The Man Who Thought He Could Keep AI Safe

Sebastian Mallaby

At first glance, Demis Hassabis, a co-founder of the DeepMind AI lab, seems a familiar type: the missionary entrepreneur and out-of-the-box scientist who emerges as the right person for a particular moment. In this case, that moment was when hardware and software and data aligned to make superintelligence possible. But Hassabis is hardly a conventional figure. He has devoted his life to creating a technology that he thinks has the potential to destroy the world.

Hassabis agreed to talk with me about his quest, because he believes that societies will never trust inventors unless they understand what makes them tick. For almost three years, as I worked on my book *The Infinity Machine*, we met regularly at a pub near his home, in North London. We would climb a shabby wooden staircase to a room on the second floor, sit with cappuccinos under a once-grand chandelier, and spend two hours talking: me with an obsessively detailed list of topics to get through; Hassabis with his sparky riffs on intelligence and life, computer science and neuroscience, philosophy and movies. Through every conversation, the question of motivation hung in the air like the image of the mushroom cloud over Los Alamos.

Hassabis is fluent in the full gamut of AI doom scenarios. He met one of his DeepMind co-founders, Shane Legg, at a lecture on AI safety. He buttonholed his first financial backer, Peter Thiel, at a Singularity summit, where futurists shared visions of machines that outsmart people. Once, Hassabis impressed a bumptious rocket builder, Elon Musk, by telling him that space colonization would not ensure human survival. Superintelligent systems could also build rockets, Hassabis observed. If the systems turned out to be malign, a colony on Mars would offer no protection.

Raffi Krikorian: The validation machines

When Google bought DeepMind in 2014, Hassabis showed how seriously he took the risks of his mission. He made the sale conditional on Google's agreement to controls over his technology, refusing to allow a for-profit corporation to have the final say on AI's rollout. He insisted on an outside board of independent sages to oversee the process, and said that military applications of AI would be strictly prohibited. Ever since he was a teenager, Hassabis had been determined to build powerful AI. He could justify his life's work only if he ensured its safety.

Google's negotiators had never encountered demands like these. A small start-up in London was suggesting that Google pay hundreds of millions of dollars for an asset that it wouldn't control completely. But Hassabis had assembled the world's best AI brain trust. Google accepted his conditions.

Hassabis also had ideas about the physical setting in which superintelligence should be invented. He imagined absconding with his top researchers to a clandestine bunker, hidden from enemies and removed from worldly distractions. Philosophers and scientists from academia would join the effort, and a dream team from around the world would midwife powerful AI on behalf of all humanity. Implicit in this vision was the so-called singleton scenario. A single scientific outfit, presumably with Hassabis at its command, would ensure safe superintelligence.

One day, I met a researcher who had joined DeepMind a bit after the sale to Google. During the researcher's final interview, Hassabis had warned him that if he signed on, he should prepare for a climactic endgame when he would disappear into a bunker.

"Was the bunker just a metaphor?" I prodded.

No, he said. "At any stage, when I was at DeepMind, if Demis had told me to get on a flight to a secret location in Morocco, I would have felt that I had been given fair notice."

"Why Morocco?" I asked.

"Oh, the desert. I was just thinking about the Manhattan Project. That was in a desert."

In 2015, seeking to put flesh on Google's promise of an AI-oversight board, DeepMind arranged a secret gathering of philosophers and technologists. To lock in potential rivals, and to promote his singleton vision, Hassabis granted Elon Musk the honor of convening the meeting at Musk's headquarters in Hawthorne, California. But the gambit backfired. The gathering marked the moment when Hassabis's safety vision began to crumble.

Musk listened to presentations from Hassabis and his co-founders. Then he did the opposite of what they wanted. Teaming up with Sam Altman, Musk founded OpenAI, an explicitly anti-Google, anti-DeepMind venture. Later, believers in the singleton vision described this moment as the "fall": the serpent had brought evil into the garden. But the fall was inevitable, given human nature. When confronted with the prospect of a Promethean technology, people do not coalesce into a singleton effort. They are disputatious, jealous, and tribal.

At this point, Hassabis might have paused to reconsider. If the singleton scenario was naive, how could AI developers avoid a competitive race over the precipice? Besides, now that Musk had copied DeepMind's ideas, Google understandably refused to allow more rivals into the tent by holding a second oversight meeting. But rather than hit pause on his endeavors, Hassabis went into overdrive. In 2016, DeepMind produced a system that defeated a top human player at the ancient board game Go. Timelines for machine supremacy shortened.

To balance this advance, Hassabis came up with new ideas to make AI serve humanity. Together with his co-founder Mustafa Suleyman, he set out to negotiate a fresh set of governance safeguards with Google. To advance this secret “Project Mario,” he hired a top-notch legal team, secured a pledge of $1 billion from Reid Hoffman, the founder of LinkedIn, and considered spinning DeepMind out of Google if he was denied control over his technology. At the same time, Suleyman led a DeepMind effort to help Britain’s National Health Service manage acute kidney disease. If AI could be independently governed, and if it could improve health outcomes for ordinary Britons, Hassabis could feel satisfied that his quest was indeed virtuous.

Both initiatives fizzled. The governance fight with Google dragged on for three years, and illustrated the difficulty of grafting nonprofit oversight onto a for-profit company. The project to help the NHS met with a backlash from privacy campaigners, who were enraged that a subsidiary of a U.S. tech giant might get its hands on patient data. By 2019, Hassabis and DeepMind had retreated on both fronts. Suleyman was ejected from the company.

Hassabis’s next attempt at virtuous AI involved science. If DeepMind couldn’t help the NHS directly, it could aspire to accelerate drug discovery. In 2020, DeepMind unveiled an astounding system that mapped the shapes of microscopic proteins, a triumph for which Hassabis and a colleague shared the Nobel Prize in Chemistry. The award illustrated AI’s considerable upsides. But it didn’t negate its equally real dangers.

In 2022—coincidentally, just a week after I had pitched Hassabis on speaking with me for my book project—OpenAI released its conversational companion, ChatGPT, the most viral product in tech history. This presented Hassabis with another moral test. In theory, he could have left the chatbot business to his upstart rival, sticking to the high road of beneficial science. But Hassabis is the most furiously competitive person I have known, and conceding the chatbot market was in any case out of the question for Google. I remember visiting Hassabis in early 2023. “This is wartime,” he said.

The upshot has been a terrifying capitalist contest in which Hassabis’s Gemini system battles ChatGPT and a handful of other rivals. Hundreds of billions of dollars are fueling the rush, and although some contenders strive earnestly to make their models safe, neither national regulations nor corporate-governance structures prevent others from racing to the bottom. In a reversal of its earlier posture, Google is now eager to ply the American defense establishment with AI. This is the opposite of the superintelligence endgame that Hassabis wanted.

From the April 2026 issue: Inside the dirty, dystopian world of AI data centers

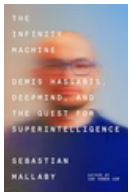
Hassabis acknowledges the setbacks, and yet he keeps racing forward. His faith in governance mechanisms now shattered, he has come to see salvation, paradoxically, in his own career advancement. He knows that his personal goal is to shape AI for the good of humanity. His new safety agenda therefore involves securing personal influence.

“Safety isn’t about governance structures,” Hassabis now says. “I mean, even if you have a governance board, it probably wouldn’t do the right thing when it came to the crunch.” Rather, his goal is to earn “a seat at the table, so when a safety issue comes up,” he can help decide the solutions.

“Things aren’t black and white, especially when you are dealing with a technology with unknown consequences,” he told me. “So you have to be adaptable. You have to move from idealist to realist, but hopefully still with your values.”

I thought at length about this verdict. On the one hand, Hassabis has compromised his original values. On the other, he is right about the need to move beyond his youthful vision. Once multiple labs in multiple countries joined the charge to build powerful AI, they found it impossible not to race. The presence of one well-meaning individual at the proverbial table offers a flimsy scaffolding of reassurance to an alarmed world. But until national governments impose restraint, well-meaning individuals may be the best comfort available.

This article has been adapted from Sebastian Mallaby’s new book, The Infinity Machine: Demis Hassabis, DeepMind, and the Quest for Superintelligence.



By Sebastian Mallaby
