Journal of Ethics and Emerging Technologies

IEET

Article

Civilizational Virtue, Civilizational Autonomy, and Existential Risks

Anders Sandberg[1]*

[1] Institute for Futures Studies, Box 591, 101 31 Stockholm, Sweden
* Correspondence: AS anders@aleph.se

Abstract: Virtues are traits or qualities that are morally good, expressed as behaviour that does what is right and avoids what is wrong. Individual virtues have been discussed in ethics since classical times. Collective or group virtues are a more recent and somewhat contested concept, attributing virtues that cannot be held without the existence of the group. Collective agents might even be virtuous without being moral patients. Could there be virtues applying to the largest groups, civilizations themselves? In existential risk scholarship human civilization is sometimes reified as a relevant actor, and virtue terms are applied to its behaviour, in particular related to long-term survival. This paper analyses whether a civilization can be virtue-apt, and what civilizational virtues may be. I argue that it makes sense to claim humanity has character, shows collective agency, something akin to free will (meeting an objection from macrohistory), and might (perhaps in the future) count as a form of truly autonomous agent. I give examples of putative civilizational virtues that are not just summative, but only makes sense as held by a civilization. It hence appears possible in principle for civilizations to be virtuous, and that there are unique civilization level virtues.

Keywords: virtue ethics; collective virtue; existential risk; civilization

1. Introduction

Virtues are traits or qualities that are morally good, expressed as behaviour that does what is right and avoids what is wrong. Normally this is considered on the level of individual persons, but it is not inconceivable that group behaviour could correctly be ascribed as virtuous or vicious. Are there virtues (and vices) on the civilization level? Toby Ord (2020) likens humanity as an adolescent that is coming into power yet still in many ways immature and unwise:

> This analogy provides us with another lens through which to assess our behavior. Rather than looking at the morality of an individual human's actions as they bear on others, we can address the dispositions and character of humanity as a whole and how these help or undercut its own chances of flourishing. When we look at humanity itself as a group agent, comprising all of us over all time, we can gain insight into the systematic strengths or weaknesses in humanity's ability to achieve flourishing. *These are virtues and vices at the largest scale—what we could call civilizational virtues and vices. One could treat these as having a fundamental moral significance, or simply as a useful way of diagnosing important weakness in our character and suggesting remedies.* [Italics mine]

Here humanity as a whole is seen as an agent that can act prudently, with self-discipline, or showing patience. Ord cites Stewart Brand (2020) for another example of civilizational virtue thinking:

Citation: Sandberg, Anders. 2025. Civilizational Virtue, Civilizational Autonomy, and Existential Risks. *Journal of Ethics and Emerging Technologies* 35: 2. https://doi.org/10.55613/jeet.v35i2.170

Received: 02/03/2025
Accepted: 02/03/2025
Published: 10/03/2025

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

> Ecological problems were thought unsolvable because they could not be solved in a year or two... It turns out that environmental problems are solvable. It's just that it takes focused effort over a decade or three to move toward solutions, and the solutions sometimes take centuries. Environmentalism teaches patience. Patience, I believe, is a core competency of a healthy civilization.

He also considers civilizational virtues related to our relationships with the wider world: mistreatment of our animals and our environment may be flaws in our compassion and stewardship not just individually, but collectively. These civilizational virtues may be generalisations of our individual virtues to a wider sphere, but also emergent results of what our civilization actually does as a whole.

Nobody really desires climate change, nor is anybody individually able to cause it.
There has also been talk about civilizational crimes. Victory Hugo said: "Peace is the virtue of civilization. War is its crime." (Hugo 1878) Fred Hoyle (writing about poverty and population): "I suspect the same thing for our whole species: if we insist on always following the easy path we could end up as a criminal species." (Hoyle 1966) Here the choice is between allowing overpopulation and poverty causing persistent unrest and suffering, or costly and difficult remedies.[1] If there can be civilizational crimes and vices, civilizational virtues seem plausible.

There is no shortage of claims that particular human societies or civilizations exhibit particular virtues (or laments that they don't). This commonly either claims that their members often behave according to virtues that are valued by their culture ("Romans are virtuous"), or that they have particular virtue other societies are deficient in or even unable to possess ("Only Romans have true *Virtus*"). More rarely, but relevant for this paper, is the claim that their overall collective behaviour exhibits the virtue ("Rome expands in a virtuous way"). This can be a result of the prevalence or supremacy of the societal virtues affecting the collective[2] (a summativist account) or in an emergent fashion from institutions, culture and individual actions that produce virtuous behaviour on the largest scale (an anti-summativist account) (Fricker 2010).
This paper aims at exploring the nature, types, and impact of civilizational virtues. As discussed below these are not the "virtues of civilization" (properties or benefits of being "civilized"), but the kinds of large-scale behavioural dispositions that could properly be described as right or wrong applied to the largest social structures.

[1] Freeman Dyson contrasts a verdant galaxy of carefully expanding civilization with rapidly growing technological cancer (Shenker 1972) and elsewhere discusses the possibility of civilizational insanity (Dyson 1981). Here the issue is framed more as health than vice, but his proposals all involve civilization-level coordination, restraint and wisdom. Indeed, he defines sanity as "nothing more tan the ability to live in harmony with nature's laws", a virtue ethics definition.

[2] For example, "The Chinese traditional cultural values of harmony, benevolence, righteousness, courtesy, wisdom, honesty, loyalty, and filial piety are embodied in China's diplomacy through the concept of harmony, the most important Chinese traditional value." (Lihua 2013)

One reason to investigate this is that ethical systems often change in interesting ways as they are scaled to very large domains. Virtue ethics is traditionally very individual focused, particular, and often time- and culture-bound. Hence it is interesting to see what happens here.

Another reason is that some recent scholarship on existential risk, the future of humanity, and civilizational trajectories (Baum et al. 2019) tends to reify human civilization as a relevant actor -- certainly composed of individuals and institutions, but having a somewhat independent or emergent existence from the parts --- and hence understanding in what sense we can speak of ethical behaviour of civilizations becomes relevant. One can clearly consider consequentialist and deontologist frames for e.g. existential risk reduction (also covered in Ord (2020) and elsewhere), but virtue ethics seems underapplied in this domain.

2. Virtue Ethics

In the following I will speak about virtues but that does not necessarily mean endorsing virtue ethics. Virtue ethics argues that virtues are foundational and that other normative notions can be grounded in them. Meanwhile consequentialists may define them as traits that typically yield good consequences, and deontologists as traits held by people who reliably fulfil their duties (Hursthouse & Pettigrove 2023). We can hence still use the term, with some caution, even if we do not endorse the core claims of virtue ethics. Normally virtue is described as belonging to a person (Hursthouse & Pettigrove 2023):

> A virtue is an excellent trait of character. It is a disposition, well entrenched in its possessor—something that, as we say, goes all the way down, unlike a habit such as being a tea-drinker—to notice, expect, value, feel, desire, choose, act, and react in certain characteristic ways. To possess a virtue is to be a certain sort of person with a certain complex mindset. A significant aspect of this mindset is the wholehearted acceptance of a distinctive range of considerations as reasons for action.

Generally, virtues, like personality, are seen as persistent traits rather than present states or actions. They also develop over time. They depend on various cognitive, emotional and moral capabilities like noticing morally salient facts or desiring to behave in a moral fashion. Virtues are held to a degree, and one can fail at perfection in various ways (e.g. internal conflict making virtuous action happen but effortfully, or lack in practical wisdom making the selection of the right reasons for actions fail). Virtues are fundamentally skills at doing the right thing. However, virtue ethicists will emphasize that the virtue is more than a habit or skill. It is done because it is a virtue.

Virtues can be good in themselves (the virtue ethics core claim), good for the virtuous being (the Aristotelean idea of eudaimonia), or means for achieving a good goal (the consequentialist take on virtue).

Some virtues are about others, some about the self. The latter may seem more promising for a civilizational virtue where the “self” is the civilization. Some are about caring for or behaving

relative to your future self. The acts intended to get to the intended end-state have to be selected well: there are trajectories that reach the endpoint yet do so in a vicious way (e.g. to be virtuous the desire for athletic excellence needs to fit in with a sense of fair play).

The literature is replete with lists of virtues (and even more vices), often strongly culturally anchored. Often virtue ethicists try to find the core commonality that makes something a virtue rather than just a trait. The modern virtue theorist Alasdair MacIntyre came up with an initial, tentative definition that may be useful (MacIntyre 1981):

> A virtue is an acquired human quality the possession and exercise of which tends to enable us to achieve those goods which are internal to practices and the lack of which effectively prevents us from achieving any such good.

He stresses that they are about achieving internal goods rather than external rewards, and that they are linked to particular practices. These practices are different in different contexts and cultures but have as an end what constitutes the good of a whole human life. One can argue that virtues are skills whose execution is constitutive of the good life and good societies.

Safeguarding our future can be motivated by virtues for individuals such as gratitude (to past generations), compassion and fairness (toward future generations), and love, unity or solidarity toward the rest of humanity. Jonathan Schell (1982) considers love in the sense of a generalization of parental or procreative love: the love with which we bring others into the world. This love expands to encompass humanity's relationship with the Earth, future generations, and all life. This acts as his moral foundation for arguing against allowing nuclear existential risk to be realized: it threatens to end love itself, and it threatens that which we love. Similarly, Samuel Scheffler (2018) argues that current generations experience meaning from and hold preferences for the wellbeing of future generations for many reasons, including reasons of love. While Schell is directly motivated by an individual virtue generalized outward, Scheffler deals with self-reflection of what gives current life significance. However, this could easily be turned into an individual virtue of trans-temporal care, and if accepted by a community a collective virtue.

3. Collective Virtues

Plato ascribed collective virtues to the state in *The Republic*, but they appear to have been either an analogy to a human or directly derived from the members: a wise state is wise because the ruling class is wise (Mulgan 1968). The latter is a summativist view where a group holds a virtue if most members have the virtue. Anti-summativist views hold that the truth of whether the group has a virtue is accounted for only by reference to group dependent properties. Such virtues cannot be held without the existence of the group.

Groups can have properties that are not present in members, e.g. being a committee or show groupthink behavior, so strict summativism doesn't have to hold and appears weak. However, virtuous behavior requires that the good conduct happens because of good motive and skill rather than random chance. Not all emergent behavior causing doing what is right is virtuous. What is

needed may be Margaret Gilbert's (1992) concept of the "plural subject" where individuals jointly commit to a given action or belief under common knowledge: this looks more virtue-apt.

People often form group plural subjects simply to manage their cognitive load. Managing and navigating a Naval vessel requires abilities beyond any individual (Hutchins 1996). Groups, the normative structures they embrace, and extended cognition across society help solve hard coordination problems (Holm, Sandberg & Fisher 2025). One can speak of “mental institutions” that act as extended cognition for their members (Gallagher & Crisafi 2009), and by the parity principle actually does constitute cognitive processes (Gallagher 2013). Hence, they may be virtue-apt.

Note that it is essential that an explicit agreement to form a collective subject may not be needed. Were it needed for a collective subject to come into existence, human civilization would presumably not be such a subject since we do not have the choice to not take part. Yet being explicitly aware of one’s humanity and participation in a joint civilizational endeavor allows for individual and collective reflection.

Collective, non-summativist virtues have been defended by Miranda Fricker (drawing on Gilbert’s plural subjects), especially in terms of institutions that have a formal and procedural structure (Fricker 2010). Byerly formulates her account as:

> A collective C has a motive-based/skill-based[3] virtue V just when the members of C, qua members of C, commit to achieving the end of V because it is good, and they reliably achieve this end.

The structure and procedures act as a skeleton on which the flesh of the people who animating the institution moves, enabling (or hindering) group virtue. Collective virtues often manifest through institutional frameworks that structure and constrain individual choices. Ryan and Meghan Byerly (2015) further argue that these group virtues are multiply realizable, not depending on which members are part of the group as long as they adhere to the joint motives or skills. They give their own disposition-based account of collective virtue:

> A collective C has a virtue V to the extent that C is disposed to behave in ways characteristic of V under appropriate circumstances.

(however, they also suggest a more individual centered account where “members of C are disposed, qua members of C, to behave in ways characteristic of V”). Beggs (2003) similarly argues that a group stably manifesting some virtuous behavior, when none of the committed members can be credited with the virtue themselves, can be ascribed a group virtue. Here internal motivations are disregarded.

[3] I merge two statements here.

Do virtue-apt entities have to be moral patients? Normally being a moral agent implies being a moral patient, and lack of moral patienthood is used to argue for lack of moral agenthood.

Corporations can be viewed as artificial intelligent agents, but it is not clear that they are conscious. This has been used to argue that they are not moral agents by Ben Kuipers (2012):

> Without the ability to feel things like pain, or fear, or shame, or guilt, the concept of "taking responsibility" cannot mean for a corporate agent anything like what it means for an individual human being. Therefore, a corporate entity, as such, does not have a conscience: the ability to understand, feel, and regret what they have done wrong.

Yet anti-summativist accounts can certainly ascribe virtues and vices to the corporation – it might be socially aware, transparent, or environmentally callous. It might not feel anything in the sense a human does, but the emergent behavior not easily ascribable to individual members still shows patterns and come about in ways that make a virtue description meaningful. Fricker (2010) argues that there is no need to ascribe metaphysical existence of mental faculties to corporate agents, just as-if faculties due to the pooled faculties of members.

One can also bite the bullet on consciousness in spatially distributed group agents, as in (Schwitzgebel 2014), although he does not take a stand on issues of moral patienthood or agenthood there.

Can markets be virtuous? The Invisible Hand of the Market, the emergent interactions from self-interested agents, may produce desirable or morally relevant outcomes such as bringing about general prosperity, but is the internal process the right kind of process to be virtue-apt? Here there may not be any real reflection among the agents or subsets of the market about achieving the good states. They merely come about because they are game-theoretic equilibria, not due to the collective desire of the agents for bringing about joint prosperity (as famously espoused by Mandeville's *Fable of the Bees*). The market itself is not reflecting in any sense on its activities, it just is. On Beggs account it is virtuous, but neither Fricker's or the Byerlys' account would agree.

Hence, even if we accept anti-summativism on collective virtue, its subjects might not be moral patients, and the internal structure of a group may prevent it from being virtue-apt. There could perhaps exist conscious, patient corporations or a market that has a structure (or motivated traders) making it suitably organized to be virtuous – but it does not look like there is a *necessary* link between virtue-aptness and moral patienthood for collective agents.

4. What Does It Take for a Civilization to Be Virtue-Apt?

4.1 What Do I Mean by "Civilization"?

The term civilization is somewhat contentious and often used in several forms. It has been applied to particular societies with certain properties (irrigation, urban areas, social stratification, symbolic

systems of communication, division of labor etc.). This is the plural meaning: there are civilizations, more or less distinguishable across time and space, but with certain shared properties.

Civilization in the singular sense is used to denote the idea that some societies are civilized in contrast to the barbarous or primitive because they have the above properties, often with implied "virtues". This singular sense is typically set up so the author's own society is the height of Civilization.

I am not interested in either sense, since they are too narrow (having irrigation is beside the point) and intrinsically normative (begging the question of whether civilization is virtuous).

There are also spatial and temporal perspectives that treat civilizations as cultural entities or phenomena located in certain regions and eras, potentially in processes of migrating or developing. I propose a tentative definition as:

> Civilization: a cohesive, long-range (social) structure with a high degree of coordination across time and space. It changes its environment in an organized way for systematic reasons that persist over long periods.

There are shared values, institutions, and cultural practices that persist for a long duration (e.g. the imperial examination system and the concept of the mandate of heaven in Chinese civilization, or the Orthodox church and Justinian legal code in the Byzantine empire). This persistence is not always immutable: Egyptian civilization replaced institutions dynasty by dynasty, and the Japanese Meiji Restoration transformed society yet, arguably continued the same civilization. What makes them civilizations is not perfect stability of values or institutions, but rather their capacity to adapt these elements while maintaining sufficient coherence to act upon and transform their environment in systematic ways that persist over long periods.

In particular we can imagine civilizations on the largest possible scale as the joint human endeavor ("human civilization") or a global phenomenon. Obviously one can contest whether there is currently enough coordination to count, but it is not too implausible given growing globalization and interconnectedness that even if there is no human civilization right now it may come into existence in the future.

Civilizations have memory, and are path dependent. The reasons for acting in particular ways can change due to internal reflection among the members, spread through the intra-civilizational discourse, and become a fixture.

In fact, it is not unreasonable to say they perform distributed cognitive functions. Whether this also carries over to distributed emotions, consciousness or other mental properties is already contentious on the group level, but may not matter much for civilizational virtues as we have seen.

4.2 Civilizational Virtues

We may start with a tentative definition:

> A civilizational virtue is an acquired trait of a civilization that produces good outcomes or is good in itself.

The virtue is contingent: the civilization is able to learn it or turn away from it, so it is not intrinsic to the civilization. While a civilization by definition must have some long-range planning, a patient civilization has long-range plans that are due to its understanding of its situation and ability to limit short-term impulses among its members or as a whole. Here the understanding can be seen as z form of Fricker's pooled faculties, the result of the combined understanding of the members.

A key issue is how the virtue comes about and is applied. In individuals, virtue ethics demand that the virtuous behavior comes about for the right internal reasons rather than by accident or habit. We may hence care about how the civilization generates its collective behavior and why.
Human civilization can be said to have become greener by individual scientists observing the world and noticed certain environmental risks, activists and intellectuals bringing these concerns to public discourse, national and international organizations deliberating them and responding among other things with laws and treaties, as well as funding and organizing projects to remedy the matter, as well as programs to inform people about the situation. The pooled understanding results in an institutional framework guiding future action.

To whom is the good beneficial? Even though the civilization may not be a moral agent, subject, or patient it might just be that the virtue makes things good for the actual moral subjects, or produces good by its exercise. If one bites the collective mind bullet it is entirely possible to have the civilization as beneficiary too.

If we were to calque more strongly on MacIntyre's definition we may add good outcomes within the practices of the civilization. Civilizations do certain things (maintain themselves, change their culture, have interactions between their parts etc.) and some of these may be regarded as practices in his sense (e.g. a civilization engaged in environmental stewardship). We may hence ask what practices civilizations engage in in order to find candidate virtues.

One can note that civilizations are not obviously moral patients: they do not appear to have a first-person phenomenal experience of the world and it is hard to imagine a civilization suffering despite its members doing fine (but see Schweitzgebel 2014). However, it may well be a moral agent since it is composed of individual moral agents able to coordinate their own behavior on a large scale (indeed, this coordination potential is part of our definition of it being a civilization).

4.3 Does It Make Sense to Talk About the Character of Humanity?

Virtue is tightly associated with character. But for civilizational virtues to function conceptually, must there not be a "character of humanity"?

People throughout philosophical history have referred to humanity as having a collective character, often exploring its universal traits or lamenting its moral failings. Aristotle's conception of humans as "political animals" implies a shared nature that shapes our collective endeavours and allows us to meaningfully ask, "What is an excellent civilization?" without committing a category error. Similarly, Enlightenment thinkers emphasized human autonomy and self-governance as foundational principles, yet the epistemic, emotional, and practical limitations of individuals inevitably shape the kinds of societies that can emerge. These limitations imbue civilization with a distinct character, even when alternative configurations might be logically possible.

Even if one denies that humanity as a whole has a unified character, historical civilizations have displayed distinct "personalities" in how they pursued their goals. For instance, Viking colonization was characterized by opportunistic wintering and settlement, while the classical Greeks established colonies through deliberate planning in the polis. In contrast, the Chinese empire expanded primarily through integration and governance rather than colonization. These differences reflect variations in parameters such as discounting future gains, risk aversion, epistemic approaches, and the ways in which individual contributions are synthesized into collective action. Hence global civilization plausibly has some form of character, which could be different.

The opposing case - that all sufficiently mature civilizations become similar in overall behavior – represents a bold sociological and ethical claim. This perspective assumes that universal principles or constraints ultimately shape the development of all societies (Coughlin 1996). However, history suggests otherwise: civilizations have exhibited significant diversity in values, priorities, and strategies, even when confronting similar challenges. It has not been true in the past at least, even if we can recognize commonalities between societies facing similar challenges (Jebari 2021). In particular, as technology advances many material limits on human lifestyle and cultural outlook become less constrained by scarcity, and inter-cultural factors such as institutions, signaling, and competition instead allow a wide range of individual and group approaches.

4.4 Does Humanity or a Civilization Have "Free Will"?

One way to argue against a moral agent account of civilizations is to argue that their behavior is deterministic, which implies that they do not have the right properties to choose alternatives based on considered reasons.

A macrohistorical argument would be: the collective behavior of humanity follows certain patterns independently of what the members intend, and these patterns are fixed by general laws of sociology, economics, game theory or ecology. Hence the resulting behavior is neither blameworthy or virtuous, but just what happens.

A common version of this goes back to Winwood Reade's *The Martyrdom of Man* (1872, Chapter II, "Religion", pp. 143-4):

> As a single atom man is an enigma: as a whole he is a mathematical problem. As an individual he is a free agent, as a species the offspring of necessity.

Many macrohistorical theories predict more or less unavoidable patterns of history, whether cycles, a Hegelian or Marxist dialectic, or inevitable collapses (Galting & Inayatullah 1997). The deep idea here is that individual free human action produces a collective behavior that has a "scale separation": the activity on the microscale leads to emergent macroscale patterns, but the details of micro-interactions are not needed for describing the behavior on the macroscale.

To make a physical analogy, the exact way gas molecules bump into each other may be individually complicated and indeterministic, yet macroscopically this averages out into properties like pressure, temperature and density that respond to changes in the environment in a deterministic way. Indeed, the success of statistical mechanics is based on the existence of scale separation in many important systems.

In the human case individual free action produce macroscopic phenomena like market equilibria, public opinion, economic growth and other things that have a dynamics independent from the individual actions.

However, this view can be criticized by a Popperian argument. Karl Popper argued in *The Poverty of Historicism* (2013) that much of social activity and the resulting historical process is due to ideas, and ideas emerge from individuals in a fundamentally unpredictable way. They then spread across society, affecting changes in behavior potentially on all scales. We logically cannot predict ideas before we have them. Hence, Popper argues, this makes historicism – the idea that we can know where history is going – impossible.

In the analogy with statistical mechanics this corresponds to a case with no scale separation: microscale events can generate cascades that reach the macroscale. This is the hallmark of systems near critical transitions (Sornette 2006), still somewhat amenable to physical analysis but usually exhibiting complex behavior and unpredictable sudden transitions.
For example, an earthquake likely starts when an individual crystal grain under increasing strain eventually slips, releasing energy and making other grains move. Eventually the cascade reaches macroscopic size and an earthquake ensues. The size distribution of earthquakes is very regular (Schorlemmer, Wiemer & Wyss 2005), yet when a quake happens and what magnitude it will be is determined by microscale factors.

What makes the human system more troublesome than most physics is that individual actions and ideas can create new structures that also persists and change the dynamics over time.

The formation of the Bretton Woods System in 1944, setting up the IMF and World Bank Group and making the US dollar the global reserve currency was due to a particular meeting of particular people with different ideas, and a fair bit of random events during the meeting (Steil 2013). Similarly, many guiding institutions that do affect the world on the largest scale, whether the United Nations,

principles for who can order a nuclear strike, the WWW, religions, etc. appear to have emerged from small groups of people or individuals.

While many collective behaviors may show scale separation, there are clear examples where microscale events can scale up and become macroscale patterns of human civilization. Hence the macrohistorical critique of civilizational agency fails.

However, one can also critique civilizational agency by arguing the opposite: there is no scale separation, and it is all an utterly unpredictable sum of individual actions. Hence the global behavior is not just unpredictable but also lacks structure. Hence there cannot be any civilizational virtues, because there is not enough regularity to generate proper reasoning and motivation for the actions actually taken – it is all essentially all ad hoc at any moment

But institutions show that humans are good at making structures that reduce uncertainty. Human societies show powerful institutions that coordinate action, yet are contingent products of individual ideas, emergent orders, and constrained by aspects of game theory, human nature etc. These institutions show complex information processing, and nontrivial behavior. It is not unreasonable to think that large institutions can produce integrated collective action on a civilizational scale (e.g. consider Catholic church in western medieval and early modern period).

4.5 Can Civilizations Count as Autonomous Agents?

This matters not only for analysis of whether it makes sense to speak of civilizational virtue, but also for the issue of how to incorporate artificial intelligence into the picture.

One of the key threats from AI may be a loss of autonomy for humanity (Holm, Sandberg & Fisher 2025) This can be loss of practical autonomy, where disempowerment of humanity by autonomous technology threatens survival or thriving, in particular if it is not value aligned with humanity. It can also be loss of moral autonomy, threatening very important values or capabilities depending on which ethical account one adheres to. Even in an error theory account human autonomy has instrumental value in enabling individual human happiness through a sense of control over one’s life.

David Copp has argued for the collective autonomy thesis, that it is possible for a collective to be morally responsible even if no member is individually responsible (Copp 2007). Animation theories postulate agency as a true emergent property of the group and its interactions, while List and Pettit on the other hand propose a methodologically individualist model of group agency relatively close to the extended cognitive systems I have focused on (List and Pettit 2011).

For autonomous agency, an entity must control its actions through rational reflection and self-determined choices. Humanity's current capacity for autonomy is questionable. While civilizational institutions enable reflection, they lack sufficient control over collective actions due to the semi-anarchic state system. However, various coordination mechanisms exist - including norms,

institutions, markets, and collective deliberation processes - that sometimes achieve global change (e.g. the International Court of Justice, the Montreal Protocol 1987, and the Nuclear Non-Proliferation Treaty 1968). Though humanity may lack the coherent structure for full autonomy now, these elements suggest potential for future autonomous agency.

Even more tantalizing, it might be possible to eventually integrate humanity enough that it does become an autonomous entity. This might well include aligned artificial intelligence as a component or a key coordinating factor (Holm, Sandberg & Fisher 2025).

5. Candidate Civilizational Virtues

Civilizations are not people, so many classic virtues do not apply. Not every individual virtue makes sense as a group or civilizational virtue. Lacking a sense of pleasure or pain -- unless certain theories of collective emotion are true -- a civilization cannot have the virtue of temperance.

Some things may change character as they are applied to larger groups. Emotionality is part of being human, and it makes sense for groups to exhibit collective or social emotions given human shared intentions and sociality. It is less clear we should wish for civilizations to be emotional - the emotions that make sense for humans given our evolutionary past and sociability may not be applicable on the civilization level.

I assume civilizations are alone in this treatment: we are already likely at the point where it makes more sense to speak of a single global civilization than several independent one, and this state is likely to remain. Unless we encounter aliens or diverge so much that different (post)human civilizations should count as separate, the civilization we consider virtues of is alone. Were we in a multi-civilization setting, social civilizational virtues such as friendliness, modesty or wittiness may make sense. Ord (2020) points out that treatment of animals and the environment may still count as civilizational compassion for non-members.

5.1 Epistemic Virtues

Epistemic virtues are obvious candidates for civilizational virtues, and most individual epistemic virtues have obvious large-scale counterparts. They represent dispositions that lead to knowledge and truth, things that are instrumentally useful and may have inherent value. A civilization that is honest values and pursues truth over self-deception. A truthful civilization maintains a correspondence between its actual state and its model of itself. A civilization that is intellectually humble recognizes its fallibility and avoids overconfident behavior and theorizing. A civilization is creative if it generates novel and valuable ideas and explanations.

Here the virtues are tied to the shared epistemic system and can be independent of member virtue or beliefs. The scientific community can overcome individual researcher bias and fallibility by an institutional framework and routines that make the overall knowledge progress (slowly) by weeding out the mistakes and selecting better theories (according to standards that may themselves be improved). Similarly, a civilization may contain various institutions and practices pursuing knowledge with varying reliability and efficiency, combining the results into a better pursuit.

It is clear that these epistemic virtues are acquired. The scientific method had to be invented, and refined. They do not come naturally from just assembling people, but by having a shared epistemic system that updates its knowledge and directs inquiry in the right way - and these behaviors can be structured and aimed in very different ways.

5.2 Caution and Bravery

Toby Ord suggested that the civilizational goal of keeping itself alive is the virtue of prudence from a civilizational perspective. In philosophy prudence in individuals is the ability to govern and discipline oneself by the use of reason, and corresponds to Aristotle's *phronesis*. Ord's usage is closer to the modern everyday meaning cautiousness. Still, seeing self-preservation as a virtue for a civilization makes sense, especially since all other virtues require its continued existence.[4]

We might see caution and bravery as virtues of risk management. A cautious civilization avoids extreme risks such as existential risk, and joint actions that may precipitate it. A courageous civilization on the other hand manages risk in a well-adjusted way.

In either case, it requires a civilization to have the knowledge that its existence is contingent and possibly can be cut short, what kind of risks exist, and a view of what value there is in maintaining itself --- as well as the actions needed to perform this maintenance, risk management, and value updating.

5.3 Patience

Stewart Brand and Fred Hoyle bring up patience. Patience is the disposition to act (or wait) to get long-term benefits even though there are near-term tempting actions. Somewhat similarly, a civilization exhibiting liberality acquires and spends resources appropriately. A civilization can through its institutions act to both prevent premature use or overuse of resources, and set up projects that will only bear fruit in the remote future. If the longtermism view is correct that most value will be achieved in the far future civilizational patience is likely necessary for realizing it.

Civilizational patience may be in tension with individual benefit, since foregoing quick improvements in order to achieve safety sustainability may leave living individual worse off (an interesting example is the “coal question” of leaving fossil fuels for future generations, see Mill (1866) and MacAskill (2022, p. 138-141). The virtue of patience in individuals is relative to their own lifespans, and does not correspond to the civilizational patience virtue.

5.4 Peace

[4] This includes Stoic preparedness for death: it is a virtue exercised during life, rather than after. A civilization acknowledging its finite existence can act accordingly, choosing what risks are worth reducing and which have to be accepted. That it will eventually end is no reason to not do meaningful things in the interim, or extend that time as is proper.

Victor Hugo suggested peace as a civilizational virtue. This is an internal virtue, perhaps similar to tranquility in individuals.

This is an example of a virtue where the way it is achieved and the reasons behind it may matter. A totalitarian state or a world of tense mutually assured destruction can be peaceful, but the reasons are not particularly good. The properly peaceful civilization lacks war because conflicts are defused in non-violent ways.

Hence global cooperation may be a necessary component virtue. It is the ability to work collaboratively across national and cultural boundaries to achieve common goals and address global challenges. This virtue involves the creation and maintenance of international institutions, treaties, and alliances that facilitate collective problem-solving and peacekeeping on a global scale. Some thinkers may also suggest pluralism or global justice are also needed component virtues.

5.5 Environmental Virtues
(Sustainability, Harmony With Nature, Preventing Suffering)

Care for the environment has become a topic of virtue ethics, and include new virtues not discussed in the premodern discourse (van Wensveen 1997). Hill (2017) argues that intrinsic value, utility, interests or divine commands are unsatisfactory for explaining why we rightly feel unease about environmental destruction, and instead argues that in order to live an excellent life one needs to love nature and that wantonly destroying it corresponds to vice. This is an individual account rather than a collective one, although by its nature environmental virtue considerations naturally tend to be highly other-regarding (Cafaro 2001).

Collective environmental virtues are important. It is not enough that there is environmental concern among people, it is generally regarded as key that these concerns are channeled into institutional structures that enable positive environmental action. For example, avoiding causing extinctions is not truly a virtue to a person since nobody can cause species extinctions personally – but as member of a society, one becomes partially responsible for its ecological conduct. As Brian Treanor (2010) notes, environmental virtue ethics may need a “virtue politics” of collective actions, traits or dispositions. Civilizational environmental virtues appear eminently feasible.

5.6 Technological Stewardship

Technological stewardship means the responsible development and deployment of technology to enhance human well-being while minimizing harm and unintended consequences. As technology becomes increasingly powerful and pervasive, global civilization must cultivate a collective sense of responsibility and ethical oversight to guide its trajectory. This involves nontrivial tradeoffs of individual freedom, near- and long-term prosperity, risks. and uncertainty.

Obvious past examples have been the handling CFCs, bans on certain environmental toxins, regulation of nuclear power and weapons, and attempts to further development of climate-friendly technologies such as renewable energy and CCS. Global technological commons like

telecommunications, seismology, meteorology and the Internet require global coordination (or other mechanisms, see Stern (2011)). The current debates about geoengineering and AI governance give other examples. The issue is not just risk management but aiming for development that is in the overall interest of humanity – and creating mechanisms making this feasible despite other incentives favoring less desirable outcomes.

Technology may also affect other virtues in a number of ways (Danaher 2024). Technology has made environmental virtues relevant. Human enhancement may be motivated by a virtue ethical view (More 1993), be subject to virtues, but also produce changes in humanity that affect civilizational virtues, e.g. moral enhancement (Persson & Savulescu 2012). Choosing technologies that affect individual and collective virtues well may be a key aspect of virtuous technological stewardship.

Are there undiscovered civilizational virtues? This seems likely: if we have only recognized environmental virtues and technological ones recently, there are likely more. Some come about because of a changed context, such as being a species changing the biosphere on a large scale or humanity leaving Earth. Others may come about because of changed internal abilities for coordination or individual-collective interactions. A world with globalized economy and identity, biomedical moral enhancement, designer people, or AI-human hybrid societies may enable – or require – new civilizational virtues.

6. Conditions

What are the conditions for civilizational virtues?

For virtues to be attributed to civilizations, there must be a sense of collective agency. This means that civilizations, through their institutions, policies, and collective actions, can be seen as entities capable of moral action. This requires collective rational deliberation and practical wisdom.

Bostrom (2019) worries that the current semi-anarchic world order lacks the structure to have the collective agency to deal with certain global risks. We might say that the semi-anarchic system of states is not right collective agency instrumentally, and hence world not virtue apt (at least not for risk-reducing civilizational virtues). His Singleton concept is a hypothetic world order that has collective agency (2006), but it is undefined what kind of rational deliberation and practical wisdom happens in this state: it might well be a profoundly unvirtuous state, even if it is instrumentally able to reduce risk well.

At least in an Aristotelean view there needs to be a telos for a being to strive excellently towards. For civilizational virtues, there must hence be an understanding of the telos of global civilization—what constitutes its flourishing or ultimate good. One could use an aggregative view and merely sum the good inside the civilization (e.g. the wellbeing of its inhabitants) and see this as a natural telos. This is highly amenable with many perspectives in longtermism and existential risk reduction

(MacAskill 2022). However, there might be aspects of civilizations that are more teleological, for example linked to choosing its trajectory (Baum et al. 2019).

One might argue that the practice of civilizations in MacIntyre's sense of practice, is – by the above tentative definition of civilization – maintaining its cohesion, ability to interact with the environment, internal reasoning and spatiotemporal structure, and hence there can be excellent ways of being a civilization. Similarly, these practices can produce internal goods for the civilization. His sense is relative to a historical and social context of practices civilizations lack, so the analogy has some weaknesses, but there is enough overlap to make sense of the statement that a civilization has e.g. a practice of maintaining its cohesion that is not just the technical ability to do it (e.g. institutions, normative frameworks) but it is subject to reflection (e.g. political philosophy, sociology and ethics) and comparison to ideals, historical precedents, and other evaluations. Together this makes the civilization virtue-apt.

7. Existential Risk and Virtue Ethics

Considering the challenges to the existential risk field raised by Sundaram, Maas and Beard (2022), the field seems to be ripe for thinking about the risks and itself from a virtue perspective.

One of the most important questions in existential risk research involves how to prioritize different risks. This is a complex tradeoff even if one takes a very straightforward Bayesian epistemology and decision theory, linked to a consequentialist ethics. However, the choice of epistemology, decision theory, and normative theory is certainly not universal. Realistically, uncertainty of different kinds (epistemic, normative), and the complexities of practical action both in researching and mitigating the risks means that there is little hope for a convenient decision-rule. Instead, we are left with weighing different forms of evidence, collaborating under disagreement, and having to act practically in the world… something that sounds much like practical wisdom.

At the very least we might hope individual existential risk researchers may show practical wisdom. However, this research is – like most research – also a communal activity. It seems entirely possible to consider whether the community shows an aggregate practical wisdom. There is shared deliberation (which can be more or less open, critical, have mutual respect, and hold diverse perspectives) with common goals and values, ideally with inclusive participation, and with institutional structures (whether research groups and publications, blogs, methods of discourse, accountability etc.), experience and learning, and contextual adaptation. Indeed, many of the challenges for ERS in (Sundaram, Maas and Beard 2022) sound like the challenges a virtuous person face: how to balance caution with lost opportunities, accountability with free expression, gaining research funding and practical influence while being independent, and so on. Indeed, they note that there is likely no single solution to any of these challenges – a very virtue ethics-style observation.

So much for the virtues of the existential research field. Civilizational virtues bring another perspective to the issue. A prudent humanity would presumably make use of the best insights and practices from this field to set policies on the largest scale.

8. How Do We Become a Virtuous Civilization/Species/Biosphere?

How do we become a virtuous civilization/species/biosphere?

Will McAskills idea of the "Long Reflection", a hypothetical period of time during which humanity works out how best to realize its long-term potential (Ord 2020, ch. 7; MacAskill 2022 p. 98-99), is perhaps an ideal of a virtuous civilization wisely considering its long-term options. But this is commonly assumed to occur after the pressing problems of existential risk have been solved. Virtue is presumably needed before that – it is not just a luxury for leisure civilizations.

A plausible pathway there is to make the tools to make the tools to make the tools – a gradual refinement of institutions, understanding and coordination mechanisms. These are useful for overcoming the challenges on the way but also for forming a collective discourse and decision mechanisms that would make the Long Reflection meaningful. Even, as McAskill notes, it might never actually happen in a pure form, something approximating it is still potentially valuable. The trajectory of our civilization is too important not to attempt to aim in a good direction (Baum et al. 2021).

Could we become autonomous about our civilizational trajectory? That seems to require again global coordination that can achieve integrated reflection on rational goals and act on them. This does not necessarily have to be a "singleton" in the sense of Bostrom (2006) as a world-order where there is a single sovereign decision-making agency at the highest level that can make decisions stick, any more than we can speak about a person being autonomous despite having conflicting internal impulses. However, the more internal reasons can be reliably reflected and acted upon according to deeply held values, the more autonomy we can speak about. This seems to suggest that humanity may be slowly gaining autonomy as a group agent. It does not currently have much, but as globalization of discourse and coordination grows this may shift – just as it has increased tremendously from the state a few centuries ago.

It is also worth noting that this is no longer a purely human system. One can argue that non-human extended cognitive systems (institutions, states, markets) are already a major part, but increasingly artificial intelligence acts as powerful complements and substitutes for human in coordinative systems. AI might extend coordination ability to the degree that it enables humanity's autonomy. Humanity on its own might be unable to be autonomous (although this has not been proven!) but requiring AI being meaningfully part of the authentic human reflection – a deeper form of AI alignment (Holm, Sandberg & Fisher 2025).

Funding: This research was funded by Effective Giving.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: This publication did not generate new raw data.

Acknowledgments: Thanks to Remi Sussan for finding the source of the Hugo quote. Further thanks for fruitful discussions to the Future of Humanity Institute community, TJ, and the participants of the "Existential Threats and Other Disasters: How Should We Address Them?" conference organized by The Center for the Study of Bioethics, The Hastings Center, and The Oxford Uehiro Center for Practical Ethics in Montenegro May 30-31 2024.

Conflicts of Interest: The author has no conflict of interest to report.

References

1. Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., ... & Yampolskiy, R. V. (2019). Long-term trajectories of human civilization. *Foresight*, *21*(1), 53-83.
2. Beggs, D. (2003). The Idea of Group Moral Virtue. *Journal of Social Philosophy*, 34(3).
3. Bostorm, N. (2006) What is a Singleton? Linguistic and Philosophical Investigations, Vol. 5, No. 2 (2006): pp. 48-54
4. Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, 10(4), 455-476.
5. Brand, S. (2000). Taking The Long View. *Time*, *155*(17), 86-86.
6. Byerly, T. R., & Byerly, M. (2016). Collective virtue. *The Journal of Value Inquiry*, 50(1), 33-50.
7. Cafaro, P. (2001). Thoreau, Leopold, and Carson: Toward an environmental virtue ethics. *Environmental ethics*, 23(1), 3-17.
8. Copp, D., 2007. The collective moral autonomy thesis. *Journal of Social Philosophy*, 38(3).
9. Coughlin, R.M. (1996) "Convergence Theories ." *Encyclopedia of Sociology*. Retrieved January 23, 2025 from Encyclopedia.com: https://www.encyclopedia.com/social-sciences/encyclopedias-almanacs-transcripts-and-maps/convergence-theories
10. Danaher, J. (2024). How Technology Alters Morality and Why It Matters [Ethics]. IEEE Robotics & Automation Magazine, 31(2), 147-148.
11. Dyson, F. *Disturbing the Universe*, Basic Books, 1981.
12. Fricker, F. (2010) Can there be Institutional Virtues? in T. Szabo Gendler & J. Hawthorne (ed.), *Oxford Studies in Epistemology 3* Oxford: Oxford University Press, 2010, pp. 235–252.
13. Gallagher, S., & Crisafi, A. (2009). Mental institutions. *Topoi*, 28, 45-51.
14. Gallagher, S. (2013). The socially extended mind. *Cognitive systems research*, 25, 4-12.
15. Gilbert, M. (1992). On social facts. Princeton University Press.
16. Hoyle, F. (1966). *Of men and galaxies* (Vol. 1). University of Washington Press.
17. Hugo, V. (1878) Discours pour Voltaire (discourse for Voltaire) given at the théâtre de la gaité, May 30 1878 for Voltaire centenary.
18. Jebari, K. (2021). Replaying history's tape: Convergent cultural evolution and the prospects of humanity after a social collapse. *Available at SSRN 3840244*.
19. Hill, T. E. (2017). Ideals of human excellence and preserving natural environments. In The Ethics of the Environment (pp. 319-332). Routledge.
20. Holm, C., Sandberg, A. & Fisher, L. (2025) *Law, Liberty, Leviathan: Limits to Individual Autonomy in an Age of Artificial Intelligence and Existential Risk*, Stockholm University Press, forthcoming.
21. Hursthouse, Rosalind and Glen Pettigrove, "Virtue Ethics", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/>.
22. Hutchins, E. (1995). *Cognition in the wild*. MIT press.
23. Kuipers, B. (2012). An existing, ecologically-successful genus of collectively intelligent artificial creatures. *arXiv preprint arXiv:1204.4116*.

24. Lihua, Z. (2013). China's traditional cultural values and national identity. Carnegie China. https://carnegieendowment.org/research/2013/11/chinas-traditional-cultural-values-and-national-identity?lang=en
25. List, C. and Pettit, P., 2011. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.
26. MacAskill, W. (2022). *What We Owe The Future*. Simon and Schuster.
27. MacIntyre, A. (1981). *The nature of the virtues*. Hastings Center Report, 27-34.
28. Mill, J.S. (1866) Speech in Parliament Tuesday 17 April 1866. https://hansard.parliament.uk/Commons/1866-04-17/debates/d4ba0459-2d9f-468f-b589-7321ecc1dfb3/Army—RegimentsInIndia#contribution-a313f5c8-cc87-4465-9bff-1f171315847e
29. More, M. (1993) Technological Self-Transformation: Expanding Personal Extropy. *Extropy* #10 (4:2) Winter/Spring 1993.
30. Mulgan, R. G. (1968). Individual and Collective Virtues in the" Republic". *Phronesis*, 84-87.
31. Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.
32. Persson, I. & Savulescu, J. (2012) Unfit for the Future: The Need for Moral Enhancement. Oxford University Press.Galtung, J., & Inayatullah, S. (1997). Macrohistory and macrohistorians. *Perspectives on individual, social, and civilizational change*.
33. Popper, K. (2013). *The poverty of historicism*. Routledge.
34. Schwitzgebel, E. (2015). If materialism is true, the United States is probably conscious. *Philosophical Studies*, *172*, 1697-1721.
35. Scheffler, S. (2018). Why worry about future generations?. Oxford University Press.Shenker, I. (1972) Will Galaxy Reveal a Technological Cancer? A Physicist Wonders. *New York Times*, Nov 27 1972.
36. Schell, J. (1982). *The Fate of the Earth*. Knopf.
37. Schorlemmer, D., Wiemer, S., & Wyss, M. (2005). Variations in earthquake-size distribution across different stress regimes. *Nature*, 437(7058), 539-542.
38. Sornette, D. (2006). *Critical phenomena in natural sciences: chaos, fractals, self organization and disorder: concepts and tools*. Springer Science & Business Media.
39. Steil, B. (2013). *The battle of Bretton Woods: John Maynard Keynes, Harry Dexter White, and the making of a new world order*. Princeton University Press.
40. Stern, P. C. (2011). Design principles for global commons: Natural resources and emerging technologies. *International Journal of the Commons*, 5(2), 213-232.
41. Sundaram, L., Maas, M. M., & Beard, S. J. (2022). Seven questions for existential risk studies. In *Managing Extreme Technological Risk* (ed. Catherine Rhodes).
42. Treanor B., Environmentalism and Public Virtue, in: *Virtue Ethics and the Environment*, P. Cafaro, R. Sandler (ed.), Dordrecht 2010, p. 9–28.
43. van Wensveen, L. (1997). *Dirty virtues*. Humanities Press Intl Inc.