Pick a transformative technology you think will be created someday. Imagine that you are in a position of responsibility for it. For example, you could be a scientist developing this technology, a government official regulating it, or a corporate executive selling it to the public. How would you approach your job to have the greatest chance of preserving humanity's potential? What are the risks you face?

Thirty years ago, Stephen Hawking warned that "the development of full artificial intelligence could spell the end of the human race."[1] Today, three years after development, Google is publicly announcing the world's first Artificial General Intelligence (AGI), *Metano.* As CEO, I had unique responsibility over both the technical aspects of *Metano's* training and its distribution. It is in this capacity as CEO, I will explain how we thus far avoided Hawking's prophecy, the rewards of AGI and our approach to preserving humanity's potential during *Metano*'s development.

The Risks and Rewards of AGI

Karnofsky defines a transformative technology as one that "precipitates a transition equivalent to (or more significant than) the agricultural or industrial revolutions."[2] We believe that AGI had, and still has, the same (or greater) promise of the industrial revolution, but also the same (or greater) potential peril of the nuclear one.

---

[1] Rory Cellan-Jones, "Stephen Hawking Warns Artificial Intelligence Could End Mankind," BBC News (BBC News, December 2, 2014), .

[2] In fact, Holden Karnofsky's definition of transformative technology was derived from his definition of transformative artificial intelligence. "Transformative Development - EA Forum," Effectivealtruism.org, 2021, https://forum.effectivealtruism.org/topics/transformative-development

As Ord writes, the development of nuclear weapons concentrated control of human potential into the hands (and precarious calculations) of a few military officials and scientists.[3] AGI threatens to spread this power to many, such as by enabling a "doomsday device" like easily manufacturable bioweapons.[4] More pressingly is the problem of AI safety, specifically alignment, given the existence of "orthogonality" and "instrumental convergence."[5] We worried that, by accelerating AGI's development, we left less time for AI safety research to mature.[6] Furthermore, we would be yet another player in a global coordination failure, a race where the *Unilateralist's curse* would favour the most bold, not cautious, developer – exposing all to risk externalities.[7]

Therefore, the first question we asked was: should we do this? In hindsight, *Metano's* impact has been unequivocally positive and transformative. Take the single sector of language translation:

> Between humans, AI-enhanced Google Translate seamlessly captures both meaning and nuance, enabling rich, cross-cultural communication. Between humans and computers,

---

[3] Toby Ord, The Precipice: Existential Risk and the Future of Humanity (Great Britain: Bloomsbury Publishing, 2020), 93-94.

[4] This creates what Nick Bostrom has termed a type 1 "vulnerable world" in which "some technology…is so destructive and so easy to use that…civilizational devastation is extremely likely." Nick Bostrom, "The Vulnerable World Hypothesis," Global Policy, September 6, 2019, 458.

[5] Orthogonality and instrumental convergence broadly means, respectively, that an AGI could conceivably pursue any goal and that could pursue any number of potentially harmful instrumental goals (such as self-preservation, goal content integrity, resource-acquisition, etc.) to achieve this.

[6] As Ord notes, "in the real world people tend to develop technologies as soon as the opportunity presents itself and deal with the consequences later." (Ord, 2020, 151)

[7] See Ord, 2020, 137-138; Bostrom et. al. pose an illustrative example: *"A little girl in a village in Azerbaijan, who has never heard about artificial intelligence, would receive her share of the risk from the creation of machine superintelligence."*
Nick Bostrom et al., "Policy Desiderata for Superintelligent AI: A Vector Field Approach the Prospect of Radically Transformative AI" (Oxford University Press, 2019), 9.

> AI-powered programming and software development has enabled widespread human creation and entrepreneurship. Finally, *Metano*, coupled with brain-image analysis, has enabled *cross-species communication,* instantly rendering meat-eating taboo and causing a Copernican shift in human values – we were no longer the centre of Earth life.

This was not simply incremental improvement but radical transformation in sectors as diverse as academic research, education, energy and transportation. However, that this positive effect would occur was not always guaranteed.

Beyond profit incentives, a few factors persuaded us to pursue *Metano's* development. We knew the largest AGI risks came from deliberate misuse or mistakes. The former directly contradicted our core belief in the social good (our unofficial mission is "don't be evil") and our significant technology lead shielded us from the latter, carelessness to gain a competitive advantage. Next, if we were successful in *safely* developing AGI, we would have a singular impact on humanity's trajectory, not only decreasing AGI's direct existential risks to humanity's potential but may also ensure our fulfilment of it. AGI could enable radical improvement in health and subjective wealthbeing, even causing cornucopia, increasing GDP by several orders of magnitude.[8] Finally, after years of research, we were as confident as we could be in our approach to AGI safety, which I will explain.

[8] Bostrom et al., 2019.

Our Approach

I advocated for a *minimax strategy* to AI safety, minimising the greatest risk to all participants. This is why we didn't dismiss safety concerns, despite many arguments from safety sceptics. Some predicted computational resources would bottle-neck any country or groups' development of AGI, limiting information hazards. While this was true initially – *Metano* itself benefited greatly from Google's hardware – it does not ensure AGI alignment or control. In fact, after *Metano's* software breakthroughs, we've transitioned into a period of hardware overhang. Others appealed to moral fallibilism, that any AGI would discover objectively true moral values that humans have not, and act accordingly.[9] Some even speculated intelligence will emerge from another paradigm, whole brain emulation (or augmentation), and AGI alignment would emerge as *empathy,* profoundly identifying with human emotional and moral experiences.[10] Unsurprisingly, both moral fallibilism and alignment as empathy have proven idealistic.

We therefore need a concrete strategy to safely develop and deploy AGI safely. This involved making careful decisions surrounding information diffusion, software architecture and training environment, early deployment areas, and long-term concerns.

Despite norms of scientific openness, we decided to keep the development of *Metano* secret. We knew of information hazards and that models of AI development showed, counterintuitively,

---

[9] This is part of the motivation behind projects such as the Allen Institute for AI's development of *Delphi* which, through its imperfect and often arbitrary decisions, ultimately, was "a reminder that the morality of any technological creation is a product of those who have built it." "Can a Machine Learn Morality?" The New York Times, November 19, 2021. https://www.nytimes.com/2021/11/19/technology/can-a-machine-learn-morality.html; We can find similar reasoning in the entries to FLI's WorldBuilding Contest such as in entry 165 or 282.

[10] For example, see entry 415 and its discussion of Brain-AI interfaces.

"increasing the information available to all the teams (about their own capability or progress towards AI, or that of the other teams) increases the risk."[11]

While I've referred to *Metano*, singular, in fact *Metano* is a collective AGI. We first trained a single AGI, then made many copies. We allowed AGIs to communicate in a limited, simulated, environment, observing emergent intelligent behaviour as a result of *multi-agent autocurricula*.[12] While all agents were created with heterogeneous objective functions, these behaviours, such as gathering resources or building coalitions, developed as a result of shaping the autocurricula towards an obedience (human instruction following) approach. This approach had important practical and safety advantages.

Early on, we suspected collective AGI would be more likely to succeed than other intelligence paradigms. Duplication is more computationally efficient than retraining, benefiting collective AGIs. Furthermore, we were aware that intelligence may have developed as part of a social arms race.[13] This would make collective AGI *superadditive*, the sum of the coordinating players greater than the individual.

Collective AGI was safer, too. We adopted an AGI "sandboxing" architecture where *Metano,* as a collective, would only work on high-level intelligence tasks (which we would communicate

---

[11] For more information about information hazards, see Ord, 2020, 137-138; Stuart Armstrong, Nick Bostrom, and Carl Shulman, "Racing to the Precipice: A Model of Artificial Intelligence Development Racing to the Precipice: A Model of Artificial Intelligence Development," 2013.
[12] Joel Z Leibo et al., "Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research," ArXiv.org, 2019.
[13] See Dunbar's social brain hypothesis. Robin I. M. Dunbar, "The Social Brain Hypothesis," Evolutionary Anthropology: Issues, News, and Reviews 6, no. 5 (1998): 178–90.

through the process of an oracle) such as fundamental physics or language problems.[14] Smaller, newly created subsets of *Metano's* agents would create more concrete plans for real-world implementation, such as developing specific translation engines, and then were promptly deleted. Firstly, this partitioning approach first allowed us for *capability control,* reducing *Metano's* ability to gain control or harm humans. Secondly, this also led to high interpretability as multi-agent interaction, through a common language, tended to be low-bandwidth. Next, there was less agency/goal-direction in any individual action, freeing us to focus all our resources on making extremely secure "sandboxes."

After this training phase, we devised two broad areas of initial deployment. First, we set *Metano* on AI alignment. Many of alignment's challenges were technical; *Metano* made significant progress through, among other things, publishing a number of successful algorithms for eliciting latent knowledge, generating adversarial training examples, developing safely interruptible agents and inventing new techniques for formal verification. Second, as Ord writes, reducing existential risk requires unprecedented institutional strength and foresight.[15] We therefore collaborated with governments to strengthen institutions, primarily by automating and augmenting government functions through making prediction, planning and administration of public services cheap, accurate and scalable. Furthermore, we expect AGI to play a greater role

[14] Richard Ngo, "Safer Sandboxing via Collective Separation - AI Alignment Forum," Alignment Forum.org, September 10, 2020.
[15] Beyond this, Ord sets a high bar for institutions. He writes that addressing long-term risks will "require institutions with access to cutting-edge information about the coming risks, capable of taking decisive actions and with the will to actually do so…[requiring] swift coordination between many or all of the world's nations." (Ord, 2020, 196)

in budget allocation, suggesting spending priorities, and creating a virtuous cycle, by encouraging governments to better prioritise existential risks.[16]

Our job isn't done. As *Metano* and AGI advances, there still remains difficult questions about broader social impact, such as around fulfilment and wealth reallocation. But we are optimistic. My predecessor, Sundar Pichai, once remarked that AI's development will be "more profound than… electricity or fire."[17] Through this announcement, we hope to spark a broader conversation among scientists, officials and the public and herald the modern era as one where AGI truly began to transform society, safely.

Word Count: 1200.

[16] For example, as Ord suggests, boosting the budget of the Biological Weapons Convention, strengthening the WHO, restarting arms reduction, and participating in multilateral agreements and organisations to decrease overall existential risk. (Ord, 2020, 202-205)

[17] Cat Clifford, "Google CEO: A.I. Is More Important than Fire or Electricity," CNBC (CNBC, February 2018), .

Works Cited

Bostrom, Nick. “The Vulnerable World Hypothesis.” Global Policy, September 6, 2019. https://doi.org/10.1111/1758-5899.12718.

Bostrom, Nick, Allan Dafoe, Carrick Flynn, Stuart Armstrong, Michael Barnett, Seth Baum, Dominic Becker, et al. “Policy Desiderata for Superintelligent AI: A Vector Field Approach the Prospect of Radically Transformative AI.” Oxford University Press, 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Policy-Desiderata-in-the-Development-of-Machine-Superintelligence.pdf.

Cellan-Jones, Rory. “Stephen Hawking Warns Artificial Intelligence Could End Mankind.” BBC News. BBC News, December 2, 2014. https://www.bbc.com/news/technology-30290540.

Clifford, Cat. “Google CEO: A.I. Is More Important than Fire or Electricity.” CNBC. CNBC, February 2018. https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html.

Dunbar, Robin I. M. “The Social Brain Hypothesis.” Evolutionary Anthropology: Issues, News, and Reviews 6, no. 5 (1998): 178–90. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8

Effectivealtruism.org. “Transformative Development - EA Forum,” 2021. https://forum.effectivealtruism.org/topics/transformative-development.

FLI Worldbuilding Contest. “Finalists - FLI Worldbuilding Contest,” May 31, 2022. https://worldbuild.ai/finalists/.

Leibo, Joel Z, Edward Hughes, Marc Lanctot, and Thore Graepel. “Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research.” ArXiv.org, 2019. https://doi.org/10.48550/arXiv.1903.00742.

Ngo, Richard. “Safer Sandboxing via Collective Separation - AI Alignment Forum.” Alignmentforum.org, September 10, 2020. https://www.alignmentforum.org/posts/Fji2nHBaB6SjdSscr/safer-sandboxing-via-collective-separation.

Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity.* Great Britain: Bloomsbury Publishing, 2020

The New York Times. “Can a Machine Learn Morality?,” 2022. https://www.nytimes.com/2021/11/19/technology/can-a-machine-learn-morality.html.