Risks and Opportunities from Artificial Intelligence

Artificial intelligence (AI) will plausibly be one of the defining technologies of the 21st century. Indeed, some predict that AI will have transformative implications within decades, having at least as profound an impact on human civilisation as the Industrial Revolution (Roser, 2023).

Like many technologies, and indeed like the Industrial Revolution itself, AI is likely to be a source of both opportunities and risks. And if AI does end up being transformative then these opportunities and risks are likely to be particularly dramatic in scale. This suggests that navigating AI well could be one of the most important tasks facing humanity in the coming decades. For this reason, GPI is interested in research that helps us to understand and navigate the largest scale risks and opportunities presented by AI.

A paradigm example of the sort of risk that our research explores is that of AI bringing about human extinction. However, our research also explores other catastrophic risks from AI, including the risk of AI disempowering humanity, leading to vast numbers of deaths, or bringing about a dystopian world. From the other direction, a paradigm example of the sort of opportunity that our research explores is the possibility of AI bringing about a utopian world, but we're also interested in a broader range of transformative opportunities presented by AI.

One motivation for this work is the thought that AI might have significant impacts on the long-term future of humanity, either by bringing about human extinction or by influencing the long term trajectory of human civilisation. See the longtermism section of our core research agenda for a more general discussion of the moral relevance of these long-term considerations.

However, we don't think that one needs to be a longtermist, or needs to be focused primarily on long-term impact, in order to be concerned by the largest scale risks and opportunities presented by AI. For example, such a focus might be justified by standard cost-benefit reasoning (Shulman & Thornley, forthcoming). More generally, the possibility of transformative risks and opportunities from AI is worth consideration from many moral viewpoints. Ultimately, we see this work as complementary to our work on longtermism but not reliant on it.

Artificial intelligence is a new focus area for us and this agenda will likely grow and change as we learn more. With that caveat in mind, here are three research areas that we're particularly interested in.

Topic 1. Catastrophic Risk from AI
AI could bring about a variety of catastrophes, including:

- *Human Extinction*. AI might bring about the extinction of humanity.
- *Human Disempowerment.* AI might disempower humanity in a way that makes the future substantially worse than it would have been if humanity had remained in control.

- *AI-Enabled Dictatorship.* AI might be used by humans to create particularly repressive and long-lived dictatorships.

GPI is interested in work that explores the risk of AI bringing about catastrophes like these. In particular, we are interested in the following strands of work.

- *Threat Modelling.* Threat models are rigorous explorations of a potential risk that might be posed by AI. Such models aim to clarify the magnitude of risks posed by AI. They also aim to clarify the details of what such a risk might look like. Note that here—as elsewhere—we're not assuming any conclusion: we're just as interested in work that carefully makes the case against these risks as we are about work that makes the case for them. Previous examples of threat modelling include Bostrom, 2014; Cotra, 2022; Grace, 2022a; Goldstein & Kirk-Giannini, 2023; Hendrycks et al., 2023; Ngo et al., 2023; Carlsmith, Forthcoming.
- *Modelling Advanced AI.* GPI is interested in models aiming to provide insight into the likely behaviour of future AI systems. For example, we might try to develop decision theoretic models to study the behaviour of individual systems (Bales, 2023; Gallow, 2023; Thornley, 2023b) or game theoretic models to explore the interaction between multiple models (Conitzer & Oesterheld, 2023). Alternatively, we might investigate the extent to which the predictive processing model of the mind can provide insight when applied to AI (Ratoff, 2021). We're also interested in the speed at which AI systems are likely to improve (Chalmers, 2010; Cotra, 2020, 2022a; Thorstad, 2022; Barnett & Besiroglu, 2023; Davidson, 2023) and whether the current deep learning paradigm is likely to yield very capable systems.
- *Characterising Alignment.* Many agree on the importance of AI *alignment*, understood broadly as ensuring that AI systems act in accordance with human values and interests. But what specifically should advanced AI systems be aligned with? Human preferences? If so, how do we handle cases where preferences differ? (Zhang & Conitzer, 2019) Some particular moral theory? If so, which one? (Barrington, 2023; D'Alessandro, 2023) How, if at all, should we account for moral uncertainty? (Korinek & Balwit, 2022)
- *Identifying Deceptive Alignment*: Plausibly, future general-purpose AI systems will exhibit an understanding of the world and their place in it. They will appear to understand (for example) that they are general-purpose AI systems being trained by humans to achieve certain goals. After all, this kind of understanding is plausibly necessary for AI systems to achieve goals effectively. But this understanding might also make possible particularly effective forms of *deceptive alignment*: AI systems merely pretending to have the goals that we want them to have early on, because doing so makes it more likely that these systems will be able to achieve their actual goals at some later time (Hubinger et al., 2019; Cotra, 2022b). Deceptive alignment is a key part of many threat models, but how likely is it? How if at all could we identify when systems are deceptively aligned? And how can we mitigate the risk of deceptive alignment?
- *Assessing and Developing Mitigation Strategies*. Authors have proposed many strategies for mitigating risks from advanced AI systems. These include both strategies for preventing accidents and strategies for preventing misuse. We are interested in assessing these strategies: where might they fall down and how could they be improved? Strategies we might evaluate include governance

proposals like slowing down AI development (Grace, 2022b; Hogarth, 2023), restricting access to hardware (Balwit, 2023), red-teaming AI systems to identify their flaws (Anthropic, 2023; Mislove, 2023), and requiring tests of AI systems' capabilities and alignment (Kinniment et al., 2023). We're also interested in assessing technical proposals including: improving our understanding of models' internals (Olah et al., 2020; Bergal & Beckstead, 2021; Nanda, 2022), developing AI systems that learn human values by observing human behaviour (Hadfield-Menell et al., 2016; Russell, 2019), creating systems that aim to be 'low-impact' (Armstrong & Levinstein, 2017), eliciting latent knowledge from AI systems (Christiano et al., 2021), training AI systems to be truthful (Evans et al., 2021), using less-advanced AI systems to help us reliably oversee the training of more-advanced AI systems (Christiano et al., 2018; Irving et al., 2018; Leike et al., 2018), using AI feedback to train AI systems (Bai et al., 2022), and designing agents that won't resist being shut down (Thornley, 2023a). In addition, we're interested in developing novel mitigation strategies.

- *Political Philosophy and AI Risk.* Some of the risks posed by AI are political in nature, including the risks posed by AI-enabled dictatorships. Other risks will inevitably involve a political dimension, for example with regulation and international agreements playing an important role in enabling or mitigating risks. For this reason, it's likely that political philosophy will be able to provide insight. Questions we're interested in include: Should AI development be left in the hands of private companies? How if at all should our political and economic institutions change if we one day share the world with digital moral patients or agents? Will AI exacerbate and entrench inequalities of wealth and power? Will AI cause mass unemployment? Will AI increase the risk of war between great powers? In each of these cases, how severe is the threat, what can be done to mitigate it, and what are the relevant trade-offs?
- *Cross Cause Comparisons.* Accurate threat models—which characterise the catastrophic risk posed by AI—don't by themselves settle whether mitigating risks from AI should be a priority. To answer this question, we also need to reflect on how moral prioritisation should proceed and on how AI risk mitigation compares to other moral causes. GPI is interested in work that engages in reflection on this question of prioritisation. For example, this might involve reflecting on how the catastrophe risk posed by AI (and the tractability of reducing it) compares with other sources of catastrophic risk (Ord, 2020). Or it might involve thinking about the circumstances under which reducing the risk of catastrophe from AI should be a priority (Shulman & Thornley, forthcoming). Or it might involve reflection on which of the risks posed by AI should be a priority. We might ask (for example) whether accident-risk or misuse-risk is more urgent, whether our focus should be on reducing the risk of extinction or on improving our prospects conditional on survival, and to what extent we should prioritise reducing risks to digital moral patients.

Topic 2. AI and the Trajectory of Civilisation

Even setting aside the possibility of catastrophe, AI might have a dramatic impact on the world and on the future of human civilisation. For example:

- *Rapid Societal and Technological Change.* AI might markedly increase the rate of economic growth (Aghion et al., 2019; Trammell & Korinek, 2020; Nordhaus, 2021), technological development, and societal change. In terms of opportunities, rapid economic growth might lift many out of

poverty and technological development might lead to novel pharmaceuticals and clean energy technologies. In terms of risks, rapid change might involve the invention of new weapons of mass destruction, potent forms of propaganda, and lie-detection and surveillance technologies. Further, rapid change might also have impacts that are harder to assess as positive or negative, perhaps giving rise to a world government or influencing whether humanity settles other solar systems. These things might in turn have large-scale impacts on the future trajectory of human civilisation.

- *Utopia and Positive Lock-In.* As a flipside of the catastrophes discussed in topic 1, AI might be able to bring about a world that is utopian by current standards, allowing for a civilisation filled with flourishing. Less dramatically, AI might lead to substantial improvements in people's wellbeing and flourishing. These improvements might be particularly important if we expect AI to induce *lock-in*, in which certain features of society persist for an extremely long time. Given lock-in, benefits that have a relatively small immediate impact might have a far larger impact once their persistence is accounted for.
- *Harms and Negative Lock-In.* Even if AI doesn't bring about any sort of acute catastrophe, it might have harmful impacts (Acemoğlu, 2022). For example, it might lead to a world of high inequality, low privacy, and low freedom. As with the positive impacts, these detrimental impacts might be particularly important given *lock-in*, which would lead these harms to persist for long periods.

GPI is interested in work that explores the impacts, both positive and negative, that AI could have on the long-term trajectory of civilisation. Among other things, we are interested in the following strands of work:

- *Lock-In.* GPI is interested in work that clarifies the nature of lock-in and the relationship between lock-in and the achievement of a desirable future. We're also interested in work that explores whether AI is likely to bring about various types of lock-in (Karnofsky, 2021; Finnveden et al., 2022). One important-seeming type is *value* lock-in (MacAskill, 2022, Chapter 4): the values instantiated by advanced AI could persist for a very long time. That suggests that it is especially important to get these values right. Unfortunately, there are also many ways in which we might get these values wrong. We might endow powerful AI with the wrong theory of normative ethics, or the wrong theory of welfare, or the wrong axiology, or the wrong population ethics, or the wrong decision theory, or the wrong theory of infinite ethics. Each of these mistakes could make the future significantly worse than it otherwise would be. With what values - if any - should we endow AI? Can we delegate this question to AI itself or otherwise wait to decide?
- *Navigating Rapid Societal and Technological Change*. As noted above, AI might lead to rapid societal and technological change. What can we do ahead of time to mitigate the risks and realise the opportunities? One idea is ensuring that powerful actors agree ahead of time to coordinate in certain ways. For example, actors might agree to share the benefits of AI and to refrain from taking actions that might be irreversible, like settling space and developing dangerous technologies. What sort of agreements would be best? Could humanity bring about and enforce agreements of this kind?
- *Shaping the Future.* If we develop safe AI, it will be important to ask how we might use it to create a flourishing future. This question will be particularly important if AI is highly capable and if lock-in is plausible. For example, this work might explore the role that democratic processes should play in

making crucial decisions about how to make use of AI. Or it might explore whether there's a role to be played by "the long reflection", an extended period during which humanity invests substantial resources into reflecting on the desirable shape for humanity's future (Ord, 2020, ch. 7; MacAskill, 2022, pp. 98–99). Or it might explore how AI could itself play a role in deciding how to make use of AI in shaping humanity's future or in answering any of the other questions in this list.

- *Pace of AI Development.* In the light of various potential benefits and costs of AI, we need to make a decision about how to proceed with AI development. Some of those concerned by risks favour pausing progress in AI until we have a better understanding of the risks and how to mitigate them. Others are sceptical that pauses truly help to address risk, and so instead favour continuing AI progress but with an appropriate focus on making AI systems safe. Meanwhile, some accelerationists believe that we should aim to make rapid progress in AI, in order to accrue the benefits as soon as possible. Which of these views is right, if any? How are we to balance risks and opportunities in deciding how to proceed with developing AI?

Topic 3. Digital Minds

So far, this agenda has largely focused on the implications that AI might have for humans and human civilisation. However, it's possible that in the future AI might itself become a moral patient deserving of our consideration. GPI is interested in work that helps us to determine the likelihood that future AI systems are moral patients and interested in work that clarifies how we should treat AI systems in light of this likelihood (Bowen & Basl, 2020; Liao, 2020; Schneider, 2020; Bostrom & Shulman, 2022). This might include the following strands of work:

- *Mind and Value.* Various issues at the intersection of value theory and the philosophy of mind might be relevant to determining whether AI counts as a moral patient and how we ought to treat AI systems if so. This might include work exploring the nature of consciousness and sentience, work exploring which mental properties are relevant to moral status, and work exploring the nature of wellbeing. See our Mind and Value research agenda for research directions we're interested in here.
- *Downstream Ethical Questions.* Regardless of whether the relevant questions in the philosophy of mind are settled, there will remain downstream ethical questions. For example, we will likely need to know what treatment is appropriate if a being's moral status is uncertain (Schwitzgebel & Garza, 2020). We're also interested in better understanding the ethical ramifications of the risk-mitigation strategies mentioned in Topic 1. Could these strategies harm or else infringe on the rights of AI moral patients? How should we respond if so?
- *The Political Philosophy of Digital Minds.* Digital minds might raise unique challenges for political philosophy. For example, digital minds might be able to duplicate themselves with relative ease, which might raise challenges for integrating them into democratic systems. How if at all should our political systems change in that case?

Bibliography

Acemoğlu, D. (2022). Harms of AI. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A.

Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.65

Aghion, P., Jones, B. F., & Jones, C. I. (2019). Artificial Intelligence and Economic Growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (pp. 237–282). University of Chicago Press. https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-and-economic-growth

Anthropic. (2023). *Frontier Threats Red Teaming for AI Safety*. Anthropic. https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Armstrong, S., & Levinstein, B. (2017). *Low Impact Artificial Intelligences* (arXiv:1705.10720). arXiv. https://doi.org/10.48550/arXiv.1705.10720

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback* (arXiv:2212.08073). arXiv. http://arxiv.org/abs/2212.08073

Bales, A. (2023). Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*. https://doi.org/10.1007/s11098-023-02023-4

Balwit, A. (2023). How We Can Regulate AI. *Asterisk*, *3*. https://asteriskmag.com/issues/03/how-we-can-regulate-ai

Barnett, M., & Besiroglu, T. (2023). *Scaling transformative autoregressive models*. Epoch AI. https://epochai.org/files/direct-approach.pdf

Barrington, M. (2023). *Absolutist AI* (arXiv:2307.10315). arXiv. http://arxiv.org/abs/2307.10315

Bergal, A., & Beckstead, N. (2021). Interpretability. *The AI Alignment Forum*. https://www.alignmentforum.org/posts/CzZ6Fch4JSpwCpu6C/interpretability

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

Bostrom, N., & Shulman, C. (2022). *Propositions Concerning Digital Minds and Society*. https://nickbostrom.com/propositions.pdf

Bowen, J., & Basl, J. (2020). AI as a Moral Right-Holder. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Carlsmith, J. (forthcoming). Existential Risk From Powerseeking AI. In J. Barrett, H. Greaves, & D. Thorstad (Eds.), *Essays on Longtermism*. Oxford University Press.

Chalmers, D. J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9–10), 7–65.

Christiano, P., Cotra, A., & Xu, M. (2021). *Eliciting Latent Knowledge*. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit?usp=sharing&usp=embed_facebook

Christiano, P., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts* (arXiv:1810.08575). arXiv. http://arxiv.org/abs/1810.08575

Conitzer, V., & Oesterheld, C. (2023). Foundations of Cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 15359–15367. https://doi.org/10.1609/aaai.v37i13.26791

Cotra, A. (2020). *Forecasting Transformative AI with Biological Anchors*. https://drive.google.com/drive/u/1/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

Cotra, A. (2022a). Two-year update on my personal AI timelines. *The AI Alignment Forum*. https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines

Cotra, A. (2022b). Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. *AI Alignment Forum*. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/

D'Alessandro, W. (2023). *Is Deontological AI Safe?*

https://drive.google.com/file/d/1UXn0OJj4kYF73spT_VEHEUBzewNXFL4q/view

Davidson, T. (2023). *What a compute-centric framework says about takeoff speeds*. Open Philanthropy. https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., & Saunders, W. (2021). *Truthful AI: Developing and governing AI that does not lie* (arXiv:2110.06674). arXiv. http://arxiv.org/abs/2110.06674

Finnveden, L., Riedel, C. J., & Shulman, C. (2022). *Artificial General Intelligence and Lock-In*. https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88/

Gallow, D. (2023). *Instrumental Convergence?* https://drive.google.com/file/d/1TH2nh5EzwvNgDrqQcMU8oQNj4n9U7x4t/view

Goldstein, S., & Kirk-Giannini, C. D. (2023). Language agents reduce the risk of existential catastrophe. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01748-4

Grace, K. (2022a). Counterarguments to the basic AI x-risk case. *The AI Alignment Forum*. https://www.alignmentforum.org/posts/LDRQ5Zfqwi8GjzPYG/counterarguments-to-the-basic-ai-x-risk-case

Grace, K. (2022b). Let's think about slowing down AI. *The AI Alignment Forum*. https://www.alignmentforum.org/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). *Cooperative Inverse Reinforcement Learning* (arXiv:1606.03137). arXiv. https://doi.org/10.48550/arXiv.1606.03137

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks* (arXiv:2306.12001). arXiv. http://arxiv.org/abs/2306.12001

Hogarth, I. (2023, April 13). We must slow down the race to God-like AI. *Financial Times*. https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems* (arXiv:1906.01820). arXiv. http://arxiv.org/abs/1906.01820

Irving, G., Christiano, P., & Amodei, D. (2018). *AI Safety Via Debate* (arXiv:1805.00899). arXiv. http://arxiv.org/abs/1805.00899

Karnofsky, H. (2021). Weak point in "most important century": Lock-in. *Cold Takes*. https://www.cold-takes.com/weak-point-in-most-important-century-lock-in/

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (2023). *Evaluating Language-Model Agents on Realistic Autonomous Tasks*. https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf

Korinek, A., & Balwit, A. (2022). *Aligned with Whom? Direct and Social Goals for AI Systems* (Working Paper 30017). National Bureau of Economic Research. https://doi.org/10.3386/w30017

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction* (arXiv:1811.07871). arXiv. http://arxiv.org/abs/1811.07871

Liao, S. M. (2020). The Moral Status and Rights of Artificial Intelligence. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0018

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

Mislove, A. (2023, August 29). Red-Teaming Large Language Models to Identify Novel AI Risks. *The White House*. https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/

Neel Nanda. (2022). *A Comprehensive Mechanistic Interpretability Explainer & Glossary*. https://neelnanda.io/glossary

Ngo, R., Chan, L., & Mindermann, S. (2023). *The alignment problem from a deep learning perspective* (arXiv:2209.00626). arXiv. https://doi.org/10.48550/arXiv.2209.00626

Nordhaus, W. D. (2021). Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth. *American Economic Journal: Macroeconomics*, *13*(1), 299–332. https://doi.org/10.1257/mac.20170105

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), e00024.001. https://doi.org/10.23915/distill.00024.001

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

Ratoff, W. (2021). Can the predictive processing model of the mind ameliorate the value-alignment problem? *Ethics and Information Technology*, *23*(4), 739–750.

Roser. (2023). AI timelines: What do experts in artificial intelligence expect for the future? *Our World in Data*. https://ourworldindata.org/ai-timelines

Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Penguin Random House.

Schneider, S. (2020). How to Catch an AI Zombie: Testing for Consciousness in Machines. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0016

Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0017

Shulman, C., & Thornley, E. (forthcoming). How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role. In J. Barrett, H. Greaves, & D. Thorstad (Eds.), *Essays on Longtermism*. Oxford University Press.

Thornley, E. (2023a). *The Shutdown Problem: Two Theorems, Incomplete Preferences as a Solution*. https://s3.amazonaws.com/pf-user-files-01/u-242443/uploads/2023-05-02/m343uwh/The%20Shutdown%20Problem-%20Two%20Theorems%2C%20Incomplete%20Preferences%20as%20a%20Solution.pdf

Thornley, E. (2023b). There are no coherence theorems. *The EA Forum*. https://forum.effectivealtruism.org/posts/FoRyordtA7LDoEhd7/there-are-no-coherence-theorems

Thorstad, D. (2022). Against the singularity hypothesis. *GPI Working Paper*, *No. 19-2022*. https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Against-the-singularity-hypothesis.pdf

Trammell, P., & Korinek, A. (2020). Economic Growth Under Transformative AI. *GPI Working Paper*, *No. 8-2020*. https://globalprioritiesinstitute.org/wp-content/uploads/Philip-Trammell-and-Anton-Korinek_economic-growth-under-transformative-ai.pdf

Zhang, H., & Conitzer, V. (2019). A PAC Framework for Aggregating Agents' Judgments. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 2237–2244. https://doi.org/10.1609/aaai.v33i01.33012237