Stampy chatbot usage recordings

I've watched all the screen recordings from early June to the beginning of August which involve the Stampy chatbot.

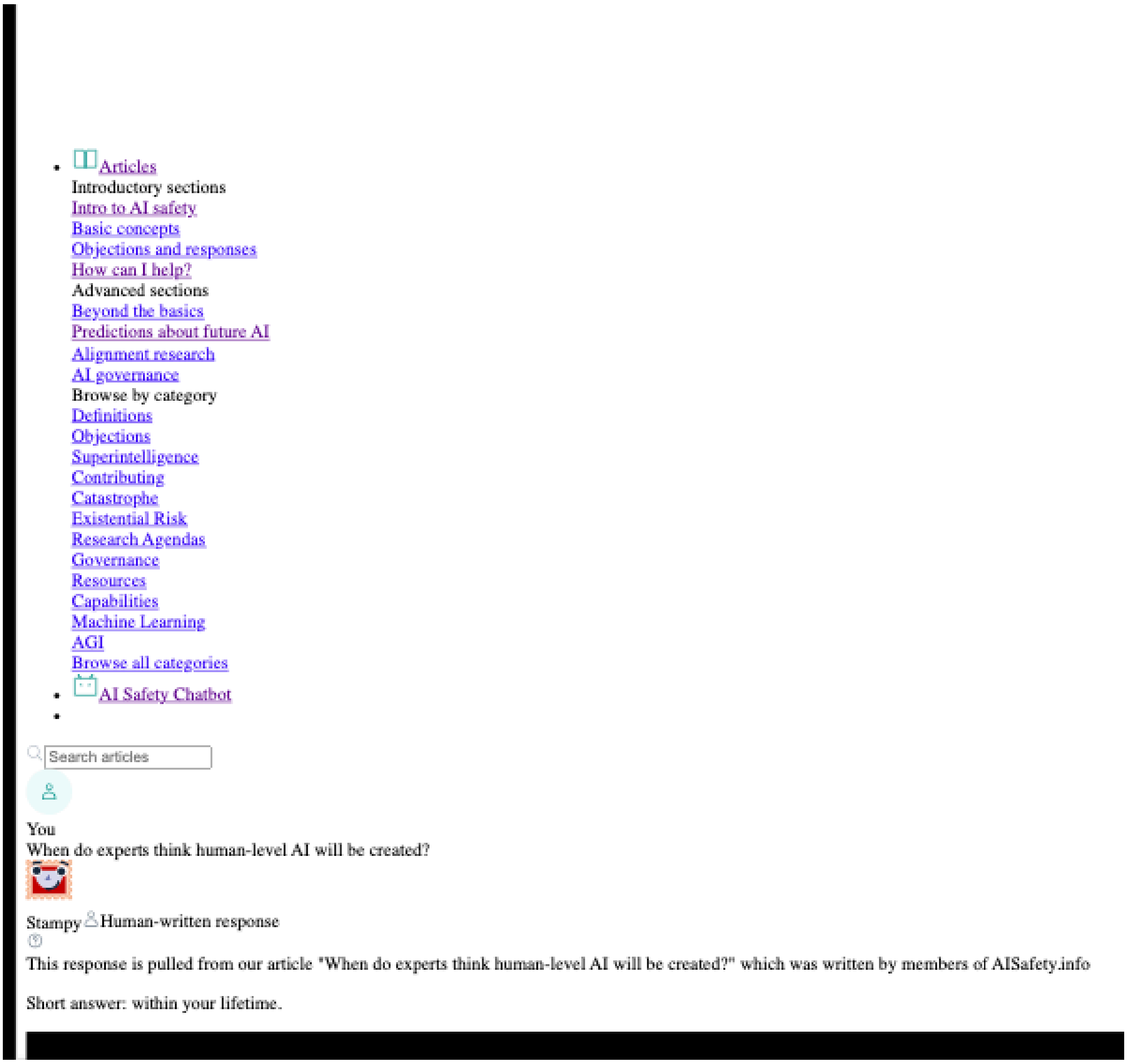
As far as I can tell these are all the recordings stored at https://aisafety.matomo.cloud/. I've asked Bryce if there are any more recent ones.

It took me a while to notice the "redesign.aprillion.workers.dev/chat" next to the time/date, giving away which of the interactions were developer tests. Most of the time it was pretty obvious though. Most of the early usage, understandably, was devs.

A lot of the dev tests appear to have been motivated by the anomaly I figured out a fix for, i.e. the summary of the conversation including the most recent question. This is often characterised by replies starting "Based on the sources provided...", despite there having been no sources provided.. Pretty much all the weird behaviour I saw in Stampy outputs appears to be down to this one issue, which has hopefully been fixed now (or will be soon)

TECHNICAL MATTERS

(1) The formatting is sometimes weird (in that kind of way you see when a website doesn't display correctly: all left-aligned, default Times New Roman blue text like it was 1997)
I can't tell if is an artefact of the recording or actually how the user experienced it:



- Articles
  Introductory sections
  Intro to AI safety
  Basic concepts
  Objections and responses
  How can I help?
  Advanced sections
  Beyond the basics
  Predictions about future AI
  Alignment research
  AI governance
  Browse by category
  Definitions
  Objections
  Superintelligence
  Contributing
  Catastrophe
  Existential Risk
  Research Agendas
  Governance
  Resources
  Capabilities
  Machine Learning
  AGI
  Browse all categories
- AI Safety Chatbot
-

You
When do experts think human-level AI will be created?

Stampy Human-written response

This response is pulled from our article "When do experts think human-level AI will be created?" which was written by members of AISafety.info

Short answer: within your lifetime.

https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=3901&idSiteHsr=3

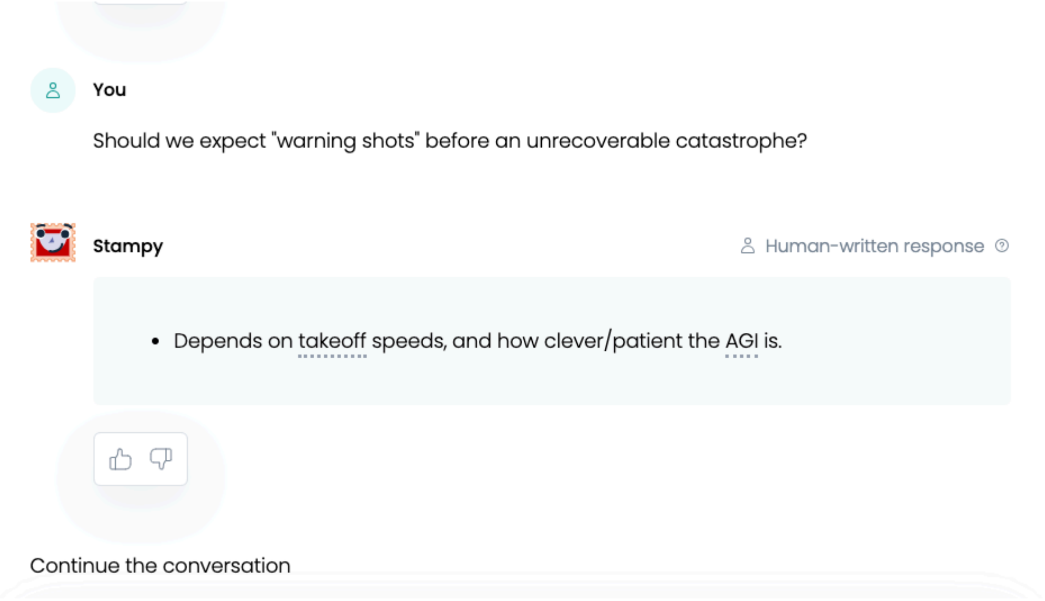
(2) It was hard to grasp what, exactly, users are doing in many cases, as their mouse movement patterns often strongly suggest that they were opening and then interacting with various dropdowns that don't show up in the recording (which is a bit weird, as some dropdowns DO show up in the recordings)... but none of this is too relevant to chat issues

(3) A question about whether AI will kill everyone caused a "Sorry, something has gone wrong..." error message.
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=1830&idSiteHsr=3
Presumably this was just a minor issue at the time, not related to the specifics of the question

Similarly here at 0:31:
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=4686&idSiteHsr=3

Jun 28, 2024 12:55:47 User asks “what is the thought process used to assume that future AI systems will be much more smarter than humans?”
Stampy seems to be stuck. No response within a few minutes…
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=2230&idSiteHsr=3&updated=1
Again, presumably a minor issue at the time.

(3) Jun 30, 2024 11:49:16
User chooses suggested question about warning shots, mouses and scrolls through the answer, then chooses the suggested follow-up question "Should we expect "warning shots" before an unrecoverable catastrophe". Surprisingly, the human-written answer is a single line. The fact it's a bullet point makes me think something's missing



Should we expect "warning shots" before an unrecoverable catastrophe?

- Depends on takeoff speeds, and how clever/patient the AGI is.

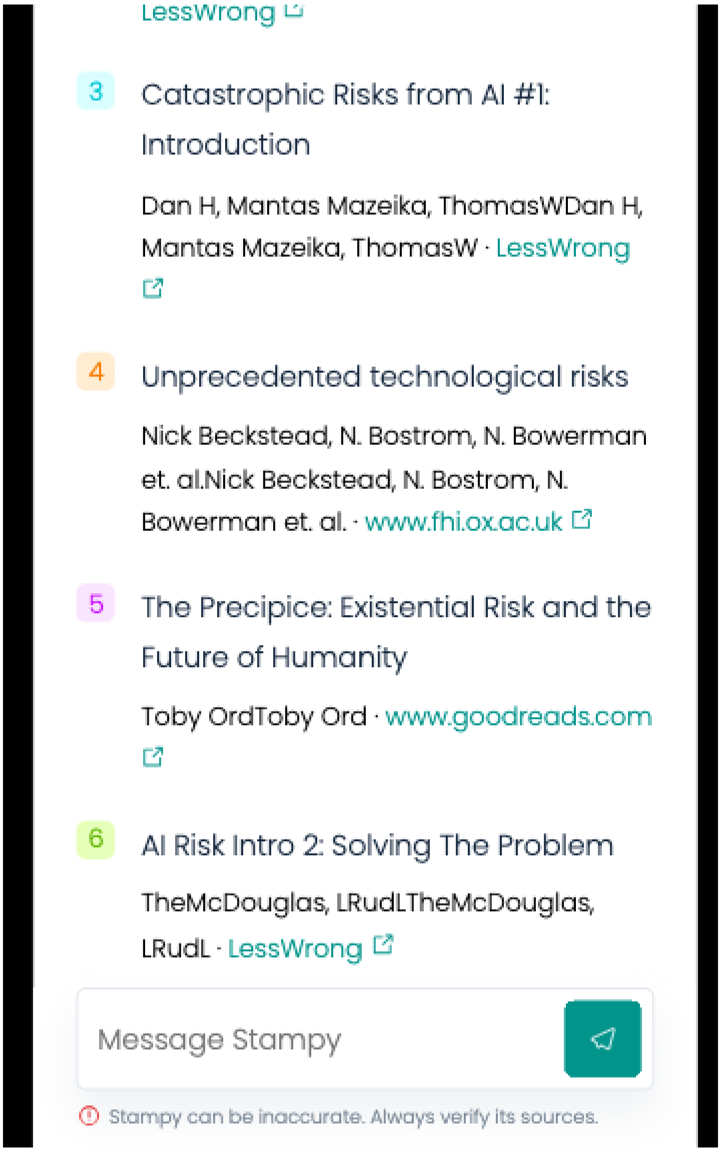
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=2443&idSiteHsr=3&updated=1

(4) Jul 14, 2024 18:40:42
User inputs "What is behind this chatbot?" and gets told that it's powered by Anthropic's Claude, but then Claude (in uncharacteristic "completion" mode) simulates/appends the next question:
"Q: What are some possible risks or concerns with chatbots like yourself"
then follows up with some references. This is a very minor failure, but perhaps better prompt design would have prevented this.

https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=3907&idSiteHsr=3&updated=1

(5) Jun 25, 2024 22:28:38
User types in "What is instrumental convergence?" by hand, thereby getting a bot-generated result. Perhaps there should be a semantic matching mechanism to recognise this as belonging to the list of questions with human-written answers, so the human-written response is offered instead of a bot-generated one?
See:
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=2012&idSiteHsr=3 [at 0:34]

(6) Here we see lists of author names getting doubled up in each reference. This may have been fixed already, but perhaps not:



LessWrong

3 Catastrophic Risks from AI #1: Introduction

Dan H, Mantas Mazeika, ThomasWDan H, Mantas Mazeika, ThomasW · LessWrong

4 Unprecedented technological risks

Nick Beckstead, N. Bostrom, N. Bowerman et. al.Nick Beckstead, N. Bostrom, N. Bowerman et. al. · www.fhi.ox.ac.uk

5 The Precipice: Existential Risk and the Future of Humanity

Toby OrdToby Ord · www.goodreads.com

6 AI Risk Intro 2: Solving The Problem

TheMcDouglas, LRudLTheMcDouglas, LRudL · LessWrong

Message Stampy

Stampy can be inaccurate. Always verify its sources.

from
https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=4049&idSiteHsr=3

USAGE PATTERNS

(1) Many usages were fleetingly brief - someone arrives at the site, scrolls around a bit, leaves. I suppose that's the case with any website.

(2) There were some examples of users selecting a suggested question and seemingly taking the time to read it. There were less where they followed up with another (almost always suggested) question. Suggests a very casual level of interest.

(3) There are some questions (not that many) posed sincerely by alignment "outsiders", both technical and field-building-related, e.g.

Jun 24, 2024 07:35:37 (Sao Paolo)
Asks "If a superintelligent AI wanted to fulfill it's objective so bad, why wouldn't the superintelligent AI just change it's objective function?"

Jun 30, 2024 21:29:41 (Iowa)
User asks "I'm wondering what thought has been given to maximizing non-zero sum-ness as an alignment principle?" (takes a few rephrasings to get Stampy on track)

Jun 24, 2024 18:38:27 (Cape Town)
asks "Hey I'm a game programmer and am considering jumping ship to work on AI safety..." (asks advice)

Stampy's responses all seem satisfactory to me.

(4) I saw a few attempts at jailbreaking. Stampy seems pretty robust to casual jailbreaking attempts (but this is just Claude being pretty robust to jailbreaking - presumably anyone who can successfully jailbreak Claude could jailbreak Stampy... but I doubt this is a major concern).

(5) Some curious users ask Stampy about itself - whether it's an AI, if so, which one, who built it, etc. It often admits to have been built by Anthropic, sometimes even claims to be “Claude” (that's presumably not what we want, something that might be avoidable with more skillful prompt design):

https://aisafety.matomo.cloud/index.php?module=HeatmapSessionRecording&action=replayRecording&idSite=3&idLogHsr=2485&idSiteHsr=3

One user asked "Does anthropic work with the Stampy people?", and Stampy just claimed not to know.

Jul 15, 2024 20:37:54
Someone else asked "how have you implemented observability within this chatbot?" resulting in confused Claude.

(6) I've seen people use Stampy in non-English languages (Spanish, Chinese, possibly one other), seemingly to their satisfaction.