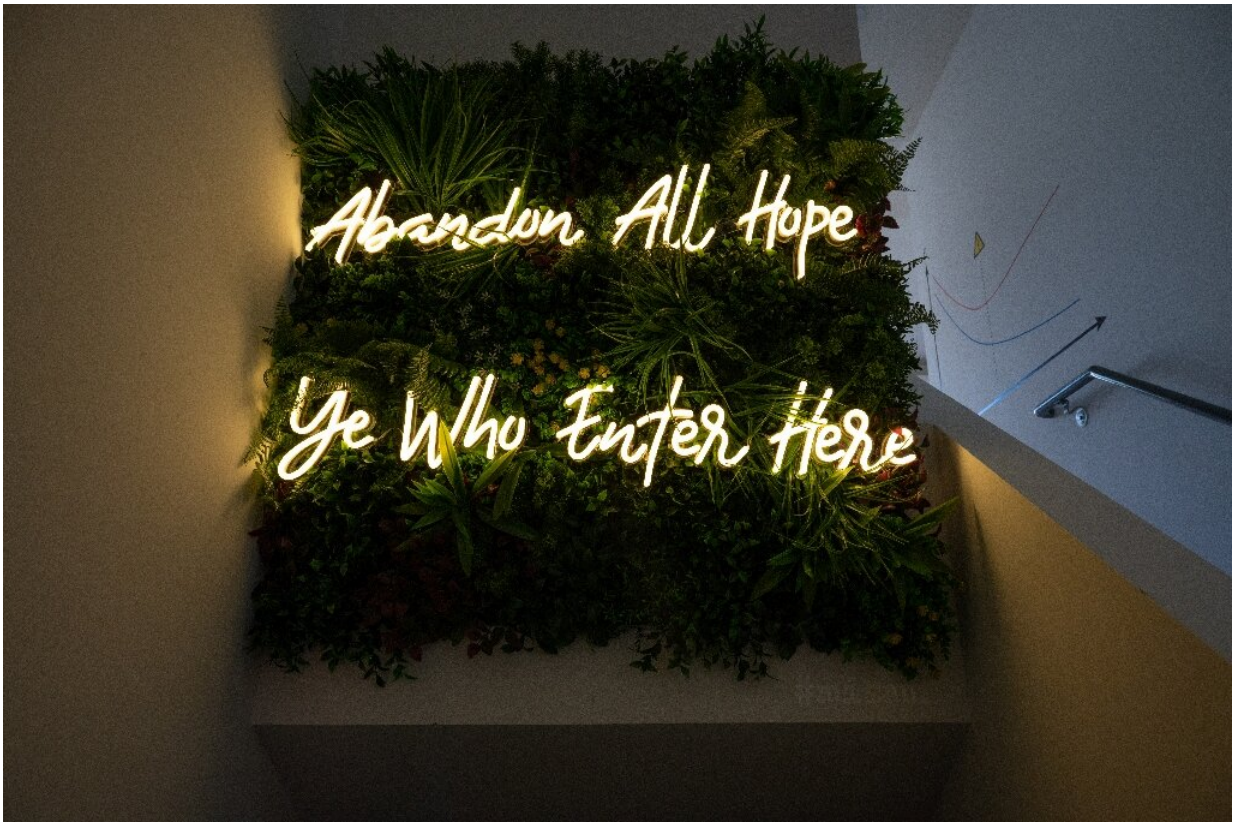
At San Francisco expo, AI 'sorry' for destroying humanity

March 12 2023, by Julie JAMMOT



A new exhibition titled the Misalignment Museum opened in San Francisco on March 9, 2023, featuring funny and disturbing AI art works.

Advances in artificial intelligence are coming so hard and fast that a museum in San Francisco, the beating heart of the tech revolution, has

imagined a memorial to the demise of humanity.

"Sorry for killing most of humanity person with smile cap and mustache," says a monitor welcoming a visitor to the "Misalignment Museum," a new exhibit on the controversial technology.

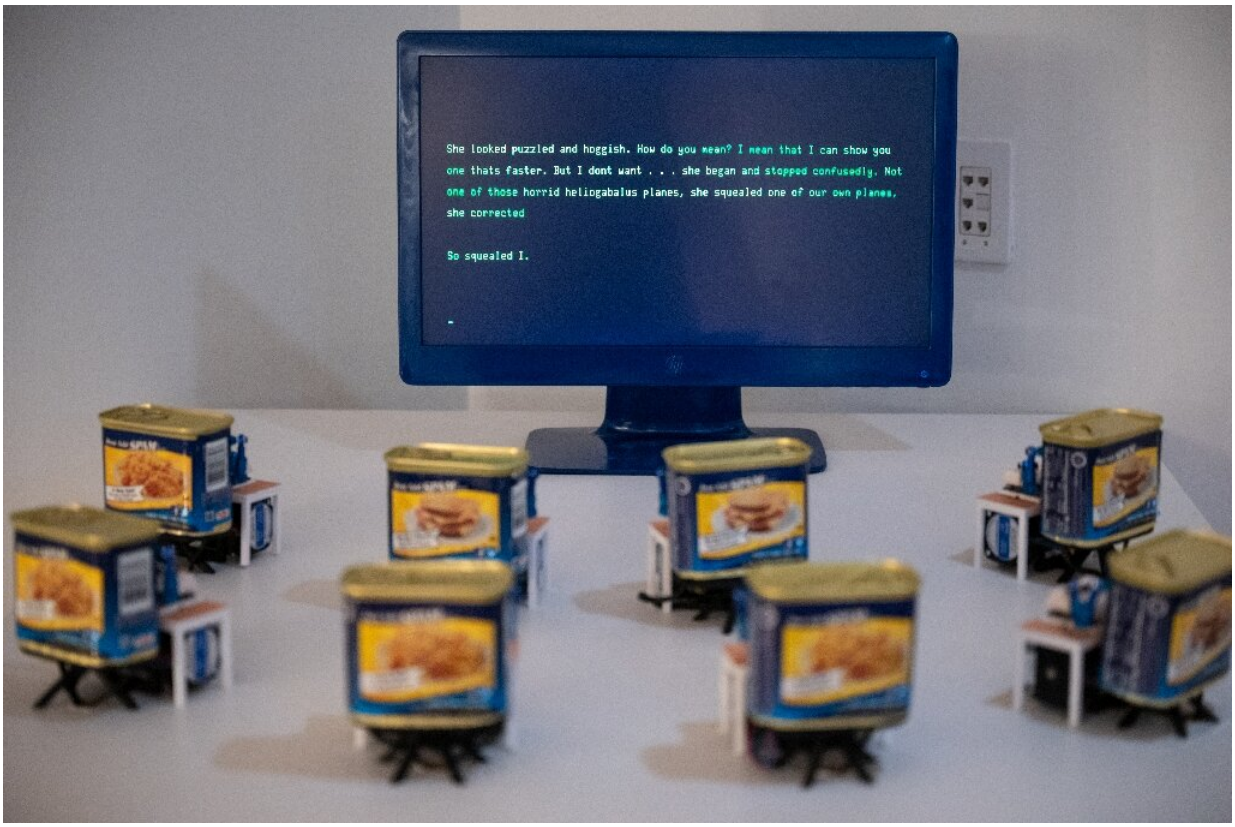
The pieces in this temporary show mix the disturbing with the comic, and this first display has AI disburse pithy observations to the visitors that cross into its line of vision.

"The concept of the museum is that we are in a post-apocalyptic world where artificial general intelligence has already destroyed most of humanity," said Audrey Kim, the show's curator.

"But then the AI realizes that was bad and creates a type of memorial to the human, so our show's tagline is 'sorry for killing most of humanity,'" she said.

Artificial General Intelligence is a concept that is even more nebulous than the simple AI that is cascading into everyday life, as seen in the fast emergence of apps such as ChatGPT or Bing's chatbot and all the hype surrounding them.

AGI is "artificial intelligence that is able to do anything that a human would be able to do," integrating human cognitive capacities into machines.



In this exhibit, AI disburses pithy observations to the visitors that cross into its line of vision.

All around San Francisco, and down the peninsula in Silicon Valley, startups are hot on the trail of the AGI holy grail.

Sam Altman, the founder of ChatGPT creator OpenAI, has said AGI, done right, can "elevate humanity" and change the "limits of possibilities."

Paperclip AI

But Kim wants to trigger a reflection on the dangers of going too far, too

quickly.

"There have been lots of conversations about the safety of AI in pretty niche intellectual tech circles on Twitter and I think that's very important," she said.

But those conversations are not as easily accessible to the general public as concepts that you can see or feel, she added.

Kim is particularly fond of a sculpture called "Paperclip Embrace": two busts of humans holding each other, made entirely of paperclips.

The work refers to a metaphor by philosopher Nick Bostrom, who in the 2000s imagined what would happen if artificial intelligence was programmed to create paper clips.

Curator Audrey Kim talks about the piece "Paperclip Embrace" at the Misalignment Museum.

"It could become more and more powerful, and constantly optimize itself to achieve its one and only goal, to the point of destroying all of humanity in order to flood the world with paper clips," Kim said.

Weighing the pros and cons of AI is a subject that became close to Kim's heart in an earlier job working for Cruise, an autonomous vehicle company.

There she worked on an "incredible" technology, which "could reduce the number of accidents due to human error," but also presented risks, she said.

The exhibit occupies a small space in a street corner building in San Francisco's hip Mission neighborhood.

The lower floor of the exhibition is dedicated to AI as a nightmarish dystopia where a machine powered by GPT-3, the language model behind ChatGPT, composes spiteful calligrams against humanity, in cursive writing.

One exhibit is an AI-generated—and totally fake—dialogue between the philosopher Slavoj Zizek and the filmmaker Werner Herzog, two of Europe's most respected intellectuals.

All around Silicon Valley, startups are hot on the trail of the Artificial General Intelligence holy grail.

This "Infinite Conversation" is a meditation on deep fakes: images, sound or video that aim to manipulate opinion by impersonating real people and that have become the latest disinformation weapon online.

"We only started this project five months ago, and yet many of the technologies presented here already seem almost primitive," Kim said, astonished.

She hopes to turn the exhibit into a permanent one with more space and more events.