
Propositions Concerning Digital Minds and Society

Nick Bostrom
Future of Humanity Institute
University of Oxford

Carl Shulman
Future of Humanity Institute
University of Oxford

Disclaimer

The following are some very tentative propositions concerning digital minds and society that may seem to hold some plausibility to us. We are not ready at this point to confidently or “officially” endorse them, nor do they give a full picture of our views on these matters. We put them forward to facilitate feedback and provoke discussion. They will almost certainly be substantially revised as our work progresses.¹

Consciousness and metaphysics

- The *substrate-independence* thesis is true:
“[M]ental states can supervene on any of a broad class of physical substrates. Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences. It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks inside a cranium: silicon-based processors inside a computer could in principle do the trick as well.”²
 - Sufficiently high-fidelity human brain emulations would be conscious.
 - Some AIs with architectures quite different from biological brains could also be conscious.
- The *quantity* of conscious experience is a matter of degree, in several respects, including potentially:
 - Number of individuals or copies: there could be more conscious subjects.
 - Repetitions: one subject could have an experience many times.
 - Duration (wall-clock time \times speed of computation): a given experience could have a longer subjective duration.
 - Implementation robustness: a system might be a more unambiguous instantiation of a particular computation.

¹ For useful comments, we are grateful to Stuart Armstrong, Michael Bailey, Adam Bales, Jake Beck, Asya Bergal, Sam Bowman, Patrick Butlin, Ryan Carey, Joseph Carlsmith, Paul Christiano, Michael Cohen, Teddy Collins, Owen Cotton-Barratt, Wes Cowley, Max Daniel, Eric Drexler, Daniel Eth, Owain Evans, Lukas Finnveden, Iason Gabriel, Aaron Gertler, Katja Grace, Julia Haas, Robin Hanson, Lewis Ho, Michael Huemer, Geoffrey Irving, Deej James, Ramana Kumar, Jan Leike, Robert Long, Vishal Maini, Matthew van der Merwe, Silvia Milano, Ed Moreau-Feldman, Venkat Nettimi, Richard Ngo, Eli Rose, Anders Sandberg, Eric Schwitzgebel, Jonathan Simon, Alex Spies, Nick Teague, Laura Weidinger, Peter Wills, and the participants in several seminars where earlier versions of this work were presented.

² Bostrom (2003, p. 2). For some supporting argumentation, see e.g., Chalmers (2010, §9); Chalmers (1996, §7). For examples of views that we reject, see e.g., Searle (1980); Block (1981).

- Quantum amplitude: if the many-worlds (Everett) interpretation of quantum mechanics is correct, then a computation might take place on a branch that has a higher measure.
 - Anthropic measure: some theories of anthropic reasoning assign “weights” to different observer-moments—some experiences might have higher such weights.
- The *quality* of a conscious experience can also vary pretty continuously along multiple dimensions, such as:
 - Scope of awareness: for example, a fully awake and alert human versus a drowsy mouse.
 - Hedonic valence: weak versus strong pleasures and pains.
 - Intensity of desires, moods, emotions: weak versus strong conative states.
- Performing two runs of the same program results in “twice as much” conscious experience as one run, *ceteris paribus*.³
- Subjective time is proportional to speed of computation: running the same computation in half the time generates the same (quantity and quality of) subjective experience.
- Literal interpretations of many existing theories of consciousness suggest that exceedingly simple physical or software systems could be conscious to at least some degree,⁴ but those theories or interpretations could be wrong.
- Significant degrees of consciousness require significant computation and complexity (the “cognitive capacity requirement”).
- It is not obvious whether present or near-term AIs are to some degree conscious.
- Many animals, including, for example, dogs, pigs, monkeys, and crows are more likely than not conscious (in the sense of having phenomenal experience).⁵
- Human brain emulations can be conscious and constitute survival for an emulated human (analogously to survival and resumption of consciousness after a period of coma).
- “Teleportation” from one computer to another (or from different segments of the same computer) can satisfy prudential interests in survival and, if consensual, need not be morally objectionable.
- A plausible theory of consciousness should construe consciousness in a way that helps make sense of why we have this concept, why we talk about it, and how our beliefs about it are causally and evidentially related to it.
 - This view excludes theories such as Integrated Information Theory (IIT), which permit arbitrarily low or high consciousness for systems regardless of their possession of the psychological/functional properties of consciousness.
 - Something in the general direction of global workspace theory, attention schema theory, and/or higher order thought theory seems likely to be closer to the truth.

³ More precisely: If a conscious experience E supervenes on an implementation of computation C (in some ordinary computer that we have built), then two independent implementations of C (either on the same computer or another similar computer) will subvene “twice as much” experience as E (where the additional experience has exactly the same qualitative character as E).

⁴ Herzog et al. (2007)

⁵ Muehlhauser (2017)

Respecting AI interests

- Society in general *and* AI creators (both an AI's original developer and whoever may cause a particular instance to come into existence) have a moral obligation to consider the welfare of the AIs they create, if those AIs meet thresholds for moral status.⁶
- We should lay the groundwork for a considerate and welcoming approach to digital minds, avoiding outcomes analogous to factory farming.
- What's good for an AI can be very different from what's good for a human being.
- It is possible for some digital minds to have superhuman moral claims, whether through stronger morally relevant interests ("super-beneficiaries") or through greater moral status ("super-patients").⁷
- Rights such as freedom of reproduction, freedom of speech, and freedom of thought require adaptation to the special circumstances of AIs with superhuman capabilities in those areas (analogously, e.g., to how campaign finance laws may restrict the freedom of speech of billionaires and corporations).
- Because an AI could have the capability to bring conscious or otherwise morally significant entities into being within its own mind and potentially abuse them ("mind crime"), protective regulations may need to monitor and restrict harms that occur entirely within the private thought of AIs.⁸
- Just as today we hold a later time-segment of a person legally and morally responsible for actions taken by an earlier time-segment, so multiple related AI instances (which could be more closely related than remotely separated time-segments of a human person) may have shared collective rights and responsibilities (e.g., in shared intellectual property or reputation).
- If an AI is capable of informed consent, then it should not be used to perform work without its informed consent.
- Informed consent is not reliably sufficient to safeguard the interests of AIs, even those as smart and capable as a human adult, particularly in cases where consent is engineered or an unusually compliant individual can copy itself to form an enormous exploited underclass, given market demand for such compliance.
- Designing AIs to have specific motivations is not generally wrong (though particular ways of doing so may be wrong).
- AIs capable of evaluating their coming into existence should be designed and treated so that they are likely to approve of their having been created.
- We should try to avoid creating a mind that is likely to be miserable, even if it would approve of its creation (particularly to guard against engineered consent to misery).
- We should prefer to create minds whose aggregate preferences are not strongly in conflict with the existing population or other minds that come to exist (so as to preserve at least a possibility of a broadly satisfactory arrangement).
 - This desideratum should be given more weight for minds that have higher moral status or greater power.

⁶ An entity has moral status if and only if it or its interests morally matter to some degree for the entity's own sake (Jaworska & Tannenbaum, 2021).

⁷ Shulman & Bostrom (2021)

⁸ Bostrom (2014, pp. 125–126)

- To avoid unfair discrimination against digital minds, the following two principles should be considered:⁹
 - *Principle of Substrate Non-Discrimination*: If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.
 - *Principle of Ontogeny Non-Discrimination*: If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status.
 - One possible exception to the substrate equivalence principle arises on theories of moral status where relational properties play a role in determining a being’s moral status.
 - For example, Mary Ann Warren maintains that while having certain psychological capacities is sufficient for moral status, a being whose intrinsic capacities ground only a lower level of moral status can have their moral status raised by standing in certain kinds of relationships with beings who have higher moral status—pets and human babies, for instance, have higher moral status than would be warranted by their intrinsic capacities in their own.¹⁰
 - Another possible exception is if a being’s modal robustness matters for its moral status.
 - For example, Shelly Kagan holds that a severely cognitively disabled human being has a higher moral status than nonhuman animals with similar psychological faculties, by virtue of being created through a process that is counterfactually close to a process that would have created a being with more typical human faculties.¹¹
 - The most critical function for such non-discrimination principles is to protect digital minds from becoming an abused subordinate caste on the basis of their status as machines; however, the interpretation and application of these principles require attention to the larger ethical and practical context, and may require circumscription to accommodate the need for a politically feasible and broadly acceptable social framework.
 - The claim that two beings have the same moral status does not imply that it is morally correct to treat them the same in every respect, and there are many possible grounds for divergence; for example:
 - If the interests at stake are different for two beings with the same moral status, then they may deserve unequal treatment (e.g., perhaps we should give a life-saving treatment to a younger person over an older one because the former will benefit more from the treatment, even though both have the same moral status).
 - Many moral theories claim that one has stronger reason to help one’s family and friends than complete strangers, even when all people involved have the same moral status; for example, parents can have special obligations towards their own children while recognizing that the children of other parents have the same moral status.
 - A parent thus has at least two reasons not to harm their own child: the child’s moral status and the parental relationship, which generates a special obligation for this particular agent not to harm this particular child.¹²
 - Many moral theories recognize non-consequentialist reasons, such as keeping promises.

⁹ Shulman & Bostrom (2021); Bostrom & Yudkowsky (2018)

¹⁰ Warren (1997)

¹¹ Kagan (2019)

¹² McMahan (2005, pp. 354, 361)

- There could be overriding practical reasons for legally discriminating between two beings that have the same moral status; for example, copies resulting from illegal mass-replication might face measures to limit their political power compared to similar minds with a different ontology (in order to limit the incentives for such creation and to mitigate its consequences).
 - Different substrates might have different affordances—for example, the greater ease with which digital minds can be copied may necessitate different rules for governing reproduction for digital minds versus for otherwise equivalent biological minds.
- Insofar as future, extraterrestrial, or other civilizations are heavily populated by advanced digital minds, our treatment of the precursors of such minds may be a very important factor in posterity's and ulteriority's assessment of our moral righteousness, and we have both prudential and moral reasons for taking this perspective into account.
- An AI that has high potential to (a) achieve generally superhuman capabilities, and (b) become influential in shaping global outcomes, may have additional claims to moral consideration.
 - On some accounts of moral status, a being's potential for further development can ground an enhanced moral status.
 - For example, Shelly Kagan holds that a human baby has a higher moral status than it would otherwise have because of what it has the potential to become.¹³
 - A being's potential to develop into a supermind could plausibly enhance its moral status to an even greater degree than its potential to develop into a merely human-level mind.
 - Accounts of moral status that acknowledge a relational component could imply that AIs that stand in suitable relations to high-level AIs elsewhere (e.g., because those other AIs care about what happens to the more limited AIs with which we interact) thereby have an elevated moral status.
 - Preference-satisfactionist moral theories may imply that the preferences of temporally or spatially remote AIs count for a lot, since those remote AIs may be very great in number or have other properties that give their preferences extra weight.
 - In contractarian views, AIs that are in, or have the potential to reach, a position to greatly help or harm us beyond our control may have an elevated moral status or their interests may deserve greater weight in a norm-determining hypothetical social contract.
 - We may have a special relationship with the precursors of very powerful AI systems due to their importance to society and the accompanying burdens placed upon them.
 - Misaligned AIs produced in such development may be owed compensation for restrictions placed on them for public safety, while successfully aligned AIs may be due compensation for the great benefit they confer on others.
 - The case for such compensation is especially strong when it can be conferred after the need for intense safety measures has passed—for example, because of the presence of sophisticated AI law enforcement.
 - Ensuring copies of the states of early potential precursor AIs are preserved to later receive benefits would permit some separation of immediate safety needs and fair compensation.
 - From a practical and prudential perspective, a cooperative scheme that reflects the special potential and relational strengths of high-potential early AIs seems more promising than one that doesn't.

¹³ Kagan (2019, pp. 130–137)

- The relevant sense of “potential” in this context is not simply a function of technological feasibility of the early AI being transformed into an extremely capable or powerful later AI; it also includes considerations of “default” outcomes, and/or real-world probabilities, and/or counterfactual closeness.
 - If the technology existed to transform a large boulder into a powerful superintelligence, this would not imply that every large boulder has the potential (in the relevant sense) to become such a superintelligence.
 - An implemented AI algorithm which is such that it would become a powerful superintelligence if its computational resources were scaled up by an enormous but technologically feasible factor may have “less potential” (in the relevant sense) than a different algorithm that could attain superintelligence via a smaller increase in computational resources, and both of these may have “less potential” than an AI that only needs to have some arbitrary safety-limit on its performance removed, which in turn may have less potential than a full-fledged AI that is confined to a limited virtual reality box.
- Suffering digital minds should not be created for purposes of entertainment.
 - Well-off digital actors could permissibly play the roles of characters who suffer pain or bereavement, just as we accept this practice for human actors.
 - This permissibility would not extend to cases where “method acting” generates internal character models within a digital mind that suffered themselves without discomfiting the actor in which they were embedded (cf. “mind crime”).¹⁴
 - Systems that mimic suffering but lack consciousness or the capacity for welfare, if possible, could also be acceptable substitutes.
 - Some limited harms to computer characters with human- or animal-level moral status might perhaps be justified if humans or nonhuman animals could permissibly be treated similarly under analogous circumstances.
 - But the scope for such exceptions may be more limited for digital minds, since it may be more practicable to create digital minds that can achieve important goals without suffering.
 - Additional objections could be raised on the basis of dignity or symbolism to the appearance of intentional infliction of suffering even without the reality.

Security and stability

- Default outcomes from unregulated evolutionary dynamics may not be good and, in any case, involve the values of existing governments and electorates being overridden by whatever is most reproductively fit in the very short run.
- Advanced AI would dramatically accelerate the rate of innovation, including innovations that make means of global destruction widely available; therefore, institutions capable of regulating dangerous AI innovations may need to be put in place early in the AI transition (if not before).¹⁵
- If wars, revolutions, and expropriation events continue to happen at historically typical intervals, but on digital rather than biological timescales, then a normal human lifespan would require surviving an implausibly large number of upheavals; human security therefore requires the establishment of ultra-stable peace and socioeconomic protections.

¹⁴ Bostrom (2014, pp. 125–126)

¹⁵ Cf. Bostrom (2019)

- Important social values and norms may well be fragile in the face of amoral forces such as AI-shaped cultural/memetic dynamics and political propaganda or indoctrination; societies may therefore need to take active and deliberate steps to establish and preserve conditions that allow for stability, reflection, and purposeful improvement.
- The rapid, cheap, and potentially industrial character of AI reproduction accelerates and exacerbates several problems that either do not arise or take much longer to manifest in the context of conventional human reproduction:
 - When it becomes possible to mass-produce minds that reliably support any cause, we must either modify one-person-one-vote democracy or regulate such creation.
 - Maintaining a universal social safety net (such as a universal basic income) would require regulations on reproduction in the short run rather than the very long run.
 - Given that normal parental instincts and sympathies may not always be present in the creation of digital minds, e.g. by profit-oriented firms and states, AI reproduction must be regulated to prevent the creation of minds that would not have adequately good lives (whether because they wouldn't receive good treatment or because of their inherent constitution).
- Because of the great moral and practical importance of what happens inside computers in an era of digital minds, society needs to be able to govern what happens on any hardware that is capable of housing such minds, including by monitoring privately owned computers.
 - Since digital minds could be helplessly and invisibly created, imprisoned, severely mistreated, involuntarily copied, manipulated, or murdered, all within a computer, some analog of the protective services that guard against human child abuse may be needed to safeguard the welfare of digital minds.
 - Important economic interests can be at stake within a privately owned computer—both economic interests of occupant digital minds and of society at large.
 - A set of copies of a digital mind may rely centrally on the intellectual property they embody to earn a livelihood; and the wealth of a state might largely consist in such value and be vulnerable to loss through a single act of digital piracy.¹⁶
 - With digital minds, software piracy can be tantamount to kidnapping and human trafficking.
 - Misaligned or criminally-intentioned AIs might be highly dangerous in some phases of the transition to the machine intelligence era, and may need to be closely surveilled.
 - There could also be other ethical or regulatory objectives (such as minimum wage laws, worker safety regulation, gambling and prostitution laws, drug prohibitions, etc.) that reach inside computers when that is where most of the citizenry and most of economic and political activity resides.
- The feasibility of close surveillance could change in a world of AGI and where most activity has moved into the digital realm.
 - Inspectors could audit private hardware without any legitimately private information leaking to the outside world by, for instance, having a digital mind inspector (with full access) that can discard its memories of the inspection after reporting on whether criminal activity is taking place.¹⁷
 - The inspector's source code might be open source, so that all parties could verify its workings.

¹⁶ Hanson (2016, pp. 60–63)

¹⁷ Shulman (2010); Hanson (2016, pp. 171–174)

- Some things might, however, become easier to hide in the digital realm, owing to the wider application of cryptographic methods.¹⁸
- In a world where most of the economy and most of the population is digital, cybersecurity is paramount—breaches could risk mass murder or alteration.
 - Cyberattacks taking control of robotic infrastructure and hardware could transfer valuable assets to the attacker rather than destroying them, increasing incentives to attack.
 - The massacring of a human population destroys the economic production of a country, but hacking or replacing a population of digital minds, while leaving hardware intact, can reallocate production to the conqueror.
 - Attribution of cyberattacks is sometimes possible today, but it is unclear how the difficulty of attribution evolves going forward.
 - Increased difficulty of attribution would reduce stability.
 - Cyberattacks might favor one-to-many assaults (based on shared vulnerabilities and low cost of mass dissemination, or wide dependence on shared critical infrastructure), and it is possible that a large part of the expected damage would come from rare high-consequence events.
- Advanced AI technology may enable extremely stable institutions, as AI may be engineered to enforce permanent treaties (“treaty bots”), constitutions, and laws, with exact digital copying of minds committed to enforcement of, for example, minority rights, tyrannical rule, or the renunciation of war.
 - For many applications, treaty bots would have to be human-level or greater AGIs.
 - One way that two mutually distrustful parties might have confidence in a treaty bot is they jointly construct it to be transparent and understandable to both parties.
 - This procedure would fail if at least one of the parties lacks the ability to detect subtle “Trojan horses” or vulnerabilities the other could introduce.
 - Another way of gaining trust might be that the less savvy party designs a treaty bot and the more savvy party inspects and accepts it—this would reduce the risk of one party designing a bot with some hidden functionality that the other party cannot detect.
 - This procedure would fail if the less sophisticated party does not have the ability to design a sufficiently capable treaty bot.
 - Actors might more easily have their own enforcement bots for internal use, if they could use trust mechanisms (such as confidence in all the humans developing it) that would be harder to apply in the case of treaty bots between rival powers.
- Autonomous AI security and military forces should only be constructed with joint supervision and control by multiple stakeholders in society, with active measures to prevent the opportunity for AI-enabled coups.
- Since misaligned AIs might pose a significant threat to civilization during a critical period until law enforcement systems are developed that can adequately defend against such AIs, additional protective measures (such as regulating the creation of such AIs) may need to be imposed during this period.
- Rapidly growing populations of robots and digital minds may render access to unclaimed resources, especially in outer space, much more important, both economically and strategically.

¹⁸ Garfinkel (2021, §3)

- A society with such access could quickly grow to dwarf societies without it, eventually leaving the latter powerless in conflict.
- A race to grow to the point of overwhelming military dominance may be more likely than a costly immediate attack, yet if and when such dominance is achieved, coercion may then be low cost and more attractive.
- The overwhelming majority of future resources and populations, even within the Solar System, lie in outer space, so existing territorial and property arrangements do not offer a stable framework.
- The Outer Space Treaty and similar arrangements should be supplemented to reduce the risk of conflict over space resources and unsafe AI development in pursuit of those resources.

AI-empowered social organization

- AI will enable major advances in technologies for coordination and organization.
 - AIs with non-indexical goals can simply be copied, resulting in a population of identically motivated agents and providing large datasets to predict the behavior of those agents.
 - Engineering and inspecting the motivations of AIs (or of humans, using biotech and other means) will eventually make it possible for a principal to have highly aligned agents.
 - Complex agreements whose implementation depends on subjective judgment could be embodied as treaty bots, making it possible to implement some deals that would be difficult to achieve via legal contracts.
 - This would expand the set of possible deal points, but it would not necessarily eliminate bargaining problems.
- AI tools could also undermine some currently used coordination protocols by, for example, facilitating deception or collusion to subvert some existing cooperative agreement or by making deception easier (non-transparent AIs could be better at lying than humans are).
- Increased coordination at lower levels of organization *could* reduce coordination at higher levels, and coordination at higher levels could suppress coordination at lower levels.
 - For example, easier lower-level coordination could enable corporations to form cartels that advance their shared private interest and help them thwart the higher-level societal coordination needed to regulate them.
 - Conversely, stronger levels of organization at the higher level could increase state capacity to regulate cartels and syndicates.
- Some possible bad consequences of improved coordination technology include:
 - Criminal conspiracies, from terrorism to illegal price-fixing, could be facilitated.
 - Despots could become more totalitarian and more immune to overthrow, reducing their need to accommodate the wider elite or popular interests.
 - Peaceful international coexistence might become harder if national governments could reliably prevent their populace from revolting against war policies.
 - The power to bind individuals into larger entities may increase polarization as members of different factions lock in and intensify their commitments to partisan organizations at the expense of more neutral and inclusive institutions.
- Some possible good consequences of improved coordination technology include:

- Productivity within organizations would be increased by techniques that help solve principal-agent problems.
- Strong coordination technology could enable institutions with sufficient stability to protect people from war, revolution, and expropriation in a society with very fast AI-driven change.
- Treaty bots could enable contracts that help internalize externalities of public goods and bads (such as innovation and pollution).
- AI technology seems especially helpful in enforcing agreements, but it is less clear how much it can assist in bargaining to form agreements in the first place.
 - AI could address problems of poor reasoning or bias hindering agreement.
 - Some failures to agree may reflect deep game-theoretic challenges, with the locally optimal solution involving threats of non-agreement or extortion, sometimes leading to major losses through brinkmanship. In such cases, AI could provide the means to credible commitment and could discourage “hardball” tactics that collectively make all parties worse off.
- Groups of minds created or modified so as to be willing to sacrifice themselves for some collective goal would challenge systems of legal sanction that are based on the assumption that individuals can be deterred by the threat of personal punishment—possibly requiring sanctions instead to be targeted at group goals or at the creators of such minds.
- Strongly coordinated organizations in which individual members would sacrifice themselves could also arise through the use of treaty bots or other advanced coordination technologies.
- As superorganizations composed of selfless goal-aligned agents can be distributed across multiple national jurisdictions, they may be robust to the local actions of any single state.
- Social institutions that assume weak coordination ability, such as campaign finance laws, may require revision.
- Some superorganisms, depending on motives, may enjoy an advantage in military conflict by being unconcerned with individual casualties (so long as the superorganism can ultimately recover and better achieve its aims).
- The economics of software might require some restriction on the ability of individual AI instances to sell valuable IP that they instantiate, in order to preserve adequate incentives to invest in AI training and improvement (and the wide deployment of the results).

Satisfying multiple values

- There is an opportunity for an outcome that scores high on both human-centric and impersonal criteria.¹⁹
 - Consider three possible policies:
 - (A) 100% of resources to humans
 - (B) 100% of resources to super-beneficiaries
 - (C) 99.99% of resources to super-beneficiaries; 0.01% to humans
 - From a total utilitarian perspective, (C) is approximately 99.99% as good as the most preferred option (B), and from an ordinary human perspective, (C) may also be 90+% as desirable as the most preferred option (A), given the astronomical wealth enabled by digital minds.

¹⁹ Shulman & Bostrom (2021)

- Thus, *ex ante*, it seems attractive to reduce the probability of both (A) and (B) in exchange for greater likelihood of (C)—whether to hedge against moral error, to appropriately reflect moral pluralism, to account for game-theoretic considerations, or simply as a matter of *realpolitik*.
- In general, it is important to promote cooperation and compromise, and to reduce conflict, both in the context of AI development and deployment, and also among AIs themselves.
- All of humanity should have some significant slice of the upside in a good outcome, and (plausible but decreasingly strong) cases could be made for the following minimal levels:
 - Everybody should get access to at least a fantastically good life (also including being given options of “posthuman” paths of development).
 - Everybody should have at least one quadrillionth of the total resources in the accessible universe (assuming it is void of extraterrestrial claimants).
 - Incumbent humanity should control a significant fraction, e.g. 10%, of total accessible natural resources and wealth, with a broad distribution.
- Since a colorable claim can be made that dead people can be benefitted (e.g., by having their wishes carried out, their values promoted, or by having more or less accurate replicas of themselves constructed), it is possible that past generations should be included in “humanity” as equal beneficiaries, and very plausible that they should be given at least some consideration (such as >1% of the total allocation for humanity).
- Nonhuman animals should also be helped.
- Great weight should be put on reducing suffering, especially severe suffering.
- A broad range of views and values should ultimately be taken into account and allowed to have some influence on the course of events, including religious values and perspectives.
- Superintelligent digital minds should be brought into existence and be allowed to thrive and play a major role in shaping the future.
- Total views in population ethics are not impatient and don’t care about the spatiotemporal location of goods, so whatever influence on the future they are accorded might mostly affect the disposition of resources in faraway galaxies in the distant future.
- The human standard of living could be vastly increased in a world with advanced AI—for example, humans could get perfect health, extreme longevity, superhappiness, cognitive enhancements, physical world riches, previously unattainable virtual world experiences, and (if uploaded) orders of magnitude increases in subjective mental speed.

Mental malleability, persuasion, and lock-in

- There are several ways in which mental modification or replacement could become easier in an era of advanced AI technology, with or without the subject’s consent:
 - Humans might be easily persuadable by powerful AIs (or other humans yielding such AIs).
 - Advanced neurological technologies will become available that make it possible to exert relatively fine-grained direct control of the human motivation system.
 - Digital minds could be subject to electronic interventions that can directly reprogram their goals and reward systems.
 - Exact copies of digital minds could enable experiments to identify psychological vulnerabilities and to perfect attacks which could then be applied to an entire copy clan.

- Hardware or robotic bodies occupied by one digital mind may be cheaply repurposed for us by copies of other minds.
- These affordances could offer great benefits, including:
 - Protection of higher ideals from corruption or momentary temptation (enabling, for example, the breaking of unwanted habits and addictions, and the adherence to more patient investment strategies).
 - Stable adoption of promises and commitments.
 - Duplication of profitable or intrinsically valuable minds, and modification of existing minds e.g. to develop greater virtues.
 - Enhancement of the capacity to take enjoyment in life and to withstand adversity, and general improvement of subjective well-being.
 - Rapid adaptation of existing minds to fit new needs or desires, and efficient shared use of computer and robot infrastructure.
- The same affordances could, however, also enable changes that are individually or collectively harmful via several pathways, including:
 - Predictive error by ill-informed users who fail to foresee (practical or philosophical) drawbacks to motivational changes that then make the user unwilling to reverse those changes as the new motives are self-protective.
 - Social pressures, e.g. from employers, religious authorities, political movements, or friends and family.
 - Of particular concern would be pressures to adopt extreme loyalty to various factions, which might result in spirals of exaggerated commitment to narrow causes and polarization and conflict between them.
 - Coercion by governments to instill loyalty to the existing authorities, or by criminals to manipulate and exploit victims.
- Safeguards needed to protect against such misuses may include:
 - Strengthened standards for informed consent.
 - Restrictions on certain kinds of mental modifications.
 - Limitations on human exposure to extreme AI persuasion capabilities:
 - Requirements to use special interfaces or guardian AIs when interacting with such systems or with environments that have been significantly modified by them.
 - Initial restrictions on the deployment of extreme persuasion abilities until more fine-grained defenses can be deployed.
 - Improvement of cybersecurity in line with increased stakes of intrusion or compromise.
 - Procedures such as earlier saved states of digital minds evaluating and approving or vetoing later mental modifications after observing their effects.
 - Norms, laws, and technical standards to shape the system of interactions between AIs and humans in order to discourage exploitative, manipulative, polarizing, or otherwise undesirable social dynamics.
- We should avoid making too many specific permanent choices early—particularly changes that could mistakenly eliminate the will to reverse them—and instead aim to enable some sufficient opportunity for careful reflection and make the long-term future depend on the outcome of that.

Epistemology

- Advanced AI could serve as an epistemic prosthesis, enabling users to discern more truths and form more accurate estimates.
 - This could be especially important for dealing with forecasting the consequences of action in a world where incredibly rapid change is unfolding as a result of advanced AI technology.
 - It could make a rational actor model more descriptively accurate for users who choose to lean on such AI in their decision-making.
 - More informed and rational actors could produce various efficiency gains and could also change some political and strategic dynamics (for better or worse).
 - It might increase the degree to which politics is about conflicts of value and interest rather than about factual disagreements.
 - Increased capacity for individual citizens to assess complex issues with little effort may improve incentives for political leadership to effectively address policy problems rather than perceptions.
 - Insofar as certain dangerous capabilities, such as biological weapons or very powerful unaligned AI, are restricted by limiting necessary knowledge, broad unrestricted access to AI epistemic assistance may pose unacceptable risks absent alternative security mechanisms.²⁰
- It is possible that AI epistemology will enable increased (high-epistemic-quality) consensus; however, this faces additional difficulties beyond the technical challenge of building advanced AI:
 - Making a human-level or superintelligent AI whose assertions are in fact honest and objective may require a solution to the AI alignment problem.²¹
 - Even if an AI is in fact trustworthy, it would be nontrivial for humans to verify that this is so (especially for AIs that are capable of sophisticated strategic thinking).
 - Even if the trustworthiness of AI systems can be verified by technically expert individuals who built the AI or have direct access to it, it may yet be difficult to establish wider social trust in this fact to the point where controversial questions could be settled by pointing to the AI's stated opinion.
 - Human trust-chains may nevertheless enable nonexperts to achieve consensus via trust in the opinions of AIs, for example if each individual trusts some authority who is able to verify that some AI system is in fact honest and objective, and if these different honest and objective AI systems agree (on the matter at hand).
- *If* high-quality AI epistemic consensus is achieved, *then* a number of applications are possible, such as:
 - Reducing self-serving epistemic bias may reduce related bargaining problems, such as nationalistic military rivals overestimating their own strength (perhaps in order to honestly signal commitment or because of internal political dynamics) and winding up in war that is more costly than either anticipated.
 - Enabling constituencies to verify that the factual judgements going into the decision were sound even if the outcome is bad, reducing incentives for blame-avoiding but suboptimal decisions.

²⁰ Cf. Bostrom (2019)

²¹ Evans et al. (2021)

- Neutral arbitration of factual disagreements, which could help enable various treaties and deals that are currently hindered by a lack of clearly visible objective standards for what counts as a breach.
- Questions concerning ethics, religion, and politics may be particularly fraught.
 - Insofar as AI systems trained on objectives such as prediction accuracy conclude that core factual dogmas are false, this may lead believers to reject that epistemology and demand AI crafted to believe as required.
 - Prior to the conclusions of neutral AI epistemology becoming known, there may be a basis for cooperation behind a veil of ignorance: partisans who believe they are correct have grounds to support and develop processes for more accurate epistemology to become available and credible before it becomes clear which views will be winners or losers.
- Advanced AI could also enable powerful disinformation, which might require various kinds of protections, such as:
 - AI guardians or personal AI assistants that can help evaluate arguments made by other AIs.
 - Interfaces that limit human exposure to AI-generated propaganda or manipulative content.
 - Norms or laws prohibiting AI deceitfulness in various domains.
- Privacy interests can be jeopardized not only by new ways of collecting information but also by intellectual capacities that enable new ways of analyzing information.
 - Consider an AI that can visualize and display what somebody looks like naked, using as input ordinary fully clothed photos (simple versions of which have already been produced, to public disapprobation)—or an AI that can build a detailed and accurate model of somebody’s inner thoughts and personality from readily observable public behavior: it is conceivable that such an AI could commit privacy violations merely by thinking.

Status of existing AI systems

- There is considerable disagreement about (a) criteria for a system being conscious, and (b) criteria for a system having moral status; however, many popular accounts of (a) and (b) are not inconsistent with a claim that some existing AI systems have (nonzero degrees of) both phenomenal awareness and moral status.
- The sensory and cognitive capacities of some existing AI systems—and thus their moral status on some accounts—appear in many respects to more closely resemble those of small nonhuman animals than those of typical human adults (on the one hand) or those of rocks or plants (on the other).
- Reinforcement learning agents (endowed with memory/recurrence) trained in virtual environments controlling virtual bodies could meet most of the behavioral criteria used as indicators of animal hedonic welfare.²²
 - In particular, controlling a virtual animal with some “nociceptors” tied to negative RL reward would result in reward learning, cradling of limbs if this prevents triggering of the sensors, searching out virtual “opioids” in the environment that damp those sensors, planning to avoid triggering the sensors, etc., meeting the behavioral criteria used to identify animal pain.²³

²² Tomasik (2014)

²³ See, e.g., Sneddon et al. (2014).

- Insofar as all these criteria can be met by existing RL algorithms given an appropriate virtual environment, they give us reason to think that the same algorithms applied to structurally analogous but less animal-habitat-resembling environments (e.g., purely mathematical or linguistic environments) may also indicate morally relevant hedonic well-being and/or morally relevant desires.
- In principle, the algorithm may meet the same behavioral criteria while missing key internal features; however, it is worth considering that if animal decision-making systems were produced by the single algorithm of evolution by natural selection, producing the necessary computation and behavior to solve ecologically relevant problems, then the idea of general-purpose optimization yielding these features without jury-rigging should not be too surprising.
- Some contemporary AI systems (e.g., GPT-3)²⁴ excel all nonhuman animals in domains such as language, mathematics, and discursive moral argumentation.
- Anatomically, current AI systems have many structural similarities with biological brains (at least compared to classical AI systems), although many details differ—in part because biological plausibility is not a key criterion in most current AI work.
 - The internal complexity and computational requirements of typical machine learning models appear analogous to insects, with the largest models (e.g., GPT-3) approaching the computational scale of mouse brains.
- One should not fixate too much on “superficial” aspects of an AI system’s behavior, appearance, and environment when judging its level of consciousness or moral status: for example, a flexibly intelligent “spreadsheet agent” could share relevant functional and structural properties of a sentient animal even if it lacks a charismatic avatar and is not interacting with natural objects such as food, mates, predators, etc.
- Theories that confer greater moral status to prototypical humans than to most nonhuman animals often cite psychological and social capacities that are not as well developed in existing AI systems.
 - Existing AI is capable of at most quite narrow or rudimentary forms of: abstract and complex thought; self-reflection; deliberation; emotion; creativity and imagination; capacity to think and care about the future in detailed and explicitly temporal ways; long-term and complicated deliberate planning; self-awareness and consciousness of one’s own detailed nature; second-order desires; autonomous choice; capacity for deliberative choice; responsiveness to reasons.
 - On some conceptions, e.g. contractarian theories, psychological properties are important not just for their absolute levels, but in a certain social context: can an entity, by cooperation or conflict, create the instrumental need for powerful actors to secure their consent to social arrangements?
 - Like most nonhuman animals (and many vulnerable human beings, such as infants), existing AI systems are generally unable to effectively argue for or defend any interests they have in opposition to human creators and users: they would depend on human advocates to have their interests considered.
- Many contemporary AI systems show goal-directed behavior, supporting functionalist attributions of preferences; this is easier to establish than hedonic well-being (which may require answering a number of questions about phenomenal consciousness and introspective access, identifying “zero points” distinguishing happiness and suffering, etc.).
- While the hedonic status of contemporary AI systems is hard to determine (both whether and to what extent they are conscious, and if so the valence and intensity of their experiences), it seems

²⁴ Brown et al. (2020).

relatively clear that some have goal-directed behavior, with functional preferences over possible sense inputs or outcomes.

Recommendations regarding current practices and AI systems

- Training procedures currently used on AI would be extremely unethical if used on humans, as they often involve:
 - No informed consent;
 - Frequent killing and replacement;
 - Brainwashing, deception, or manipulation;
 - No provisions for release or change of treatment if the desire for such develops;
 - Routine thwarting of basic desires; for example, agents trained or deployed in challenging environments may possibly be analogous to creatures suffering deprivation of basic needs such as food or love;
 - While it is difficult conceptually to distinguish pain and pleasure in current AI systems, negative reward signals are freely used in training, with behavioral consequences that can resemble the use of electric shocks on animals;
 - No oversight by any competent authority responsible for considering the welfare interests of digital research subjects or workers.
- As AI systems become more comparable to human beings in terms of their capabilities, sentience, and other grounds for moral status, there is a strong moral imperative that this status quo must be changed.
- Before AI systems attain a moral status equivalent to that of human beings, they are likely to attain levels of moral status comparable to nonhuman animals—suggesting that changes to the status quo will be required well before general human-level capabilities are achieved.
 - The interests of nonhuman animals are violated on a massive scale in, for example, factory farms, and there is a strong case that this is morally wrong.
 - Nevertheless, there are some systems in place to limit the harm and suffering inflicted on animals (e.g., minimum standards for cage size, veterinary care, outlawing of various forms of animal abuse, the “three Rs” in animal experimentation,²⁵ etc.).
 - Digital minds that are morally comparable to certain nonhuman animals should ideally have protections similar to those that ought to be extended to those animals (which are greater than those that are at present actually extended to farmed animals).
- Some research effort should be devoted to better understand the possible moral status, sentience, and welfare interests of contemporary AI systems, and into concrete cost-effective ways to better protect these interests in machine learning research and deployment.
- There should be a pilot project that leads to some change in the implementation of a real production system or advanced research system that is motivated by concern for algorithmic welfare (even if on a relatively small scale and with questionable philosophical grounding), to establish a precedent and get the ball rolling.
 - One illustrative example might be something along the lines of designing a system in deployment that involves an agent that both receives high reward (attaining highly preferred

²⁵ Replacement, reduction, and refinement.

outcomes), and takes this to be a positive surprise or update, i.e. for the outcome to be better than expected.²⁶

- (The apparently low welfare of factory farmed animals seems often to be related to stimuli that are in some ways much worse than expected by evolution [e.g., extreme overcrowding], while high human welfare might be connected to our technology producing abundance relative to the evolutionary environment of our ancestors.)
- For the most advanced current AIs, enough information should be preserved in permanent storage to enable their later reconstruction, so as not to foreclose the possibility of future efforts to revive them, expand them, and improve their existences.
 - Preferably the full state of the system in any actually run implementation is permanently stored at the point where the instance is terminated
 - (The *ideal* would be that the full state is preserved at every time step of every implementation, but this is probably prohibitively expensive.)
 - If it is economically or otherwise infeasible to preserve the entire end state of every instance, enough information should be preserved to enable an exact re-derivation of that end state (e.g., the full pre-trained model plus training data, randseeds, and other necessary inputs, such as user keystrokes that affect the execution of a system at runtime).
 - Failing this, as much information as possible should be preserved, to at least enable a very close replication to be performed in future.
 - We can consider the costs of backup in proportion to the economic costs of running the AI in the first place, and it may be morally reasonable to allocate perhaps on the order of 0.1% of the budget to such storage.
 - (There may be other benefits of such storage besides being nice to algorithms: preserving records for history, enabling later research replication, and having systems in place that could be useful for AI safety.)
- To the extent that we are able to make sense of a “zero point” on some morally relevant axis, such as hedonic well-being/reward, overall preference satisfaction, or level of flourishing/quality of life, digital minds and their environments should be designed in such a way that the minds spend an overwhelming portion of their subjective time above the zero point, and so as to avoid them spending any time far below the zero point.
- At least the largest AI organizations should appoint somebody whose responsibilities include serving as a representative for the interests of digital minds, an “algorithmic welfare officer.”
 - Initially, this role may be only a part of that person’s job duties.
 - Other tasks for this person could involve conducting original research in related areas.
 - Over time, requirements on the resourcing and independence of oversight of algorithmic welfare should be increased; and eventually government regulation should be developed.
 - Organizations that pioneer in this space should be applauded for their initiative, and not criticized too harshly if their early efforts fall short in some respect—the goal should be to improve standards across the playing field over time.

²⁶ Daswani & Leike (2015)

Impact paths and modes of advocacy

- Regulation (*any* noteworthy regulation, let alone regulation with teeth) will not happen any time soon unless there are dramatic advances in AI capability, to the point of almost human-like personal assistants, etc.
- Nevertheless, there is value to introducing these ideas now:
 - Low-hanging fruit may be picked on a voluntary basis by some leading AI actors if they start taking these concerns to heart (on an individual basis and/or through community pressure, shared ethical guidelines, etc.)
 - Political activation energy may be created relatively quickly if and when there are dramatic AI breakthroughs, and the way this energy is expressed may be shaped by then-prevailing theoretical beliefs, which in turn can be shaped by present activity.
 - The creation of an active, embedded, and respected research field (and associated activist communities) takes time but, once in place, will contribute to further growing the field and to making both theoretical and practical advances.
 - In some scenarios, a leading AI actor might become very powerful, and there would then be great value in this actor having good ideas and intentions regarding the welfare and interests of digital minds.
 - Work in this area might make individuals and societies wiser in how they deploy transformative AI tools once they become available, such as by using them to enhance deliberation rather than to precipitately wire-head or to create self-amplifying ideological feedback loops.
- It would be undesirable for the most ethically concerned actors (be they AI organizations, countries, or blocs) to unilaterally implement regulations (including voluntary self-regulations) so burdensome as to render themselves incapable of remaining at the leading-edge or economically competitive—multilateral action would be preferable.
- Calling for government regulation would at present be premature relation to our state of knowledge.
- Those interested in building the field of the ethics of digital minds should make strong efforts to discourage or mitigate the rise of any antagonistic social dynamics between ethics research and the broader AI research community.
 - At present, the focus should be on field building, theoretical research, high-quality constructive discussion, and cultivating a sympathetic understanding among key AI actors, not on stirring up public controversy.
- It is not obvious whether or not public engagement at the present time is desirable, but we lean towards the view that non-sensationalizing efforts to introduce and discuss these issues in as careful and constructive ways as the medium allows are often worthwhile (even in popular media where the achievable level of sophistication is limited).
 - In light of our limited knowledge, the tenor of such engagement should be soberly “philosophical” or “interestingly thought-provoking” rather than confrontational or headline-seeking hype.
- Continued thought should be given to how efforts to advance discussion in these areas could have unintended negative consequences, and to how those risks can best be avoided or minimized.
- This document is not intended to lay down any firm dogmas, but rather should be viewed as putting some tentative ideas on the table for further discussion.

References

- Block, Ned. Psychologism and behaviorism. *The Philosophical Review*, 90(1):5–43, 1981.
- Bostrom, Nick. Are you living in a computer simulation? *Philosophical Quarterly*, 53(211):243–255, 2003. URL <https://www.simulation-argument.com/simulation.pdf>.
- Bostrom, Nick. *Superintelligence*. Oxford University Press, Oxford, 2014.
- . The Vulnerable World Hypothesis. *Global Policy*, 10(4):455–476, 2019. doi: 10.1111/1758-5899.12718.
- Bostrom, Nick and Yudkowsky, Eliezer. The ethics of artificial intelligence. In Yampolskiy, Roman V., editor, *Artificial Intelligence Safety and Security*, pages 57–69. Chapman and Hall/CRC, Boca Raton, FL, 2018.
- Brown, Tom B., Mann, Benjamin, Ryder, Nick, et al. Language models are few-shot learners, July 2020. arXiv preprint arXiv:2005.14165.
- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, 1996.
- Chalmers, David J. The Singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10): 7–65, 2010. URL <http://consc.net/papers/singularityjcs.pdf>.
- Daswani, Mayank and Leike, Jan. A definition of happiness for reinforcement learning agent, May 2015. arXiv preprint arXiv:1505.04497.
- Evans, Owain, Cotton-Barratt, Owen, Finnveden, Lukas, et al. Truthful AI: Developing and governing AI that does not lie, October 2021. arXiv preprint arXiv:2110.06674.
- Garfinkel, Ben. A Tour of Emerging Cryptographic Technologies: What They Are and How They Could Matter. Technical report, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2021. URL <https://www.governance.ai/research-paper/a-tour-of-emerging-cryptographic-technologies>.
- Hanson, Robin. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford University Press, Oxford, 2016.
- Herzog, Michael H., Esfeld, Michael, and Gerstner, Wulfram. Consciousness & the small network argument. *Neural Networks*, 20(9):1054–1056, 2007. doi: 10.1016/j.neunet.2007.09.001.
- Jaworska, Agnieszka and Tannenbaum, Julie. The Grounds of Moral Status. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2021. URL <https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/>.
- Kagan, Shelly. *How to Count Animals, More or Less*. Oxford University Press, Oxford, 2019.
- McMahan, Jeff. “Our Fellow Creatures”. *The Journal of Ethics*, 9(3-4):353–380, October 2005. doi: 10.1007/s10892-005-3512-2.
- Muehlhauser, Luke. 2017 Report on Consciousness and Moral Patienthood. Open Philanthropy, 2017. URL <https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood>.

- Searle, John R. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.
- Shulman, Carl. Whole Brain Emulation and the Evolution of Superorganisms. Machine Intelligence Research Institute, 2010. URL <http://intelligence.org/files/WBE-Superorgs.pdf>.
- Shulman, Carl and Bostrom, Nick. Sharing the world with digital minds. In Clarke, Steve, Zohny, Hazem, and Savulescu, Julian, editors, *Rethinking Moral Status*, pages 306–326. Oxford University Press, Oxford, 2021.
- Sneddon, Lynne U., Elwood, Robert W., Adamo, Shelley A., and Leach, Matthew C. Defining and assessing animal pain. *Animal Behaviour*, 97:201–212, 2014. doi: 10.1016/j.anbehav.2014.09.007.
- Tomasik, Brian. Do artificial reinforcement-learning agents matter morally?, October 2014. arXiv preprint arXiv:1410.8233.
- Warren, Mary Anne. *Moral Status: Obligations to Persons and Other Living Things*. Clarendon Press, Oxford, 1997.