Longtermism and Cultural Evolution[1]

Aron Vallinder

Abstract

In this chapter, I argue that the field of cultural evolution can usefully inform attempts to understand and influence the long-term future. First, I provide an overview of cultural evolution, covering what it means for culture to evolve, the mechanisms by which it happens, the crucial importance of cumulative cultural evolution for hunan history, and how cultural evolution (and in particular intergroup competition) has driven the rise of large-scale cooperation. Second, I draw out some possible lessons from cultural evolution for thinking about the long-term future. In particular, I suggest we should be careful not to prematurely "lock in" certain values or other cultural features, and instead aim for a society open to variation and competition. I also consider whether the future will bring greater selection pressure for particular kinds of values, such as patience.

Keywords: cultural evolution, value change, lock-in, persistence

1 Introduction

Humans are rather unique animals. We occupy a wider range of habitats than any other large land animal (Boyd and Richerson 2009). We have modified more than a third of the Earth's land area (Vitousek et al 1997). We have used science and technology to radically transform and increasingly understand our environment. We write poetry and compose symphonies. We have religions, social norms, international agreements, courts of law, and a wide range of other institutions. What makes us unique is in large part our capacity for cumulative cultural learning. That is, by observing others we are able to learn skills, norms, beliefs, and behaviours that have taken shape over several generations. This means that individuals don't have to reinvent the wheel, but can instead rely on insights and innovations that have accumulated over time. Indeed, most

[1] Forthcoming in Jacob Barrett, Hilary Greaves, & David Thorstad (eds.), *Essays in Longtermism*, Oxford University Press. For helpful discussion and feedback, I'm grateful to Pablo Stafforini, Peter Hartree, an anonymous referee, and the editors of this volume.

of the things we take for granted in contemporary society would be impossible without cumulative culture: no single individual could discover them on their own.

According to the field of cultural evolution, cultural traits are subject to Darwinian evolution just like genes are, and can be studied using similar types of models. This process has been a major driver of human history so far. Unless something unprecedented happens, we should expect it to continue to shape our trajectory into the long-term future. For this reason, the field of cultural evolution provides a set of tools and insights that can inform our thinking about the long-term future. For example, it can help us identify various possible cultural selection pressures an intervention aimed at influencing the long-term future must be able to survive in order to persist over time. The field of cultural evolution should therefore be of significant interest to longtermists, i.e. those who (as I will here understand the term) believe that our impact on the long-term future should be a major moral consideration today. Moreover, some have proposed cultural change—and in particular values change—as a potentially promising way of having significant impact on the long-term future (MacAskill 2022, Anthis and Paez 2021). The tools of cultural evolution can help us evaluate the feasibility and desirability of such interventions. In the next section, I provide an introduction to cultural evolution: how it operates on an individual level, the importance of cumulative cultural evolution and how it came about, and how intergroup competition has shaped our history, including in the emergence of large-scale cooperation. After this, in section 3, I explore how cultural evolutionary considerations can inform our thinking about the long-term future. I suggest that such considerations provide reason against "locking in" certain values or other features of society indefinitely, or that at the very least should make us wary of doing so prematurely. I argue that cultural evolution supports an increase in experimentation and variation. I also examine how cultural selection pressures may constrain the range of feasible scenarios for the long-term future. Overall, I hope that these explorations will demonstrate that cultural evolution is an underappreciated source of insights relevant to the project of trying to steer the course of the far future and that further research in this direction would be useful.

2 The Basics of Cultural Evolution

"Culture" in the relevant sense is a broad category: any information that is transmitted via social learning mechanisms such as teaching and imitation. This includes languages,

hunting practices, cooking techniques, programming languages, religious beliefs and rituals, as well as social norms, values, preferences and a range of other things. How do cultural traits arise, spread, and change over time? For evolutionary adaptation to happen, three conditions must be met:

1. *Variation*. Entities vary in their characteristics.
2. *Inheritance*. Characteristics that contribute to differential fitness are heritable (i.e. there is transmission of information).
3. *Differential fitness*. *Entities* with different characteristics have different rates of survival and reproduction.

Genes and cultural traits are two types of entities that meet these conditions. In the case of culture, consider for example the length of arrows used by hunters in some community. Likely there will be variation in this cultural trait: not all hunters use the exact same length. Second, some arrow lengths will lead to greater hunting success than others, increasing survival rates for hunters using them. In turn, this success makes other individuals more inclined to copy them, further increasing their spread. Or consider norms around food sharing. Communities with more cooperative and generous food-sharing practices may experience increased social cohesion and support, and ultimately higher survival and reproduction rates. As a result, these food-sharing norms may become more prevalent.

2.1 The Mechanisms of Cultural Micro-Evolution

As the examples of arrow length and food sharing norms illustrate, cultural evolution differs from genetic evolution in some key ways. For example, genes are almost always inherited from one's biological parents (vertical transmission), whereas cultural traits can also be acquired from a range of other sources, such as other members of the parental generation (oblique transmission), other members of one's own generation (horizontal transmission), or members of later generations. Horizontal transmission allows for adaptive cultural traits to spread faster than does vertical transmission (Cavalli-Sforza and Feldman 1981, p. 351-357). In small-scale societies, oblique and horizontal transmission typically happens on a one-to-one basis, but industrial society allows for one-to-many transmission via mass media, which also has the effect of speeding up cultural evolution.

Moreover, genes are discrete units of information whereas many cultural traits (such as arrow length) are continuous. Genetic information is copied and replicated, but cultural evolution does not always require a notion of replication. For example, suppose that an individual decides on arrow length by averaging the arrow lengths of the three most successful hunters. There's no obvious sense in which anything was replicated, but it's nevertheless clear that something was inherited (Boyd 2002).

Within this framework, several factors affect how cultural traits evolve over time. First, consider the sources of variation. New traits can be introduced through either random variation or guided variation. In random variation, cultural traits mutate just like genes. These variations can arise from mistakes in learning or the transmission of information, or from spontaneous innovations. In guided variation, by contrast, individuals actively modify, adapt, or innovate cultural traits based on their own experience, knowledge, or reasoning before passing them on to others.

Second, consider how the distribution of existing cultural traits changes over time. Even without cultural learning, cultural traits that increase the fitness of their bearers will increase in frequency as a result of natural selection. The distribution of cultural traits can also change over time as a result of cultural drift, i.e. random fluctuations in the frequency of cultural traits due to sampling errors in the transmission process. In small populations, cultural drift can have a significant impact on the distribution of cultural traits, leading to the loss of some traits and the fixation of others over time. But perhaps the most important force behind changes in frequency over time is cultural selection, or biased transmission. Cultural learners don't learn at random, but rather show preferential adoption of certain cultural traits based on factors such as frequency, model, or content. Let's consider these factors in turn.

In the most common type of frequency bias, conformist transmission, we tend to preferentially copy the most common trait. Suppose that you want to buy a new pair of headphones. To help you decide, you check out five different product recommendation websites and find that three of them recommend the same pair of headphones. If you would pick the product recommended by three out five websites more than 60% of the time, you are engaging in conformist transmission. Models of conformist transmission

have shown it to be adaptive in a wide range of circumstances. In a model by Richerson and Boyd (1985), the environment varies spatially, so that which cultural trait is adaptive depends on one's location. In each environment, however, learning mechanisms cause the adaptive cultural trait to be more widespread than the other. At the same time, migration introduces less favoured variants. In this setting, individuals predisposed to acquire the most common trait will be more likely to acquire the favoured trait. Boyd and Henrich (1998) extend this result, showing that conformist transmission is also adaptive in environments that are both temporally and spatially variable. Nakahashi, Wakano, and Henrich (2012) point out that many models of conformist transmission focus on the case where there are only two cultural variants, and argue that conformist transmission only becomes more adaptive the more traits there are. Conformist transmission has also been experimentally well-documented (Muthukrishna, Morgan, and Henrich, 2016). Boyd and Henrich (1998) show that conformist transmission supports the emergence of stable group differences that persist over time.

In model-based bias, learners preferentially adopt traits from certain kinds of individuals. If the skill in question is easy to assess, you can simply try to emulate the most successful individual. This success bias is well-established, and has been observed both in laboratory settings and in the real world. (Henrich and Gil-White 2001, Henrich and Broesch 2011, and Henrich and Henrich 2007, chapter 2.) However, in many cases there may not be an unambiguous measure of success or any other way to easily discern the most highly-skilled individual. In those cases, one useful strategy can be to observe whom others pay attention to, defer to, and imitate. Since other people have faced the same challenge of figuring out who to learn from, you can take advantage of the efforts they have already made. Prestige bias refers to this tendency to preferentially learn from and imitate others who are perceived as having high status, success, or prestige within a social group. This bias guides individuals to acquire cultural knowledge, skills, and behaviors from those who are considered to be the most knowledgeable, experienced, or influential. Prestigious individuals benefit from increased social standing and influence, and those who defer to them benefit from preferential access affording them greater learning opportunities. While prestige bias is often adaptive, we also have a tendency to overimitate, such as when deciding what product to buy based on celebrity endorsement.

In content-biased transmission, there is a preference for acquiring specific types of information or cultural traits based on their inherent characteristics. Only some cultural information enhances fitness. Natural selection has therefore favored paying greater attention to certain kinds of cultural information, such as information about kinship, social norms, reputation, and animals and plants (Chudek, Muthukrishna, and Henrich, 2015).

2.2. Cumulative Cultural Evolution

Boyd and Richerson (2005, ch. 1) construct a model to explore when social learning (i.e. learning from other individuals rather than on one's own) is adaptive. Suppose some population lives in an environment that can be either wet or dry. If the environment is dry, hunting and gathering is the best strategy. If on the other hand it is wet, farming is the best strategy. In deciding which strategy to pursue, individuals have two sources of information to consider: individual learning (i.e. their own observations) and social learning (i.e. copying a member of the previous generation). Neither source of information is perfect. Individual learning leads to the right answer on average, but sometimes leads astray. For example, even a dry environment might see a series of consecutive rainy years, leading those relying only on individual learning to mistakenly adopt farming. Social learning can correct for these errors in individual learning. However, if you have moved to a new environment, imitating the previous generation may lead you astray. If individuals do not frequently move between different environment types, the best learning strategy is to mostly imitate, only relying on individual learning when it is highly accurate. On the other hand, if individuals move so frequently that their environment is effectively random with respect to that of the previous generation, social learning adds no value, and the best strategy is to rely on individual learning alone. In between these extremes, some mix of individual and social learning is best.

Many other animals are capable of social learning. For example, chimpanzees have been observed using tools, such as sticks, to extract termites from their nests. They learn this behavior by watching and imitating other chimpanzees (Boesch and Boesch 1990). Bottlenose dolphins have been documented using marine sponges to protect their noses

while foraging on the ocean floor. This behavior is learned from their mothers and passed down through generations (Mann et al 2012).

But a crucial feature of human cultural evolution is that it is cumulative. Cumulative cultural evolution is when culturally transmitted traits become so complex that no single individual could design them on their own, via asocial learning alone (Boyd and Richerson, 1996). While there is evidence of some cumulative cultural in non-human animals like chimpanzees (Yamamoto, Humle, Tanaka 2013) and New Caledonian crows (Hunt and Gray 2003), no other species relies on a complex body of accumulated cultural information to survive to the extent that humans do. Henrich and Muthukrishna (manuscript) argue that this cumulative cultural evolution is what explains humanity's dominance. Our position is not due to the intelligence of the individual, but rather to the cultural knowledge that has accumulated over generations.

The cumulative nature of cultural evolution is vividly illustrated by the many stories of lost European explorers reported by Henrich (2016). For example, in 1860 a small group of explorers travelling across the interior of Australia ran out of provisions and were forced to live off the land. Eventually, they made contact with a local Aboriginal group who shared food with them, including bread made from the nardoo plant. After this encounter, the explorers managed to find the plant themselves, pound the seeds, make flour, and bake nardoo bread. Initially it seemed like they had come across a reliable source of calories, but progressively they became weaker, with some of them dying from starvation. It turned out that nardoo is indigestible and mildly toxic unless properly processed. In order to make it edible, the Aboriginal individuals followed an elaborate procedure of preparation. This procedure had taken shape through trial-and-error over many generations. The European explorers, by contrast, did not have access to this culturally evolved knowledge, and instead faced what turned out to be the insurmountable task of figuring it out for themselves.

The power of cumulative cultural evolution has also been explored in laboratory studies. Muthukrishna et al (2014) asked participants to carry out a difficult task that they had no previous experience with. Participants were arranged into ten generations of five people each, with information sharing between generations. In one treatment, participants in generations 2 to 10 had access to guidance from all participants in the

previous generation. In the other treatment, participants only had access to information from one of the five participants of the previous generation. The treatment with more cumulative cultural evolution led to significantly greater performance.

If cumulative cultural learning provides such an adaptive advantage, why is it not more widespread? Why did it only emerge in the past couple of million years? Various researchers have proposed that for cumulative cultural evolution to get off the ground, certain cognitive, behavioral, or other preconditions are necessary. For example, Tomasello (1994) argues that high-fidelity transmission is necessary for traits to persist over time, that this high accuracy is cognitively demanding, and therefore is only possible in animals of sufficient cognitive development. While some dispute the importance of high-fidelity transmission for cumulative cultural learning (e.g. Sterelny 2021, p. 9), most researchers still agree that complex cognition is nevertheless a necessary precondition for one reason or another. Consistent with this hypothesis, average encephalization (brain size relative to body size) in mammals has increased over the past 66 million years (the Cenozoic era). In the human lineage, the expansion has been particularly fast over the past few million years. Five million years ago, our ancestors had a brain volume of around 350cm^3, compared to the 1350cm^3 of modern humans. Most of this increase (from 500cm^3) happened in the past two million years. What accounts for this fast development? Boyd and Richerson (2005, ch. 4) suggest that changing climatic conditions played a crucial role. Over the past several million years, the climate became much more variable than previously, with the effect of making existing habitats to change and become less stable. This led to increased selection for abilities to cope with more variable environments, which includes more complex cognition.

Heyes (2018) agrees that advanced cognitive capacity is a precondition for cumulative culture, and suggests two additional factors. First, relative to other primates, humans are remarkably peaceful and tolerant of others, including strangers. This creates an environment where individuals are exposed to more potential models to learn from. Second, we are endowed with various attentional biases that guide cultural learning. Almost from birth, these biases guide us to look at human faces and listen to human voices, thereby facilitating teaching and learning.

Whatever the exact set of preconditions are, once cumulative cultural evolution got started, it is likely that cultural learning harnessed various other of our psychological traits, and that these traits were in turn shaped to facilitate greater and more efficient learning. For example, Tomasello (2000) and Boyd and Richerson (2005) suggest that theory of mind, i.e. our ability to infer the beliefs, desires, and intentions of others, proved useful for learning skills and behaviours from others. If I know what someone is trying to do, I can more easily copy them. Theory of mind may initially have evolved because it allowed people to better predict the behaviour of others in their social group. Once it emerged, it was able to support observational learning and cumulative cultural evolution.

2.3 Large-Scale Cooperation and Intergroup Competition

Cumulative cultural evolution likely emerged sometime in the last few million years. For most of this time, innovations were slow to spread. The earliest known stone tools date from 3.3 million years ago. So called Ondowan stone tools date from around 2.5 million years ago, and the more advanced Acheulian tools are from 1.8 million years ago. It is only in the past few hundred thousand years that innovations take less than one hundred thousand years to become established (Sterelny 2021). In particular, starting around 12,000 years ago, the pace of cultural evolution picked up dramatically. As agriculture emerged, humans began cooperating in larger and more hierarchical groups. Over time, there has been a dramatic increase in the scale of cooperation among humans. From nuclear families to nation states and beyond, how did this happen?

There are some genetic mechanisms that foster small-scale cooperation, such as kin-based altruism ("help your relatives") and direct reciprocity ("if you help me I help you"). However, to scale up cooperation further, other mechanisms are needed. Evolutionary theorists have studied how mechanisms involving reputation, punishment, and signalling can support the emergence of large-scale cooperation. For example, cooperation can be sustained in models of diffuse punishment, where those who defect can be punished (at some cost) by any punishers in the group. However, this creates a second-order free-rider problem: who will punish punishers that refrain from punishing to evade the cost? Diffuse punishment can also serve as a way for punishers to signal their prosociality (cooperativeness and trustworthiness), thereby increasing their own chances of favorable social interactions in the future. However, it turns out

that these mechanisms can support any equally costly action even if it doesn't benefit anyone (Boyd and Henrich 2009:3283). Therefore, while these mechanisms can explain how cooperation is sustained over time, they cannot explain how that behaviour (and not some other equilibrium) arises in the first place.

Boyd and Richerson (2009) argue that intergroup competition played a crucial role in the emergence of large-scale cooperation. To understand how intergroup competition works, we can think of individuals as belonging to nested hierarchies of social groups. For example, they might belong to nuclear families, which are united into clans, which are in turn united into tribes. Nuclear families that unite into clans tend to outcompete independent nuclear families, for example by being at an advantage should any violent conflict arise. Groups with social norms that are more conducive to large-scale cooperation will be more successful, and such norms will therefore spread. Sometimes the interests of lower-level groups may not be aligned with those of the larger unit, such as when clans compete for power and influence within a tribe. Greater cooperation at lower levels (e.g. nepotism) can be deleterious for cooperation at higher levels. As a result, groups that better manage to suppress damaging low-level cooperation may enjoy a competitive advantage.

Henrich (2014) identifies five important mechanisms of intergroup competition: violent conflict, varying group survival rates, migration, fertility rates, and prestige-biased group transmission.

1. *Violent conflict.* Violent conflict is perhaps the most vivid form of intergroup competition. War, raiding, and other violent conflict can result in the elimination or assimilation of weaker social groups by others who have norms and institutions that are more conducive to cooperation, or have other competitive advantages. Some have argued that warfare facilitated transitions to larger scales of cooperation and social complexity (Choi and Bowles 2007, Morris 2014, Turchin 2016).
2. *Varying group survival rates.* In hostile environments, only groups with a sufficient level of cooperation and sharing will be able to survive and grow, and those norms that promote such behaviour tend to become more prevalent. For example, Stark (1996) argues that norms of care gave Christians higher survival

rates than the rest of the Roman population during both the Antonine plague (AD 165 to 180) and the plague of Cyprian (AD 249 to 262), thereby contributing to Christianity's rise from AD 40 (1,000 Christians) to AD 350 (34 million Christians), growing at about 40% per decade for three centuries.

3. *Migration*. Given that social norms can create groups with higher well-being and quality of life, many people will want to emigrate from less successful groups to more successful ones. Immigration can also serve to increase cultural variation, potentially spurring greater innovation.
4. *Fertility rates*. Social norms, such as religiously prescribed pro-fertility norms, can affect a group's fertility rate. Given that children will typically come to share their group's norms, cultures with pro-fertility norms will become more widespread over time. Stark (1996) argues that another important factor in its rise was the Christian ban on female infanticide, a widespread practice in the Greco-Roman world. Together with generally higher birth rates, this made the Christian population grow at a substantially faster pace than the rest of the Roman world. More recently, Kaufmann (2010) has suggested that although birthrates are generally declining across the world, some religious groups appear to be resisting the trend. More specifically, he argues that groups like Mormons, the Amish, Hutterites, Salafist Muslims, and Haredi Jews not only have very high birthrates, but also sufficiently high rates of retention that they are growing at substantially faster rates than other groups. However, a subsequent assessment found that growth rates may be declining for some of these groups (Juniewicz, 2022) .
5. *Prestige-biased group transmission*. People tend to pay greater attention to individuals from more successful groups, e.g. groups with higher living standards. For example, new nations may take inspiration from the constitutions or broader set of institutions of successful countries.

Intergroup competition has plausibly played a major role in shaping the course of human history. For example, Scheidel (2019) argues that one crucial reason why the Industrial Revolution happened in Europe and not in e.g. China was the fact that since the fall of the Roman Empire, Europe—unlike most other parts of the world—was never united into a single empire, instead consisting of several smaller units of roughly equal power. This made intergroup competition a much more important selection pressure,

driving further development and innovation. Mokyr (2016) also emphasizes the fragmented nature of Europe in his account of how a cultural shift in the early modern period facilitated the Industrial Revolution. Fragmentation meant that European rulers were competing with one another for the most skilled citizens, be they painters, artisans, musicians, or engineers. Such competition between states also made it more difficult for those defending conservatism to coordinate their attempts to suppress intellectual innovators. Those who were persecuted in one state could often set up shop in another one instead. This was in part made possible by Europe's unusual combination of political fragmentation with intellectual and cultural unity—an integrated market for ideas. This unity came from Europe's classical heritage, the use of Latin as lingua franca, and the Christian Church. It allowed for the emergence of The Republic of Letters, a transnational community of scholars who disseminated ideas and corresponded with one another, giving intellectual innovators a much larger audience than they could otherwise have had. It also provided a set of institutional incentives that encouraged academic superstars and allowed heterodox scholars to spread their original ideas in the hope of gaining prestige. Among these scholars the idea that it was both possible and desirable to understand, manipulate, and improve upon the natural world began to take hold. Mokyr argues that this "culture of growth" played a crucial role in enabling the Industrial Revolution.

3 Longtermist lessons from cultural evolution

How can the study of cultural evolution help to guide the project of trying to steer the course of the far future? At the most general level, for as long as competition between cultural units (nations, religions, ideologies, firms, subcultures, etc.) remains a potent force in shaping the trajectory of Earth-originating life, the tools of cultural evolution can help us gain a better understanding of what the long-term future may look like and what, if anything, we can do to influence it. For some particular change to persist over long time spans, it must be able to successfully compete and survive the process of cultural evolution over that time span. In this way, cultural evolution can help us assess the feasibility of various proposed longtermist interventions.

Consider for example the suggestion to evaluate longtermist interventions based on the significance, persistence, and contingency of the states of affairs those interventions are likely to bring about (MacAskill 2022; MacAskill, Thomas and Vallinder 2022). In this

framework, the significance of a state of affairs is its average value per unit time. To calculate its total value, we also need to know its persistence, i.e. how long it lasts for. When evaluating longtermist interventions we also care about contingency, i.e. to what extent the state of affairs can be traced back to some particular decision or other originating event. If an intervention brings about change which, though highly persistent, would have happened soon after even without the intervention, its longtermist value is correspondingly smaller. Cultural evolution can inform our thinking about these factors, particularly persistence and contingency. To persist, a trait must be able to survive the process of competition. To be contingent, it must be the case that competitive pressures would not have brought it about sooner or later anyway.

With this framework in mind, many proposed longtermist interventions fall into two broad categories. One set of interventions aims to reduce the risk of human extinction, whether by unaligned AI, engineered pandemics or some other global catastrophe. Given some assumptions (e.g. that survival is a net positive, and that the expected lifespan of humanity conditional on this risk reduction is not very short), it's clear how such interventions may score highly across all three dimensions. Assuming that the risk reduction happens in the near term, we don't need to consider the dynamics of cultural evolution in order to explain how persistent influence is possible.[2]

Another set of interventions aims to increase the value of the long-term future conditional on a long future containing a large number of sentient and intelligent beings. In this case, the path to long-term impact is less clear than it is for extinction risk mitigation. One may reasonably worry that the effects of any such intervention will eventually wash out, with no impact on the long-term future. Moreover, even if it does have some long-term impact, how sure can we be that it is in fact for the better? In response, Greaves and MacAskill (2023) suggest that there may be certain *persistent states* such that once the world enters a persistent state, it will remain in that state for a very long period of time (in expectation at least). If we can influence which persistent state the world enters, we can thereby have predictable impact on the long-term future. Human extinction is one clear example of a persistent state: if humanity went extinct, it

[2] At least so far as the first-order effects are concerned. In theory it could be the case that interventions that reduce risks today have the further effect of making future generations less inclined to reduce those risks, but it's unclear whether we have any reason to think that this is ever true in practice.

is plausibly unlikely that another species that could realize humanity's potential would evolve. But there may be other persistent states as well.

3.1 Artificial General Intelligence and Lock-In

One salient possibility is that artificial general intelligence (AGI) that greatly exceeds human performance in most areas of interest could allow for indefinite value lock-in. Finnveden et al (2022) argue that AGI will enable precise preservation of goals into the far future, the creation of institutions that intelligently pursue those goals, and the prevention of any disruption to its pursuit, be it natural catastrophes or other agents. Let's consider these in turn.

1. *Preserving information.* Digital error correction can ensure that information describing the goals can persist into the long-term future, and storing this information redundantly in several places further increases persistence.
2. *Executing intentions*. Ensuring that the goals are pursued as intended requires solving the AI alignment problem. Many have claimed that this is an exceptionally difficult problem (e.g. Bostrom 2014, Ngo 2020, Cotra 2021), but assuming it can be solved, we would presumably end up with a system very well-equipped to execute programmed intentions over the long-term future. Moreover, if we fail to solve the AI alignment problem, one plausible scenario is that we still end up with value lock-in, only that now the locked-in values are those of an unaligned AI rather than any intentionally programmed goals.
3. *Preventing disruption.* If AGI is in the hands of a state or other global actor with uncontested economic dominance, that actor could use AGI to make its dominance persist into the far future.

Another useful angle is to consider what the sources of values change are today, and whether they would necessarily remain operant in a world with AGI. Finnveden et al (2022) discuss the following sources:

- *Intergroup competition*. As we have seen, warfare and other forms of competition between different states has been a major driver of value change in human history so far. However, we may yet see the emergence of a world government. AGI could even make such an outcome more likely, by providing whoever first

develops it with a decisive strategic advantage. Thus if a stable world government arises, competition between states would no longer be a source of values change.

- *Aging and death*. When the leader of an authoritarian regime dies, future leaders may steer things in a different direction, causing values and goals to change over time. Caplan (2008) argues that this problem of succession was the greatest cause of ideological change within the Soviet Union and communist China. In democratic countries too, generational replacement is a source of values change. AGI would not be subject to aging or death, and could therefore continue to pursue values unchanged by this process.
- *Technological or societal changes favoring new values*. In the past, technological or other societal changes have favored new values. For example, the adoption of agriculture lead to a change from egalitarian values to values more accepting of hierarchy. Morris (2015) argues that such values were more adaptive for agricultural societies that relied on more large-scale cooperation and long-term planning. Similarly, because it required more physical strength than other methods of harvesting, plough use encouraged greater gendered division of labor, the effects of which can still be observed today in the form of more unequal gender norms in societies that traditionally relied more heavily on the plough (Alesina, Giuliano and Nunn 2013). If we reach what Bostrom (2013) calls "technological maturity," i.e. a level of technological advancement that gives us close to maximum capacity for economic productivity and control over nature , this source of change would no longer be in play. However, it might stop operating before that, when there are no remaining technological changes sufficiently transformative to overcome the will of a powerful, dominant world government.
- *Internal rebellion*. In the past, coups and revolutions have been a frequent source of values change. In many cases, such regime change has had to rely on support from the military or some other critical state institution. However, if these institutions were no longer reliant on humans, instead being automated by AGI aligned with regime goals, such support would no longer be possible. Moreover, attempts at regime change without the support of key government institutions can also be prevented by AGI.

If AGI-enabled lock-in is feasible, what forms could it take? It could be that the AGI is controlled by some particular state which thereby gains a decisive strategic advantage and becomes able to impose its will globally. In the most extreme case, one could imagine a global totalitarian state where whoever controls the state is able to impose their values around the world. But one could also imagine that a democratic world government only locks something in after extensive public debate and oversight, perhaps also making sure that whatever is locked in will be sensitive to how the will of the people changes in the future. In between these extremes, of course, is a range of different scenarios.

3.2 Lock-In and Cultural Evolution

It's clear that we want to avoid lock-in by AGI-powered global totalitarianism. But there might be more benign forms of lock-in, such as locking in values chosen in accordance with some democratic process, or locking in procedural elements (i.e. letting the system continue to evolve, but only in accordance with the evolving will of the people etc., perhaps subject to some further constraints such as human rights, free speech, etc.). Would some lock-in of this kind be desirable? And if so, which particular features would it be desirable to lock in?

When some feature of society gets locked in, that feature is no longer subject to the usual competitive pressures that drive social and cultural change. If we take seriously the idea that it was not our individual intelligence but rather cumulative cultural evolution—the often gradual improvements that have accrued over generations of trial and error—that gave humanity a decisive advantage, we might worry that putting an end to that process risks locking in a suboptimal future. Even supposing that the locked-in feature is currently optimal, it may not remain so as the environment continues to change. This suggests we should be wary of locking in social institutions prematurely. Similarly, today we find at least some of the values of almost any previous historical era to be defective if not outright horrifying. We should expect that future generations will look back on some of our values today in much the same way. For this reason, locking in the specific values we have today might be unwise.

Given that we are bad at intentionally planning and designing effective institutions, Henrich (2015, p. 331) suggests we ensure that there is sufficient variety and

appropriate selection mechanisms, so that different alternatives can compete and evolve. This way, superior arrangements can emerge and spread. We find some support for this idea in the cultural evolution of innovation. Muthukrishna and Henrich (2016) argue that three key factors behind rates of innovation are sociality, transmission fidelity, and cultural variation. Their starting point is that innovation is more often the result of recombination, gradual improvement or just pure luck, as opposed to revolutionary leaps by individual geniuses. With this in mind, consider that individuals in populations that are larger and more interconnected will be exposed to a broader range of ideas and practices. If it's not too difficult to learn these new cultural traits (i.e. if transmission fidelity is sufficiently high), some of them will begin spreading through the population. People will use various cues like success and prestige to decide which other people to pay attention to. Assuming that people are at least somewhat able to discern improvements, those improvements will spread to a larger share of the population. For all of this to work, there must be sufficient variation among the ideas and practices that people are exposed to. Too little cultural variation could mean that some superior solution will never be discovered because it can't easily be reached via recombination of current ideas.

MacAskill (2022, p. 97) similarly suggests we toward a morally exploratory world. This would mean keeping our options as open as possible, by delaying both large-scale and small-scale lock-in, so as not to risk prematurely ruling out desirable alternatives (Ord 2020, p. 158). It would also mean a political experimentalism of the kind Henrich gestures at, where people are encouraged to try out different and new ideas. Finally, and crucially, it would require arranging things so that the process of cultural evolution globally guides us toward more desirable arrangements.

What mechanisms could be used to ensure that values, norms, and institutions that are in some appropriate sense better have greater chance to survive and spread? MacAskill suggests it would involve support for free speech so that a broader range of ideas get a fair hearing, relatively free migration so that people can vote with their feet, international norms or laws preventing any one country from achieving decisive military and economic dominance and unilaterally locking in its goals.

While free speech may encourage a broader range of ideas to be considered, it is hardly a guarantee. Some worry that the global elite may converge on a fairly narrow set of values, practices and policies. As MacAskill (2022, p. 96) notes, the relatively small range of global policy responses to the COVID-19 pandemic (e.g. not a single country allowed for human challenge trials of the vaccines developed in early 2020) suggests some such convergence. This trend may only become more pronounced if governance becomes increasingly global. Perhaps even a democratic world government risks leaving too little room for competition.

Thus on this proposal, we should only aim to lock in features that prevent further, undesirable lock-in and guide us toward desirable outcomes we may not have discovered through intentional design. Of course, it may be that future technological developments eventually allow design to outperform cumulative cultural evolution. But we should be wary of taking ourselves to have reached that point prematurely.

3.3 Influencing persistent values

If feasible, AI-enabled lock-in represents the clearest mechanism by which having persistent, predictable influence on the long-term future might be possible. However, cultural evolution suggests there may be other ways of having such influence. Suppose that some set of values provides a sufficiently large competitive advantage in terms of influence over the long-term future. We should then expect those values to become increasingly prevalent over time, eventually coming to dominate. Plausibly, this is what drove the emergence of large-scale cooperation, as those who were able to more effectively organize into larger units outcompeted others. Are there any other such values that have not yet come to dominate, but will plausibly do so eventually? Hanson (2018) claims that caring explicitly about the long-term future is one such value. He argues that over time, planning and taking action over much longer time frames will become increasingly feasible. Therefore, those who are relatively more inclined to take such actions (rather than actions that are more motivated by short-term concerns) will increasingly have the means to exercise a greater influence on the long-term future, until they eventually come to dominate it.

There are two key steps of this argument that could be questioned. First, why should we expect long-term planning and execution to become increasingly feasible? Second,

do we have reason to think that, once long-term planning is possible, taking action with the long-term consequences explicitly in mind will provide a sufficient competitive advantage with respect to long-term influence? Let's consider these in turn. Hanson claims that history so far can be seen as a competition between various kinds of units (organisms, genes, cultures) to control the distant future. So far this has not been very explicit or intentional, because we are not good at planning and taking action over very long time spans. However, he claims that there has been a trend toward more capable long-term planning, and that we should expect this trend to continue. Predators and prey developed the ability to plan for at least part of the duration of a chase. Some animals, like chimpanzees and ravens, are able to plan tool use over several hours (Mulcahy and Call 2006, Kabadayi and Osvath 2017). With farming, humans became able to plan on the scale of a year (e.g. by saving grains to eat in winter and seeds to sow in spring). Today institutions and organizations are able to make some plans on the scale of a few years. This is admittedly a rather small number of examples on which to base our extrapolation, but we can imagine future technological developments that would enable it, like the ones discussed earlier in relation to lock-in. Second, consider now the question of whether explicitly caring for the future will provide a sufficiently large competitive advantage with respect to the long-term future to eventually achieve domination. There is evidence that patience strongly correlates with development, e.g. per capita income and the accumulation of physical capital, human capital, and productivity (Sunde et al 2022). This suggests that in at least some environments"long views" may indeed confer a competitive advantage.

If we accept that patient values will eventually come to dominate, what are the practical implications? Hanson suggests that one intervention longtermists might consider is to speed up the arrival of these long views. One might think that, if long views will come to dominate eventually anyway, speeding up their arrival will only have a limited impact on the future. However, as care for the distant future becomes dominant, we will begin investing more in efforts to mitigate extinction risks (assuming that care for the distant future goes together with a belief that continued existence would be good). Therefore, by speeding up the arrival of long views, we reduce the total extinction risk facing us in the future.

How might one work to hasten the arrival of long views? Some possibilities are promoting greater concern for the long-term future in general, making existing cultural units more inclined to care for the long term, or working to make long-term planning more feasible (e.g. by improving our predictive capacities). Further research in this direction might prove useful. What about influencing the broader package of values that go with caring about the long-term future? There are many different ways of caring about the long-term future. Presumably, not all of these are equally good, and one might therefore think we should work to make better ones more likely.

Hanson further argues, along similar lines, that future entities (whether biological or artificial) will eventually come to directly and explicitly value having as many descendants as possible. So far we mostly care about descendants in indirect ways. However, again if long-term planning becomes more feasible those who invest more in taking long-term action will have greater influence on the future. Given such abilities, those who directly plan for having as many long-term descendants as possible will in fact have more long-term descendants than those who don't. This suggests it might be worthwhile to invest further effort in identifying other traits that should reasonably be expected to become dominant in the future. We should then look for a clear way in which the trait provides sufficient long-term advantage, an explanation for why it has not yet come to dominate, and an account of how it may come to do so in the future. This way, we can get a clearer picture of the future landscape of cultural selection pressures that longtermist interventions have to contend with.

4 Conclusion

I have argued that the tools of cultural evolution can inform our thinking about the long-term future. At the most general level, I suggested that for as long as competition between different cultural units remains a relevant force in shaping history, an understanding of cultural selection pressures will be crucial for understanding what the long-term future may look like, and which interventions may be successful. I also claimed that considerations from cultural evolution may support continued experimentation and variation over lock-in and centralization. But the main takeaway I want to convey is that cultural evolution remains an underexplored source of insights relevant to the project of trying to understand and steer the course of the far future. Further work in this direction may well reveal new crucial considerations.

5 Bibliography

Alberto Alesina, Paola Giuliano and Nathan Nunn (2013). On the Origins of Gender Roles: Women and the Plough. *Quarterly Journal of Economics* 128(2):469-530.

Jacy Reese Anthis and Eza Paez (2021). Moral circle expansion: A promising strategy to impact the far future. *Futures*

Christophe Boesch and Hedwige Boesch (1990). Tool use and tool making in wild chimpanzees. *Folia Primatologica* 54(1-2):86–99.

Nick Bostrom (2013). Existential Risk Prevention as Global Priority. *Global Policy* 4(1):15-31.

Nick Bostrom (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Robert Boyd (2002). On Modeling Cognition and Culture: Why cultural evolution does not require replication of representations. *Journal of Cognition and Culture* 2(2):87-112.

Robert Boyd and Peter J. Richerson (1985). *Culture and the Evolutionary Process*. London: University of Chicago Press.

Robert Boyd and Peter J. Richerson (1996). Why culture is common, but cultural evolution is rare. In W.G. Runciman, J. M. Smith, and R. I. M. Dunbar (eds.), *Evolution of social behaviour patterns in primates and man*. Oxford: Oxford University Press.

Robert Boyd and Peter J. Richerson (2005). The Origin and Evolution of Cultures. Oxford: Oxford University Press.

Robert Boyd and Peter J. Richerson (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of The Royal Society B* 364:3281–3288.

Bryan Caplan (2008). The Totalitarian Threat. In Nick Bostrom and Milan M. Cirkovic (eds.), *Global Catastrophic Risks*. Oxford: Oxford University Press.

LL Cavalli-Sforza and Marcus W. Feldman (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Jung-Kyoo Choi and Samuel Bowles (2007). The coevolution of parochial altruism and war. *Science* 318(5850):636-40.

Maciej Chudek, Michael Muthukrishna and Joseph Henrich (2015). Cultural Evolution. In *The Handbook of Evolutionary Psychology*, edited by David M. Buss. Vol. 2. 2nd ed. John Wiley and Sons

Ajeya Cotra (2021). Why AI alignment could be hard with modern deep learning. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/

Lewis G. Dean, Gill L. Vale, Kevin N. Laland, Emma Flynn, and Rachel L. Kendal(2014). Human cumulative culture: a comparative perspective. *Biological Reviews* 89:284-301.

Lukas Finnveden, Jess Riedel and Carl Shulman (2022). Artificial General Intelligence and Lock-In.

Hilary Greaves and William MacAskill (2023). The Case for Strong Longtermism.

Robin Hanson (2018). Long Views Are Coming. *Overcoming Bias*. https://www.overcomingbias.com/p/long-views-are-cominghtml

Joseph Henrich (2015) *The Secret of Our Success*. Princeton, NJ: Princeton University Press.

Joseph Henrich and Robert Boyd (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior* 19(4):215–241.

Joseph Henrich and James Broesch (2011). On the nature of cultural transmission networks: evidence from Fijian villages for adaptive learning biases. *Philosophical Transactions of The Royal Society B* 366(1567):1139-48.

Joseph Henrich and Francisco J Gil-White (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior* 22(3):165–196.

Joseph Henrich and Michael Muthukrishna (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology* 72:207-240.

Joseph Henrich and Michael Muthukrishan (manuscript). What makes us smart?

Natalie Henrich and Joseph Henrich (2007). *Why Humans Cooperate: A Cultural and Evolutionary Explanation.* New York: Oxford University Press.

Cecilia Heyes (2018). *Cognitive Gadgets*. Cambridge, MA: The Belknap Press of Harvard University Press.

Can Kabadayi and Mathias Osvath (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science* 357(6347): 202-204 doi: 10.1126/science.aam8138.

Eric Kaufmann (2010). *Shall the Religious Inherit the Earth? Demography and Politics in the Twenty-first Century*. Profile Books.

William MacAskill (2022). *What We Owe The Future*

William MacAskill, Teruji Thomas and Aron Vallinder (2022). The Significance, Persistence, and Contingency Framework. *GPI Technical Report*

Janet Mann, Margaret A. Stanton, Eric M. Patterson, Elisa J. Bienenstock and Lisa O. Singh (2012). Social networks reveal cultural behaviour in tool-using dolphins. *Nature Communications* 3(980).

Joel Mokyr (2016). *A Culture of Growth*. Princeton, NJ: Princeton University Press.

Ian Morris (2014). *War! What Is It Good For? Conflict and the Progress of Civilization from Primates to Robots*. Macmillan.

Ian Morris (2015). *Foragers, Farmers, and Fossil Fuels: How Human Values Evolve*. Princeton, NJ: Princeton University Press.

Nicholas J Mulcahy and Josep Call (2006). Apes save tools for future use. *Science* 312(5776):1038-40. doi: 10.1126/science.1125456.

Michael Muthukrishna, Michael Doebeli, Maciej Chudek, and Joseph Henrich (2018). The Cultural Brain Hypothesis: How culture drives brain expansion, sociality, and life history. *PLoS Computational Biology* 14(11).

Michael Muthukrishna, Thomas Morgan, and Joseph Henrich, (2016). The when and who of social learning and conformist transmission. *Evolution and Human Behavior* 37:10-20.

Michael Muthukrishna, Ben W. Shulman, Vlad Vasilescu and Joseph Henrich (2014). Sociality influences cultural complexity. *Proceedings of The Royal Society B* 281: 20132511.

Wataru Nakahashi, Joe Yuichiro Wakano, and Joseph Henrich (2012). Adaptive Social Learning Strategies in Temporally and Spatially Varying Environments. *Human Nature* 23:386–418.

Richard Ngo (2020). AGI Safety from First Principles. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Toby Ord (2020). *The Precipice: Existential Risk and the Future of Humanity*.

Walter Scheidel (2019). *Escape from Rome: The Failure of Empire and the Road to Prosperity*. Princeton, NJ: Princeton University Press.

Rodney Stark (1996). *The Rise of Christianity*. Princeton, NJ: Princeton University Press.

Kim Sterelny (2016). Contingency and History. *Philosophy of Science* 83(4):521-539.

Kim Sterelny (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford: Oxford University Press.

Uwe Sunde, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman and Gerrit Meyerheim (2022). Patience and Comparative Development. *Review of Economic Studies* 89(5):2806-2840.

Michael Tomasello (2000). Two Hypotheses About Primate Cognition. In Cecilia Heyes and Ludwig Huber (eds.) *The Evolution of Cognition*. Cambridge MA: MIT Press.

Peter Turchin (2016). *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth*. Beresta Books.