."

Three participants cited overcommitment as a challenge in fulfilling their values. One of these participants said:

> "I'm pulling harder towards EA than I've pledged too, which brings a little bit of tension into other areas… Part of the challenge is, in one sense EA is a source of meaning, and without EA, I would have drastically less meaning, but it's not enough to keep me going without other things in my life be going well, so I still have to make sure I'm socializing, exercising, stuff like that… It's motivating to pull in the direction of EA, but that's really not the direction that you need to pull in for the long-term sustainability for acting in an EA manner, or having the capacity to do things."

One mentioned overcommitment as a threat to the EA movement as a whole:

> "I think that there's a common pattern observed with religions… the first converts will die for the thing, and then over a few generations it becomes something that can fit into an otherwise ordinary life. And I think there are people who believe in very imminent x-risks who believe that the best way to have EA go is to have a few people who will die for the thing because otherwise the human species is doomed, but looking more from a perspective of the less existential stuff, it seems like we don't want martyrs, we want people who can be more sustainable."

A couple of participants, on the other hand, seemed to think that EAs might be too concerned about sustainability:

> "A lot of the concern has been about burnout, but that doesn't seem like much of the problem… there certainly doesn't seem to be evidence that we need to be concerned about people doing too much."

More broadly, others talked about people in the movement becoming frustrated as they struggle to meet their goals early on in their involvement in the movement, and some lose motivation because of this:

> "I've seen people find that it's difficult for them, like they try some path to having an impact and it ends up being difficult and it doesn't work, and they end up not feeling good about it. They try looking for direct work and don't end up getting hired by any organizations and aren't sure what they're supposed to do, and so they stay with their traditional job and just kind of give up on it."

Another participant noted observing something similar:

> "I feel like value erosion mostly actually happens following in the footsteps of ambition values the most. People who have a dream or vocation for a certain career or they're going to be a brilliant writer or something, and then reality sets in. I think that's the kind of value that often drifts the most."

Seven participants discussed mental and physical health as a constraint on their, or others', ability to fully live out their values. However, participants make peace with this conflict, realizing that maintaining mental and physical health is important to effectively being able to do good. As one participant said:

> "Part of me was disappointed when I came to the conclusion that I may have been over-optimistic in the past about my ability to pursue highly competitive careers while still maintaining my mental health, and in that choice, the right choice is to prioritize my mental health... But I would have liked to have the ability to do more and to commit myself more in whatever ways. But it's whatever. The way that I think about it to make it not a big deal is, yeah, I wish I had a billion dollars to donate, too. No use sitting around wishing all day. I wish that I could solve poverty by snapping my fingers."

No participants specifically cited burnout as a reason that they might leave the effective altruism movement in the future. Two mentioned it as a reason others they know have left the

effective altruism movement, and as previously mentioned, one person said burnout led them to withdraw from effective altruism for some time.

6. Personality

Finally, some personality traits seemed to be prevalent among people who do not experience value drift. One trait that seems to be associated with less value drift is a natural inclination towards effective altruism and utilitarianism. Six participants expressed having held EA beliefs before coming across EA for the first time. As one participant said: "I was basically trying to do EA before I knew the movement existed. I think that I would be doing it on a desert island."

One participant mentioned that people who tend to value drift seem to lack an initial alignment with EA values:

> "I've had friends who say that in principle they think that donating some fraction of their income is the right thing to do, and yet they don't really seem to believe that on a gut level or on every level, so that prevents them from taking action on that belief, and eventually they'll just decide that no, I don't actually care about it that much, I'm not going to do it. And that kind of seems to me like value drift. There's also a sense that the initial belief that they had wasn't fully EA about it. Like they had some doubt in their mind all along, and they seem to acknowledge that that was there or something. It doesn't seem like their values were actually drifting. Maybe they're just mistaken on what they actually believe."

One local group organizer noted a relationship between early commitment and lack of value drift:

> "The people that I know closely haven't really drifted, but those are more committed EAs as well. That relationship between how committed you are at the beginning and how much you drift is pretty negative. The more committed you are, the less you drift is pretty much what the evidence says."

The same participant said that their value stability is, in part, a result of their strong EA identity:

> "I'm good at being really committed to something. Also, it just defines me so much. When it's a very core part of your identity, and you're getting positive reinforcement by being around a lot of EAs, then it will continue to be part of your identity, I feel like for me it would decrease my utility if I were to abandon this, I would feel purposeless, and, yeah, so it's both personally positive for me and I just like the idea."

Another trait that seems common among those who are less prone to value drift is a tendency to be value stable throughout life. Four participants cited that, based on their prior experiences, they felt they were naturally less inclined to experience value drift than most people. One participant mentioned: "I'm autistic, so I think that makes me predisposed to not like change, and therefore when I find something that works, I tend to stick with it."

Another participant specifically said that "a cornerstone of my personality is not having value drift, and not experiencing value drift, and I work really really hard to sort of stay aligned with all versions of myself from the past." When asked about reasons they might leave the effective altruism movement in the future, they said: "I can't think of anything that isn't, just, disaster. Like I hit my head really hard, or I had a schizophrenic break and decided that existence is bad in itself. It would take something I have a hard time predicting." While all other participants were able to identify reasons that their values might change that do not involve

changes, several expressed that they did not think the reasons they cited were likely, in part because of their tendency not to experience value drift throughout life.

Discussion

Perceptions of Value Drift

In general, participants in this study seem to think that value drift is worthy of concern among the human population as a whole, less worthy of concern among the effective altruism community, and even less worthy of concern in themselves. Participants may be telling the truth here; as many participants seem to believe, EAs and others involved in action-oriented social movements may legitimately be less prone to value drift because they are more aware of their values and have goals related to them. Research has supported this, showing that setting goals can motivate someone to accomplish them (Locke and Latham 2002). Further, given the skew of this sample, participants may legitimately be less prone to value drift than the average EA. So, to some extent, participants' beliefs about who is most likely to value drift may be somewhat accurate. Still, this finding may also indicate that EAs, and perhaps others involved in social movements, tend to underestimate the risk of value drift among themselves and the community because of overconfidence bias. This overconfidence might also extend to others in the effective altruism movement; if the effective altruism community as a whole is part of the participants' in-group, the qualities generalized about EAs might reflect on themselves as well, and people generally try to maintain a positive self-image, when possible. EAs and others involved in morally-motivated social movements might also be less likely to notice their own cases value drift due to moral licensing – that is, they may use their morally motivated actions, such as donating to EA charities, to excuse themselves from partaking in other morally motivated actions, such as volunteering for an EA organization, even if the latter is more impactful than the

former[10]. Though this study does not indicate how prevalent value drift is among EAs, people in the community should, perhaps, assume that they are more likely to be vulnerable to value drift than they intuitively think they are because of the biases they hold.

Worth noting is that these perceptions of value drift are further complicated by the subjectivity of value drift; what one person may see as value drift, another may see as a value-aligned change. For example, one person might think that someone who moves from social justice movements to the effective altruism movement experienced value drift because they became less interested in social justice, while another might say that the person is simply further pursuing their value of doing good and helping others in a different way. Part of this distinction comes down to whether values are defined abstractly or concretely, as discussed previously, but it might also come from personal biases and perceptions about social justice and effective altruism. In sum, self-reported value changes, both in ourselves and others, are difficult to define and measure objectively, and hence, readers should not place too much weight on responses reports of their tendency to value drift and that of others.

Most EAs express complicated feelings about whether value drift should be embraced or avoided. They seem to recognize cases in which value drift can be morally good, morally bad, or morally neutral, both in their own lives and in others. Hence, the moral consequences of value drift, based on participants' experiences, seem to be dependent on the circumstances. From the standpoint of morality, then, perhaps people should not attempt to entirely prevent value drift, or to entirely embrace value drift, because each change in values can be harmful, helpful, or neither. As such, people should judge the morality of each value change individually. Unfortunately,

[10] This is only an example; cases in which donating to an organization is more impactful than volunteering exist.

people tend to be poor judges of their own morality, so more research on how to critically observe our own value changes may be useful for this purpose.

Value Hierarchies

Competing values were frequently cited as a reason for moving away from effective altruism, both among the participants and in their observations of others, suggesting that internal value drift is relatively common among EAs. However, most people likely have little desire to prevent this type of value drift; our values, by definition, exist because we believe they are important, so giving additional weight to an already important value in one's value hierarchy is more likely to be perceived by the person holding that value as morally positive. Intentionality seems to be important to preventing morally negative value drift from conflicting values; being aware of one's own hierarchy of values and of how certain values might change in the face of distractions, and making key decisions based on those values, may lead to better value-aligned outcomes, rather than making decisions based on the easiest or most convenient options.

Social Networks

As expected, engagement with the EA community and EA activities seems to be correlated with less value drift. The direction of the causality of this relationship – or whether it is causal at all – is unclear from this research alone. Perhaps people who are willing to commit to EA more are less likely to value drift, or perhaps getting more involved in the EA community causes less value drift. Previously discussed research about the effects of our behaviors on our values suggests the latter (Cooney 2011; Chaiken and Baldwin 1981; Reinders and Youniss 2006), though the relationship could potentially work in both directions. For EAs who want to avoid value drift, engaging with the EA community through local group meetups and conferences, maintaining positive relationships with people in the EA community, and taking

responsibility in the EA community through mutually-beneficial volunteering or EA-aligned career paths are all likely to help avoid value drift. For EA organizations wanting to prevent value drift among the EA community as a whole, improving welcomingness and controlling rates of EA movement growth such that too many non-value-aligned people do not enter all at once might be effective approaches to take.

However, in light of the association between value stability and involvement with like-minded communities, the risk of groupthink exists. For many highly-involved EAs who get along well with the community, the risk of groupthink is probably even higher than the risk of value drift from involvement with non-value-aligned communities. Groupthink may cause EAs to give too much weight to existing popular perceptions about how to do the most good in the world and may inhibit EAs' ability to be open to causes that may be highly effective, or even *more* effective than the most popular EA cause areas. Hence, getting EAs involved in the effective altruism community, while preventing them from being too insulated by the community, is a careful balance that must be made here. Maintaining diverse communities with diverse beliefs within the effective altruism community may help alleviate this, but does not solve it completely, as people might still choose to interact with communities that share their own beliefs. Further research could explore the risks of both groupthink and value drift more deeply.

Getting Involved

Participants' experiences support existing research suggesting that our behaviors can affect our values. People who are more involved in the movement seem to be more value stable. This relationship appears to be causal, as people tend to report feeling more aligned with effective altruism ideas as they get more involved, but the possibility that people who are value

stable also happen to tend to get more involved because of other overlapping factors in their personality still exists. In any case, providing opportunities for people who want to get more involved with effective altruism to get more involved with the movement seems like an effective way to prevent value drift.

Sustainability

Participants acknowledged the importance of living a lifestyle that was sustainable over the long-term to maximize their long-term impact. Somewhat surprisingly, participants did not seem tremendously concerned about burnout and did not frequently see it in others. Burnout is still a possible cause of value drift, but in general, people seem to be more at risk from becoming under-involved rather than over-involved. Still, taking care of one's self, addressing physical and mental health issues, and ensuring that one's goals are in-line with their abilities all seem to be effective ways that individuals can sustain altruistic values.

Open-Mindedness

Participants who did not report experiencing value drift seemed to reflect Colby and Damon's findings regarding experiencing high amounts of concrete value drift guided by altruistic abstract values. Most participants had changed cause areas several times throughout their time as an EA and otherwise mentioned an openness to new ideas. However, there did not seem to be a significant difference between those who reported experiencing significant value drift and those who did not. Further, we might expect people in the effective altruism movement, regardless of whether or not they are prone to value drift, to be open to new ideas, given that effective altruism is a relatively new idea that has not taken off in the mainstream yet. So, this data does not provide strong evidence suggesting that openness to change inversely affects value drift, but it does not disprove the idea either.

Wealth

As mentioned, two of the three people who reported significant external value drift were on earning-to-give paths at the time they experienced it, but neither attributed their high incomes to their experience with value drift. The third person reported experiencing value drift for the opposite reason: not having enough money to continue sustainably donating to charity. Other participants reported observing value drift in others as a result of wealth occasionally, but wealth does not seem to be a prevalent cause of value drift (perhaps because wealth itself, at the level in which it inhibits morality, is relatively uncommon), though it still may be an important factor.

Most value drift participants observed in others related to wealth is also related to social pressures; wealthy people tend to have more ties with other wealthy people, which may lead them to feel pressured to follow social expectations associated with wealth, many of which conflict with effective altruism ideas and morality more generally. In addition, people on earning-to-give paths have to take more time outside of work to connect with the effective altruism community, which may make it more difficult for them to stay in-line with effective altruism ideas. Those who work in EA-related careers have can have an impact while maintaining interaction with EA ideas and the community all during their workday, then can go enjoy other hobbies and participate in other activities after hours. Those who earn-to-give, however, must go to work, then connect with the EA community outside of work hours to stay engaged with the community and its ideas. Hence, people who earn-to-give might be more at risk of falling out of engagement with the effective altruism community, and more at risk of overcommitting and experiencing burnout, if they spend too much time engaging with the effective altruism community while keeping up with a full-time job. So, wealth, contrary to what

previous research has suggested, might not be a direct cause of value drift, but it could perhaps impact one's ability to maintain their values.

Personality

Many participants report feeling that they are less likely to value drift because of their personality. This result might suggest that personality plays some role in people's tendency to experience value drift, or it might further reveal how people engage in self-serving biases. More research on the extent to which our tendency to experience value drift is determined by our personality might be help determine the extent to which participants' perceptions are true.

If personality plays a role in the likelihood of experiencing value drift, EA organizations interested in preventing value drift should strategically engage broadly with people who are most naturally inclined to become and stay involved with the effective altruism movement, and more deeply with those who are less inclined to stay involved. (Reaching out to people who are likely to be sympathetic to effective altruism is probably more effective – and ethical, for those who are against exploiting others – than simply reaching out to people who tend to be value stable.) Research on what determines whether someone is likely to become and stay involved with effective altruism may be useful to support these efforts. For individuals, evaluating one's personal history of changing values may help reveal one's tendency to value drift; however, even the most value-stable people may experience value drift in response to drastic changes in their environment, so personality alone likely cannot prevent value drift.

Limitations

As previously mentioned, the sample used in this study was generally more involved than the average EA, so the experiences detailed above primarily reflect those of highly-involved EAs. However, these people might have a deeper understanding of the EA community and have

deeper insight than those who are less involved. Also, the sample consisted of only eighteen people, so these results may not hold up among the general population. Further research on these themes would be useful in discovering more significant results.

In addition, I (the author and interviewer) work at an organization that promotes effective altruism, and some of the participants expressed interest in working for an EA organization in the future; hence, some participants might have altered their answers to paint themselves in a more positive light for a potential employer. At the same time, I think my familiarity with effective altruism helped facilitate deeper conversations and facilitate trust among participants. There were also some participants that I knew personally. I do not expect the dynamics of these relationships to have had a significant impact on the results, but readers should be aware of them as they interpret these results.

Finally, this study relies on self-reported data about experiences over several years, which can often be unreliable. I have tried to note where self-reporting may have resulted in biased answers, but this limitation should also be considered in these results.

Conclusion

In this paper, I have explored factors that influence the experience to value drift and the implications of these factors on how we ought to live our lives. I have used the effective altruism movement as a case study and examined if and how EAs alignment with the effective altruism movement might change over time. Based on this research, it appears that, though most EAs have complicated feelings about the morality of value drift, they are typically not concerned about value drift, for better or for worse. I have found that social networks seem to play a significant role in the stability of and changes in people's values; that conflicting values can often cause other values to be de-prioritized; that getting involved in the effective altruism

movement can lead to one feeling more aligned with its ideas; and that traits like natural inclination towards effective altruism and a history of value stability might be correlated with less experience with value drift.

The question of whether EAs, and humanity as a whole, ought to prevent value drift remains complex. Further research on the direct causes of value drift among a wider population, interviews with people who have left the EA movement, and studies on value drift in other social movements more generally could be useful contributions to this field. Learning more about value drift is and will continue to be important, as our values shape how we interact with and care for the world. By understanding value drift more deeply, we can ideally become and stay better people in the long-run, help others to do the same, and make the world a better place.

References

Agarwalla, Vaidehi. 2019. "A Sociological Model for Understanding Effective Altruism Coordination Challenges." Poster presentation at Effective Altruism Global London 2019.

Alfrink, Toon. 2019. "Against Value Drift." *Effective Altruism Forum.* Retrieved from https://forum.effectivealtruism.org/posts/qYLqgoa7anKFEByJ6/against-value-drift.

Centre for Effective Altruism. 2016. "Introduction to Effective Altruism." *Effective Altruism.* Retrieved from https://www.effectivealtruism.org/articles/introduction-to-effective-altruism/.

Centre for Effective Altruism. n.d. "The Funnel Model." *Centre for Effective Altruism.* Retrieved from https://www.centreforeffectivealtruism.org/the-funnel-model/.

Chaiken, Shelly and Mark W. Baldwin. 1981. "Affective-Cognitive Consistency and the Effect of Salient Behavioral Information on the Self-Perception of Attitudes." *Journal of Personality and Social Psychology, 41*(1), 1-12. Retrieved from https://www.mcgill.ca/social-intelligence/files/social-intelligence/Chaiken_Baldwin_1981.pdf.

Colby, Anne and William Damon. 1992. *Some Do Care: Contemporary Lives of Moral Commitment.* New York, NY: The Free Press.

Cooney, Nick. 2011. *Change of Heart: What Psychology Can Teach Us About Spreading Social Change.* New York, NY: Lantern Books.

Dullaghan, Neil, 2019. "EA Survey 2018 Series: How welcoming is EA?" *Effective Altruism Forum.* Retrieved from https://forum.effectivealtruism.org/posts/eoCexTGET3eFQz3w2/ea-survey-2018-series-

how-welcoming-is-ea?fbclid=IwAR0MtdJ9DlLFN5F6zIoPKfIB3HOeNapmY5M3pBB5F7VI8K2UZ6yFNFh81rA.

Gambacorta, Daniel. 2019. "Value Drift & How Not to Be Evil Part II." *Global Optimum*. Retrieved from http://globaloptimum.libsyn.com/value-drift-how-to-not-be-evil-part-ii.

Gino, Francesca and Lamar Pierce. 2009. "The abundance effect: Unethical behavior in the presence of wealth." *Organizational Behavior and Human Decision Processes, 109*(2), 142-155. Retrieved from https://www.sciencedirect.com/science/article/pii/S0749597809000247.

Gouveia, Valdiney V., Katia C. Vione, Taciano L. Milfont, and Ronald Fischer. 2015. "Patterns of Value Change During the Lifespan: Some Evidence From a Functional Approach to Values." *Personality and Social Psychology Bulletin, 41*(9), 1276-1290. Retrieved from https://journals.sagepub.com/doi/full/10.1177/0146167215594189.

Hurford, Peter. 2019. "EA Survey 2018 Series: How Long Do EAs Stay in EA?" *Effective Altruism Forum.* Retrieved from https://forum.effectivealtruism.org/posts/bGcKJiBt4HSSScF76/ea-survey-2018-series-how-long-do-eas-stay-in-ea.

Locke, Edwin A. and Gary P. Latham. 2002. "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey." *American Psychologist, 57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705.

Kleinman, Carol S. 2006. "Ethical Drift: When Good People Do Bad Things." *JONA'S Healthcare Law, Ethics, and Regulation, 8*(3).

Mazar, Nina and Chen-Bo Zhong. 2010. “Do Green Products Make Us Better People?” *Psychological Science, 21*(4), 494-498.

Nesov, Vladamir. 2010. “Re: Hedging Our Bets: The Case for Pursuing Whole Brain Emulation to Safeguard Humanity’s Future.” [Forum comment]. *LessWrong*. Retrieved from https://www.lesswrong.com/posts/v5AJZyEY7YFthkzax/hedging-our-bets-the-case-for-pursuing-whole-brain-emulation#uYLHbiG7qzz9dgLsv.

Reinders, Heinz and James Youniss. 2006. “Community Service and Civic Development in Adolescence: Theoretical Considerations and Empirical Evidence. *Citizenship Education, Theory, Research, and Practice* (195-208). Retrieved from https://www.researchgate.net/publication/273654686_Community_Service_and_Civic_Development_in_Adolescence_Theoretical_Considerations_and_Empirical_Evidence/link/55081e160cf26ff55f80009e/download.

Savoie, Joey. 2018. “Empirical Data on Value Drift.” *Effective Altruism Forum*. Retrieved from https://forum.effectivealtruism.org/posts/mZWFEFpyDs3R6hD3r/empirical-data-on-value-drift.

Tie, Ylona Chun, Melanie Birks, and Karen Francis. 2019. “Grounded theory research: A design framework for novice researchers.” *SAGE Open Medicine 7*, 1-8. Retrieved from https://journals.sagepub.com/doi/full/10.1177/2050312118822927.

Todd, Benjamin. 2018. “Should you play to your comparative advantage when choosing your career?” *80,000 Hours.* https://80000hours.org/articles/comparative-advantage/.

Tong, Kowk Kit, Eva P. W. Hung, and Sze Man Yuen. 2010. “The Quality of Social Networks: Its Determinants and Impacts on Volunteering in Macao.” *Social Indicators Research, 102*(2), 351-261. Retrieved from https://search-ebscohost-

com.ezproxy.loyno.edu/login.aspx?direct=true&db=edsjsr&AN=edsjsr.41476486&site=eds-live&scope=site.

Whetstone, Laura. 2018. “EA Survey Series 2018: Community Demographics and Characteristics.” *Effective Altruism Forum*. Retrieved from https://forum.effectivealtruism.org/posts/S2Sonawxz2cY4YdXK/ea-survey-2018-series-community-demographics-and.

Wiblin, Robert. 2017. “How to compare different global problems in terms of impact.” *80,000 Hours.* Retrieved from https://80000hours.org/articles/problem-framework/.

World Health Organization. 2019. “Burn-out an "occupational phenomenon": International Classification of Diseases.” *World Health Organization.* Retrieved from https://www.who.int/mental_health/evidence/burn-out/en/.

Appendix: Interview Guide

I. Effective Altruism Involvement

   a. How did you first hear about effective altruism? How did you get involved with the movement? In what ways are you involved now?
   b. How long have you been involved in the effective altruism movement?
   c. What is your current job or career?
      i. Do you enjoy your work?
   d. What are your future career plans and goals?
   e. Do you donate to charity? What charities do you donate to? Have you always donated to those charities?
   f. Have you taken the Giving What We Can pledge? If so, have you maintained your commitment?
   g. Are you involved in a local effective altruism group?
   h. Have you changed cause areas during your involvement in effective altruism?
   i. On a scale of one to ten, how strongly do you identify as an EA?
   j. How would you describe your experience with the effective altruism community?
      i. What percentage of your social circles are EA or EA-aligned?
   k. How confident are you in the ideas of effective altruism?

II. Value Drift

   a. How would you define value drift?
   b. Overall, do you feel like you've experienced a significant amount of value drift…
      i. In your time as an EA?
      ii. Throughout your life?

iii. Compared to others in your social circles?

c. What challenges do you face in upholding your values?

d. How has your alignment with EA values changed over time? Have there been any points since you first became involved that you felt more or less aligned with EA values?

e. Are you concerned about experiencing value drift in the future?

f. What do you think causes value drift in general?

g. Have you noticed others' experiencing value drift?

h. Do you think value drift is a threat to EAs? Or the EA movement?

i. Are there any changes in the values of the EA movement that would cause you to leave? How likely do you think it is that this might happen?

j. If you found out that the person you will be in ten years is not an effective altruist, what would you expect to be the cause of that?

k. Do you do anything in particular to prevent value drift?