Inside the Feud Over OpenAI

Continued from page B1

calling OpenAI "mendacious" and saying "these facts suggest a pattern of behavior that I've seen often from Sam Altman."

The divide between the two giant AI companies—both are valued at more than $300 billion—has its roots in debates that Amodei and others had in a San Francisco townhouse a decade ago. As both companies hurtle toward an IPO, the philosophical and personal differences between their leadership have metastasized.

At an AI event in New Delhi in February, Indian Prime Minister Narendra Modi and the assembled tech leaders closed by joining hands and raising them above their heads. Amodei and Altman opted out, awkwardly touching elbows.

This account of the rift between the leaders of Anthropic and OpenAI is based on interviews with current and former employees at both companies and people close to the leaders.

The house on Delano Avenue

The tension in the relationship between the founders of OpenAI and Anthropic began in 2016 at a group house on San Francisco's Delano Avenue.

Dario Amodei lived there with his sister, Daniela Amodei. The siblings had grown up in the Bay Area. After getting a doctorate in biophysics, Dario Amodei was working as an AI researcher at Google. Daniela was a young executive at payments startup Stripe.

Brockman, a prolific programmer and one of OpenAI's co-founders, was friends with Daniela and began hanging around the house. They had met at Stripe where he was one of the earliest employees after dropping out of Harvard and then MIT. Brockman tried unsuccessfully to recruit Dario for OpenAI's founding team when it launched in 2015.

Daniela's fiancé, Holden Karnofsky, also lived in the house. Karnofsky was the founder of a philanthropy that promoted effective altruism, a movement that was one of the first communities to take the potential power, and danger, of AI seriously. Through Karnofsky, Brockman became interested in some of those ideas.

One day in early 2016, Brockman, Dario and Karnofsky were sitting around the house, debating the right way to build AI. Brockman argued if the technology was indeed going to change everyone's life as much as they all thought it might, its makers needed to inform Americans about what was coming.

Sam Altman, center, and Dario Amodei did not hold hands at a February event with Indian Prime Minister Narendra Modi. Below, Daniela Amodei.

Dario and Karnofsky said it might not be a good idea to broadcast the most bullish views of what AI might be about to do. Dario argued that when it came to sensitive topics like how fast AI was developing, it would be better to tell the government first.

Brockman took away from the exchange the belief the duo didn't want to tell the public about what was happening. Years later, Brockman came to think the exchange illustrated a core difference in the philosophies of OpenAI and Anthropic.

Early traumas

By mid-2016, impressed by OpenAI's talent roster, Dario had joined the lab, staying up late with the famously nocturnal Brockman training AI agents to solve videogames. By 2017, one of its early projects, called Universe, which aimed to train AI agents to play games and use computers like humans, was floundering.

Musk, OpenAI's then principal financial supporter, asked Brockman and Chief Scientist Ilya Sutskever to make a spreadsheet listing every employee and what contribution they had made—a classically Muskian precursor to staff cuts.

Dario was horrified as he watched his colleagues be fired one by one, which he considered needlessly cruel. In the end, between 10% and 20% of OpenAI's staff of 60 lost their jobs, including one who would go on to co-found Anthropic.

In the fall of 2017, Dario hired an ethics and policy adviser who gave a presentation to OpenAI leadership about how the nonprofit lab could be a coordinating entity among other AI companies, and ultimately between those companies and the U.S. government, to get something like an international coordination regime for advanced AI.

Brockman saw within the presentation the seed of a fundraising idea: OpenAI could sell artificial general intelligence to governments. When Dario asked which governments, Brockman said it would be to the nuclear powers that made up the UN Security Council so as not to destabilize the world order. The idea was briefly batted around the organization.

The notion of selling AGI to rival powers such Russia and China struck Dario as tantamount to treason, and he considered quitting.

Conflicting promises

In early 2018 Musk exited OpenAI. Altman stepped into the leadership void. He met with Dario, and together they agreed that the lab's employees didn't have faith in Brockman and Sutskever's leadership, given the layoffs.

Dario agreed to stay so long as Altman promised Brockman and Sutskever wouldn't be in charge. Altman agreed.

Dario soon learned Altman had made a promise he felt conflicted with that agreement. During a meeting about reporting structure, Brockman mentioned Altman told him and Sutskever they could fire Altman if they ever thought he was doing a bad job.

Brockman recruited Daniela to OpenAI, where she became a jill-of-all-trades, working on engineering management and recruiting. She had married Karnofsky, who was on the OpenAI board, the prior year.

Tensions flared after OpenAI researcher Alec Radford laid the groundwork for large language models and the lab's Generative Pre-Trained Transformer, or GPT, series. Brockman wanted a role in this new direction of research, but Dario, who was research director at the time, wanted him nowhere near it. Brockman appealed to Altman, who lobbied to allow Brockman to work on the project.

Daniela, who was coleading the language project with Radford, told Brockman he couldn't work on it. When Altman asked her if there was any way to make it work, she offered to step down as head of the project rather than allow Brockman onto it.

During a subsequent staff meeting, Dario listed numerous reasons why Brockman shouldn't be allowed to work on the language project, including that Radford didn't want to work with him. Radford was mortified. Brockman and Altman reluctantly agreed to keep Brockman off the language project.

Fight after fight

Dario's profile at OpenAI grew as he and his team launched GPT-2 and GPT-3, but he didn't always feel properly recognized for his contributions.

One such slight came in 2018. Brockman asked Dario to double-check a fact on one of his slides for an important meeting. Dario asked who the slides were for. When Brockman said he and Altman were going to meet former President Barack Obama, Dario got angry he had been left out of the loop.

The following year, Dario asked for a promotion. Altman agreed and sent an email to the board saying Dario would report directly to him and receive equal PR treatment to co-founders.

Altman's November 2019 email also told the board that Dario agreed "not to denigrate projects he doesn't believe in that others want to bet on."

The truce was short-lived. A few months later, the OpenAI leaders had a blowup in the office.

Altman called Dario and Daniela into a conference room and accused them of plotting against him by encouraging colleagues to send negative feedback about him to the board. The brother and sister denied it.

Altman told them he had heard about it from a top OpenAI executive. Daniela called that executive into the room, who said she had no idea what Altman was talking about. Altman then denied that he had said it, prompting the Amodeis to begin shouting angrily.

The final split

Toward the end of 2020—with Covid having pushed everyone into video chat boxes—a group coalesced around Dario to break off and form their own company.

Altman went over to Dario's house to ask him to stay. Dario said he would accept nothing less than reporting directly to the board. He also said he couldn't work with Brockman.

During his final weeks, Dario—who had become known for his lengthy technical memos—wrote a long memo outlining two types of AI companies: Market companies and public-good companies.

Market companies like OpenAI thought they would make the world better by building and selling products to benefit people, including, eventually, AGI.

The public-good company, he argued, would conduct safety research and address various dangers and opportunities of AGI.

Dario wrote that the ideal mix would be 75% public good and 25% market.

Weeks later, Dario, Daniela and nearly a dozen other employees had left OpenAI. Within five years, they would be lining up banks for Anthropic, racing to IPO before their former employer.

LUDOVIC MARIN/AFP/GETTY IMAGES; DAVID PAUL MORRIS/BLOOMBERG NEWS