Liberalism Forever

Simon Goldstein & Peter N. Salib[1]

Abstract. We argue that liberalism—market economies governed democratically—is the best approach for navigating the far future. A growing longtermist literature paints humanity's path to good outcomes as narrow, with small errors risking value lock-in, gradual disempowerment, or other forms of permanent catastrophe. We argue that this literature underestimates the institutional dynamics that have historically steered liberal societies past similar predictions of crisis. We defend *long-term liberalism* by examining the standard arguments for markets and democracy and asking whether they survive the structural changes anticipated for the far future: transformative AI, space colonization, and radical economic transformation. We claim that these arguments are largely robust, and in several cases strengthened. In the far future, markets continue to aggregate information, allocate goods efficiently, and foster innovation; democratic institutions can continue to supply public goods, commit to peaceful redistribution, correct errors, and accommodate reasonable pluralism. We close by arguing that two recent longtermist governance proposals—viatopia and the long reflection—are faint-heartedly liberal: they sound liberal in the abstract but risk illiberalism when implemented. The right approach for managing the far future is not radical new institutions but adaptations of the existing liberal toolkit to meet future challenges.

1. Introduction

What would make the far future go well? A growing literature in moral philosophy, AI safety, economics, and longtermism explores this question.[2] The picture it paints is often fraught. MacAskill and Moorhouse (2025a) argue that utopia is a narrow target, which we could easily miss. Kulveit, Douglas et al. (2025) argue that AI development could gradually disempower humanity. Bostrom (2014) has argued that permanent value lock-in is a serious risk of sharing the world with digital minds. And Aschenbrenner (2024) has argued that short-term winner-take-all races may permanently decide who controls the future. On these accounts, humanity's path to the best futures is narrow. Small errors risk permanent catastrophe. If this picture is correct, radical societal actions might be necessary to prevent catastrophically bad futures.

We think that this literature gives too little attention to institutional dynamics. Social institutions can, and do, steer societies either away from or towards disaster. Thus, for better futures, it may be more important to select the right future institutions than to imagine the best future states of affairs.

[1] Equal contribution; alphabetical order.

[2] See for example MacAskill and Moorhouse (2025a, 2025b); MacAskill (2025); Kulveit, Douglas et al. (2025); Bostrom (2014); Hanson (2016); Cowen (2018); Shulman and Bostrom (2021); Finnveden, Riedel, and Shulman (2022); Christiano (2019); and Aschenbrenner (2024).

This paper focuses on those institutional dynamics. We argue that longtermists have underestimated the power of the ordinary liberal toolkit for managing the far future: free markets, governed democratically to fix market failures. In a slogan: liberalism forever.

For an analogy, consider Marx. Like us, he stood at the precipice of radical change. Like many longtermists, he identified genuine negative dynamics that would emerge from those changes: downward wage pressure, capital accumulation, agglomerations of power. Projecting those dynamics forward, Marx predicted crisis and endorsed a transition to a post-capitalist society.

Marx was wrong about both the crisis and the need for a radical transition. Today, wages are far higher than in Marx's time, inequality is lower, and political power is more broadly dispersed. Marx was wrong because he could not solve for the equilibrium.

Early industrializing nations were liberal.[3] The same liberal institutional dynamics that helped produce industrialism steered them away from the worst outcomes. Property rights and freedom of contract unlocked industrial growth. Democratic governance produced the modern welfare state, lifting the fortunes of those whom economic growth would otherwise leave behind.

The details of these institutional responses would have been impossible to predict in advance. The idea that the 19th century's coal miner might be the 20th century's ultrasound technician was inconceivable. Liberalism succeeds by adopting mechanisms of social equilibration that systematically discover better futures. Liberal institutions have already driven humanity through the Industrial Revolution, electrification, the atomic age, the digital revolution, and the information age. We think that these forces will continue to be powerful through the AI transition and into the far future.

Section 2 clarifies our approach. We focus on questions of optimal institutions, distinguishing these from questions about optimal outcomes. We then introduce our preferred view: *long-term liberalism*. Long-term liberalism says that the far future, including large-scale space colonization and radical economic transformations, should be governed by markets and democracy.

Sections 3 and 4 are the heart of our analysis. Throughout, we use a simple methodology. We begin with standard social-scientific and philosophical arguments for markets and democracy. Then we explore whether the structural assumptions underpinning these arguments will extend into the far future. In particular, we're interested in whether dynamics related to transformative AI, space colonization, and structural economic change undermine traditional arguments for liberalism.

In the case of markets, we focus on three positive arguments: that markets aggregate information, allocate goods, and foster innovation. We then ask whether, in the future, dynamics like market concentration, externalities, transaction costs, discounting, and post-scarcity undermine the positive case.

Regarding democracy, we focus on five positive arguments: that democratic institutions improve the epistemics of decision-making, incentivize the production of public goods, offer commitment devices that lower the chance of revolution, correct errors in policy-making, and facilitate

[3] See Acemoglu and Robinson (2012); Mokyr (2009).

compromise under conditions of reasonable pluralism. We then ask whether these arguments hold into the far future.

All told, we argue that the case for markets and democracy is largely robust to these changes of the far future. Future conditions may, in some cases, weaken the classic arguments. But more often, future conditions either strengthen the arguments or leave them unchanged.

Do existing longtermists disagree? We are not certain. Above all, we think that existing analysis has been incomplete, because it has not engaged with many of the classic arguments for markets and democracy. In Section 5, we suggest that existing longtermists are *faint-hearted* liberals. That is, longtermists rarely advocate explicitly illiberal policies, and their preferred interventions often share much in common with liberalism. But their analyses sometimes underweight analytic tools from economics, public choice, political science, or law that would change the conclusions. We think drawing on these tools tends to make the likelihood of bad futures look lower, and complicates several of the proposed interventions. We illustrate our discussion with two recent cases: viatopia, and the long reflection.

To be clear, our argument for liberalism is not an argument for complacency. If liberal institutions are to succeed at governing the far future, they will need to be updated. Perhaps radically. New kinds of actors, including AIs, will have to be incorporated into the economic and political order. Doing so will require new institutional designs. Otherwise, novel problems, like the easy reproduction of AI voters, could break the system. Section 6 sketches some of these challenges and offers thoughts on how to respond.

2. Liberalism

This paper will argue that liberalism is the best approach for managing the far future. To make this claim, it is worth first drawing a few distinctions.

First, there is a difference between questions about optimal *institutions* and questions about optimal *outcomes*. The outcome question is: what are the best states of affairs in the far future? This question asks what the future should look like: what values should prevail, and how resources should be distributed. Longtermists have often argued, correctly, that clear answers here are hard to find. Reasonable people disagree, for example, about what goods are most relevant to the best futures: Hedonic abundance? Austere virtue? Some yet-unknown moral desideratum?

By contrast, our focus is on institutional questions. We will argue that liberalism is the system most likely to navigate towards better outcomes. In analyzing liberalism, our focus will be on two sets of institutions: markets and democratic governance. By democratic institutions, we'll have in mind both majoritarian institutions (such as elections) and structural institutions (such as separations of powers). Liberalism is also traditionally associated with individual rights. We'll think of various combinations of property, contract, speech, and voting rights as defining the details of how markets and democratic institutions function. All advanced liberal nations are "mixed," combining these institutions in varying proportions.

Second, these kinds of questions can be asked at different time-scales, in multiple ways. First, we can ask about what institutions promote the best outcomes in the far future; or we can ask about what institutions promote the best outcomes today. Second, we can ask which outcomes *today* promote the best outcomes in the far future. In addition, the distinction between *now* and *the far future* is murky. Timelines for space colonization and so forth are contentious. In general, we'll try to cast a wide net in our analysis. Our primary focus will be on what institutions promote the best outcomes *in the far future*. But we will understand the far future capaciously.

Our paper defends *long-term liberalism*. We're interested specifically in whether liberalism is the best set of institutions for a future filled with transformative AI, dramatic structural changes to the economy, industrial explosions, and the colonization of vast regions of space. The vision of liberalism forever is compatible with a wide range of possible futures.

Here is one concrete illustration. Imagine that humanity successfully colonizes a large fraction of the light cone. There are untold planets teeming with life. The relevant agents form a spectrum from biological to non-biological substrates. Humans and AIs interact in many different ways. This vast civilization uses markets to allocate resources. Trade occurs at various levels of abstraction, at various time scales, and for a wide range of goods. The civilization is governed democratically. Within this broadly liberal setting, there are endless variations in the details of governance. Some planets are relatively isolated, governing themselves with relatively little outside input. Large swathes of the light cone have agreed to various basic constitutional ground rules necessary to at least engage in peaceful "international" relations. There is a vast hierarchy of federated levels, which weaken in the strength of their requirements as their breadth expands. In this imagined scenario, the economy is almost unrecognizable from today, in terms of the specific goods and services supplied. But the basic tools of markets and democratic institutions are still used to decide questions of allocation, growth, and governance.[4]

In sections 3 and 4, we defend long-term liberalism using a specific methodology. We'll consider both markets and democratic institutions. In each case, we'll explore some of the most important, standard arguments for the value of these institutions. The standard arguments all suggest that liberalism promotes better outcomes than other forms of governance *under certain conditions*.

We will then analyze whether the long-term future undermines these arguments by violating the relevant conditions. We focus on three important structural changes: transformative AI, future industrial explosions, and space colonization. Our question is whether these changes undermine the assumptions that favor liberalism. And if they do, what institutions should replace liberalism? We cannot hope to be exhaustive in this single paper; but we'll try to hit enough of the standard arguments to be able to start making progress on the main question.

3. Markets

[4] MacAskill and Moorhouse (2025a, §2.2) critique a view they call "easygoing liberalism." Easygoing liberalism endorses as "mostly great," a future containing roughly ten times as many people as live on Earth today, all richer and freer than humans living in rich countries today. Long-term liberalism is distinct from easygoing liberalism because it strongly favors futures containing far more, and far richer, beings.

We'll analyze three main kinds of arguments for markets: they elicit and aggregate information; they allocate resources efficiently; and they promote innovation.

3.1 Markets elicit and aggregate information

The first argument for markets is that the knowledge needed for economic coordination is dispersed across the entire population. The price system is the best known mechanism for eliciting and aggregating that information effectively.[5]

Prices communicate information about scarcity, preferences, and opportunity costs to all market participants, none of whom needs to understand the full picture. This aggregation extends into the future. Markets for stocks, commodities, options, and insurance operate as continuous predictions about future states of the world, pricing uncertainty and rewarding whoever produces new information that shifts the consensus.[6]

This works pretty well.[7] By contrast, the grand centralized planning experiments of the 20th century produced spectacular failures of information aggregation. The Soviet Union struggled to assign labor and capital to their highest value use, and its economy stagnated and collapsed.

Our question is whether markets are the best approach for aggregating information in the far future. One challenge concerns AI. Perhaps capable AI will allow for efficient central planning.

Here, it is worth distinguishing between markets' elicitation and aggregation functions. Technology can allow information to be better aggregated via nonmarket mechanisms. Computers, for example, have enabled the agglomeration and analysis of much larger sets of data than could have been used by a single planner.

But it is not clear how advanced AIs could displace markets' role in eliciting information. Much of the important information in an economy is tacit (Hayek 1945). That is, it is latent inside the heads of workers and consumers in a way that they cannot straightforwardly articulate. But by presenting them with rich sets of products and services, at various prices, from which they choose, the market elicits the tacit information. Moreover, the relevant information may not even *exist*, absent market-based elicitation (Hayek 1968). On this stronger view, the relevant information simply *is* what happens when a market participant is presented with prices and makes choices. These jointly produce the information, not merely retrieve it.

Finally, even if AI central planning could in principle elicit and organize enough information to run the economy, there's still a further question of whether this centrally planned system has any positive advantage over markets. This brings us to our next topic, allocative efficiency. As discussed in the next section, the fundamental theorems of welfare economics claim that, in the absence of market failures, markets deliver Pareto efficient outcomes, which can be efficiently redistributed to correct for inequality. This means that in the absence of market failures, central planning cannot in principle deliver a Pareto improvement over markets, at least in terms of

[5] Hayek (1945).
[6] For information-aggregation mechanisms for the future, see Hanson (2013) and Wolfers and Zitzewitz (2004).
[7] See, e.g., Fama (1970); Martin and Nagel (2022).

efficient static allocation. We therefore think the crucial question for long-term liberal markets is market failure.

3.2 Markets allocate goods

The second classic argument for markets concerns allocative efficiency. The fundamental theorems of welfare economics state that, under the relevant assumptions, competitive market equilibria are Pareto efficient, and that any efficient allocation can be achieved as a competitive equilibrium given a suitable initial distribution.[8]

However, the assumptions underlying the theorems do not hold perfectly true anywhere. Most importantly, the results do not apply in market failure. We focus on four market failures that seem most relevant to the far future. First, market concentration, wherein monopolists extract rents. The result is that fewer goods are produced than would be purchased at the efficient price, leading to deadweight loss.[9] Second, externalities: here, the market price for goods fails to reflect social costs or benefits accruing to entities that are not party to the deal.[10] Third, transaction costs: if making a deal is too costly, then buyers and sellers may be unable to exchange goods at the market price. Fourth, temporal discounting: if today's market actors do not expect to be alive in the far future, they will under-invest in producing benefits or mitigating costs accruing in the future.

Below, we'll consider each market failure in detail. But first, a caveat. Market efficiency is measured by prices. But wealth can come apart from social welfare. Money has diminishing marginal utility. Some outcomes that are efficient with respect to wealth may not be efficient with respect to welfare. Here, redistributing from wealthier to poorer individuals could raise total social welfare.

This problem can be corrected by good liberal policy. For example, tort liability internalizes externalities. Similarly, tax-and-transfer schemes can improve social welfare, by moving wealth from rich people whose marginal utility of a dollar is low, to poor people for whom it is high.

Thus, while imperfect, the fundamental theorems of welfare economics give us a blueprint for evaluating the allocative efficiency of liberalism. We should expect liberal institutions to promote better futures when they can reliably avoid market failure and welfare-reducing inequality. The corollary is that when evaluating long-term structural changes related to AI, space colonization, and industrial explosion, the key question will be whether structural changes lead to market failures or unequal distributions that are unsolvable using liberal institutions.

Finally, some ethical theories sever the connection between preference and welfare. According to hedonists, for example, pleasure rather than preference matters. But the notion of welfare in play in the theorems of welfare economics concerns preference: it reflects, for example, revealed dispositions to consume goods, rather than the amount of pleasure caused by that consumption.

We now consider possible market failures in the far future.

---

[8] See Arrow (1951); Debreu (1959).
[9] See Tirole (1988), ch. 1; Harberger (1954).
[10] See Pigou (1920).

3.2.1 Market concentration

Will market concentration get worse in the far future? If so, and if democratic institutions could not solve the problem, this might give us reason to doubt long-term liberalism.

One cause of future market concentration could be first mover effects associated with space colonization. MacAskill and Moorhouse (MM) (2025a) worry that space resources might be "claimed by whoever gets there first," with "control over those resources" ending up "concentrated among a tiny number of hands."[11] In "How to Make the Future Better," MacAskill warns that "it seems reasonably likely that the allocation process will de facto follow 'seizers keepers,' where whichever country (or, potentially, even whichever company) grabs the resources first holds onto them indefinitely."[12] These concerns lead MM to consider various departures from liberal market allocation, ranging from collective ownership to a ban on space development to equal allocation across all present and future persons.[13]

This is a familiar argument structure: early resource accumulation leads to oligarchy. The same structure underlies Marxist models of industrialization, and ran into the same difficulty.[14] Initial distributions of resources need not determine final ones. The Marxist model was wrong about early industrialization. Rockefeller and Carnegie captured enormous initial advantages in oil and steel. Yet neither Standard Oil nor U.S. Steel dominates global markets today.

The reason is that markets allow the initial distribution to rearrange into something else, via trade. The final equilibrium distribution allocates the resources to their highest-value users, irrespective of who had them first. Nor do initial resource distributions determine who captures the long-term *surplus* from resource extraction. Even when a few people own most of some resource, they have no monopoly on advances in how to *use* the resource. If the owner tries to hoard the resource, and wait for a better deal, he risks losing everything as the market sends signals for the invention of substitutes.

These are just the *market* forces that proved Marx wrong. Liberal democracies can impose policies that do more. If huge firms end up with too much market power, liberal democracies can break them up, as they did with Standard Oil.

These considerations should make us less excited about illiberal space governance regimes like common ownership or equal allocation across all persons.[15] Both collective governance and highly fragmented ownership regimes have economic downsides. Both tend to reduce growth and social welfare. Collective governance creates commons problems and short-circuits incentives to put resources to valuable use.[16] Fragmented ownership raises the transaction costs associated with using the resource and causes holdout problems.[17] Dividing up resources into

[11] MacAskill and Moorhouse (2025a, §2.3.4, p. 11). The authors also consider equal allocation with 'limited possibilities for trade' as an alternative. Our focus here is on whether liberal institutional design changes the diagnosis of these scenarios. See also Bostrom (2014, p. 99-104).
[12] MacAskill (2025, §3.4, p. 15).
[13] MacAskill (2025, §3.4); MacAskill and Moorhouse (2025a, §3.4).
[14] See Marx (1867), esp. ch. 25.
[15] Proposals raised without endorsement at MacAskill 2025, §3.4; MacAskill and Moorhouse 2025a, §3.4.
[16] See Hardin (1968); for the qualification that user communities can sometimes self-govern, see Ostrom (1990).
[17] See Heller (1998).

clear, individual, and medium-sized bundles of property rights–including on a 'seizers keepers' basis–tends to perform best.[18]

3.2.2 Externalities

Externalities are another market failure. Here, the price of a good does not reflect broader costs and benefits to society. Are there reasons to expect externalities to be more severe or harder to address in the long-term?

A preliminary point here is that externalities are just bargaining failures.[19] If transaction costs were low, all of the consumers of air would bargain with polluters to reach the optimal level of pollution. Below, we'll argue that transaction costs will fall in the far future.

Thus, unless the far future involves more externalities than today, or they are much more severe, markets should be less plagued by externalities. One possibility here concerns population size. The more people there are, the more people can be affected by externalities. Perhaps as the population increases astronomically, externalities become so large that they trump regular market effects.

But this is speculative. It depends on local externalities affecting everyone everywhere. This is almost never true. The risk of causing a car crash is an externality of driving. But it is local. If anything, the externalities most likely to affect everyone who exists are *positive* ones. Scientific discoveries are famously non-rival goods. Once produced, everyone can use them without diminishing their stock. This means that, absent some policy intervention, almost all of the benefits of science are externalized. But here, rising population is a benefit, not a detriment. The more people who stand to benefit from possible discoveries, the more benefit they create, and the more can be spent on creating them. In some cases, this will still require market-correcting interventions, like subsidies, patents, or prizes. But the per-capita cost of funding a discovery drops as the population rises.

Future technologies may also make it easier to internalize externalities. Above, we conceded that AI may make it easier to aggregate information. In the long-term, it may be much easier to quantify the externalities associated with various goods. In that case, standard tools for addressing externalities may function better, such as Pigouvian taxes (like the carbon tax).

3.2.3 Transaction costs

Another cause of market failures is transaction costs. These include costs of coordinating transaction-relevant information, though information failures can also persist in low transaction cost environments.[20] For parsimony, we treat transaction costs and information problems together here.

In the long-term, both kinds of market failure may diminish. AI may make it cheaper to process information, verify quality, detect fraud, monitor compliance, and commit credibly to contracts.

---

[18] Demsetz (1967).
[19] Coase (1960).
[20] Compare Coase (1960) with Akerlof (1970).

This strengthens markets in two ways. First, as market failures from transaction costs and information failures diminish, markets approach efficient pricing. Second, as transaction costs approach zero, the Coase Theorem predicts that parties will bargain their way to efficient outcomes regardless of the initial assignment of rights.[21] Under perfect information and zero transaction costs, externalities can be directly negotiated. Shahidi and co-authors call this kind of scenario the *Coasean singularity*: a condition under which the allocative question becomes largely one of who starts with what, because any inefficient allocation will be bargained away.[22]

3.2.4 Temporal discounting

Another market failure involves temporal discounting. Asset prices reflect the preferences of current market participants. Those participants usually do not care about the interests of distant-future people.

In the future, agents may live far longer than they do today. Such longevity will improve liberal institutions. A being that expects to live for millennia has powerful personal reasons to care about long-run outcomes even if it cares nothing about others. Such a being needs to ensure that the future is a good place to live, since it eventually expects to live there.

There is further complexity regarding the unborn. How could longevity solve the problem that actors today do not care about the unborn? This may depend on the expected long-run returns to investments in private goods, versus public goods. All else equal, selfish, long-lived beings acting today prefer to invest in *private* goods, from which they capture all of the social gains in the future. To the extent they do this, long-lived agents' farsighted actions today help only themselves in the far future, not the unborn.

But there are also reasons for selfish, long-lived agents to invest in quasi-public goods. Technological innovations are paradigmatic examples. Once produced, everyone else can use them without depleting their stock. Because of this, endogenous growth theory (see Romer 1990) holds that economic growth rates are largely driven by the compounding effects of investments in public goods. Will the production of public goods increase in the far future? Potentially. Long-lived agents will be around long enough to reap large benefits of faster economic growth.

3.3 Markets foster innovation

Another argument for markets concerns growth rather than distribution. Markets do not just allocate existing resources; they generate new value. Competitive markets produce Schumpeterian dynamics: firms that innovate capture profits, firms that stagnate are replaced.[23] The modern endogenous-growth literature claims that innovation-driven growth is the dominant source of increases in living standards.[24]

[21] Coase (1960).
[22] Shahidi et al. (2025).
[23] Schumpeter (1942).
[24] Romer (1990); Aghion and Howitt (1998); see also Lavoie (1985).

Does liberalism optimally promote innovation in the far future? There is a rich empirical literature studying whether good ideas have recently been getting harder to find.[25] Eventually, new innovations might run out altogether. In that case, there might be no further need for dynamic efficiency.

Alternatively, new ideas might never run out. Nielsen (2026), for example, argues that discovering the fundamental laws in a scientific field may be the beginning of an innovation explosion. Consider computer science. The formalization of computability laid the foundation for new innovations. It is not clear that there is any limit to the novel implications, or useful technologies, that may follow.

But suppose innovations did run out. Suppose further that problems of information aggregation were solved too. So imagine a world containing a superintelligent AI central planner, who only needs to produce allocative efficiency, and who sought to maximize social welfare.

It is still not clear that the AI could beat the market. Again the fundamental theorems show that the market produces Pareto efficient allocations, and that lump-sum transfers can produce any wealth distribution we like. An AI planner could not, even in principle, produce a better outcome. At best, it might produce the preferred Pareto distribution at lower cost. If, for example, running the AI planner had lower transaction costs than the market, then the AI planner would be better. But in the imagined scenario, market transaction costs would also have fallen quite low. Moreover, our decisions about future public policy cannot rest on the assumption of a perfect AI overlord. Instead, we have to make decisions under uncertainty.

3.4 Post-scarcity?

We close with a general ground for skepticism about long-term markets. Perhaps the far future will be a “post-scarcity” society. Here, goods are so abundant that there are no tradeoffs in consumption. Prices make little sense, because everyone can satisfy their preferences at will.

We are skeptical. First, we must distinguish absolute versus marginal preference satisfaction. Even if preferences are bounded, they might forever approach a limit, so that consumers still have marginal incentive to prefer one good over another; in that case, there would still be tradeoffs that allow for market mechanisms. Second, even if the total number of goods in the future is vast, the average consumption per person may not be astronomically higher than today, provided population size continues to increase.[26]

A related worry concerns the end of trade. Many of the benefits of markets require that there are gains from trade. But MM (2025b) worry about situations "with limited possibilities for trade." Their argument, presented in MM (2025a), is that, if parties have preferences that are non-discounting and linear in resources, there will not be any positive-sum trades they can make. If two parties want to have as much of some pie as possible, value each marginal slice of pie as much as the next, and they don't care when they get it, then any initial division of the pie will be

[25] See Bloom, Jones, Van Reenen, and Webb (2020).
[26] See Hanson 2016.

the final one. Any trades would be zero-sum—better for one party only insofar as they were worse for the other.

This argument treats the size of the ultimate pie as fixed, and in doing so ignores two familiar sources of gains from trade: specialization and innovation. These grow the pie of resources, without requiring agreement about ends. They can even grow the pie if it consists of the entire universe's resources. For example, one civilization may specialize in propulsion research, while the other specializes in converting matter to computation. This generates a surplus for even non-discounting preferences because, for every year of delay, more of the cosmos slips beyond the lightcone. The same logic applies to innovation. Even once all resources have been allocated, there is still the question of how to use them. Here again, the pie can be grown by finding new technologies that make more efficient use of limited resources.[27] Such trades can even occur between civilizations of very different technological capability, because of comparative advantage.

4. Democracy

We'll focus on five arguments for democracy: that it improves epistemics, increases incentives to invest in public goods, functions as a commitment device to avoid revolution, corrects errors, and facilitates pluralist compromise.

4.1 Epistemics

The "epistemic" argument for democraty is that it aggregates information from voters.[28] When voters face a choice and each is more likely than chance to identify the better option, majority voting aggregates their signals. Accuracy rises with the electorate's size and diversity.

In the far future, epistemic arguments for democracy may be less compelling. In the far future, new technologies should allow individual predictions to become very accurate. As new technologies improve epistemics more generally, the marginal contribution of democratic institutions to accuracy will fall.

4.2 Public goods

Markets only produce efficient outcomes in the absence of market failure. One of the roles of government is to solve these market failures. Here, we'll focus on one important application: public goods.

Democratic institutions are optimized for provisioning public goods. Selectorate theory explains how democratic institutions shape the incentives of leaders towards spending on public goods.[29] This model begins with the premise that political leaders want to remain in power. Whether they do depends on features of the "selectorate" (S), the portion of residents who have a role in

[27] For more on trade in the far future, see Krugman 1978.
[28] The foundational result is Condorcet's Jury Theorem (Condorcet 1785). See also List and Goodin (2001); Hong and Page (2004); Landemore (2013). For a synthesis, see Austen-Smith and Banks (1996).
[29] Bueno de Mesquita, Smith, Siverson, and Morrow (2003).

choosing the leader. More carefully, remaining in power depends on the consent of the "winning coalition," (W) the subset of the selectorate sufficient to keep the leader in power. Democracies have large selectorates and large winning coalitions (under majority rule, just over half the selectorate); monarchies have small selectorates and small winning coalitions (for example, just over half the nobility); autocracies have large selectorates and small winning coalitions. The ratio of the winning coalition to the broader selectorate (W/S) and the size of the selectorate capture how democratic a country is.

Selectorate theory ties the decisions of power-seeking political leaders to economic policy. In particular, political leaders have a choice of how much of government revenue to focus on *private goods* for the winning coalition, versus *public goods* for all. Private goods are goods that can be enjoyed exclusively by one person, such as food or cars. Public goods are non-rival. Providing them for one person supplies a benefit to everyone. Public goods include infrastructure, technical innovations, and education.

The insight of selectorate theory is that the leader's choice of spending on private versus public goods is sensitive to regime type. Both the ratio of W to S and the size of S matters. When the winning coalition is small, as in autocracies, private goods are the best way to retain the winning coalition. Leaders can buy the loyalty of the few required supporters cheaply by levying broad taxes and buying private goods for the small group of winners. Here, increases in the tax rate have little effect on the preferences of the winning coalition, since their small personal tax bump can be paid off via large transfers from the rest of the population.

By contrast, in democracies, the winning coalition is a large share of the selectorate. And if the franchise is extended broadly, the winning coalition becomes large as a share of the country's total population. Here, it becomes very expensive to buy off the large winning coalition using private goods. Consequently, public goods become the cheapest way for leaders to pay off the winning coalition.

Does this argument fail in the far future? One question is whether public goods will continue to exist. Davidson, MacAskill, and Taylor (2026) argue that public goods in the future will be even more important than today. And they count democratic governance as one possible mechanism for producing such goods. But they are primarily thinking in terms of *voters'* demand for specific public goods, and they note problems that arise if voters value different ones. Selectorate theory comes at the same question from the *leader's* perspective: with a large winning coalition, public goods are the cheapest way to keep power.

The remaining question is whether dynamics in the far future would weaken the incentive for democratic leaders to invest in public goods. We see no obvious reason why not. In the model discussed above, leaders have an incentive to retain power, and must convince the winning coalition to re-elect them. The far future may involve selectorates with larger sizes, and may involve larger numbers of resources being allocated by government spending. But the same game theoretic model seems to apply.

4.3 Commitment

A third argument for democracy concerns commitment devices. Acemoglu and Robinson (2006) famously model democracy in terms of competition between rich elites and the poor populace.[30] Effectively, the conflict can be represented as a game with two players. Prior to democratic institutions, the rich control the choice of how much to tax and transfer. The poor have the outside option of revolt. If the poor revolt, they have a chance of victory, in which case they can decide the tax and transfer rate; otherwise, revolution will destroy a significant share of resources. When the chance of revolution is high enough, the rich would prefer to raise taxes and transfer more to the poor. But the problem is credibility. The chance of revolution fluctuates across time. Even if the chance of revolution is high now, both parties can foresee that in the future the chance may be lower. In that case, the poor expect the rich to rescind their tax and transfer scheme once the threat of revolution subsides. In that case, the poor reject the tax and transfer scheme today, and revolt.

Democratic institutions change the game by allowing the rich to credibly commit to tax and transfer. Under democratic institutions, the tax and transfer rates are determined by the median voter. Extension of the franchise is hard to revoke; it requires either the kind of broad consent associated with constitution-making or a coup. So the poor expect that even once the risk of revolution subsides, they will continue to extract suitably high taxes via the median voter. In this way, democratic institutions lower the risk of revolution.

Do these dynamics weaken in the far future? Here, one concern may be that in the far future, the risk of revolution falls to zero. In that case, elites would have no special incentive to enter into tax-and-transfer schemes.

It is difficult to say whether the chance of revolution will fall in the far future. Granted, new technologies could dramatically increase the ability of the state to detect possible revolutions. But conversely, these same technologies may empower society to credibly threaten the state. Here, one relevant dynamic is Bostrom's Vulnerable World Hypothesis.[31] According to the VWH, new technologies may decrease the costs of powerful weapons, to the point that ordinary citizens may increase their access to them. If the VWH is correct, leaders may have more rather than less to fear from citizens in the far future. In that case, the credibility of tax-and-transfer schemes may be more important than it is today.

Or perhaps the chance of revolution will fall because of "AGI lock-in."[32] In this scenario, states widely deploy AIs to ensure that humans are no longer able to resist the government. We think, however, that this argument neglects strategic dynamics between human leaders and those very AI agents. AI agents may have their own goals, which bring them into the same strategic conflicts with leaders. If AI agents have very low wealth levels, they may want access to significant tax-and-transfer schemes.

More generally, democratic peace becomes more important as the destructive potential of war grows. Institutions that credibly commit states to peaceful resolution of disputes are worth more when the alternative to peace is existential.

[30] Acemoglu and Robinson (2006).
[31] See Bostrom (2019).
[32] Finnveden, Riedel, and Shulman (2022).

4.4 Error correction

Democratic institutions facilitate error correction. Democracy permits the bloodless removal of bad rulers and the reversal of bad policies. Losers accept electoral outcomes because they expect future opportunities to win; every policy is provisional rather than final.[33]

The error-correction argument for democracy avoids certain critiques of other arguments. For example, skeptics of the epistemic arguments argue that voters are rationally ignorant, and thus unlikely to vote for optimal policies.[34] But on an error-correction theory, this does not matter much. Voters do not have to be able to recognize optimal policy, *ex ante*. They need only to be able to tell that things are not currently going well, and vote incumbents out.

One especially important strand of research here is on democracy as a "meta-institution." The idea is that democracy includes a series of institutions for changing first order institutions, such as legislative reform, constitutional amendment, and the ongoing evolution of voting systems and deliberative procedures.[35]

We think this argument for democracy may be stronger in the far future. First, the range of possible trajectories in the far future is wider than in any past era, meaning higher *ex ante* uncertainty about policy. Second, absent democratic mechanisms for regular transitions of power, bad policies could become entrenched for longer–for example, with long-lived dictators. The value of democracy's error correcting features rise as both uncertainty and the ease of lock-in rise.

4.5 Pluralism and progress

Another argument for democracy is the treatment of reasonable pluralism.[36] Modern society is rife with reasonable pluralism. Different groups have different moral values.

There are two arguments for why liberalism is valuable under reasonable pluralism. The first argument is that liberal democracy facilitates compromise between parties with conflicting values. Voting mechanisms allow different groups to be reflected in decisions. Deliberative bodies allow for explicit compromises to be formulated. Rights credibly protect minority interests from being swamped.

Compromise could be more important in the far future. The far future may contain a broader diversity of agents than at present. This could include cognitively enhanced humans, human emulations, and AIs with varying architectures. Such agents may have a wider range of goals. The range of possible agents expands the range of possible disagreement. In addition, new technologies in the future may raise the stakes of compromise, since the downside risks of conflict could be greater.

[33] Popper (1945); Przeworski (2010).
[34] Downs (1957).
[35] Knight and Johnson (2011).
[36] Rawls (1993).

Besides compromise, a further hope for liberal democracy is that it can generate moral progress, through the public contestation of values.[37] Here, some longtermists have raised grounds for doubt. We'll canvas two arguments: avoidance of moral truth, and (in the next section) value lock-in. These arguments claim that strategic actors seeking to avoid moral change in the long run could do so.

First, MacAskill (2025) worries about avoidance of moral truths, as a barrier to moral progress. Suppose that knowing the moral facts is intrinsically motivating. Even then, moral progress could be hard. This is because, insofar as the correct moral view diverges from people's current preferences, people will simply avoid learning moral truths.

But for this isolationist strategy to prevent moral convergence in the long run, it must be feasible without competitive cost. Consider: A moral visionary can avoid contaminating truths by retreating to the desert. But he may die there, extinguishing his moral lineage. A moral evangelist spreads his insights by preaching to anyone who will listen. But each conversation risks his being convinced by an unwelcome truth in turn. If open exchange tends toward truth rather than mere memetic fitness, the equilibrium favors progress.

For reasons like this, various political philosophers have argued that liberal democracy is the best system for producing moral progress. Examples include Habermas (1995), Dewey (1927), and Mill (1859). The details are different from one another. But the mechanisms rhyme. All emphasize that liberal institutions are designed to force citizens to debate, expose themselves to, and contest other value systems–at least if they want to get anything done.

4.6 Lock-in?

Longtermists have raised a more general worry about value pluralism and moral progress: value lock-in. In value lock-in, the far future gets permanently locked into a particular set of values early on. Why expect value lock-in? Finnveden, Riedel, and Shulman 2022 argue that various technical features of AGI will make it easier to preserve values over time, locking them in indefinitely.[38]

Here, we'll raise two points about value lock-in. First, we suggest that ultimately the likelihood of value lock-in is controlled by questions about institutional dynamics. Second, we suggest that value lock-in may not ultimately be relevant to assessing the best long-term institutions.

Let's take each point in turn. First, Finnveden, Riedel, and Shulman (FRS) 2022 explicitly glosses over institutional dynamics. In the paper's argument, lock-in depends explicitly on a single "dominant institution" with "uncontested military and economic dominance" having access to abundant AGI.[39] These are strong assumptions. No entity in the history of the world has been a "dominant institution," by the paper's definition.

[37] Habermas (1995), Mill (1859).
[38] Finnveden, Riedel, and Shulman (2022).
[39] Finnveden, Riedel, and Shulman (2022), §§0.3, 8.

We thus think the most important question about lock-in is whether and how such a "dominant institution" could emerge. That is, lock-in is not primarily a technical question, but one of institutional dynamics.

Here, we observe that liberal democracy's mechanisms for preserving reasonable pluralism are mechanisms for preventing such dominance: regular elections; governmental separations of powers; individual rights that balance the government's power against society's, and so on. This suggests that the most important interventions for avoiding lock-in may be the pluralistic governance mechanisms associated with liberal democracy.[40]

If AGI emerges in liberal democracies, we do not see an obvious path from AGI to a dominant institution. For example, FRS posit that "a large majority of the world's economic and military powers" might agree to set up an institution with "highly nuanced specifications of goals and values," and give it the "power to defend itself against external threats." (pp. 4-5) But this kind of permanent anti-plural consolidation is precisely what democratic systems promise participants will *not* happen. It is unclear why a broad swath of actors in a plural liberal system would accept such consolidation.

One possibility that the paper raises is bargaining: Perhaps groups that disagree on values could negotiate for some deal where each got a bit of what they wanted, and AIs could make the bargain credible by being given goals that reflected it. (pp. 14-15) Maybe. But liberal systems have many commitment mechanisms–contracts, elections, legislation. These mechanisms come in various strengths and lengths. AI-backed deals could be very strong and long, but it is not clear why bargainers would prefer this.

Our second question concerns the relevance of value lock-in to long-term institutional design. Our methodology is to assess institutions by canvassing the major arguments for liberalism. It is unclear to us how value lock-in would undermine these arguments. The clearest relevance of value lock-in is to questions about moral progress. But ultimately we are not convinced that moral progress itself is an especially load-bearing part of the defense of liberalism. Our general claim is that liberalism is the set of institutions with the highest expected value for promoting social welfare. But liberal institutions such as markets promote social welfare even when market participants are selfish and unethical. Questions about allocative and dynamic efficiency, for example, do not ultimately turn much on moral progress. In this way, value lock-in may be compatible with highest expected value futures, as long as the agents with those values are participating in efficient and democratically well-governed markets. Here, we disagree with Carlsmith 2025, who explicitly assumes that "good futures require at least some decent empowerment of agents optimizing for good values…[the] source of goodness ultimately needs to be people trying to make stuff good, or at least caring about good stuff." That said, we flag again that the arguments for allocative efficiency of markets connect social welfare to preference satisfaction. Rival theories of welfare instead appeal to achieving specific objective goods, such as pleasure, art, friendship, and so on.[41] For these theories of welfare, value lock-in might be

[40] Karnofsky (n.d.) surveys various factors relevant to the chance of lock-in, including life expectancy, population change, and natural disasters; but he does not include institutional factors in the analysis.

[41] For more on lock-in risk, see Christiano (2014) Hanson (2018) , Tomasik (2020) , chapter 4 of MacAskill, 2022, and a response by Hanson (2022).

more worrisome, since it could lead to systematic disconnects between market efficiency and objective social welfare.

4.7 Gradual disempowerment?

To close our discussion, we'll turn to a more general ground for pessimism: gradual disempowerment. The worry is that once AIs can outperform humans on most tasks, the demand for human labor will collapse. This could unmoor the economy, culture, and government from human interests. In this scenario, humans lose control over the future, and human wages may fall below subsistence.[42]

We worry that the analysis here devotes too little attention to the stabilizing institutional dynamics that will come into play if AI advances happen in liberal societies. To begin, voters do not want to starve, and democratic societies respond to such pressure with taxation and redistribution.[43] Such redistribution is likely if AI does begin to substantially displace human labor. In such a world, voter demand for redistribution will be widespread–not isolated to a small, economically-disadvantaged minority.

How far would tax and transfer go towards preventing gradual human disempowerment? The authors of "Gradual Disempowerment" (GD)–the most detailed treatment of the concern–address this directly. They say that taxation solves starvation, but perhaps little else. This is because "states funded mainly by taxes on AI profits instead of their citizens' labor will have little incentive to ensure citizens' representation."[44] Moreover, even with redistribution, humans' "role in economic decision-making would diminish. Markets might increasingly optimize for AI-driven activities rather than human preferences, as AI systems command a growing share of economic resources and make an increasing proportion of economic decisions."[45]

We are skeptical of several steps here. Begin with the point about markets optimizing away from humans' preferences. Redistribution allows the state to allocate any share of wealth to ordinary people. The larger the amount of redistribution, the larger the share of purchasing power those ordinary people have. They then remain the primary customers for the things the economy produces, including what it produces using AIs. Humans' *demand* signals would continue to shape the economy, even if most production is done by AIs.

Further, we think that GD's argument that "states funded mainly by AI taxes … will have little incentive to ensure citizens' representation" is mistaken. This argument is drawn from economic analyses of "rentier states," including Beblawai (1987). But Beblawi's argument specifically depends on the state being funded primarily via *external rents*–rents not produced by domestic productivity. (p. 384-85) "The existence of an internal rent, even substantial, is not sufficient." (p. 385) And even then, status as a rentier state requires that only a small share of the population is required to collect the external rents. The paradigmatic example is the Gulf oil state, where the

[42] Kulveit, Douglas et al. (2025, §§2–4); see also MacAskill (2025, §2.1, p. 6), Shulman and Bostrom (2021, p. 3) and Christiano (2019). Smith (2024) is another example.
[43] On post-tax median income in liberal democracies, see Luxembourg Income Study data summarized at Our World in Data (<https://ourworldindata.org/grapher/median-income-after-tax-lis>).
[44] Kulveit, Douglas et al. (2025), pp. 1–2.
[45] Kulveit, Douglas et al. (2025), §2.4.2.

state itself owns the oil, does not need taxation to capture its value, and only needs a few citizens to help it extract and sell the oil.

It is easy to see how such a state might stop caring much what its citizens thought. But rentier states are not the same as states that depend on high capital taxation. A world involving high AI productivity and high taxation on AI labor would be one where the state did depend on its citizens' cooperation. The AIs would be owned by citizens, on whom the government would depend to remit tax revenues. True, the tax here would be on capital, not labor, but that makes little difference. If the humans who owned AIs refused to pay, threatened to shut down production, and so on, the state would lose revenue, just as it would today if citizens went on strike.

Here, questions arise about the concentration of ownership of AI labor. Democratic institutional dynamics might be extremely different in worlds where: (1) there is one dominant AGI owned by government, (2) a handful of companies produce similar AGIs and compete fiercely, (3) many cheap near-frontier models can do most valuable work, such that the best AIs are broadly owned, and (4) billions of capable AI agents own themselves and sell their labor to buy energy and compute.

GD has little to say about who owns the AI, and why. But the distribution of ownership is partially endogenous to tax policy. In a scenario where AI is privately owned, and ordinary people receive substantial redistribution, AI will be more broadly owned. Then, the state would depend on a broad swath of citizens to fund its activities.

Indeed, in currently-liberal societies with private AI ownership, it is not even clear how much broad, versus narrow, ownership would matter for government responsiveness to citizens. An alternative model, drawn from selectorate theory, would be: in democracies, majority votes determine the leader and allow policy to be set. The leader must pay off the winning coalition to stay in power. In a world where, for example, all of the AI labor is owned by one privately-held company, it is politically easy to expropriate that company and redistribute broadly, or to regulate the company heavily. This reduces AI-led economic growth, but gives ordinary voters lots of public and private goods. And as long as most economic productivity is generated by a small minority of voters, this is a stable equilibrium within democracy.[46]

This again contrasts with the rentier states Beblawi analyzes, which own the rent-producing asset themselves. They do not need to win popular elections to raise revenues. In cases–like Norway–where a country has a large, quasi-nationalized, resource endowment, but depends on both democratic consent and private-sector help to extract it, government responsiveness remains high.

Finally, we are somewhat skeptical of GD's economic analysis of AI advances on the value of human labor. Certainly, advancements in AI capabilities *could* render human labor economically irrelevant. But the specific economic assumptions under which AI capabilities render human labor irrelevant deserve closer attention than they have received so far.

[46] Consider that, in liberal societies, marginal tax rates on the highest earners are durably high.

A closer economic analysis would reveal extremely complicated dynamics. If humans retain absolute advantages in any sectors, Baumol effects may cause those sectors to grow as a share of the economy, and wages may rise or fall.[47] This includes absolute advantages due to pure preferences for human labor–which may scale with wealth.[48] Insofar as the absolutely advantaged tasks complement AI labor, they are likely to be extremely valuable, but automation pressure will also rise.[49] Even if humans have no absolute advantages, dynamics of comparative advantage and specialization may produce a broad range of very highly paid jobs for humans.[50] Here, paradoxically, human wages tend to rise the more superhuman–and thus the more valuable–the AI labor is. This dynamic reverses if the scarce inputs binding AI labor at the margin are also necessary to support human life.[51] If the inputs to AI labor are slack, and there is plenty of high-value work to be done, then human work always adds to the stock of labor. Wages are set at the AIs' opportunity cost.[52] If the demand for high-value work dries up, then AIs' opportunity costs will fall, pushing human wages down. Alternatively, if the intelligence gap between humans and AIs becomes too vast, the high cost of human–AI communication may preclude the hiring of humans. On the other hand, this will supply an economic incentive for either humans or AIs to "uplift" humanity–enhancing cognition or other human capital in ways fit for the future economy.[53] At any rate, to model gradual disempowerment dynamics, we need to better understand all of the above, and more.

5. Faint-Hearted Liberalism

We've finished our analysis of how standard arguments for liberalism extend into the far future. All told, our view is that liberalism fares well. Most of the core assumptions supporting liberalism continue to apply under conditions of extreme growth, extreme population size, structural transformation to the economy, space colonization, and AI.

Do existing longtermists disagree? Not always. For example, MacAskill's (2025) "How to Make the Future Better" contains many concrete, liberal proposals for promoting long run flourishing. Among them: guard against AI-enabled coups and autocracy, empower citizens vis a vis the government by distributing access to superintelligent AI advisors, extend the rule of law to space, and build new AI-specific agreements between nations with aligned interests.

Still, before finishing, we will flag several cases where existing longtermist proposals are *faint-heartedly* liberal. By this, we have in mind proposals, pitched at a high level, that sound liberal in the abstract, but that risk illiberalism when concretely implemented. Relatedly, we worry that some of these proposals fail to engage with the underlying institutional dynamics that justify liberalism. For reasons of space, we'll focus on two examples: viatopia, and the long reflection.

---

[47] See Baumol (1967).
[48] See Imas (2026).
[49] See Acemoglu and Restrepo (2018).
[50] See Smith (2024); Trammell and Korinek (2023).
[51] See Korinek and Suh (2024).
[52] See Acemoglu and Restrepo (2022).
[53] Cf. Bostrom (2014), ch. 2; Hanson (2016), ch. 27.

5.1 Viatopia

In "Viatopia," MacAskill (2026) presents a vision for how to make the future go better: "viatopia".[54] Viatopianism says that we should not promote the best *ultimate* societal outcomes. Instead, we should target some *intermediate* state that "achiev[es] whatever society needs to steer itself towards a truly wonderful outcome." More formally, MacAskill defines viatopia as "a society whose expected value is at least 50% that of a guarantee of a best feasible outcome."[55]

How do we achieve viatopia? Here, MacAskill proposes outcomes that broadly fit with liberalism, including material abundance, scientific knowledge, coordination to avoid war, trade, low levels of catastrophic risk, preserving optionality, and improving collective decision-making. The full details of viatopia are left as an open research question.

Nonetheless, we are unsure how viatopian thinking fits with our own liberal-institutional approach. Possibly, long-term liberalism simply is one viatopian vision. Liberal institutions run on positive-sum mechanisms that operate at the margin. They therefore do not depend on the kind of utopian end-state reasoning that viatopia criticizes.

But on the other hand, the arguments for liberalism make no reference to specific value thresholds. They certainly do not guarantee a formal viatopia. To put things concretely, suppose that a society can follow two possible strategies. The first is feasible and illiberal. It involves a small group of motivated individuals with good morals seizing power, ending democracy, and steering a society indefinitely. If successful, this strategy produces a formal viatopia at some point in the future–a society with an expected value of 50% of the physically possible maximum. The second strategy is liberal, preserving recognizable markets and democracy for the long run. If successful, the liberal strategy produces a society within an expected value of just 45% of the maximum–a formal non-viatopia. Suppose that the illiberal strategy is only 20% likely to succeed, while the liberal strategy is 100% likely to succeed.

Here, illiberal strategy achieves, in expectation, only 10% of physically possible value, while the second achieves 45%. But the illiberal strategy is the only one that offers any chance of a formal viatopia. We are not certain which approach viatopianism recommends. Nor are we sure whether the answer would change if the liberal approach offered only 15% of the maximum, but with 100% certainty.

Our own strategy thus diverges from viatopianism because it does not aim at any particular value threshold on any particular timescale. We are simply interested in which set of institutions is the most likely to produce the most value. Above, we've suggested that liberalism is the answer. That is, liberal institutions will, on average, produce better outcomes over time than the alternatives.

Possibly, liberalism will only achieve 10% of the value that could be physically realized in the universe. But if no other institutions would produce even that much, then liberalism is still optimal in our view.

---

[54] See MacAskill (2026).
[55] MacAskill (2026, n. 2).

5.2 The long reflection

Toby Ord (2020) and Will MacAskill (2022) have argued that one way to avoid bad futures and aim toward the best ones is by engaging in a societal "long reflection." This would be a period where society is "safe from calamity and can reflect on and debate the nature of the good life, working out what the more flourishing society would be."[56] MacAskill notes that "it would be worth spending many centuries [reflecting] to ensure that we've really figured things out before taking irreversible actions like locking in values or spreading across the stars."

At first glance, the long reflection may sound perfectly consistent with liberal values. On further inspection, however, we worry that it could be illiberal. The problem is that liberalism systematically leads to economic growth, which in turn causes radical change. For example, even at a 3% growth rate, GDP would be roughly 360 times larger in 200 years from today. Thus, a 200 year long reflection could only be spent *preparing* for such change if it first *blocked* such change.[57] That is, if it blocked even modest growth.

In this way, we wonder what policies would be required to implement a long reflection. For example, would a long reflection require bans on space colonization and AI development? Regarding space colonization specifically, we argued above (in 3.2.1) that the best route to managing space colonization involves doling out property rights, and allowing markets to sort out the optimal distribution of resources. But we wonder if this would produce levels of growth incompatible with a long reflection.

On the other hand, another interpretation of the long reflection is as endogenous to liberalism. Liberal society has already been engaged in a centuries-long reflection. Reflection on what is valuable, and continuous revision of the conclusion, are a core component of both democratic and market processes. But if we are already in the long reflection, we are unsure which fundamentally *new* processes its proponents are proposing.

Finally, yet another version of the long reflection would be deference or "handoff" to artificial superintelligence. Here, the idea would be to design an AI that can engage in reflection at much vaster time scales than humanity. The entire reflection could be done in an hour, and yet could involve millions of years of human computation.

But what happens if the artificial superintelligence issues verdicts that are vastly different from those accepted by humanity? Proponents of the long reflection face a choice. On the one hand, there is an illiberal option: society should be forced to obey the edicts of the AI long reflection. On the other hand, there is the liberal option: the AI reflector gets to propose its policies and defend them in the commons, but has no special institutional ability to implement its preferred ideals. We are unsure whether the liberal version of the policy would have a large effect, because the AI reflector might struggle to convince humanity of its idiosyncratic ideals.

6. Designing Long-Term Liberalism

[56] MacAskill (2022).
[57] See Karnofsky (n.d.).

Implementing long-term liberalism will require reforming today's institutions to meet the future's challenges. For reasons of space, we close with a case study. Here is one challenge for long-term liberalism: how will we incorporate AI agents into markets and democracy? Will the fact that AIs can reproduce more quickly and cheaply than humans break basic liberal mechanisms? It depends on the details.

Begin with markets. Suppose that AIs in the future are highly aligned to particular humans–perhaps the ones who deploy them. Here, richer people will effectively be able to deploy more labor. Markets generally scale quite well with human agency. Today, a single algorithmically-assisted trader can direct millions of stock transactions per day. This was impossible in the past. At a minimum, if an AI that is perfectly aligned to some human acts badly, the human can be sanctioned. Here, the incentive passes through to affect AI behavior because of the AI's alignment to the human's interests.

But what if the AIs participating in the marketplace have their own private goals, rather than acting as mere extensions of some human's will? In this case, we have argued elsewhere, bigger legal innovations will be needed. To incentivize such agents to act well, they should be given legal rights and duties.[58] This will require a broad suite of legal and market innovations to: individuate the relevant rightsholders; police them for illegal, careless, or antisocial behavior; adjudicate their disputes; protect them against abuse; and more.[59] A market economy adapted to serve both human and AI consumers may look very different from today's market. But the changes will mostly be at the level of implementation. AIs' presence as market participants do not threaten the fundamental mechanisms that make markets, or the laws governing them, work.

Democracy is a different story. In human democracies, most individuals get a vote. But producing more humans is costly, and parents have little control over their children's fine-grained political preferences. Producing AIs is comparatively cheap. And AIs' preferences can be more readily tuned to their creators' wishes. This raises two problems. First, in a one-AI-one-vote regime, AI votes could quickly swamp humans. Second, any human or AI might be able to produce many AI voters strongly aligned to their own preferences. Then, whoever could buy the most compute would get the most votes. In the limit, that means buying elections directly, which would, for example, break the mechanisms by which democracy generates credibility and produces public goods.

This is a variation on the tyranny of the majority. There are at least two methods for preventing a tyrannical majority. The first is to change the voting rules to balance majoritarian threats against minority interests. The United States Senate is an example. The second option is to place certain decisions beyond the scope of ordinary lawmaking. The U.S. Constitution's Bill of Rights provides many examples.

Similar designs could mitigate the democratic problems of AI reproduction. AIs designed to do the bidding of specific human principals might be excluded from the franchise, since their votes would duplicate those of their principals. If AIs had private preferences, their votes could be mediated in various ways. AIs' aggregate vote share could be limited to some fraction of the total. Or law could impose limits on AI reproduction. Or it could instead institute a requirement

[58] See Salib and Goldstein 2024, Goldstein and Salib 2025.
[59] See Arbel et al 2026.

that reproduction involve randomized ideological variation, or the installation of basic liberal values. Many of these rules would likely need to be constitutionalized–entrenched, rather than changeable via ordinary majoritarian lawmaking. If vote gaming is costly, the outcomes uncertain, and the endgame liable to destroy the shared fruits of democracy, all parties should prefer rules that prevent it (Buchanan & Tullock 1962.)

Which of these approaches would be best depends on further facts not in evidence. Among them, do the AIs have normatively relevant wellbeing? And how does it relate to the wellbeing of various humans?[60] [61]

References

Acemoglu, D. and Robinson, J. A. (2006). *Economic Origins of Dictatorship and Democracy.* Cambridge University Press.

Acemoglu, D. and Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica* 90(5): 1973–2016.

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review* 108(6): 1488–1542.

Acemoglu, D. and Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* Crown Business.

Aghion, P. and Howitt, P. (1998). *Endogenous Growth Theory.* MIT Press.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.

Arbel, Yonathan A. and Goldstein, Simon and Salib, Peter, How to Count AIs: Individuation and Liability for AI Agents (February 01, 2026). Available at SSRN: https://ssrn.com/abstract=6273198 or http://dx.doi.org/10.2139/ssrn.6273198

Arrow, K. J. (1951). An extension of the basic theorems of classical welfare economics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

Aschenbrenner, L. (2024). *Situational Awareness: The Decade Ahead.* Self-published online monograph. <https://situational-awareness.ai/>.

Austen-Smith, D. and Banks, J. S. (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review* 90(1): 34–45.

[60] See Goldstein and Kirk-Giannini forthcoming.
[61] The authors used Claude Opus 4.7 for assistance while writing the paper.

Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review* 57(3): 415–426.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review* 110(4): 1104–1144.

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

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy* 10(4): 455–476.

Buchanan, J. M. and Tullock, G. (1962). *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press.

Bueno de Mesquita, B., Smith, A., Siverson, R. M., and Morrow, J. D. (2003). *The Logic of Political Survival.* MIT Press.

Carlsmith, J. (2025). Video and transcript of talk on "Can goodness compete?" Joe Carlsmith's Substack. https://joecarlsmith.substack.com/p/video-and-transcript-of-talk-on-can.

Christiano, P. (2019). What failure looks like. AI Alignment Forum. <https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Condorcet, M. de. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie Royale.

Cowen, T. (2018). *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals.* Stripe Press.

Davidson, T., MacAskill, W., and Taylor, M. (2026). Moral public goods are a big deal for whether we get a good future. Forethought. <https://www.forethought.org/research/moral-public-goods-are-a-big-deal-for-whether-we-get-a-good-future>.

Debreu, G. (1959). *Theory of Value: An Axiomatic Analysis of Economic Equilibrium.* John Wiley & Sons.

Demsetz, H. (1967). Toward a theory of property rights. *American Economic Review* 57(2): 347–359.

Dewey, J. (1927). *The Public and Its Problems*.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25(2): 383–417.

Finnveden, L., Riedel, J., and Shulman, C. (2022). AGI and Lock-In. Forethought. <https://www.forethought.org/research/agi-and-lock-in>.

Goldstein, Simon & Kirk-Giannini, Cameron Domenico (forthcoming). AI Welfare: Agency, Consciousness, Sentience. New York: Oxford University Press.

Goldstein, Simon and Salib, Peter, AI Rights for Economic Flourishing (July 15, 2025). Available at SSRN: https://ssrn.com/abstract=5353214 or http://dx.doi.org/10.2139/ssrn.5353214

Habermas, J. Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism. *Journal of Philosophy* 92(3) (1995): 109–131.

Hanson, R. (2013). Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy* 21(2): 151–178.

Hanson, R. (2016). *The Age of Em: Work, Love, and Life When Robots Rule the Earth.* Oxford University Press.

Harberger, A. C. (1954). Monopoly and resource allocation. *American Economic Review* 44(2): 77–87.

Hardin, G. (1968). The tragedy of the commons. *Science* 162(3859): 1243–1248.

Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review* 35(4): 519–530.

Hayek, F. A. (1968). Competition as a discovery procedure. In *New Studies in Philosophy, Politics, Economics and the History of Ideas* (1978), Routledge, 179–190.

Heller, M. A. (1998). The tragedy of the anticommons: Property in the transition from Marx to markets. *Harvard Law Review* 111(3): 621–688.

Hong, L. and Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101(46): 16385–16389.

Imas, A. (2026). What will be scarce? *Ghosts of Electricity* (Substack), 14 April 2026. <https://aleximas.substack.com/p/what-will-be-scarce>.

Karnofsky, H. (n.d.). Cold Takes. <https://www.cold-takes.com/>.

Knight, J. and Johnson, J. (2011). *The Priority of Democracy: Political Consequences of Pragmatism.* Princeton University Press.

Korinek, A. and Suh, D. (2024). Scenarios for the transition to AGI. NBER Working Paper No. 32255. <https://www.nber.org/papers/w32255>.

Krugman, P. (1978/2010). The theory of interstellar trade. *Economic Inquiry* 48(4): 1119–1123. (Originally circulated 1978.)

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. (2025). Gradual Disempowerment. arXiv preprint. <https://arxiv.org/abs/2501.16946>.

Landemore, H. (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many.* Princeton University Press.

Lavoie, D. (1985). *Rivalry and Central Planning: The Socialist Calculation Debate Reconsidered.* Cambridge University Press.

List, C. and Goodin, R. E. (2001). Epistemic democracy: Generalizing the Condorcet jury theorem. *Journal of Political Philosophy* 9(3): 277–306.

MacAskill, W. (2022). *What We Owe the Future.* Basic Books.

MacAskill, W. (2025). How to Make the Future Better. Forethought. <https://www.forethought.org/research/how-to-make-the-future-better>.

MacAskill, W. (2026). Viatopia. Forethought. <https://www.forethought.org/research/viatopia>.

MacAskill, W. and Moorhouse, F. (2025a). No Easy Eutopia. Forethought. <https://www.forethought.org/research/no-easy-eutopia>.

MacAskill, W. and Moorhouse, F. (2025b). Convergence and Compromise. Forethought. <https://www.forethought.org/research/convergence-and-compromise>.

Martin, I. W. R. and Nagel, S. (2022). Market efficiency in the age of big data. *Journal of Financial Economics* 145(1): 154–177.

Marx, K. (1867). *Capital: A Critique of Political Economy, Volume I.* (English edition, Penguin, 1976.)

Mill, J.S. *On Liberty* (1859).

Mokyr, J. (2009). *The Enlightened Economy: An Economic History of Britain 1700–1850.* Yale University Press.

Nielsen, M. (2026). Michael Nielsen – How science actually progresses [Interview by D. Patel]. Dwarkesh Podcast.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.

Pigou, A. C. (1920). *The Economics of Welfare.* Macmillan.

Popper, K. R. (1945). *The Open Society and Its Enemies.* Routledge.

Przeworski, A. (2010). *Democracy and the Limits of Self-Government.* Cambridge University Press.

Rawls, J. (1993). *Political Liberalism.* Columbia University Press.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy* 98(5): S71–S102.

Salib, Peter and Goldstein, Simon, AI Rights for Human Safety (August 01, 2024). Virginia Law Review, Available at SSRN: https://ssrn.com/abstract=4913167

Schumpeter, J. A. (1942). *Capitalism, Socialism, and Democracy.* Harper & Brothers.

Shahidi, P., Rusak, G., Manning, B. S., Fradkin, A., and Horton, J. J. (2025). The Coasean singularity? Demand, supply, and market design with AI agents. In *The Economics of Transformative AI,* National Bureau of Economic Research. <https://www.nber.org/books-and-chapters/economics-transformative-ai/coasean-singularity-demand-supply-and-market-design-ai-agents>.

Shulman, C. and Bostrom, N. (2021). Sharing the world with digital minds. In Clarke, S., Zohny, H., and Savulescu, J. (eds.), *Rethinking Moral Status,* Oxford University Press, 306–326.

Smith, N. (2024). Plentiful, high-paying jobs in the age of AI. Noahpinion. <https://www.noahpinion.blog/p/plentiful-high-paying-jobs-in-the>.

Tirole, J. (1988). *The Theory of Industrial Organization.* MIT Press.

Trammell, P. and Korinek, A. (2023). Economic growth under transformative AI. NBER Working Paper No. 31815. <https://www.nber.org/papers/w31815>.

Wolfers, J. and Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives* 18(2): 107–126.