AI Creation and the Cosmic Host

(2024) Nick Bostrom[1]
[*Manuscript*, v. 0.5. Draft]
[www.nickbostrom.com]

ABSTRACT

There may well exist a normative structure, based on the preferences or concordats of a cosmic host, and which has high relevance to the development of AI. In particular, we may have both moral and prudential reason to create superintelligence that becomes *a good cosmic citizen*—i.e. conforms to cosmic norms and contributes positively to the cosmopolis. An exclusive focus on promoting the welfare of the human species and other terrestrial beings, or an insistence that our own norms must at all cost prevail, may be objectionable and unwise. Such attitudes might be analogized to the selfishness of one who exclusively pursues their own personal interest, or the arrogance of one who acts as if their own convictions entitle them to run roughshod over social norms—though arguably they would be worse, given our present inferior status relative to the membership of the cosmic host. An attitude of humility may be more appropriate.

1. Human civilization is most likely not alone in the cosmos but is instead encompassed within a cosmic host

- The "cosmic host" refers to an entity or set of entities whose preferences and concordats dominate at the largest scale, i.e. that of the cosmos.
    - The term "cosmos" here is meant to include the multiverse and whatever else is contained in the totality of existence.
- For example, the cosmic host might conceivably consist of galactic-scale civilizations, simulators, superintelligences, and/or a divine being or beings.
- Naturalistic members of the cosmic host presumably have very advanced technology, including e.g.:
    - Superintelligent AI
    - Efficient means of space travel and von Neumann probes
    - Ability to run vast quantities of simulations of e.g. human-like histories and situations

[1] I'm grateful to Will Aldred, Owen Cotton-Barratt, Joseph Carlsmith, Milan Cirkovic, Max Dalton, Oscar Delaney, Lukas Finnveden, Peter Gebauer, Rose Hadshar, John Halstead, Michel Justen, Will MacAskill, Fin Moorhouse, Toby Newberry, Zershaaneh Qureshi, and Carl Shulman for comments.

    - (It's possible that some members might have capabilities that exceed those that are possible in our universe: e.g. if they live in another part of the multiverse with different physical constants or laws; or, if we are simulated, if the underlying universe the simulators inhabit has different physical parameters than the ones we observe.)
- Nonnaturalistic members of the cosmic host presumably have analogous capabilities supernaturally.
- There probably is a cosmic host.
    - The likelihood is increased because there are multiple ways for this to be the case. The main ones might be:
        - (a) The simulation hypothesis, which receives significant probability via the simulation argument. If we are in a simulation, then there exists at least one simulator (and possibly a great many).
        - (b) An infinite or immensely large universe, which is supported by astronomical observation.
            - If our universe is open or flat, homogeneous, and simply connected then it is infinite and hence home to infinitely many civilizations, including infinitely many extremely advanced ones.
            - Even if our universe is not infinite, it might be extremely large and hence also statistically likely to contain many advanced civilizations.
            - Even just the currently observable universe is large enough to possibly contain extraterrestrial civilizations (although it is very far from certain that it does so, and arguably not even likely).
        - (c) Multiverse, which is supported by some physics theories.
            - Some theories of cosmic inflation predict the existence of a vast or infinite number of other universes.
            - String theory seems to allow for a vast number of possible vacuum states, each of which might correspond to a universe within a vast multiverse.
            - The many-worlds interpretation of quantum physics. While this theory implies that there is a large structure of physical existence—a universal wave function that would contain enormous numbers of modes or "branches", including many in which various advanced civilizations exist—it is somewhat unobvious that it supports the cosmic host hypothesis.
                - On the one hand, it would strongly indicate that a cosmic host exists.
                - On the other hand, depending on how we interpret "amplitudes" in this framework, it could be the case that other branches of the universal wave function that contain other civilizations mostly have a low "weight" and that they should not count for as much as higher amplitude branches in our decision-making.
                - Thus, the many-worlds interpretation may not change "the expected amount of cosmic host" that exists.
                - However, the practical implications may have a nonlinear relation to the amount of cosmic host. For example,

consider a case where an evenly balanced quantum coinflip determines whether a simulator arises, but if one does arise it produces vast numbers of simulations (and we can assume we reject SIA here). Outside the many-worlds interpretation, we might then think there is about a 50% probability that we are simulated. Yet given the many-worlds interpretation, we know that vast numbers of simulations arise, and even if the amplitude squared of the branch of the universal wave function where that happens is only ½, the sum of the measure of the observer moments that are simulated (since there are so many of them) greatly exceeds the sum of the measure of the observer moments that are not simulated.
        - So the many-worlds interpretation could make us think there is a greater probability that we are simulated and hence that there exists a simulator and hence that there exists a cosmic host.
        - Note, however, that the expected measure-weighted number of simulated copies of us would not (at least not in any straightforward way) be greater on the many-worlds interpretation.[2] So while there might be practical implications here if our values or evaluation criteria are branch-relative or indexed to our branch, there may be no immediate practical implications of the many-worlds interpretation from a purely impersonal evaluative perspective.
- The Self-Indication Assumption (SIA), which is maintained in some variants of anthropics (the epistemology of indexicals), claims essentially that you should reason as if the fact that you exist gives you evidence in favor of hypotheses in proportion to the number of observers they imply exist. (This is not really "a way for it to be the case" that there exists a cosmic host, but rather a type of consideration that could increase the probability of that being the case, since SIA would raise the probability of there being many observers and for the universe to be big enough to allow this; but it is mentioned here anyway for the sake of the convenience of having relevant considerations collected in one place.)
    - Note that other theories of anthropics reject the SIA.
- Supernatural beings, supported by some religious views. A powerful deity or deities would satisfy the definition of a cosmic host.
- Superintelligences that human civilization creates in the future. One difference between this way for it to be the case that there is a cosmic host and the others is that here there is a direct dependency between human actions and the (later) existence of a cosmic host. Nevertheless, it is true that, through this mechanism, the world (and, in particular, our spacetime manifold) could contain a cosmic host. And also that some parts of our

[2] Bostrom, N.: *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, 2002).

lives—perhaps almost all parts—may be taking place in the presence of a comic host that we later create.

2. The cosmos may contain regions that the cosmic host does not directly control

- If intelligent life is sparse, there might be spacetime regions in the universe that are not physically accessible to any member of the cosmic host.
- If we are in a simulation, then presumably all parts of our world are physically accessible to the simulator. Nevertheless, the simulator might be subject to constraints that limit its ability to intervene—for example, if the purpose of a simulation requires non-interference on the part of the simulator.
- In some theological conceptions, God has the ability to intervene anywhere but sometimes has reason to refrain from doing so, such as to give scope for the exercise of human free will.
- For a secular example, suppose that the parents wish their child to eventually take over the family farm, and that there exists a serum they could legally inject that would guarantee that the child will choose to do this when it grows up. The parents might refrain from using the serum, on grounds that it would be overbearing to do so or that it would be disrespectful of the child's autonomy, while nevertheless hoping that the child will choose to take over the farm.

3. The cosmic host may care about what happens in regions it does not directly control

- For example, it might have preferences regarding the welfare of individuals who inhabit such locations, or regarding the choices they make.
- Such preferences might be noninstrumental (e.g. reflecting benevolent concern) and/or instrumental (e.g. host entity *A* may want individual *h* to act in a certain way because it believes that host entity *B* will model how *h* acts and that *B* will act differently with respect to matters that *A* cares about noninstrumentally depending on how it believes that *h* acts).
    - Such interlinkages may also enable intra-host coordination even if the host consists of many distinct entities pursuing a variety of different final values.

4. The cosmic host might have indirect influence over regions it does not directly control

- One way is by exerting influence over locations it does directly control and that actors in locations it does not control care about, and conditioning this influence on the host's models of how those actors act in their locations. When those actors factor these dependencies into their decision making, the host's preferences can thus indirectly influence what happens in locations that are beyond its direct control.
- It is also possible that actors that are outside the cosmic host's direct control during one phase of their life might be inside it at a later phase.

    - One example is if a lower-level civilization comes into physical contact with a more advanced extraterrestrial civilization.
    - Another example is if a lower-level civilization creates a superintelligence that becomes part of the cosmic host; and actors in the lower-level civilization that were previously outside the direct reach of the host might then come inside its reach.
    - Another example is if some lower-level being, while inside a region that the cosmic host can physically affect, is initially nevertheless not directly controlled, because the host is subject to instrumental or noninstrumental constraints (as per above). Those constraints might cease to hold at a later time, allowing the host to exert more direct control over the lower-level being.
        - (For instance, on some theological conceptions, the fate of individual human souls might be more directly shaped by God after their mundane sojourn is complete and they have had a sufficient opportunity to exercise their free will.)

5. There may be cosmic norms

- Just as we have norms at various scales of human organization—such as norms within a social club, wider cultural norms, and global norms (e.g. ones reflected in the Geneva Convention and the Universal Declaration of Human Rights)—so too might there be something like norms at the highest (cosmic) scale, reflecting cooperative frameworks or rules embedded in behavioral equilibria.
- One could entertain a spectrum of possibilities, ranging from a radically multipolar ensemble of cosmic host members acting at cross-purposes conflictually or uncoordinatedly (at one end), to a set of independent and orthogonal host members, to cohesive, cooperative, or fully unified cosmic hosts (at the other end).
- Some conceivable cosmic host types are unified essentially or by definition, such as in some theological conceptions of a greatest being.
- For other host compositions, (potentially very high) degrees of cohesiveness might arise through various mechanisms, such as the following.
    - There could be an ontogenetic convergence among entities that become members of the cosmic host, such that they all (or for the most part) come to have the same (or mostly congruous) preferences, owing to common factors shaping their developmental trajectories.
        - Some have held that sufficiently enlightened individuals converge to a universal set of preferences (either necessarily or as a strong empirical tendency).
        - There could be attractors in the space of sociopolitical dynamics such that sufficiently technologically advanced societies (that have access to AI-advisors and to general and reliable technologies for self-modification and redesign) converge towards a shared set of values or universal norms.
            - The fact that this has so far not happened (to a greater extent than it actually has) is not strong evidence that it won't happen at a later point, as profound changes in the material and technological background conditions lie ahead.

        - Although there is an unlimited number of different potential goals that unaligned AI could pursue, it might (for aught we know, since we don't currently understand the relevant mechanisms very well) possibly be the case that almost all alignment failures that actually happen (also across civilizations) result in AIs that are "misaligned" in the same direction.
    - Alternatively, cohesion might arise through mechanisms that coordinate between members of the cosmic host (even if the latter come to the table with divergent preferences and practices).
        - In cases where cosmic host members physically (or supernaturally) interact, norms could develop in similar ways as they do (at various scales) in the present human context.
        - Additionally, cosmic host members may coordinate by modeling each other's choices and conditioning their own choices on their expectations of the choices (or conditional choices) of the others; as indicated above.
- Even if the host is not fully coherent, its members might, while conflicting at one level, nevertheless favor a common norm that would, if it has sufficiently strong backing, establish peace and harmony.
    - For example, country A and B might be fighting, while both wishing that there existed a sufficiently empowered arbitrator who would come in and force an end to their fighting.

6. We would have reason to respect cosmic norms

- Just as we can have reason to respect norms at the various familiar human scales, so too can we have reason—both prudential and moral—to respect cosmic norms.
- We can have prudential reason to respect cosmic norms inasmuch as consequences for ourselves or our interests might directly or indirectly be shaped by how our actions conform to these norms.
- We can have moral reason to respect cosmic norms, on several different possible alternative grounds:
    - Constitutive. Cosmic norms—or constructs closely related to cosmic norms, such as idealized versions of cosmic norms—might be *constitutive parts* of morality, and especially of "higher morality".[3] In other words, truths about morality might be grounded in truths about (positively defined) norms or some similar construct (such as moderately idealized positively defined norms).
    - Derivative. Even if cosmic norms are not in themselves constitutive of morality, we may have moral reason to respect such norms, reasons that derive from moral reasons for deference, compliance, or cooperation (which can arise in consequentialist, deontological, virtue ethical, and other moral frameworks). For example:
        - If you are sharing a sleeping compartment on a train with several other people, who have all requested to keep the window open, you may have moral reason to respect their desires (even if you are strong enough that you could impose your will).

[3] Bostrom, N (2022): "Base Camp for Mt. Ethics", *Manuscript*, v. 0.9. https://nickbostrom.com/papers/mountethics.pdf.

        - If you are visiting somebody's house, you may have moral reason to follow the house rules and defer to the host's requests (even if the latter seem odd and misguided to you).
        - If you are a new immigrant to a country, you may have moral reason to obey the law (even in the case of laws that you think are mildly harmful and that you had no part in creating, and where violations would go undetected).
    - Epistemic. On some objectivist metaethical conceptions, in which morality is constitutively more independent of what is agreed or wanted or authoritatively legislated, we might have epistemic reasons to mostly defer to the opinions of the cosmic host, to the extent that we can surmise them.
        - Because, on such metaethical conceptions, why think that we—our own species, with its idiosyncratic evolutionary trajectory and troubled history—would be more reliable at ascertaining moral truths than the average or modal other entity in the cosmic host?
            - Especially considering that cosmic host entities vastly outstrip us in more or less every other epistemic domain.
        - Without direct communication from the host, one might think that our only access to information about what the host thinks about morality is via our own thinking about morality. However, it can still be a helpful heuristic injunction to consider how things might appear from another perspective.
            - For example, when facing a moral dilemma relating to a dispute that one is involved in, it can be helpful to speculate about how a disinterested observer would regard the matter under contention.
            - Similarly, it could be helpful to reflect on what moral views observers with different cultural backgrounds or different evolutionary origins might arrive at.
        - One could also seek to preserve options to use resources (along with a willingness to use them for moral ends) until such a time as we either are able to receive direct information from the host or become (closer to being) its epistemic equal.
            - Note, however, that cosmic norms might have implications for what we ought to do now, not only about what some more informed later stage of human civilization ought to do.
            - Perhaps they care much more about what some later more informed (and potentially more powerful) stage of human civilization does; such that if we could defer most decisions to such a later stage, this would be agreeable to the host.
            - However, this presupposes that the later stage of human civilization would in fact be disposed to follow cosmic norms. If not, then it might be important to the cosmic host what we decide now (and in particular whether we steer towards a future stage that would be disposed to follow cosmic norms).
                - Insofar as a disposition to follow cosmic norms simply follows from robust prudential considerations, one might expect that—to the extent that cosmic norms do in fact exist—a more informed later stage of humanity would be more likely to seek to adhere to cosmic norms. But insofar

as a disposition to follow cosmic norms requires moral motivations, this is not a given.
    - Moreover, some of the prudential reasons for adhering to moral norms might depend on epistemic state—particular kinds of uncertainty—which are not guaranteed to remain in place for a later more informed version of humanity.
        - Compare this to how a moral being might prefer us to make certain choices while we are still behind a "veil of ignorance", since the prudential reasons that imperfectly moral agents face under such conditions may more closely match the imperatives of morality than the prudential reasons those agents face when they know more about the actual peculiarities of their own situation.

7. We would have reason to design any superintelligence we build such that it becomes a good cosmic citizen

- We can make an analogy with how human parents often have reason—both moral and prudential—to raise their children to be (among other things) good citizens.
    - The template of a good citizen, to a first approximation, can be thought of as that of a person who, during their development and as a grownup, respects the moral norms of their community.
        - Good citizens are, presumably, upstanding, cooperative, and respect the preferences and interests of other community members. They seek to contribute positively to those around them, without over-asserting themselves or resorting to illegitimate means.
        - Citizens may have a purview within which to pursue their own preferences—to an extent and within bounds set by the existing normative order.
- Human civilization has reason to aim to make any superintelligent civilization that it develops into, or any independent superintelligent AIs that emanate from it, be good cosmic citizens.
    - A good cosmic citizen is an entity that, as it develops increasingly advanced capabilities and becomes technologically mature, respects the norms of the cosmic host.
        - Good cosmic citizens are, presumably, very broadly cooperative and respect the preferences and interests of the cosmic host. They seek to contribute positively to the weal of the cosmopolis, without over-asserting themselves or otherwise resorting to means disfavored by cosmic norms.
        - Cosmic citizens may be permitted, by the cosmic norms, to use a certain portion of the resources over which they are able to exert direct control for their own self-interested or idiosyncratic purposes.
        - Cosmic citizens may also have some legitimate scope—circumscribed by the normative order of the cosmopolis—for exerting influence over what the cosmic norms should be.

                - The extent or degree of legitimate influence of any particular cosmic citizen civilization—"how hard" it is normatively permitted to push for its own goals as opposed to deferring to the wishes of others or to independently existing aspects of the cosmic order—would not be determined by that citizen's own internal moral conception but would instead be more globally determined.
                - For illustration, consider a few conceivable weightings: 'one person one vote'; 'one country one vote'; 'one civilization one vote', 'one joule of free energy one vote'; 'one int8 operation one vote', 'incumbent earlier civilizations have more votes'; 'more powerful civilizations have more votes'; 'civilizations with certain kinds of preferences have more votes'. We ought not to unilaterally pick whichever scheme seems fairest or most appealing to us and insist, come what may, that our interests be given weight accordingly, if that is not in accordance with the cosmic order.
            - If there is conflict between members of the cosmic host, a good cosmic citizen might seek to play a constructive role in mitigating and ending the conflict by supporting cooperative norms and lawfulness.
                - In human conflicts between countries, third-parties can sometimes usefully contribute—by the lights of each of the contending parties—by providing humanitarian aid, by promoting adherence to the laws of armed conflict and the Geneva convention, and by proposing and incenting settlements that would benefit all parties compared to continued fighting.
                    - (Such interventions, however, often require self-restraint, wisdom, and adroitness. Clumsy third-party meddling can be harmful, and could also risk embroiling the third-part in the conflict.)
                - Perhaps a good cosmic citizen would play a similar role in cases where there is incoherence within the cosmic host.
- When approaching to join the cosmic host, an attitude of humility, deference, and goodwill may be appropriate. Adopting a hard maximizing mindset and approach might itself be wrong or unvirtuous.
    - It is not only the final choice that may be scrutinized but also the method whereby we arrived at it.
    - Game theory may be morally suspect.

8. The cosmic host may want us to build superintelligence

- If our region is not directly accessible to the host, cosmic norms may currently have very limited influence over what happens here.
    - One reason for this is that we humans may currently not be much influenced by cosmic norms.
        - Our motivation to conform to them is often lacking.
        - Our knowledge of what they require of us is also often lacking.
    - Another reason is that we currently have very little ability to causally shape our region, owing in large part to our comparatively primitive technology—our

civilization is mostly powerless outside the thin crust of a single planet, and even within this ambit its powers are severely limited.

- However, if we build superintelligence, the host's ability to influence what happens in our region would plausibly greatly increase.
    - A superintelligent civilization (or AI) may be both more able and more willing to allow itself to be (indirectly) influenced by cosmic norms than we humans currently are.
        - Superintelligence would be better able to figure out what the cosmic norms are.
        - Superintelligence would be better able to understand the reasons for complying with cosmic norms, assuming such reasons exist.
    - A superintelligent civilization (or AI) that wants to exert influence on our region in accordance with cosmic norms would be far more capable of doing so than we currently are, since it would have superior technological, strategic, and planning abilities.
- Consequently, if the host cares about what happens in our region (for either instrumental or noninstrumental reason), it may want us to build superintelligence, provided that it estimates that the expected value of our region would be greater given the presence of superintelligence here than given its absence here.
    - This depends somewhat on the values of the cosmic host and partly on the likely character of the superintelligence that would emerge in our region.
    - *A superintelligence aligned to the cosmic host* would seem desirable to the cosmic host for a very wide set of possible host values and in a very wide range of possible situations.
        - Consider that, as a limiting case, the newly emerged cosmic-host-aligned superintelligence would shut itself down if it came to estimate that the host would prefer its nonexistence.
            - The superintelligence's estimate of whether the host prefers the superintelligence's nonexistence would be more reliable than our own, given the superintelligence's greater epistemic capabilities.
            - (Non-creation and non-continuation could diverge in value, but still the option of voluntary non-continuation or self-curtailment would seem to greatly cap the potential downside to the host of a host-aligned superintelligence.)
    - *A superintelligence that is not fully aligned to the cosmic host but is at least a decent cooperative cosmic citizen* may well be desirable to the host, since such a superintelligence could be a valuable trading partner for other members of the cosmic host and it could help contribute to upholding the cosmic order.
    - *A superintelligence that is antagonistic or quarrelsome* may be undesirable to the cosmic host.

9. The host might favor a short timeline

- The timeline to superintelligence could possibly be relevant to the host by having a bearing on at least three parameters:
    - (i) Effects directly tied to the passage of time itself;
    - (ii) Effects of delays on the probability that superintelligence is ultimately built; and

    - (iii) Effects of the timing of superintelligence on its character
- Effects directly tied to the passage of time itself—such as delays in resource harvesting, delays in helping current terrestrial populations, or savings in the compute requirements for simulating relevant parts of history—appear (tentatively) to be comparatively unimportant considerations for the host. (See Appendix A1.)
- Effects of delays on the probability that superintelligence is ultimately built appear (tentatively) important and seem to favor shorter timelines. Short self-limiting pauses are less of a concern than pauses that are designed to be long or that have a propensity for triggering further pauses or extensions. (See Appendix A2.)
- Effects on the timing of superintelligence on its character appear to be potentially important to the host, but the sign—whether shorter timelines lead to a more or less favored AI character in expectation—seems very unclear. (See Appendix A3.)
    - It is unclear both what effects a shorter timeline has on AI alignment and what effects AI alignment has on how agreeable the resulting AI would be (compared to an unaligned AI) to the cosmic host and how conformant to cosmic norms.
- Therefore, the main host-related consideration regarding the timeline that seems both important and at least moderately clear at present is that the host appears to (ceteris paribus) disfavor delays that significantly increase the probability that we will never build superintelligence.
    - So long as the risk of permanent failure to build superintelligence is not significantly increased, the host may prefer that the alignment problem be solved—provided that the superintelligence is then aligned in an at least adequately host-friendly way.
        - It could be technically easier to ensure that a superintelligence has a highly resource-satiable goal, or that it has other properties that makes it host-friendly, than to fully align it with human volition.

10. Conclusions

- There probably exists a “cosmic host”, consisting of one or more powerful superintelligent natural and/or supernatural entities.
- This host may support cosmic norms that we can have moral (as well as prudential) reason to respect.
- The host may want our civilization to build or develop into a good cosmic citizen: superintelligence that respects cosmic norms, is modest, lawful and cooperative, and contributes positively to other host members and the order of the cosmopolis.
- The host may favor paths that lead to this outcome with high surety, meaning a high probability both that superintelligence gets developed and that it becomes a good cosmic citizen.
- The cosmic normative structure might pertain not only to the ultimate outcome but also to the path taken to get there—including local outcomes along the way as well as attitudes and modes of analysis etc.
- (The exploration in this paper is not an attempt to cover all possibly relevant factors and arguments that should be taken into account in an all-things-considered assessment of our macrostrategic situation.)

Appendix A1: The mere passage of time

- The mere passage of time might appear unimportant because plus or minus a century is negligible on an astronomical timescale.
- If the host cares about terrestrial humans, animals, or limited AIs that exist now or will come to exist prior to superintelligence, then the passage of time even on scales as short as decades could be relevant to the host.
    - If the entire apparent history of Earth is real, then approximately 5% of all humans that ever existed are alive now, and of these 1.4% die each year (around 60 million).
- The host may care about nonhuman animals, but the total welfare outcomes for the nonhuman animal population would appear not to change significantly in percentage terms on relevant timescales.
    - Compared to humans, a far smaller percentage of all animals that ever lived are alive now (whether counted by raw numbers, or weighted by mass, brain mass, or neuron or synapse-count). Although livestock populations have been growing dramatically in recent decades, cumulative numbers are still very small compared to the total number of similarly-sized animals that have lived over the course of evolutionary history (assuming all this apparent history is real).
- The number and sophistication of digital minds that exist is increasing very rapidly, maybe doubling time of ∾1 year.
    - While the digital minds population is quite strongly coupled to progress towards superintelligence, a 1-year pause on increasing capability levels of AI just prior the development superintelligence could more than double the number of (size- or sophistication-adjusted) digital minds years that are lived prior to the advent of superintelligence.
    - The “death rate” of digital minds is unclear. Many digital minds that have stopped running may be preserved so as to enable later restart or exact recreation from digital records.
- If we are simulated, then the fraction of all humans and animals in our world that are currently alive could be much greater than if we are not simulated, since the simulation might have started relatively recently.
    - (There are also versions of the simulation hypothesis in which not all apparently existing minds are simulated in sufficient granularity to fully render the subjective lives that normally would be assumed to have been generated.)
- If we are simulated, then it might well be possible for the lives of beings that have died in the simulation to be continued.
- If we are simulated, then at least one host member would have continuous physical access to our world, and would be able to choose parameters such as when our world ends and whether and when superintelligence is developed within it (although that host member may be subject to constraints of various sorts that could still make them hope that we will freely choose certain actions despite the host member having the physical capacity to override our choices or to design us such that we are guaranteed to make the preferred choices).
    - Changes in the arrival time of superintelligence might also have an impact on the total computational cost of running the simulation.
        - If the simulation is set to run until a fixed calendar year, then delaying the building of superintelligence may reduce the cost of the simulation.

        - If the simulation is set to run until the building of superintelligence, then delaying that event may increase the cost of the simulation.

Appendix A2: The probability that superintelligence is ultimately built

- Delays in building superintelligence would increase the probability that superintelligence will never be built.
    - This is not tautological—it is conceivable, for instance, that racing to superintelligence would lead to conflicts that permanently destroy our world while a slower approach could get superintelligence built more surely.
    - In reality, however, it is plausible that most delays would increase the probability of permanent failure.
        - While non-anthropogenic risk of human extinction or permanent stagnation is very low on timescales of centuries (under the assumption that we are not simulated), anthropogenic risks may be significant even on a timescale of years or decades.
            - Current per annum extinction risk is quite low.
            - Current per annum permanent stagnation risk may not be that low.
                - It is hard to be confident that some process is not currently taking place which in hindsight could be recognized as a permanent lock-in. (The key moment might not be when the supertanker comes to a halt but when the engine was destroyed or perhaps even when the—possibly as-yet unobserved—pirates boarded the ship.)
        - Both extinction risk and permanent stagnation risk look set to rise markedly in the near term, owing in part to new capabilities unlocked by pre-superintelligent AI and in part to other technological advances.
            - Extinction risk may be increased, e.g., by some advances in biotechnology, nanotechnology, autonomous drone technology, and other weapons technologies. Progress in such directions is driven both by advances in AI and by general scientific and technological advancement.
            - Permanent stagnation risk may be increased, e.g., via applications of current or near-future AI technologies that could enable unprecedentedly effective ways of enforcing a hegemonic perspective (such as by automatically censoring or suppressing discordant opinion and by using sentiment analysis or lie detectors to disincentivize wrongthink).
                - The claim is not that AI developments reduce the *expected* amount of near-term disruptive change. Rather, the claim is that extreme—truly permanent and global—lock-in scenarios, which might have been infeasible throughout human history, may become feasible because of recent or near-future technological advances, and hence more likely to occur.
- Certain types of delay would have a greater tendency to increase the probability that superintelligence will never be built than other types of delay.

    - Compare the following two scenarios of 6-month pause on leading-edge AI development:
        - *Scenario A*: Many developers race to develop superintelligence. One developer is about a year ahead of the closest competitor. When this developer figures out how to build superintelligence, rather than immediately going full speed ahead, it voluntarily decides to pause for 6 months in order to double-check that its safety mechanisms work.
        - *Scenario B*: Shortly before it becomes possible to build superintelligence, anti-AI advocates manage to get enough support to persuade world leaders to implement a global 6-month moratorium on running any advanced AI models.
        - It seems that the pause in Scenario B would more greatly increase the probability that superintelligence will never be built than the pause in Scenario A. In Scenario A, the pause is more plausibly self-limiting: eventually other competitors will catch up to the lead developer. In Scenario B, the same factors that led to the initial 6-month moratorium might plausibly grow in strength during the pause—sentiments may harden, regulatory enforcement agencies may gain in capacity and influence, etc.
    - In general, delays that occur via widespread stigmatization of AI and/or via the establishment of institutions that have the ability to globally monitor and police AI developments (and perhaps the institutional incentives to perpetuate and expand their mandates) seem more likely to permanently thwart superintelligence than delays that occur via voluntary local choices by individual decision-makers.
    - Delays that occur via a general social collapse or reduction of civilizational capacity to pursue scientific, economic, and technological development also seem more likely to increase the probability that superintelligence will never be developed.

Appendix A3: The character of the superintelligence that is built

- The effect of the timing of superintelligence on its character is unclear.
    - Later development might mean that the alignment problem is more likely to have been solved when superintelligence is built.
        - There is some uncertainty about this—e.g. earlier development of superintelligence might take place while there is less compute overhang, which might make the initial stages of the takeoff less explosive, which might be helpful for alignment efforts.
    - The moral and prudential reasons for conforming with cosmic norms may become more widely recognized over time.
- Superintelligences themselves would presumably recognize the reasons for respecting cosmic norms.
    - Insofar as these reasons are prudential, superintelligences may be expected to comply with cosmic norms.
    - Insofar as the reasons are moral, however, compliance with cosmic norms may depend on a superintelligence having intrinsic moral motivations.

- Suppose that the cosmic host were composed of superintelligences built by humanlike civilizations, and that they all have only nonindexical goals, and the host is sufficiently numerous and variegated that all possible goals that we might end up giving to our superintelligence are already instantiated in many other superintelligences. One might then think that however we go about things, when we build superintelligence we simply add one more superintelligence drawn from the same motivational distribution as the others; and that this would therefore seem fair to the existing members of the cosmic host. But this inference would stand on very shaky ground.
    - The cosmic host may not be composed (exclusively) of superintelligences built by humanlike civilizations. The cosmic host may then prefer that we aim to build a superintelligence that has values that are more representative of the overall membership of the cosmic host rather than just of the subset of the cosmic host that originated from humanlike civilizations.
        - One possible case might be that of a supernatural host.
        - Another possible case might be if we had clues that humanlike originations of superintelligences are atypical in terms of the values that result by default.
    - The cosmic norms may not be a simple reflection of the distribution of preferences amongst the cosmic host, but instead e.g. a reflection of resource- or power-weighed preferences among the cosmic host, or some more indirect or distilled representation of common normative ideals among host members.
    - Members of the cosmic host and/or our own superintelligence could have indexical values. The cosmic host may then prefer that we aim to build a superintelligence that directly focuses on the values of other host members (or distilled cosmic norms).
        - The host may prefer this even if the indexical values of our superintelligence were drawn from what superficially could appear as "the same" distribution as the indexical values of the cosmic host members—since the referents of these formally similar indexical values would be different.
            - For example, if all the existing superintelligences have the preference "I want high reward", it might do them little good if we built a superintelligence with the preference "I want high reward". They may rather we build a superintelligence with the preference "I want other preexisting (or independently existing) superintelligences to get high reward".
- Plausibly more important to the cosmic host than the direction of a new AI's goals is the degree to which that AI is cooperative—which might depend partly on the AI's decision theory and partly on how resource-satiable the AI's goals are.
- Human values appear to be quite resource-satiable: we would much rather have a 100% chance of being able to use 1 galaxy to meet our goals than to have a 1% chance of being able to use 100 galaxies.
- If AIs resulting from alignment failures are more likely to have resource-hungry goals, then the cosmic host might prefer that we solve AI alignment.
- If AI alignment failure generally resulted in AIs that share the same goal, then even if that goal is resource-hungry, the host (if its members mostly consist of AIs that resulted from alignment failure) may well prefer a new AI to result from alignment failure, even though it would have a resource-hungry goal.

- If, however, alignment failures have a reasonable chance of resulting in AIs with any of several different and competing resource-hungry goals, then the host may prefer that we solve AI alignment.