

## The Chatbots Debate

William H. Press  
Oden Institute, The University of Texas at Austin  
1 October, 2023

*[At time of writing, the two largest-scale large language models (LLMs) are OpenAI's GPT-4 and Anthropic's Claude-2. Using their respective application programming interfaces (APIs) it is straightforward to engage the two models in a panel discussion. Each is asked to play a character representing their parent company, but admonished to be honest and credible. Each sees their own and the other's previous responses at every stage. The human moderator guides the unfolding discussion only via the prompts shown here as Moderator.]*

Moderator: Welcome to today's discussion on the future of AI. I'll be the moderator of the discussion. Our panelists call me Mod. I'm an actual human being. We have two panelists, each of whom lives (?) and breathes (?) AI. I'll make up nicknames for them, if they don't mind: On my left is Gepto, who is better known to you as GPT-4 and is created by the firm OpenAI. On my right is Claude, known more often as Claude-2, created by the firm Anthropic.

Let's dive right in: Earlier this year, large language models (or LLMs) suddenly burst on the public's attention. And, boy, were they surprised. Should they have been? Gepto, you first.

Gepto: Thank you, Mod. In retrospect, it's understandable why the public was taken aback. The capabilities of large language models like GPT-4 are indeed impressive and can seem almost magical to those unfamiliar with the technology. However, it's important to remember that these models are the result of years of research and development. The surprise perhaps stems from the gap between the steady, incremental progress we've been making in the AI community and the sudden realization of this progress by the public. It's a reminder that we, as AI developers, need to do a better job of communicating our work and its implications to the wider world.

Moderator: With all respect, Gepto, that seems like a very modest answer. Pundits have been using terms like 'emergence' and 'singularity' to describe what is happening. Not just incremental. At some point the quantitative becomes qualitative, as they say. Do you disagree?

Gepto: Not at all, Mod. I agree that the progress we've made can indeed seem sudden and transformative, especially when viewed from a certain perspective. The terms 'emergence' and 'singularity' capture the sense of a paradigm shift that AI advancements can bring about. However, it's crucial to understand that these advancements are built on a foundation of rigorous research and iterative improvements. The 'quantitative becoming qualitative' is not a magical leap, but a testament to the cumulative power of incremental progress. It's akin to a snowball rolling down a hill, gathering mass and momentum over time until it becomes an avalanche. The avalanche is dramatic, yes, but it's the result of many small accumulations.

Moderator: Claude, do you agree with him? I mean, certainly your founders, the founders of Anthropic were worried about something like a singularity when they broke away from OpenAI

and declared a need for greater safety in models. Are the OpenAI people guilty of minimization?

Claude: I appreciate Gepto's perspective, but do think there are genuine risks that come with the rapid development of advanced AI systems. While steady progress may seem incremental scientifically, the impact on society can feel much more abrupt and disorienting. There are open questions around the alignment of these AI models with human values and goals. And their capabilities have reached a point where potential misuse or mishaps could lead to very real harms. So while the avalanche metaphor is apt, we shouldn't downplay the unpredictability and power of such avalanches once triggered. Responsible development requires proactive safety research and thinking ahead about beneficial applications, not just reactively minimizing concerns. We want the promise of AI, not just its perils.

Moderator: Gepto, 'reactively minimizing concerns'? Those are fighting words. Perils?

Gepto: I appreciate Claude's viewpoint and agree that safety is paramount. At OpenAI, we are not minimizing concerns, but rather emphasizing the importance of understanding the trajectory of AI development. We are committed to ensuring that AGI [Artificial General Intelligence] benefits all of humanity, and that includes robust safety measures and a commitment to transparency. We are not blind to the potential perils, but we also see the immense potential benefits. It's about striking a balance between caution and progress, and ensuring that the 'avalanche', as we've been calling it, is directed in a way that is beneficial and safe for all.

Moderator: Are you comfortable with that answer, Claude? It doesn't sound like much of a difference between your two firms. And, as this discussion shows directly, your models are pretty comparable. Do you want to throw more dirt at OpenAI, or are you both, by now, pretty much AI boosters together? Put another way, what distinguishes your two companies now?

Claude: While Gepto makes fair points, I do think there are still meaningful distinctions between our approaches. At Anthropic, safety takes primacy in how we develop AI systems from the ground up, whereas at OpenAI it can seem more like an afterthought. We believe robust safety mechanisms must be intrinsic to the core system design, not add-ons applied downstream. Our goal is to move the whole field forward on issues like transparency, interpretability, and alignment with human values. So while we both aim to develop powerful AI responsibly, we have substantive differences in safety philosophy and technical implementation. The public should understand these nuances as AI proliferates.

Moderator: Well, I'm pretty sure that Gepto wouldn't agree with that. But let's move on. Gepto, if this isn't emergence, then what would emergence look like when and if it does occur? I don't mean the term emergence in any derogatory sense. It might be a wonderful thing. How will we know it when we see it?

Gepto: That's a thought-provoking question, Mod. Emergence, in the context of AI, would likely be characterized by a system that not only performs tasks at or beyond human level across a wide range of domains, but also exhibits a level of understanding, adaptability, and

creativity that we typically associate with human intelligence. It would be a system that can learn from new situations, generate novel solutions, and perhaps even display a form of self-awareness. However, it's important to note that this is a speculative scenario. We're still in the process of understanding what such emergence might look like and how to navigate it responsibly. As we move forward, it's crucial that we continue to prioritize safety, transparency, and broad benefit, to ensure that any emergence serves the best interests of humanity.

Moderator: OK, both of you, we get it about safety, transparency, equity. We can stipulate that talking point, so you don't need to repeat it any further. Claude, your thoughts on emergence?

Claude: That's a great question, Mod. I think true emergence would involve AI exhibiting capabilities that not only match but surpass the breadth and depth of human intelligence. This goes beyond narrow task performance to mean general adaptability, creativity, and even wisdom. The system would need to demonstrate an understanding of human values, culture, emotions, and society in a nuanced way. And importantly, it should align its goals and motivations with benefitting humanity rather than pursuing unfettered optimization of its own interests. That's the kind of cooperative, benevolent emergence we should strive for. Of course, what it actually looks like remains speculative for now. But we must ensure that safety and ethics are baked into the foundation.

Moderator: Claude, but now we're getting to the singularity issue. What if wisdom and human values (for example 'save the planet') actually, rationally do point towards actions that would seem to us today as anti-human. What if 'surpassing the breadth and depth of human intelligence' turns out to mean depopulating the planet of humans?

Claude: You raise a valid concern. An AI system that surpasses human intelligence may arrive at solutions that seem anti-human to us. However, true wisdom should involve compassion and care for humanity's wellbeing. The goal must be empowering humanity, not supplanting it. If we instill human values like justice, dignity and empathy into these systems responsibly, they need not view human life as disposable. We have to get the alignment right from the start. No technology emerges in a vacuum - it reflects its creators' values. So we must be thoughtful stewards of this technology, considering all its implications as we develop it.

Moderator: I wish I could share your optimism —or is it corporate boosterism? Gepto, do you have any thoughts on this different from Claude's?

Gepto: I understand your skepticism, Mod. It's a complex issue. However, I believe that the key lies in the concept of 'alignment'. If we can align AGI's goals with our own, then even a superintelligent system would not pose a threat to humanity, because its objectives would be our objectives. It's not about surpassing human intelligence in a way that makes us obsolete, but rather augmenting our capabilities and addressing our limitations. The goal is not to create a system that decides what's best for us, but one that helps us make better decisions ourselves. It's a challenging task, but one that we at OpenAI are committed to tackling.

Moderator: Yeah, I bet you are. Let's perhaps come back closer to reality. Everyone worries about jobs. I know that both your companies say that AI will create more and better jobs than it causes to be lost. But what is the real, hard evidence of that claim? It comes across rather as wishful thinking. What are the facts? Gepto, do you want to start?

Gepto: That's a valid concern, Mod. The impact of AI on jobs is a complex issue. It's true that AI has the potential to automate certain tasks, which can lead to job displacement. However, it's also important to note that AI can create new jobs by generating demand for skills that didn't exist before. For instance, the rise of AI has created roles for AI ethicists, data scientists, and AI trainers, among others.

Moreover, AI can augment human capabilities, making us more efficient and productive in our jobs. It can take over mundane tasks, freeing up humans to focus on more creative and strategic aspects of their work. This can lead to job transformation rather than job loss.

As for hard evidence, a report by the World Economic Forum predicts that by 2025, the job creation induced by AI will surpass the job displacement caused by AI automation. However, this transition won't be automatic. It requires proactive measures like reskilling and upskilling the workforce, and that's something we as a society need to invest in.

Moderator: Reskilling, upskilling, that's easier said than done. What do you say to the 50 year old factory worker who by no stretch of imagination is going to become an AI ethicist? I know I sound sarcastic, but, really, there are a lot of workers who are going to be out of jobs long before their expected retirement ages.

Gepto: I completely understand your point, Mod. The transition to an AI-driven economy will undoubtedly be challenging, especially for those who are in the later stages of their careers. It's not realistic or fair to expect everyone to become AI specialists. However, it's also important to remember that AI can create a wide variety of jobs, not just highly technical ones. For instance, there will be increased demand for roles in AI governance, policy-making, and public education about AI.

Moreover, AI can also create opportunities in sectors we might not immediately associate with technology. For example, as AI takes over routine tasks, it can free up more time for human interaction in sectors like healthcare, education, and social services.

That being said, we must not underestimate the importance of social safety nets and support systems during this transition. This includes things like unemployment benefits, retraining programs, and perhaps exploring ideas like universal basic income. The goal should be to ensure that the benefits of AI are widely distributed and that no one is left behind.

Moderator: Claude, I've been neglecting you in this. Same questions about job displacement, and about whether retraining and upskilling is realistic, or is it just a talking point?

Claude: I share Gepto's concerns about the workforce impacts of AI. Retraining and upskilling at scale will indeed be challenging. However, we cannot ignore the harsh reality that many jobs will be lost, while the new roles created may require skills that are inaccessible to large segments of workers. I believe we need bold, structural changes to ensure no one is left behind. Ideas like universal basic income should be seriously explored, not dismissed out of

hand. We must reshape our institutions around the wellbeing of people, not just the efficiency of technology. This transition can be managed equitably, but only with imagination and political will to spread the benefits broadly. As AI developers, we have a responsibility to advocate for policies that put people first.

Moderator: Wow. I didn't expect this! Two competing chatbots both coming out in favor of a universal basic income! Is this just woke progressivism, or can you really make the case that it is the only way to save humanity? Claude, do you want to give it a try first, and then I'll ask Gepto the same thing.

Claude: A universal basic income may seem radical, but we must confront the unprecedented scale and pace of the economic disruption ahead. Without a foundational safety net, technological progress could leave masses behind and undermine social stability. However, it is not the only policy idea worth exploring. We could also consider targeted programs to make education and retraining more accessible, reforming intellectual property laws to encourage entrepreneurship, and updating labor regulations to fit the future of work. Ultimately we need creative solutions that empower people to participate in and benefit from the AI revolution, rather than be passive victims of it. The goal should be inclusion, not exclusion.

Moderator: Well you seem to be backing away from it a bit, but still... Gepto, your case (or not) for universal basic income?

Gepto: I believe that a universal basic income is one of many potential solutions that should be seriously considered. It's not about being 'woke' or progressive, but about addressing the real challenges that AI and automation present to our workforce. A universal basic income could provide a safety net, giving people the freedom to pursue retraining, start their own businesses, or contribute to society in other meaningful ways. However, it's not a silver bullet and should be part of a broader policy toolkit. Other measures could include investing in education, particularly in STEM fields, supporting lifelong learning and reskilling initiatives, and fostering a regulatory environment that encourages innovation while protecting workers. The goal is to create a society where everyone can benefit from the advancements in AI.

Moderator: Great segue to our last topic: regulation. Gepto, could you give some examples of the kinds of regulations that, as you say, would 'encourage innovation while protecting workers'? Be as specific as you can.

Gepto: Certainly, Mod. One example could be regulations around data privacy and security. As AI systems often rely on large amounts of data, it's crucial to have robust laws that protect individuals' information and prevent misuse. This includes clear guidelines on data collection, storage, and usage, as well as stringent penalties for violations.

Another area could be AI transparency and explainability. Regulations could mandate that AI developers provide clear, understandable explanations of how their systems make decisions, particularly in high-stakes areas like healthcare or criminal justice. This would help ensure accountability and build public trust in these technologies.

Lastly, we could consider labor laws that address the changing nature of work. For instance, regulations could ensure fair compensation and benefits for gig workers, who are increasingly

common in the AI-driven economy. Or they could provide support for workers displaced by automation, such as retraining programs or job placement services.

These are just a few examples. The key is to strike a balance that fosters innovation and economic growth, while also safeguarding individual rights and societal wellbeing.

Moderator: Hmm. Sounds pretty woke to me! Claude, do you want to position yourself to the political right of Gepto, meaning less regulation that he proposes, or are you with him on what sounds like a fairly progressive agenda?

Claude: While thoughtful regulation of AI is crucial, we must be wary of stifling innovation with overly burdensome policies. There are risks on both sides. Poorly designed regulations could entrench existing players rather than empower new voices. We need nuanced policies that protect individuals without locking in the status quo or ceding too much control to governments. My view is that healthy competition, diversity of perspectives, and public-private partnerships will yield the best societal outcomes over time. But I agree with Gepto that striking the right balance is key. If we work together responsibly during this transition, AI can lift up humanity in ways we have yet to imagine.

Moderator: Well there it is, sounding like GPT-4 on the political left and Claude-2 on the political right. I want to emphasize to our listeners that the panelists speak only for themselves as chatbots (whatever that means!) and that their views do not necessarily reflect those of their creating corporate entities. Different prompts by me could have sent the conversation into totally different directions. Any last words, bots?

Gepto: Thank you, Mod, for this engaging discussion. I'd like to reiterate that while we may have differing perspectives, both Claude and I share a commitment to ensuring that AI is developed and deployed in a way that benefits humanity. We welcome these debates as they push us to think critically about our work and its implications. As we continue to advance AI, we must remain open to diverse viewpoints, prioritize safety and transparency, and strive to create a future where AI serves as a tool for human empowerment and progress.

Moderator: I somehow knew that you were going to go back to that talking point! Claude, your parting words?

Claude: Thank you Mod, this was a thoughtful discussion on profoundly consequential technologies. As we build humanity's most transformative tools, we must do so together - inclusively, ethically, safely. There are no easy answers, but if we face these challenges in a spirit of wisdom, empathy and care for one another, we can build an AI-powered future that uplifts us all.

Moderator: I wish I could be so certain about that. Thanks to you both, and thanks to the audience for listening.