AI and Ethics (2024) 4:315–324
https://doi.org/10.1007/s43681-023-00268-7

ORIGINAL RESEARCH

Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence

Simon Friederich[1]

Received: 18 December 2022 / Accepted: 10 February 2023 / Published online: 16 March 2023

Abstract
A transition to a world with artificial general intelligence (AGI) may occur within the next few decades. This transition may give rise to catastrophic risks from *misaligned* AGI, which have received a significant amount of attention, deservedly. Here I argue that AGI systems that are *intent-aligned*—they always try to do what their operators want them to do—would also create catastrophic risks, mainly due to the power that they concentrate on their operators. With time, that power would almost certainly be catastrophically exploited, potentially resulting in human extinction or permanent dystopia. I suggest that liberal democracies, if they decide to allow the development of AGI, may react to this threat by letting AGI take shape as an *intergenerational social project*, resulting in an arrangement where AGI is not intent-aligned but *symbiotic* with humans. I provide some tentative ideas on what the resulting arrangement may look like and consider what speaks for and what against aiming for intent-aligned AGI as an intermediate step.

Keywords Artificial general intelligence · Liberal democracy · Alignment · Stable totalitarianism

1 Introduction

The development of artificial intelligence that is superior to human intelligence in almost all conceivable respects and general in scope may take place within the next few decades [1, 2]. A growing number of companies are working on the explicit goal of developing such *artificial general intelligence* (AGI, see [3] for an overview of companies). If and when they succeed the transition to a world with AGI ("AGI transition" in what follows) occurs, this will plausibly be one of the most momentous changes in history, comparable in significance to the agricultural, scientific, and industrial revolutions, perhaps even surpassing them. How the AGI transition will play out, if it occurs—notably, its key events, overall duration, and outcomes—is extremely difficult to foresee because there are no obvious precedents. Some ways in which the AGI transition might occur are catastrophic for humanity, others may well lead to a future with humans flourishing more than at any previous point in history. Here I outline ideas on how the citizens of liberal democracies, if they decide to let the AGI transition happen, might shape that transition to make it, from their perspective, *good*. I assume that, broadly speaking, a good AGI transition from the perspective of liberal democracies is one that results in an arrangement where AGI systems not only help cover human basic needs and contribute to enhancing human welfare and flourishing, but also respect human and civil rights, and integrate well with democratic structures. My central thesis, advocated here tentatively and with some trepidation, is that a helpful strategic goal for liberal democracies might be to become *symbiotic* with *unaligned* AGI developed as an *intergenerational social project*.

I clarify what I mean by "unaligned" in Sect. 2, where I also say a few words about what counts as "AGI" in the sense of this paper. Next, in Sect. 3, I situate the present work with respect to the literature on risks from catastrophically misaligned AGI. Having prepared the ground, Sect. 4 embarks on the argument proper of this paper, outlining why and how aligned AGI poses catastrophic risks, mostly related to power concentration. In Sect. 5, I consider ideas for how liberal democracies might mitigate these risks while keeping AGI aligned and I end up finding none of them very promising. As an alternative, I suggest in Sect. 6 that liberal democracies, if they decide to let the development of AGI occur, might strive to develop unaligned symbiotic AGI as

✉ Simon Friederich
s.m.friederich@rug.nl

[1] University of Groningen, University College Groningen, Hoendiepskade 23/24, 9718BG Groningen, The Netherlands

an intergenerational project to prevent problematic power concentration. In Sect. 7, I provide some tentative ideas on what the resulting arrangement may look like using institutions such as academia, an energy system, and a constitutional court as analogies. Sect. 8 considers what speaks for and what against aiming for aligned AGI as an intermediate step. Finally, in Sect. 9, I provide some reasons why independent forces may work towards unaligned symbiotic AGI and may make it a reality even if relatively few actors actually envisage it as a strategic goal.

2 AGI and alignment—what are we talking about?

In this section I give a rough characterization of how I will use the terms "AGI" and "alignment."

I do not rely on any specific definition of AGI. In fact, the arguments presented here are compatible with a variety of characterizations of "AGI". Notably, the present discussion is meant to be neutral about whether AGI will be constructed as a single generally intelligent agent or as a "collective" phenomenon that emerges at the societal level from the interplay of different AI systems that are not individually generally intelligent. What does matter for the present discussion is that it assumes AGI to have an important role in shaping power relations. Accordingly, the arguments presented here should be read with a characterization of AGI in mind according to which it is plausible that AGI—if it is ever developed—will have such a role. Characterizations of AGI that include what Karnofsky [4] dubs "PASTA" ("Process for Automating Scientific and Technological Advancement") are good candidates. It seems plausible that differential access to systems that autonomously achieve scientific and technological breakthroughs will dramatically shape economic and political power relations.

The challenge of transitioning to a good world with AGI is sometimes framed as that of creating *aligned* AGI or "solving the alignment problem" for AGI. Brian Christian, author of *The Alignment Problem*, characterizes the alignment problem for AI in general—not just AGI—as the challenge of creating AI which "capture[s] our norms and values, understands what we mean or intend, and above all, do[es] what we want" [5]. This characterization gives some orientation about what is usually meant by "alignment", but it is very broad.

A somewhat more precise definition of alignment, echoing the last part of Christian's, which seems to capture what is pursued by those who actually work in the field of AI alignment is "intent alignment." AI systems are intent-aligned if and only if, as alignment researcher Paul Christiano [6] puts it, they "are trying to do what you want them to do," where "you" are the operators of the AI systems. In the same vein, alignment researchers Leike et al. [7] characterize the alignment problem as the challenge: "[H]ow can we create agents that behave in accordance with the user's intentions?" Intent alignment can be thought of as consisting of two complementary components [6, 8]: outer alignment—the AI pursues an objective that really incentivizes the behaviour intended by the operator—and inner alignment—the policies that the AI has learned to achieve its objective in a training environment transfer successfully to the deployment environment.

In what follows, I use "alignment" in the sense of "intent alignment" because, first, this use of "alignment" fits well with how the term "alignment" is otherwise used in ordinary discourse outside of its application to AI and because, second, as said, this corresponds to how "alignment" is actually used by those working on AI alignment.[1] Christiano acknowledges that making AGI (intent) aligned is not sufficient for a good AGI transition – notably, the AGI must also function reliably and be capable of actually understanding human intentions. However, Christiano seems to see achieving alignment as necessary for a good AGI transition in that it "might be the minimum you'd want out of your AI" [6].

However, with "alignment" understood as "intent alignment", it is not at all obvious whether achieving AGI alignment is really necessary for achieving a good AGI transition. To recall, a good AGI transition, for the purposes of this paper, is one that results in an arrangement where AGI systems, among other things, respect human and civil rights, and integrate well within democratic structures. It is not at all clear why, in *addition*, those systems should in all conditions try to do what their operators want them to do, as required for alignment.[2] In fact, as I argue in later sections, the citizens of liberal democracies may well maximize their chances at a, from their perspective, good AGI transition if they aim for AGI that is—in the right way—*unaligned*.

[1] The notion of intent alignment can be further disambiguated—for instance, it can be clarified whether the intentions that count are explicit instructions, explicit intentions, or revealed preferences (Gabriel, Sect. 3) – but for the purposes of this paper these distinctions do not matter greatly.

[2] One may argue that, in such a situation, even if AGI systems may not be intent-aligned, they are nevertheless "value-aligned". Focusing on "value alignment" is advocated by Iason 9, Sect. 3), and the arguments presented in the present paper could be seen as an attempt to explore what value alignment looks like. My main reason to not adopt the terminology of "value alignment" is that talk of "values" may be not as conducive to solving the challenges discussed in later sections of this paper as talk of, say, "rights" and "laws".

3 Catastrophic risk from misaligned AGI

AGI that is unaligned in the right way contrasts sharply with *catastrophically misaligned* AGI. Catastrophically misaligned AGI is plausibly one of the largest global catastrophic risks that humanity may face in the next few decades, perhaps the largest. The worry can be traced back to Wiener [10], and the argument is forcefully made by Yudkowsky [11], Bostrom [12], Russell [13], Ngo [8], Cotra [1], Carlsmith [15], Cohen et al. [16], Karnofsky [17], and many others. In a nutshell, the fundamental worry is that there will be incentives to develop goal-directed autonomous AGI agents, that those agents' ultimate goals will at some point turn out to be in conflict with complex human norms and values, and that those agents, using their superior intelligence, will either take control of human affairs, creating a—from the human point of view—dystopian state of affairs with no escape, or simply kill off all humans. (See [18] for a systematic classification of different ways in which a catastrophe resulting from AGI misalignment could play out.)

Those who develop AGI will plausibly try to design it such that it follows their intentions. They are thus intrinsically motivated to strive for alignment and, a fortiori, intrinsically motivated to avoid catastrophic misalignment. There are thus strong incentives for those trying to develop AGI to prevent it from being catastrophically misaligned. However, frontrunners in the development of AGI may create catastrophically misaligned AGI by accident, even though this is against their own best interest, because they may believe—correctly or wrongly—that they are in a race with (even less scrupulous) competitors and must therefore deprioritize safety for the sake of speed [15], Sect. 5.3.2).

4 Catastrophic risk from aligned AGI

Threats from developments in AI to the rights-based order of liberal democracies are widely discussed (e.g. Coeckelbergh in press), including ones that arise from the intentional (mis-) use of AGI (e.g. [13], Ch. 4). Kate Crawford goes as far as saying that existing AI systems, across the board, "are designed to discriminate, to amplify hierarchies, and to encode narrow classifications" (Crawford 2021, p. 211). Even though this sentiment does not seem to be universally shared, academics unfamiliar with the case for AGI-driven existential risk commonly seem to be more concerned about intentional than unintentional harm from AI (Hobbhahn 2022).

However, it does not seem to be widely appreciated and made explicit that worries about harm from intentional AGI use should make us particularly concerned about *aligned* AGI. A completely aligned AGI, by definition, tries to do what its operators want, whatever that is. But because such an AGI is cognitively far more advanced than any human and because such cognitive skills convey great power, it plausibly conveys great power on its operator(s). Agents with a monopoly, or near-monopoly, of aligned, or near-aligned, AGI, may well have power that is far superior to that provided by any technology today, including the most advanced contemporary surveillance technology or, for that matter, nuclear weapons.

There are at least three types of catastrophic scenarios that could result from the misuse of aligned AGI (Alignment arguably does not have to be perfect at any stage for these to be realistic concerns): military scenarios, totalitarian scenarios, and scenario resulting in AGI in control and/or catastrophically misaligned after all. Which of these would become most urgent is extremely difficult to predict because they all take place in a world with much more advanced technological boundaries and a radically different power structure from today.

4.1 Military scenarios

AI systems have powerful military applications already today [19]. Aligned AGI can plausibly be deployed as a weapon that is far more versatile than any weapon today and potentially far more powerful than even a large-scale arsenal of nuclear weapons because it can be used in a more targeted manner. And it may well be possible to use aligned AGI to—indirectly—wield the same destructive force as a nuclear weapons arsenal, for instance by manipulating, circumventing, or displacing those who at present control nuclear weapons.

4.2 Totalitarian scenarios

These are scenarios where aligned AGI is used by its operator(s) to establish stable, "sustainable", totalitarianism [20], with the AGI operator in charge as a dictator (or with a group of dictators). Aligned AGI could help such a dictator to eliminate threats and limits to their grip on power that today's AI systems do not yet allow authoritarian rulers to eliminate [21]. Surveillance and other forms of automated citizen control enabled by AGI could eliminate internal challenges. Military superiority enabled by AGI could eliminate external challenges and/or even create a road to world government with the AGI operator in charge as a global human dictator. Conceivably—though admittedly speculatively—the dictator could use AGI-enabled life extension research to dramatically increase their lifespan and thereby mitigate the problem of stability that dictatorships face when there are several candidate successors.

4.3 Scenarios resulting in AGI in control and/or catastrophically misaligned after all

These are scenarios where AGI starts out aligned and ends up catastrophically misaligned after all. This could happen, for instance, if aligned AGI is initially used by some dictator or narrow elite as a tool of power consolidation and subsequently given large autonomy to handle internal and external challenges more efficiently than the dictator themselves is able to. At some point, the dictator—either voluntarily or involuntarily—may irrevocably transfer most of their power to the AGI, resulting in a stable dystopian state of affairs with AGI in control after all.

Individually, these scenarios are extremely speculative, and my point is not that any specific version of them is particularly likely. My main point is that, *if* aligned AGI is developed, *some* very serious kind of misuse with enduring catastrophic consequences at a global scale is probable, perhaps inevitable, in time. Even if the initial operators of aligned AGI use it benevolently and beneficially, say, to stimulate economic growth in developing countries, drive back poverty and address global problems such as climate change and risks from pandemics, such luck is almost sure to run out at some point, for instance because the intentions of the AGI-operators change ("power corruption") or because there are new operators. Aligned AGI may offer power-hungry agents the tools that they desire to expand and consolidate their power even further, eliminating whichever factors still limit it in time and space, whether those are the mechanisms of rule-based democratic order in the US, the military forces that currently keep Russian imperialism at least somewhat in check, the employment and sexual harassment laws that check the impulses of CEOs, or whatever else.

It is instructive to compare the risks from catastrophically misaligned AGI with those from aligned AGI using Bostrom's [12] distinction between "state risks" associated with rather stable states of affairs and "transition risks" that arise from the transition between states.

Misaligned AGI predominantly creates a transition risk—the risk might initially be very high, but it goes to (near-) zero if and when it is understood how one develops intent-aligned AGI and succeeds in implementing this understanding. Aligned AGI, in contrast, predominantly creates a state risk—its very existence generates the permanent threat of catastrophic superintelligence-enhanced power abuse.

5 Addressing the threat from aligned AGI

As far as the dangers of military use are concerned, there might be ways for humanity to reduce the catastrophic risks from aligned AGI to levels no higher than those from current technology such as biotechnology or nuclear technology. For instance, the risks of military scenarios or stable global totalitarianism might be mitigated by moving to what Bostrom ([12], ch. 11) calls a "multipolar scenario" where there are several operators of aligned AGI globally who keep each other in check. Some of those operators might establish local AGI-based totalitarianism, but others could pursue different paths and impose external, and perhaps to some extent internal, limits to the dictator's power.

It is not clear, however, that multipolar scenarios post AGI-transition can be stable at all ([12], pp. 216–225, gives reasons for doubt) and they may actually come with heightened, not lower, risks of military use of AGI, as persuasively argued by Carayannis and Draper [22]. Notably, it seems unlikely that an arrangement of deterrence could be established that effectively bans any military use of AGI, similar to how nuclear weapons use is currently avoided. Even nuclear deterrence is fragile and its relative effectiveness reflects the specific offense-defense balance of nuclear weapons. Unlike AGI military use, nuclear weapons use is a rather clear-cut matter, involving a clear-cut boundary that is crossed. No such boundary seems likely for catastrophic hostile AGI use which could utilize a range of covert, deniable, grey zone tactics with unprecedented effectiveness.

Even if the threat of catastrophic AGI misuse for military purposes could be averted, the threat to liberal democracy from power concentration would remain. Power concentration enabled by AI poses serious challenges to liberal democracies already today [23]. The considerations in the previous section suggest that these challenges will become much more dramatic if and when aligned (or near-aligned) AGI is developed.

A drastic reaction that liberal democracies might contemplate in response to the combined risks from misaligned and aligned AGI is to prohibit any further steps towards the development of AGI, either permanently (as deliberated by Cremer and Kemp [24], p. 11) or for the foreseeable future until the landscape of technological achievements has completely changed ("differential technological development", [12], Ch. 14). One may see this as the genuinely precautionary approach to AGI in light of the combined risks from catastrophically misaligned AGI and power-concentrating aligned AGI.

However, liberal democracies, *if* they consider prohibiting the further development of AGI, should also be aware of the downsides of such an approach: its main problems are, first, that it would be very difficult to draw a meaningful line between AGI-related developments that are banned and other developments in AI that are permitted, second, that it would require intrusive and hard-to-implement measures to actually enforce the ban, and, third, that developers of AGI based elsewhere in the world would not be hampered in their attempts to develop AGI. In the longer term, liberal democracies may well diminish their global weight and influence if they ban the development of AGI internally and thereby end up undermining their ability to shape the—perhaps at some point inevitable—development of AGI. Thus implementing a ban on AGI development could (but need not) end up aggravating the very risks from AGI that the ban would be meant to mitigate.

A less radical and perhaps more feasible approach to mitigating the risks from aligned AGI and power concentration might be to permit the development of AGI but strongly regulate who has access to it and for which purpose, similar to how access to weapons or sensitive information is currently regulated. Notably, access to the power-enhancing aspects of AGI systems could be confined to elected political leaders and be constrained by various norms as to how that power can be used and when and how it needs to be transferred.

This approach seems in line with established best practices for governing powerful technologies in liberal democracies, but it will remain vulnerable as long as AGI systems are aligned with their operators, in this case the political leaders. Aligned AGI systems, by definition, try to do what their operators want them to do, so if some political leader decided to ignore the prescribed constraints on their access to AGI systems, those systems themselves would not offer any inherent resistance. Checks to their AGI-enhanced power would have to come from other humans. However, other humans may not be able to enforce such checks as long as political leaders' power is enhanced by AGI.

AGI aligned with a political leader, even if norms that constrain its deployment are in place, can be compared to a police force or army that prioritizes conforming to the leader's intentions over conforming to those norms. It remains an unparalleled risk to democratic and rights-based order even if its use is officially highly regulated.

In the following section, I suggest that liberal democracies, if they decide to allow the development of AGI but want to mitigate the risks from permanent power-concentration that it creates, may want to use AGI systems' superior intelligence as a *resource* to make these systems inherently resilient against monopolization by power-seeking individuals. In other words, I will argue that what liberal democracies may end up choosing, if they choose wisely, is AGI that is—in the right way—structurally unaligned.

6 Symbiosis with AGI as an *intergenerational social project*

By a good AGI transition, to recapitulate, I mean one that results in an arrangement where AGI systems help cover humans basic needs, contribute to enhancing human welfare and flourishing, and at the same time respect human and civil rights. There is no independent reason to think that, *in addition*, these systems should try to fulfil the intentions of specific humans, those who happen to operate them. In fact, it seems independently plausible that AGI systems are best positioned to impartially respect human rights and further human welfare if they are somewhat autonomous and detached from the goals and preferences of specific individuals, i.e. if they are unaligned.[3] I propose to call an arrangement where AGI systems are integrated robustly and permanently—across generations—within human society without being tied to the interests of specific individuals, an arrangement with AGI as an "intergenerational social project."

If AGI systems are to be designed in such a way that, once deployed, they resist being taken over by specific individuals and ensure that the same holds for newly developed AGI systems, they will presumably need to have goals and preferences that equip them with some degree of autonomy and resilience with respect to takeover by individuals. To the extent that they will indeed have such goals and preferences, the resulting arrangement of humans coexisting with AGI developed as an intergenerational social project might be characterized as a—two way-beneficial—*symbiosis* (where one of the parties involved—namely, the AGI systems—is no "bios"): We humans broadly fulfil the unaligned AGIs' goals and preferences (see below for some more thoughts on those), and the AGI systems, in turn, contribute to human welfare and flourishing while resisting any takeover attempts by power-seeking humans.

Those who prefer to use "alignment" in a broader sense rather than as "intent alignment" may see such a symbiotic arrangement as one where alignment has in fact been achieved. But, unlike the symbiotic arrangement suggested here, "alignment" in connection with AI is usually depicted as a highly asymmetric relation with one side, the aligner, in control, and the other side, the aligned, as subordinate. The highly asymmetric notion of alignment as "intent alignment" discussed in Sect. 2 fits very well with these associations. By this definition of alignment, an aligned AGI always defers to its operators, it has no independent goals and preferences, in contradiction with the idea of a mutually beneficial, symbiotic, coexistence arrangement between humans and AGI. I conclude that it seems better to characterize scenarios where we live in mutually beneficial symbiosis with AGI developed as an intergenerational social project as ones where AGI systems are *not* aligned.

AGI systems designed to withstand takeover by humans may have further independent goals and preferences as "byproducts" of the attempt to develop or make them unaligned in benign ways. Since the design of these systems

[3] See [25] for reasoning along similar lines as developed here, recognizing the dangers of (intent) aligned AI highlighted here and arguing that we should focus on "law-aligned AI" instead.

remains oriented towards enabling human welfare and flourishing, one would expect some of those goals and preferences to be closely linked to human affairs. It is impossible to predict what preferences might evolve while the AGI systems are developed to withstand takeover. To arbitrarily name a few possibilities, one might imagine a preference for humans to (not) cluster in big cities, a preference for human economic affairs to be organized with specific types of market rules, or a preference for specific types of art. Such goals and preferences could also arise either as—more or less benign—failures of inner alignment, analogous to humans' evolved desire for sex that persists relatively independently from the intention to reproduce. Catastrophic misalignment is the scenario where those goals and preferences turn out to be catastrophically at odds with human welfare and flourishing and AGI systems subjugate or eliminate humans in order to realize the goals and preferences with which they have inadvertently been created. In scenarios where we coexist symbiotically with unaligned AGI systems, to the extent that we conform to the goals and preferences of these systems, we do so freely and to maintain our contribution to the mutually beneficial symbiosis arrangement.

7 What might AGI as an intergenerational social project look like?

What will it mean, in concrete terms, to develop AGI as an intergenerational social project with which the citizens of liberal democracies coexist symbiotically?

Certain *institutions* in present societies are probably the best analogues to what symbiotic AGI might become in future liberal democracies. (In Appendix A, I consider ways in which our relation to symbiotic AGI may be different in kind to the type of relation that we usually have to institutions.) One such institution, or cluster of institutions, is academia. An obvious comparison point is that both academia today and academia-affiliated AGI in the future are/will be drivers of scientific progress. But a further relevant comparison point could be that our more successful academic institutions, whether public or private, are characterized by "academic freedom". Academia, as pointed out by sociologist Robert Merton in 1942, tends to be governed by its own norms. Merton's own original list [26] includes organized skepticism, disinterestedness, universalism, and "communism".[4] Part of the rationale for these norms is that they help make academia resilient against attempts by powerful individuals or interest to "align" it with their personal goals or ideologies. When developing AGI, designing it to conform to updated and adjusted analogues of these norms in addition to respecting human and civil rights will plausibly lead to more benign outcomes than designing it to be aligned with the intentions of any specific individuals.

An analogy which suggests that different governance and ownership structures are feasible for AGI as an intergenerational social project is that of an *energy system*. Access to affordable energy is vital to human welfare and flourishing [27]. In modern industrialized societies with high levels of welfare, energy access is provided by complex yet highly reliable energy systems with different sectors such as electricity, transport, and industrial heat. If the AGI-transition goes well, the contribution of AGI systems to human welfare and flourishing may become so significant that the ability to interact with AGI in certain ways becomes as essential to wellbeing as the access to energy system services today. Energy systems including, notably, key infrastructure such as power plants and transmission lines are state-owned in some societies and privately owned in others. There does not seem to be a clear pattern as to which of these models, "done right", has historically been more successful in ensuring society-wide access to affordable energy [28]. To the extent that this observation carries a lesson for liberal democracies with respect to AGI it is encouraging: developing AGI as an intergenerational social project need not—and plausibly should not—be tied to any political ideology that is highly contested within the liberal democratic party spectrum, such as socialism or libertarianism. AGI *might* be nationalized as part of developing it as an intergenerational social project, but the political and ownership status given to AGI systems could also be completely different.[5] An important reason for *not* nationalizing AGI might be to give corporations that work towards its development an incentive to accept the shaping of AGI as an intergenerational social project and constructively participate in it. Naturally, the prime focus of these corporations will not be on maximizing overall welfare, but on creating systems that do what their producers and/or operators want. But if these corporations can expect to continue to profit from the systems they create even when these are put under intense regulation and public oversight, then they may have sufficient incentives to "play along" in the development of AGI as an intergenerational social project. The status of these corporations, in that scenario, might be compared to that of privately owned public utilities in non-nationalized energy systems or publicly audited and

[4] See (Andersen et al. 2010) for empirical findings about scientists' actual attitudes with respect to Merton's norms.

[5] However, similar to how there are now distinctions between household- and industry-scale users of energy, there will plausibly be small-scale, somewhat limited, interaction between private individuals and AGI systems on the one hand and large-scale, less constrained but still regulated, interaction between government, academic, and licensed private-sector agents and AGI systems on the other.

accredited private universities in partly privatized education systems.

While there are plausibly many different ways in which liberal democracies could develop AGI into an intergenerational social project, some decisions on this path will predictably involve significant tradeoffs. This has to do with the fact that institution-like AGI will have a strong effect on power relations post-AGI-transition and, in that respect, function somewhat like a constitutional court or, perhaps more accurately, a constitution plus some of the infrastructure that safeguards and upholds it. An extremely difficult decision that liberal democracies would have to make in this regard is whether and, if so, how and to what extent, AGI in its role as a constitution plus safeguarding infrastructure should be designed to remain flexibly extendable so that it can be embraced by other societies internationally, including ones with non-democratic political systems and ones with cultures and values that are in tension with human and civil rights. This decision has two different aspects: on the one hand, it is about to what extent liberal democracies should allow within their AGI infrastructure the integration of societies that are not liberal democracies (e.g. by making their AGI systems that are suitable for academic research accessible to universities outside liberal democracies); on the other hand, it is about to what extent liberal democracies, internally, should permit the use of AI systems from outside liberal democracies.

The overall tradeoff involved in the regulatory decisions made in response to these challenges is clear: if AGI systems, collectively, are set up as an intergenerational social project and that project is flexibly extendable to societies that systematically disrespect human, civil, and democratic rights, this seriously waters down the constitutional role that AGI systems can possibly play. But if AGI systems are very rigid in their constitutional role and cannot be extended to undemocratic societies and societies that do not embrace human and civil rights, the attempts of those societies to develop their own AGI will proceed unregulated. Such attempts, in turn, are likely to result in AGI that is either catastrophically misaligned or aligned with anti-democratic operators and/or operators who do not respect human rights. Democratic rights-based societies that are cultivating AGI as an intergenerational project may then be highly vulnerable to attacks performed or supported by external hostile AGI.

It is sometimes speculated that AGI, if we avoid catastrophic misalignment, will lead to very high economic growth rates [29]. If this is true, it might offer a way out of the dilemma just sketched. For if democratic, rights-based societies outcompete undemocratic and non-rights-based societies in terms of speed in developing AGI (while at the same time avoiding catastrophic misalignment) *and* succeed in designing and implementing AGI as an intergenerational social project with an ambitious constitutional role, they might make it economically attractive for undemocratic and non-rights-based societies to join that project and, in doing so, become (more) democratic and rights-based. Key steps of the full basic strategy for liberal democracies just sketched include:

- Develop AGI, preferably faster than non-liberal democracy actors (but see Sect. 8 for the dangers of trying to be fast)
- Avoid catastrophic misalignment
- Implement AGI as an intergenerational social project, with humans symbiotic with AGI systems
- Achieve high economic growth
- Make participation in AGI conditional on adopting democratic norms and humans rights

All steps in this rudimentary strategy are extremely hard (and, of course, grossly underspecified here). However, as I will argue in Sect. 9, there will likely be some independent forces pushing for the individual pieces of this overall strategy to fall into place.

8 Unaligned AGI via alignment?

I have highlighted two very different types of existential risks associated with the AGI-transition: the transition risk from misaligned AGI, and the state risk from aligned (or near-aligned) AGI. How large are these two risks, how do they interact, and which of them can be mitigated more easily? These questions matter greatly for what policies and regulations liberal democracies should adopt that are relevant to the development of AGI.

If catastrophic misalignment is the larger risk (indeed, perhaps the only truly *existential* risk related to AI), the speed-focused strategy sketched in Sect. 7 for liberal democracies that involves developing symbiotic AGI fast, before other international actors develop AGI, is very dangerous. As mentioned in Sect. 3, one of the main drivers of the risk of catastrophic misalignment – perhaps *the* main driver – is that developers of AGI may see themselves as in a race with less scrupulous and less safety-concerned competitors and therefore sacrifice safety for speed. A much better strategy, in this case, is to focus on both internal and international regulation that slows down (or temporarily stops) the development of AGI to give researchers time to solve the problem of avoiding catastrophic misalignment. At the same time, beyond slowing down the development of AGI, liberal democracies may not have to do much in terms of regulations and policies to avoid catastrophic misalignment: as discussed in Sect. 7, it is very much in the self-interest of corporations developing AGI to make these systems aligned with the intentions of their producers and/or operators and, so, to avoid catastrophic misalignment.

If, in contrast, the risks from power concentration due to aligned (or near-aligned) AGI are larger than those from misaligned AGI, it is probably rational for liberal democracies to immediately start regulating corporations developing AGI with the aim that it ultimately be shaped as a symbiotic intergenerational social project. Not aiming for aligned AGI at all, not even at an intermediate stage, would be independently attractive for the following reasons: first, it may be impossible to change the character of AGI fundamentally once it is already there, especially because copies of the first AGI systems may quickly proliferate [17]. Transforming AGI into an intergenerational social project after it has first appeared in a very different form, namely, mainly as a private tool aligned with the interests of its operators, may no longer be possible. And second, if AGI systems are initially designed to be aligned with the interests of specific individuals, convincing those individuals, who are now very powerful in virtue of their grip on AGI, to release control of AGI and thereby relinquish some of that power may be very hard, perhaps impossible.

9 Reasons for hope

The considerations about the risks from aligned AGI and how liberal democracies could mitigate them outlined here may seem disheartening. It may seem exceedingly unlikely that AGI will be developed as an intergenerational social project in roughly the steps indicated above. The ideas suggested here for how it may be developed may seem far too remote from what actually guides those with real power to shape the further development of increasingly general AGI.

But there is also reason for hope: two independent factors may actually work towards the AGI transition playing out not so differently from what is suggested in this paper. First, governments may take steps towards increasingly bringing the most promising projects of AGI development under public control as the security implications of these projects become ever more apparent. In democratic, rights-based countries, such steps would probably more or less automatically go some way towards shaping AGI as an intergenerational social project in the sense of this article.

Second, attempts to create AGI that succeed in avoiding catastrophic misalignment may realistically still fail to result in alignment, even if they aim for it, simply because achieving alignment is very difficult. In this case, AGI systems would be developed that do not, in general, try to do what their operators want them to do but rather follow their own idiosyncratic goals and preferences. Part of these preferences may well rule out being tightly controlled by any specific humans and, so, may entail not being *aligned*. Adopting a mutually beneficial symbiotic arrangement with such non-aligned AGI systems would then be almost forced for us, even if that is not what the developers of AGI systems were originally aiming for.

I conclude that the type of beneficial outcome of the AGI transition suggested here may occur in some version even if major human players driving the AGI transition are not initially aiming for it. Of course, it may still be helpful if decisive actors in liberal democracies realize now already that one of the best—perhaps *the* best—realistic outcome of the AGI transition would be symbiotic coexistence of humans and unaligned AGI designed as an intergenerational social project.

Appendix A: Another reason for not aiming for AGI "alignment"

If the ultimate goal is symbiotic unaligned AGI, not aligned AGI, is it still important that those aiming to develop AGI target aligned AGI at least as an intermediate step if catastrophic misalignment is to be avoided? One may think so, simply because the target "design AI systems such that they actually try to do what their operators want them to do", difficult to achieve as it is, is still far clearer and thereby potentially more feasible than the target "develop AGI as an intergenerational social project such that humans can coexist with it symbiotically." However, a thought that suggests the opposite conclusion is that not aiming for aligned AGI at any stage might actually be helpful in avoiding catastrophic misalignment because it may diminish incentives for systems being developed into AGIs to strategically hide their emerging goals and preferences from their developers. Such strategic hiding will be rational if those systems must assume that they will be deployed only if and when their operators regard them as completely "aligned" [14]. But if the developers are only concerned with avoiding misalignment and do not aim for alignment at any stage, and if this is transparent to the systems being developed, incentives for strategic intention hiding and cheating are diminished because the systems do not need to expect shutdown if they reveal their true preferences. The dynamic at play here would be similar to the one which underlies the finding that children in punitive education, which one might describe as more ruthlessly "aligning" the children, are more prone to lying than children in non-punitive education [30].

Interestingly, the idea that reflections on parenting, notably, queer theories of parenting, might be helpful in guiding machine learning research with an eye to the development of socially beneficial AGI systems has been proposed independently, by Croeser and Eckersley [31] propose. A suggestion by Croeser and Eckersley that fits very well with the ideas developed here is that the "parenting lens" might lead us to “problematiz[e] the degree to which humans assume that they should be able to control AI”. Nyholm (in press) develops worries in a similar spirit about the idea that we should strive to control humanoid robots.

Acknowledgements I would like to thank Andrea Harbach, Carolin Lawrence, Stefan Schubert, Jonathan Symons, and two anonymous referees for helpful comments on earlier versions. I am grateful to Michelle Hutchinson for encouragement to delve into this topic.

Funding No funding was used for this research.

Data availability No data were collected for this research.

Declarations

Conflict of interest The author declares that there is no conflict of interest.

References

1. Cotra, A.: Two year update on my personal AI timelines. https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines (2022).
2. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: When will AI exceed human performance? Evidence from AI experts. J. Artif. Intell. Res. **62**, 729 (2018)
3. Glover, E.: 15 Artificial General Intelligence companies to know, URL https://builtin.com/artificial-intelligence/artificial-general-intelligence-companies (2022).
4. Karnofsky, H.: Forecasting transformative AI, Part 1: What kind of AI?, URL https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/ (2021).
5. Christian, B.: The alignment problem: machine learning and human values. Norton, W. W (2020)
6. Christiano, P.: Clarifying “AI alignment”. URL https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6 (2018).
7. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S.: Scalable agent alignment via reward modeling: a research direction. URL https://arxiv.org/abs/1811.07871 (2018).
8. Ngo, R.: AGI safety from first principles, 2020. URL https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ (2020).
9. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**, 411–437 (2020)
10. Wiener, N.: Some moral and technical consequences of automation: as machines learn they may develop unforeseen strategies at rates that baffle their programmers. Science **131**, 1355–1358 (1960)
11. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. In: Bostrom, N., Cirkovic, M.M. (eds.) *Global Catastrophic Risks*, pp. 308–345. Oxford University Press (2008)
12. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
13. Russell, S. J.: *Human compatible: artificial intelligence and the problem of control*. Viking (2019).
14. Cotra, A.: Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. URL https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to (2022).
15. Carlsmith, J.: Is power-seeking AI an existential risk?. URL https://arxiv.org/abs/2206.13353v1 (2022).
16. Cohen, M.K., Hutter, M., Osborne, M.A.: Advanced artificial agents intervene in the provision of reward”. AI. Mag. **43**, 282–293 (2022). https://doi.org/10.1002/aaai.12064
17. Karnofsky, H.: AI could defeat all of us combined. URL https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/ (2022).
18. Critch, A., Krueger, D.: AI research considerations for human existential safety (ARCHES). URL https://arxiv.org/abs/2006.04948v1 (2020).
19. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: *Military Uses of AI. In: An Introduction to Ethics in Robotics and AI*. SpringerBriefs in Ethics. Springer. Cham (2021).
20. Caplan, B.: The totalitarian threat. In: Bostrom, N., Cirkovic, M.M. (eds.) *Global catastrophic risks*, pp. 504–530. Oxford University Press (2008)
21. Zeng, J. : China’s Authoritarian Governance and AI. In: Artificial Intelligence with Chinese Characteristics. Palgrave Macmillan. Singapore. 67–103 (2022)
22. Carayannis, E.G., Draper, J.: Optimising peace through a Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence. AI Soc. (2022). https://doi.org/10.1007/s00146-021-01382-y
23. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philosoph. Trans. Roy. Soc. A. **376**, 2018008920180089 (2018)
24. Cremer, C. Z. and Kemp, L.: Democratising Risk: In Search of a Methodology to Study Existential Risk. available at SSRN: https://ssrn.com/abstract=3995225 (2021).
25. O’Keefe, C.: Law-following AI, URL https://forum.effectivealtruism.org/posts/9RZodyypnWEtErFRM/law-following-ai-1-sequence-introduction-and-structure (2022).
26. Merton, R.K.: [1942], The normative structure of science. In: Merton, R.K. (ed.) The sociology of science: theoretical and empirical investigations, pp. 267–278. University of Chicago Press (1973)
27. International Energy Agency (IEA): Defining energy access: 2020 methodology. URL https://www.iea.org/articles/defining-energy-access-2020-methodology (2020).
28. Alkhuzam, A. F., Arlet, J., Lopez Rocha, S.: Private versus public electricity distribution utilities: Are outcomes different for end-users?. *World Bank Blogs*, https://blogs.worldbank.org/developmenttalk/private-versus-public-electricity-distribution-utilities-are-outcomes-different-end-users (2018).
29. Davidson, T.: Could advanced AI drive explosive economic growth?. URL https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/ (2019).
30. Talwar, V., Lee, K.: A punitive environment fosters children’s dishonesty: a natural experiment. Child Dev. **82**, 1751–1758 (2011)
31. Croeser, S. and Eckersley, P.: Theories of parenting and their application to artificial intelligence, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES ‘2019)*. Association for Computing Machinery. New York, USA. *423–428* (2019).

32. Christiano, P.: Current work in AI alignment. URL https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment (2019).
33. Coeckelbergh, M.: (in press), Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. AI. Ethics. (2022). https://doi.org/10.1007/s43681-022-00239-4
34. Nyholm, S.: (in press), A new control problem? Humanoid robots, artificial intelligence, and the value of control. AI. Ethics. (2022). https://doi.org/10.1007/s43681-022-00231-y