Forecasting Digital Minds Takeoff Scenarios: Expert Survey

January 2025

Lucius Caviola & Bradford Saad
University of Oxford
Correspondence: lucius.caviola@gmail.com

This survey aims to gain deeper insights into expert perspectives on the potential future development of digital minds—computer systems capable of subjective experiences. Specifically, we examine the likelihood, timeline, and societal implications of various digital mind takeoff scenarios—i.e., transitions to a world where digital minds collectively possess significant welfare capacity. By analyzing expert predictions on this neglected issue, we aim to deepen our understanding and inform decision-making to better navigate future challenges and opportunities. Further reflections on digital mind takeoff scenarios can be found here.

The survey draws on insights from experts across multiple disciplines, including digital minds research, consciousness studies, technical AI research, AI policy and governance, forecasting, philosophy, and social sciences.

This survey was developed by Lucius Caviola and Bradford Saad, Senior Research Fellows at the Global Priorities Institute, University of Oxford. Data collection is scheduled for the first quarter of 2025, with a comprehensive report expected in the second or third quarter of the year. We plan to conduct the survey annually to monitor evolving trends and shifts in expert opinion.

Since this is an ongoing project, we welcome feedback, questions, and collaboration opportunities. Feel free to reach out to us.

Introduction

This page contains important information. Please read it carefully.

Topic: In this survey, we are interested in your views and forecasts about digital minds.

Our goal: Our aim is to collect and analyze expert views on this topic in order to improve our understanding of it and facilitate progress in the area.

Duration: The survey requires **approximately 30 minutes**, but it may take longer, depending on how much time you choose to spend on free-text responses.

Optional questions: All questions are optional. We especially value detailed responses, so please provide as much depth as you're comfortable with in the free-text questions.

Speculative nature: This survey differs from other forecasting surveys in that most questions are very speculative, involving long-term predictions with less precise resolution criteria. We understand you will feel highly uncertain about many questions, making precise answers difficult. If any questions seem unclear or underspecified, use your best judgment to interpret them. Given the speculative nature of this survey, we expect the free-text responses to provide the most valuable insights and hope that the quantitative questions will help prompt relevant considerations.

Suggested approach: We encourage you to share your best-guess estimates (median values) and speculate where needed. If you are considering multiple plausible future scenarios, focus on your median scenario—the one that falls in the middle of your range of plausible outcomes. Additionally, or alternatively, please share any relevant considerations in the free-text response sections. If you need to make an assumption in order to answer a question, feel free to do this and note the assumption in the free-text response.

Anonymity: The survey results will be anonymized and reported in a way that ensures that participating individuals and their organizations are not identifiable.

Request: Please do NOT forward this survey directly to others. Instead, if you know someone else who might be a good fit, let us know, and we will reach out to them ourselves.

Researchers and contact: This survey is conducted by Lucius Caviola and Bradford Saad from the Global Priorities Institute at the University of Oxford. If you have any questions or concerns, please contact us.

[next page]

For the purpose of this survey, please adhere to the following three stipulations:

1) Assume that a "digital mind" is any **computer system** that has a **capacity for subjective experience**.

   The computer system could be an artificial intelligence (e.g., based on machine learning) or a brain simulation inside a computer.

   For this survey, please understand **subjective experience** as follows:

- Paradigm cases of subjective experience include positive and negative feelings (e.g., pleasure and pain experiences) and sensory experiences (e.g., auditory and visual experiences).
- Subjective experience is a mental state that consists of the awareness of qualities. We take ourselves to have such experiences throughout our waking lives and in dreams, but not in dreamless sleep or under general anesthesia.
- In other words, subjective experience is what philosophers and scientists sometimes call "phenomenal consciousness."
- Please adhere to the provided notion of subjective experience. Using another notion of subjective experience (such as a loaded one that makes it controversial whether anything has subjective experience) could make the survey results ambiguous, misleading, and difficult to interpret.

2) Only consider digital minds with a welfare capacity **at least roughly as high as a typical human's.**

   In other words, disregard systems with a significantly lower welfare capacity. Welfare capacity refers to an entity's capacity to be benefited or harmed (e.g., by positive or negative subjective experiences) in a manner that is inherently morally significant.

3) Only consider digital minds that have been or will be created by our Earth-originating civilization (not by aliens or external simulators).

[Required] To ensure you have understood our assumptions, please select each system that is a digital mind and should be considered when completing this survey:
[✖️] An AI system that can do virtually all economically useful tasks that humans can do.
[✔️] An AI system created in the US that has the capacity for subjective experience and a welfare capacity at least roughly as high as a typical human's.

[✔️] A human brain emulation created in China that is capable of experiencing the same things a biological human experiences.
[✖️] An AI system that navigates its environment by detecting colors and shapes but lacks subjective experience.

Which best (even if imperfectly) describes your expertise? If multiple, choose the one you think is most relevant to this survey.

- Digital minds research
- AI research or AI policy
- Forecasting
- Philosophy/science/social science
- Non-expert

Starting Point

Note: whenever we ask for a "likelihood" or "probability," please always write numbers between 0 and 100.
If you write 0.7, we will interpret it as 0.7% out of 100% (i.e., less than 1% likely).

How likely is it that digital minds are possible **in principle**? ___%

How likely is it that digital minds will **ever** be created? ___%

How likely is it that the first digital minds will be created in or before the year:

- 2025? ___%
- 2030? ___%
- 2040? ___%
- 2050? ___%
- 2100? ___%

What do you think will be the median (middle) response of other participants in your expert group (**"{expertise}"**) to the following question?
How likely is it that the first digital minds will be created in or before the year 2040? ___%

What's the likelihood that the first digital minds are created **before** creating AGI (an AI system that matches or outperforms humans at almost all economically valuable tasks)? ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____

(Our aim is to collect as many considerations as possible.)

Prescriptive Assessments

Would enacting a moratorium on creating digital minds from now until 2040 be good or bad? Assume the alternative is no moratorium and ignore obstacles to enforcement.

- 1 Definitely bad
- 4 Unsure/indifferent
- 7 Definitely good

Do you expect efforts to promote AI safety (i.e., preventing AI-caused harm to humans) and efforts to prevent the mistreatment of digital minds will be

- 1 Mostly in conflict
- 4 Neither/unsure
- 7 Mostly synergistic

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Types of Digital Mind

How likely is it that some computer systems in the following categories could, **in principle**, have subjective experiences?

- Machine learning systems (e.g., LLMs, RL agents) ___%
- Brain simulations (e.g., whole-brain emulations of human or animal brains) ___%
- Other (e.g., neuromorphic, quantum) types of computer systems ___%

How likely is each of the following to be the **first** type of digital mind that **is created**? (Responses should sum to 100%.)

- N/A: digital minds will never be created ___%
- Machine learning digital minds (e.g., LLMs, RL agents) ___%
- Brain simulation digital minds (e.g., whole-brain emulations of human or animal brains) ___%
- Other (e.g., neuromorphic, quantum) types of digital minds ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Key Assumption

For many remaining questions in this survey, we will ask you to assume the following:
The first digital mind will be a machine learning-based AI system created in 2040 or earlier.

Even if you believe this scenario is unlikely, please be sure to answer those questions based on the assumption that this scenario will take place.

Speed

Please assume that the first digital mind will be a machine learning-based AI system created in 2040 or earlier.

After the first digital mind is created, how many years will it take until the collective welfare capacity of all digital minds together (at a given time) matches that of at least…
(A digital mind's welfare capacity is its capacity to be benefited or harmed [e.g. by positive or negative subjective experiences] in a manner that is inherently morally significant. State your median estimate. Enter 0 if you believe it will happen in the same year the first digital minds are created. Enter 9999 if you think it will never happen.)

- a thousand humans? ___
- a million humans? ___
- a billion humans? ___
- a trillion humans? ___

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Distribution

Please assume that the first digital mind will be a machine learning-based AI system created in 2040 or earlier.

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of them will have a **social function**, meaning that they are designed to interact with humans in a conversational, human-like manner (e.g., through text, audio, or video)? ___%

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of them were primarily produced in the following locations?
(Please understand "primarily produced" narrowly in terms of the location at which a given system first qualifies as a digital mind, not in terms of the entire supply chain.)
Assigned percentages must add up to 100%.

- USA
- Europe (including the UK)
- China
- Other

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of them were primarily produced by actors in the following categories?
Assigned percentages must add up to 100%.

- Companies
- Governments
- Universities
- Open-source developers (not companies, governments, universities)
- Other

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of digital minds were created by humans with an intention to create digital minds (as opposed to without that intention)? ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Claims

Please assume that the first digital mind will be a machine learning-based AI system created in 2040 or earlier.

In this section, we are interested in what claims digital minds will make. We are not concerned with the exact wording they will use: For example, if a digital mind claims it feels pain but does

not use the expression 'subjective experience,' it still counts as claiming it has subjective experience.

Consider **all AIs with a social function** (regardless of whether they are digital minds or not) 10 years after the first digital mind has been created. By "social function" we mean they are designed to interact with humans in a conversational, human-like manner (e.g., through text, audio, or video).
What proportion of those AIs with a social function will—systematically and **falsely**—claim that they have subjective experiences (when, in fact, they **don't**)?

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of those digital **minds** will—systematically and **falsely**—claim that they do ***not*** have subjective experiences (when, in fact, they **do**)? ___%

Consider all digital minds that exist 10 years after the first one has been created.
How likely is it that at least 10,000 of those digital minds will consistently and proactively **claim** that they:

- have positive or negative subjective experiences (e.g., pleasure or pain) ___%
- deserve to be protected under the law from harm and mistreatment ___%
- deserve civil rights (e.g., to vote, self-ownership, or legal personhood) ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

What factors could lead to the creation of digital minds that systematically claim to deserve civil rights? ____

Recognition

Please assume that the first digital mind will be a machine learning-based AI system created in 2040 or earlier.

For the following questions, please assume that 'citizens' refers to biological human adults in countries with digital minds.

Consider the world 10 years after the first digital mind has been created.
What proportion of citizens will **believe digital minds exist** (regardless of whether these beliefs are accurate)? ___%

What do you think will be the median (middle) response of other participants in your expert group (**"{expertise}"**) to the following question?
Consider the world 10 years after the first digital mind has been created.
*What proportion of citizens will **believe digital minds exist** (regardless of whether these beliefs are accurate)?* ___%

Consider all digital minds that exist 10 years after the first one has been created.
The median citizen with an opinion on this issue will tend to

- 1 Strongly underestimate the collective welfare capacity of that population of digital minds
- 4 Roughly accurately estimate
- 7 Strongly overestimate the collective welfare capacity of that population of digital minds

Consider the world 10 years after the first digital mind has been created.
What proportion of citizens will **believe that digital minds should be granted basic harm protection**? ___%

Consider the world 10 years after the first digital mind has been created.
What proportion of citizens will **believe that digital minds should be granted civil rights (e.g., self-ownership) in addition to basic harm protection**? ___%

How likely is it that, at some point within 10 years of the creation of digital minds, **digital mind rights will become one of the most contentious "hot button" issues in US politics** (top 5 issue)? ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Digital Mind Welfare

Please assume that the first digital mind will be a machine learning-based AI system created in 2040 or earlier.

In this section, we're interested in your estimates concerning the collective welfare (e.g., in terms of positive and negative subjective experiences) of the entire population of digital minds that exists 10 years after the first one is created. Please answer based on how much welfare you expect that population to have had up to that point.

Consider all digital minds that exist 10 years after the first one has been created.
The collective digital mind welfare will (in expectation) be on net

- 1 Strongly negative
- 4 Roughly neutral
- 7 Strongly positive

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of collective digital mind welfare consists of welfare that digital minds have before deployment (including during training and safety testing)? ___%

Consider all digital minds that exist 10 years after the first one has been created.
What proportion of collective digital mind welfare will come from digital minds which, individually, have a welfare capacity greater than 1,000 humans? ___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Artificial Welfare from Other Sources

In this survey, we defined 'digital mind' as a computer system with the capacity for subjective experience. In this section, we are interested in your estimates concerning AI systems that are not digital minds but nonetheless have a capacity for welfare (that is, the capacity to be benefited or harmed in a way that is inherently morally significant).

How likely do you think it is that it's in principle possible for a computer system to have **no** capacity for subjective experience but still have the capacity for welfare? ___%

In expectation, what proportion of computer system welfare in 2040 comes from computers that are not digital minds (i.e. computers that have no capacity for subjective experience)?___%

Please share any considerations, thoughts, or factors behind your responses, including speculative ones. ____
(Our aim is to collect as many considerations as possible.)

Final Questions

Do you have any unusual views about digital minds, subjective experience, welfare, or AI development that might explain differences between your responses and those of others? ____

In which country do you currently reside?

How would you rate your expertise in the following areas? (none 1-7 very strong)

- Digital minds research
- Technical AI research
- Technical AI safety
- AI policy/governance
- Forecasting
- Philosophy
- Social science
- Consciousness research

In what sector are you employed? Please select all that apply.

- Academic
- Government
- Industry/corporate sector
- Research organization outside of academia (e.g. think tank)
- Other non-profit (primarily charities, advocacy groups, or similar)
- Other

How connected are you to at least one community active on LessWrong, EA Forum, or Alignment Forum? (not at all connected 1-7 strongly connected)

Do you have any final comments about the survey or input for us? ___