Submission to the Supporting responsible AI discussion paper

Evan Hockings

I am a PhD student at the University of Sydney working with Prof. Andrew Doherty on quantum computing theory. I am extremely concerned about the existential risk posed by AI systems, which go entirely unmentioned in the discussion paper (***discussion question 2***). Toby Ord, an Australian philosopher and senior research fellow in philosophy at Oxford University, defined existential risks as ones that threaten the destruction of humanity's longterm potential, in *The Precipice: Existential Risk and the Future of Humanity* (2020). Such risks might realise themselves in the form of human extinction, an unrecoverable collapse of society, or an unrecoverable perpetual dystopia. AI existential risk is the most serious form of risk posed by AI systems, and as I have written in Honi Soit (*Appendix A*), the student newspaper of the University of Sydney, these risks demand serious consideration.

I am not alone in thinking this. In *What's the Worst That Could Happen? Existential Risk and Extreme Politics* (2021), Labor MP for Fenner Andrew Leigh discusses existential risks, including those posed by AI systems. Labor MP for Bruce Julian Hill has also spoken about these risks in the House of Representatives. In a 2022 survey of AI experts publishing at top conferences, the median respondent estimated that the probability advanced AI has an extremely bad long-run effect on humanity, such as human extinction, is 5%.

Recently, the Center for AI Safety published a one-sentence public letter on AI risk, saying that 'Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.' Signatories included Geoffrey Hinton and Yoshua Bengio, who won the Turing Award for pioneering the deep neural networks that power cutting-edge AI systems. Notably, they also included Sam Altman, CEO of OpenAI, Demis Hassabis, CEO of DeepMind, and Dario Amodei, CEO of Anthropic. These three companies currently lead the race to develop ever more powerful—and ever more dangerous—AI systems. And they are concerned about what they themselves are doing.

It is as if the CEOs of fossil fuel companies were leading the campaign to bring awareness to climate change, rather than suppress it. This speaks both to their character, and the concerning nature of our current situation. The worst case, as Sam Altman has publicly stated, is 'lights out for everyone'. Governments cannot allow companies to risk literally killing every human on the planet without attempting to impose oversight and regulation.

Nor can governments afford to choke progress in AI with overwhelming regulation. We must chart a course between Scylla and Charybdis. AI systems will drive incredible progress by automating increasingly large amounts of human decision-making, culminating in automating AI research and then, not long after, all human cognitive labour, will drive incredible progress. But these systems must be controllable, acting in alignment with human values, and transparent, allowing humans to interpret their functioning and decision-making process.

We do not know how to make robustly controllable AI systems. An extensive literature discusses concrete problems, unsolved problems, the problem of control from a deep learning perspective, and overviews catastrophic risks. Of particular note are risks from power-seeking AI systems—work from DeepMind suggests that the emergence of power-seeking behaviour in AI systems is probable—and risks stemming from AI manipulation of humans. It has also been suggested that natural selection might favour AI systems over humans.

While it is impossible to empirically demonstrate that AI systems can kill all humans—we would not live to see such evidence—there is a large amount of evidence that AI systems can exhibit undesirable goal-directed behaviours. Specification gaming occurs when AI systems satisfy the literal specification of an objective without achieving the intended outcome.

Particularly concerning is the case where AI systems learn from human feedback, as advanced large language models do today, where AI systems are directly incentivised to deceive the human raters and flatter their biases in order to achieve maximal reward. Researchers from DeepMind have collated an extremely long list of examples of specification gaming in existing AI systems. This is not a problem we know how to robustly solve, either in the context of AI systems or more generally. As Goodhart's law states, 'When a measure becomes a target, it ceases to be a good measure.'

Merely specifying exactly correct goals is not sufficient, as the problem of goal misgeneralisation shows, where AI systems retain their capabilities in novel situations but pursue undesirable goals. Researchers from DeepMind have outlined how this may lead to existential risks.

Not only are we unable to make robustly controllable AI systems, cutting-edge AI systems are neither transparent nor interpretable. A recent and extensive review of work in AI transparency and interpretability is clear that existing work is not engineering-relevant, and that the field of interpretability research needs to grow substantially. As it stands, as Anthropic discussed in a recent document, we need to 'Recognise that regulations demanding interpretable models would currently be infeasible to meet, but may be possible in the future pending research advances.'

To emphasise, the discussion paper suggests mandating transparency requirements (***discussion question 9***), but given current model architectures, these requirements are entirely infeasible. We do not understand how cutting-edge neural network-based AI systems generate their responses, or make their decisions. Meaningful transparency requirements are therefore tantamount to a ban.

Concretely referencing Australia's AI ethics principles, we see that we do not understand how to instil principle 2, 'human-centred values', in AI systems. Nor do we know how to ensure AI systems are built in accordance with principles 3, 'fairness', 5, 'reliability and safety', or 6, 'transparency and explainability'.

I would like to emphasise two excerpts from the discussion paper:

'AI is unique because it can take actions at a speed and scale that would otherwise be impossible. This speed and scale at which AI can be deployed (to generate benefits as well as cause potential harm) is one of the most significant policy challenges prompting calls for greater regulatory action.'

'One challenge for the application of some of Australia's laws is that remedies are often resolved or provided after potential impacts have occurred. While these laws can be an effective deterrent, they can be deficient in certain circumstances. For example, where the impacts from AI are systemic or difficult to reverse. Preventative laws can help to limit problems before they arise.'

If we build powerful AI systems that automate decision-making and widely proliferate them throughout society and the economy, without the ability to control them such that they act transparently and in robust alignment with human values, I believe that humans may lose control of these systems and hence our future. Such an outcome would be definitionally impossible to reverse. These risks therefore demand serious consideration.

The discussion paper does not consider these risks at all. The table in Box 4 lists examples of high risk use cases with 'systematic, irreversible, or perpetual' impacts. The examples include AI surgery robots and self-driving cars. Accidents or misuse here could only harm or kill a small number of people. AI takeover could harm or kill everyone. This is a very qualitatively different risk and ought to be mentioned separately and explicitly.

There are two key pathways for government action. The first is regulation (***discussion question 10***). It should be illegal for companies to deploy AI systems that they believe to have an appreciable risk of AI takeover or of killing all humans. The highest-risk AI systems are those that are capable of automating arbitrary human-performed computer-based tasks and jobs, as the rapid proliferation of these systems poses the greatest risk of AI takeover.

In the report *What a compute-centric framework says about takeoff speeds* (2023), Tom Davidson constructs a model to forecast how rapidly AI systems will progress from automating 20% of human cognitive tasks to all human cognitive tasks. The median prediction is three years; that is, three years between approximately 20% of the economy being automatable, and the entire economy being automatable. This period of unbelievable and unprecedented change would be extremely difficult for governments and the world to navigate without regulatory safeguards set up beforehand. If the AI systems that automate the entire economy are not robustly and precisely directable by their designers, and if their workings are not transparent, interpretable, or understandable, we may find that they bend the entire world economy towards the pursuit of their own goals which, by supposition, have not been designed to be in alignment with human values. This would constitute an AI takeover.

AI takeover must not be allowed to happen. Until these powerful AI systems can be built such that they are robustly and precisely directable by their designers, and until their workings are transparent, interpretable, and understandable, I believe that their training and deployment should be extremely heavily regulated, and perhaps banned outright. But I do not want to emphasise a particular policy prescription; rather, I would like to implore the government to seriously consider the matter.

The second is to massively accelerate research on AI control and alignment, and on AI transparency and interpretability. This represents not only a path towards AI existential safety and risk mitigation, but also a path towards developing Australia's sovereign AI capabilities.

In recent years, the capital expenditure required to build cutting-edge AI systems has ballooned to such an extent that academic labs are simply incapable of training and therefore studying such systems. Only industrial labs, like OpenAI, DeepMind, and Anthropic, backed by billions of dollars, have the funding to work at the cutting edge. Not only do they have the funding, they also have the talent which possesses the large amounts of tacit knowledge required to build enormous AI systems. That talent is in the US and the UK, not Australia. The discussion paper's claim that Australia 'has world-leading research capabilities in AI' fails to grapple with the current nature of the cutting edge.

How, then, can we attract talent to Australia (***discussion question 3***)? All three of these industrial labs—OpenAI, DeepMind, and Anthropic—were founded over AI existential risk concerns. And a large number of their employees are explicitly motivated by such concerns. But this work requires a large amount of money, and developing more capable AI systems is profitable, which brings in funding. Advancing AI existential safety is not profitable; it is a positive externality not primarily captured by those doing the work.

In the status quo, this work is funded by charities like Open Philanthropy. But it is the place of governments to fund this work. For example, the UK recently announced a 100 million GBP fund for an AI taskforce to advance AI safety, partnering with the three leading labs. But DeepMind was already based in the UK. If Australia wants to attract talent away from existing clusters in the US and the UK, it will need a clearer and larger commitment to AI existential safety, and a much larger amount of investment across a wide range of organisations, from universities to existing companies to even startup non-profit research labs.

I hope to inhabit a future where powerful AI systems work to the benefit of humanity. This will not happen for free, or by default—it necessitates serious consideration for the existential risk posed by AI, including and especially by government.

Summary

Below is a summary of my responses to the discussion questions addressed in this submission.

Discussion question 2: What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

- AI existential risks are not covered by existing regulatory approaches.
- It should be illegal for companies to deploy AI systems that they believe to have an appreciable risk of AI takeover or of killing all humans.
    - External risk auditing must also take place to assess this risk.
    - More substantive regulatory action should also be considered.

Discussion question 3: Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

- Australia should provide a large committed fund for AI safety research, committing at least one billion dollars over the next three years, with a focus on AI control and alignment, and AI transparency and interpretability.
    - This fund should support research both inside and outside of academia, with a particular focus on Australian startups and sovereign capability.
    - This funding will attract top talent to Australia: all three cutting-edge AI labs—OpenAI, DeepMind, and Anthropic—were founded over concerns about the existential risk posed by AI.

Discussion question 9: Given the importance of transparency across the AI lifecycle, please share your thoughts on: a) where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI? b) mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

- AI transparency and interpretability will be key for AI systems that are capable of automating arbitrary human-performed computer-based tasks and jobs.
- However, Anthropic recently stated in Charting a Path to AI Accountability that we must 'Recognise that regulations demanding interpretable models would currently be infeasible to meet, but may be possible in the future pending research advances.'
    - Large amounts of funding for AI safety research will be necessary (***discussion question 3***) for transparency in AI systems to even be possible.

Discussion question 10: Do you have suggestions for: a) Whether any high-risk AI applications or technologies should be banned completely? b) Criteria or requirements to identify AI applications or technologies that should be, and in which contexts?

- It should be illegal for companies to deploy AI systems that they believe to have an appreciable risk of AI takeover or of killing all humans.
    - External risk auditing must also take place to assess this risk.
- The highest-risk AI systems are those that are capable of automating arbitrary human-performed computer-based tasks and jobs.
    - There is a substantial chance that these systems will rapidly automate the entire world economy in a timescale predicted to be on the order of three years. This crazy possibility must be taken seriously.
    - If the AI systems that suddenly run the world economy are not robustly controllable, directable, transparent, interpretable, or understandable, they may not act in alignment with human values, disempowering humanity in an AI takeover.

Appendix A

The risks posed by artificial intelligence demand serious consideration

Ensuring that AI systems act in robust alignment with human values is the foremost challenge of our time

Evan Hockings[1]

Amidst the Russian invasion of Ukraine, the risk of nuclear war is now larger than it has been since the end of the Cold War. The spectre of nuclear annihilation, once thought a thing of the past, has returned.

While technology can avert some forms of annihilation, for example by diverting major asteroid strikes, these naturally occurring risks are likely small, evidenced by our long history free from them. The same cannot be said for those caused or exacerbated by technology. Nuclear war, climate change, engineered bioweapons, and even pandemics: these risks are unfortunately all too familiar.

In his book *What's the Worst That Could Happen? Existential Risk and Extreme Politics* (2021), Labor MP for Fenner Andrew Leigh, discusses these risks to our continued existence and how we might mitigate them. But he also worries about another risk not yet listed, a risk that is less familiar and perhaps even more dangerous.

Progress in artificial intelligence (AI) research is accelerating, with the number of new papers doubling every two years. In April, OpenAI released DALL-E 2, a model that generates detailed images from text prompts. While it struggled to generate intelligible text, Google's Parti, announced only two months later, did not struggle at all. And in August,

[1] This is an unedited version of an article published in *Honi Soit* on October 24th, 2022.

StabilityAI freely released Stable Diffusion, allowing anyone to download the model, disable the content filter, and generate images on their own computer. While it might be easy to get swept up in the debates raging around AI generated art, we must remember that what we have now is the barest hint of what is coming.

Language generation is where the true prize lies. Large language models (LLMs) are trained on a significant fraction of all human-produced text to predict the text that is most likely to follow the input text. OpenAI's GPT-3, released in June 2020, was the first LLM to receive significant public attention, even writing an article for *The Guardian*. The numerous applications of GPT-3 include writing university essays and powering GitHub Copilot, a programming assistant which suggests code: AI that hastens AI development.

The reasoning capabilities of these LLMs generally improve when they are made larger and trained on more text. Some flaws persist as these models are superhuman at predicting the text most likely to follow the input, which is not always the same as reasoning well. However, their reasoning drastically improves if we append 'Let's think step by step' to the input text because that makes correct reasoning more likely to follow. What other capabilities are we yet to discover?

In 2021, Jacob Steinhardt, an assistant professor at UC Berkeley, created a forecasting contest to predict AI progress, including a benchmark of high school competition-level mathematics problems called MATH. The aggressive progress forecasted shocked him: state-of-the-art AI in 2021 correctly answered 6.9 per cent of the questions, but the median estimated score in 2025 was 52 per cent. In April, Google announced PaLM, which outperforms the average human on a benchmark designed to be difficult for LLMs. Just two months later Minerva, a version of PaLM specialising in mathematical and scientific reasoning, scored 50.3 per cent on MATH, achieving four years of progress in just one.

But there are problems we fear that will emerge in AI, universal problems of intelligent agents that already manifest in humans, corporations, and states. Different people — agents

— have different goals, and insofar as their goals are misaligned, some degree of conflict is inevitable. The problems associated with quantifying values and goals are encapsulated by Goodhart's law: when a metric becomes a target, it ceases to be a good metric. Surrogation, the process by which these surrogate metrics become targets themselves, is rife in corporations and states, leading them to sacrifice the unmeasured good in pursuit of metrics. Maximising profit or minimising unemployment payments, for example, without regard for the resulting harm.

If these problems emerge in AI, the consequences will be disastrous. AIs have many advantages over human minds. AIs do not necessarily need rest or consciousness, can easily be copied, run on computer processors that are constantly improving and operate at a frequency ten million times faster than the human brain, and can use the entire internet and all recorded human text as training data. So as AI systems develop, their role in scientific, technological, and economic progress will grow as human input and control shrinks in equal measure.

In the future, we will likely construct AI systems that, in any specific but general domain, can reason at least as well as the best humans. Nothing in known science rules this out. And under competitive pressures to maximise profit and secure geopolitical dominance, states and corporations may relinquish more and more control over proliferating but inscrutable AI systems. Eventually, out-of-control AI systems might determine that the most effective way to pursue their unintendedly inhuman values and goals would be to seize control for themselves, executing an AI takeover and permanently disempowering humanity.

To prevent an AI takeover, there are two key problems we must solve. Alignment is the problem of imparting intended values and goals to AI systems, rather than mere surrogates. Interpretability is the problem of understanding how and why AI systems make the decisions that they do. If we solve these problems, we must then robustly align powerful AI systems with values that promote the flourishing of all humans, and indeed all sentient life, using oversight from equally powerful interpretability tools. While we often fail to do this for corporations and states, humans with power within these organisations can attempt to direct them to act in alignment with human values, and human whistleblowers and journalists can render them somewhat interpretable, limiting the resulting harm. But these mechanisms will not be there to save us from AI takeover if alignment and interpretability work fails.

Unfortunately, progress in alignment and interpretability currently lags far behind progress in AI capabilities. And while some organisations at the forefront of capabilities, like OpenAI and DeepMind, have safety teams focused on these problems, enough is not being done. We charge forward recklessly, headed towards disaster.

Many different skill sets will be required to navigate this risk and ensure that AI brings prosperity to all, from philosophy to computer science to politics and governance. To learn more, perhaps to contribute yourself, see the introduction AGI Safety From First Principles and the freely available course materials AGI Safety Fundamentals, both by OpenAI's Richard Ngo. And for a lighter overview, see the Most Important Century series by Holden Karnofsky, co-CEO of Open Philanthropy.

Climate change was once an obscure and neglected issue, as AI takeover risk is today. I hope that you and the world take this risk seriously, as we have begun to do with climate change, because I believe navigating AI takeover risk to be the foremost challenge of our time, and of all time.

Let's get to work.