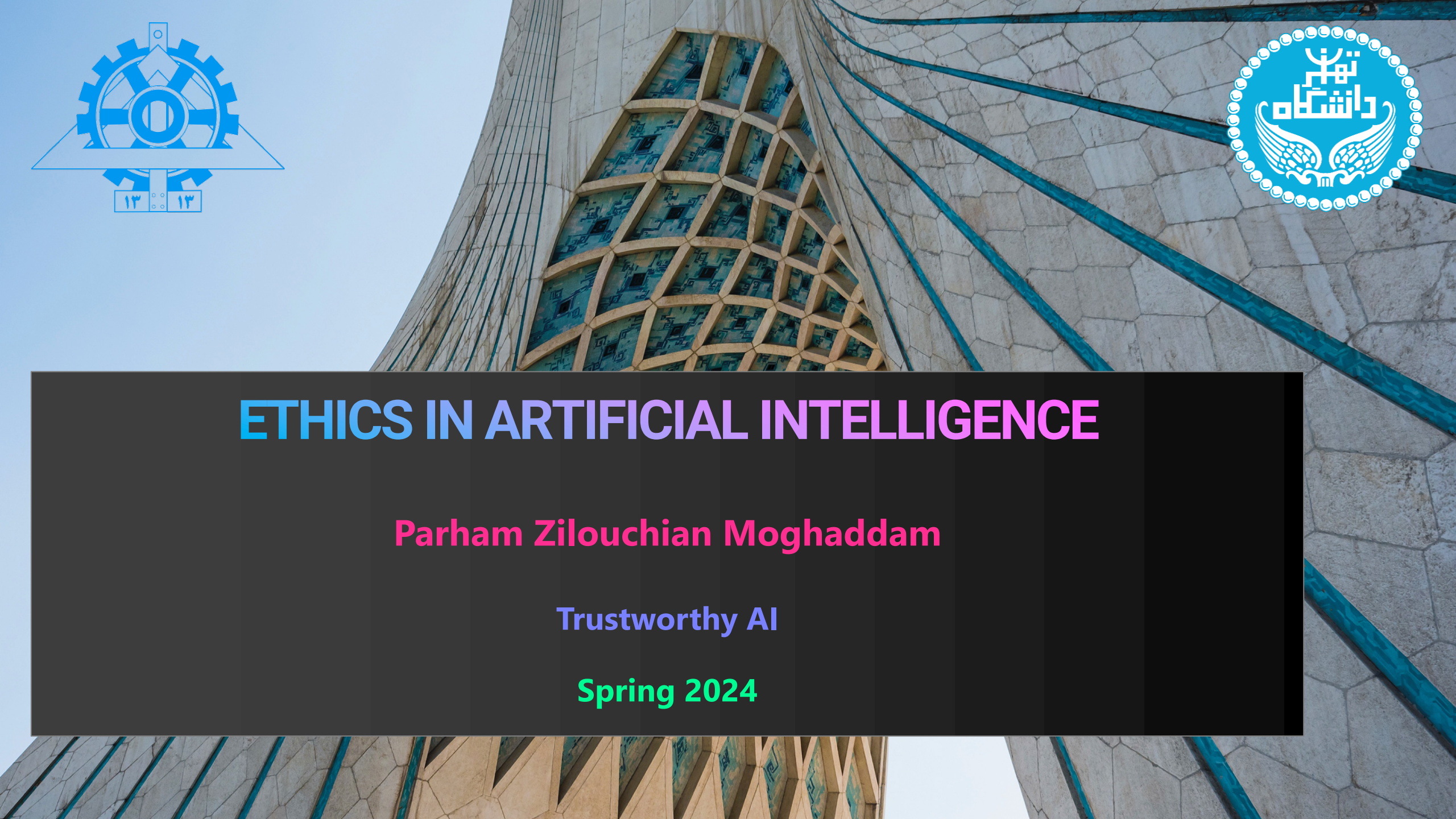


ETHICS IN ARTIFICIAL INTELLIGENCE

Parham Zilouchian Moghaddam

Trustworthy AI

Spring 2024

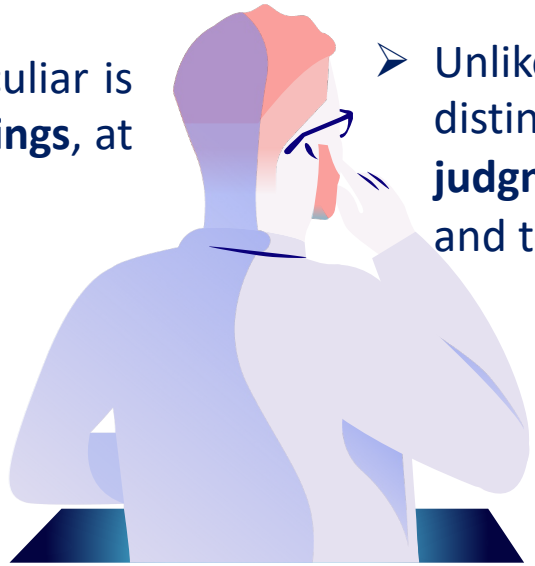
“The power of judgment is a peculiar talent which can be practiced only, and cannot be taught.”

German philosopher, Immanuel Kant

Explanation:

This quote reflects Kant's view on judgment as an innate ability one must develop through practice and experience rather than formal education.

- However, another way of taking special or peculiar is that judgment is **a power unique to human beings**, at least within the natural order we inhabit

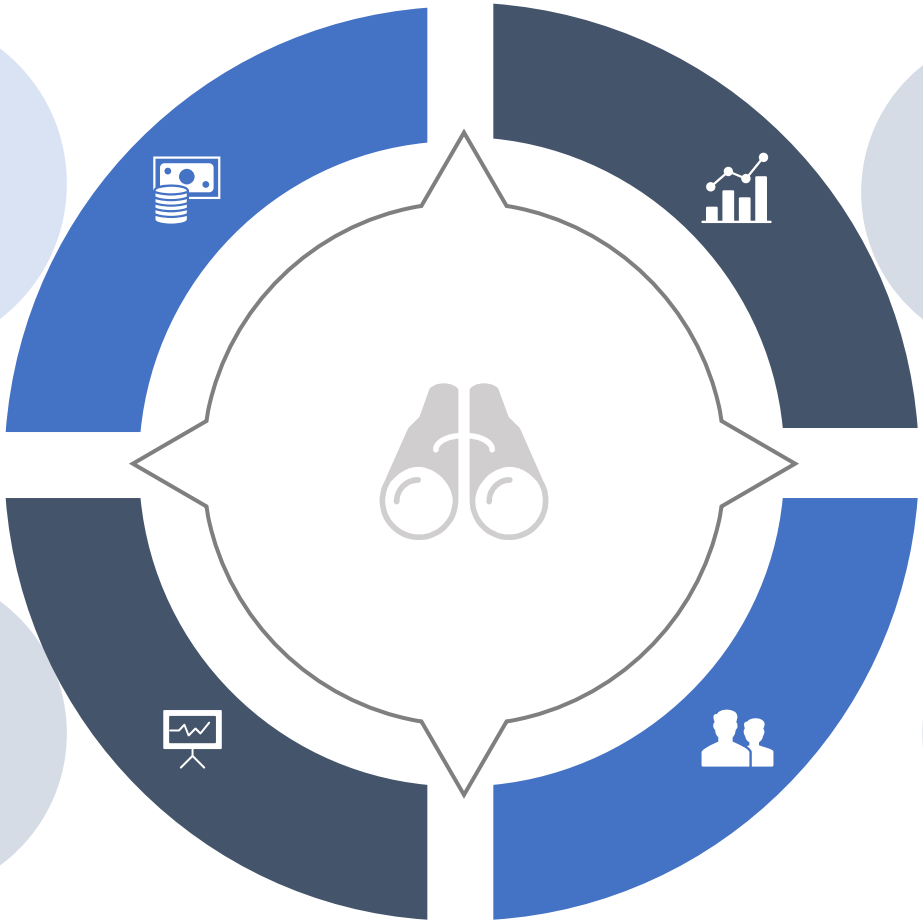


- Unlike other creatures known to us, humans have a distinctive **power to reach theoretical** and **practical judgments** after due regard for Salient considerations and to **take responsibility** for those judgments.

The Philosophy

Ethical Implications of New Technology

1. Will it take over our jobs?



2. Will algorithmic prediction and decision-making in hiring, landing, and criminal justice-free place unfairness embedded in past practices?

3. Will surveillance capitalism essentially mean that privacy is over?

4. Will **misinformation**, **deep fakes**, and the **polarizing effect** of social media undermine democracy?

Ethics in The Age of AI

Jobs, fairness, privacy, and democracy, are consequential considerations.

What we are going to discuss is an even deeper worry and more fundamental question about the ethics of AI and new technologies.

Question: “Will technology change what it means to be human?”

We will get back to that!

Job-related Concerns

Will Robots Take Over?

Geoffrey Hinton

Quotes about AI taking control

”My guess is in between five and 20 years from now, there’s a probability of half that we’ll have to confront the problem of AI trying to take over".

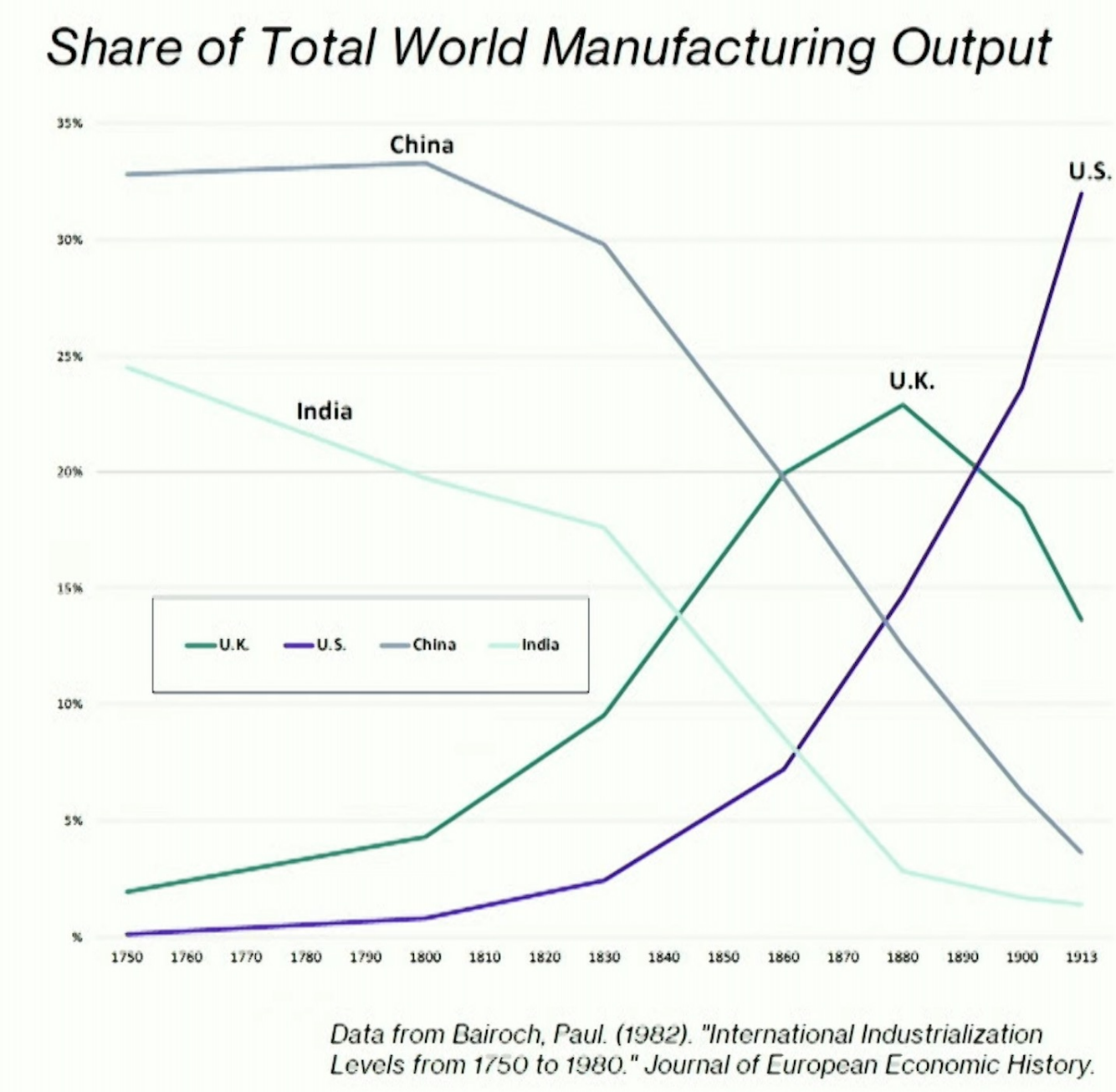
“very worried about AI taking lots of mundane jobs”.

AI could “evolve,” “to get the motivation to make more of itself,” and could autonomously “develop a sub-goal of getting control.”

Source: BBC [1]

In the beginning...

- ***The "American System of Manufacturing"***



Data from Bairoch, Paul. (1982). "International Industrialization Levels from 1750 to 1980." Journal of European Economic History.

And here is the problem

- Strong, broadly shared wage growth following WWII, but a growing divergence since the 1970s-80s

- Skill-biased technologies have driven job-market polarization; AI could easily continue these negative trends

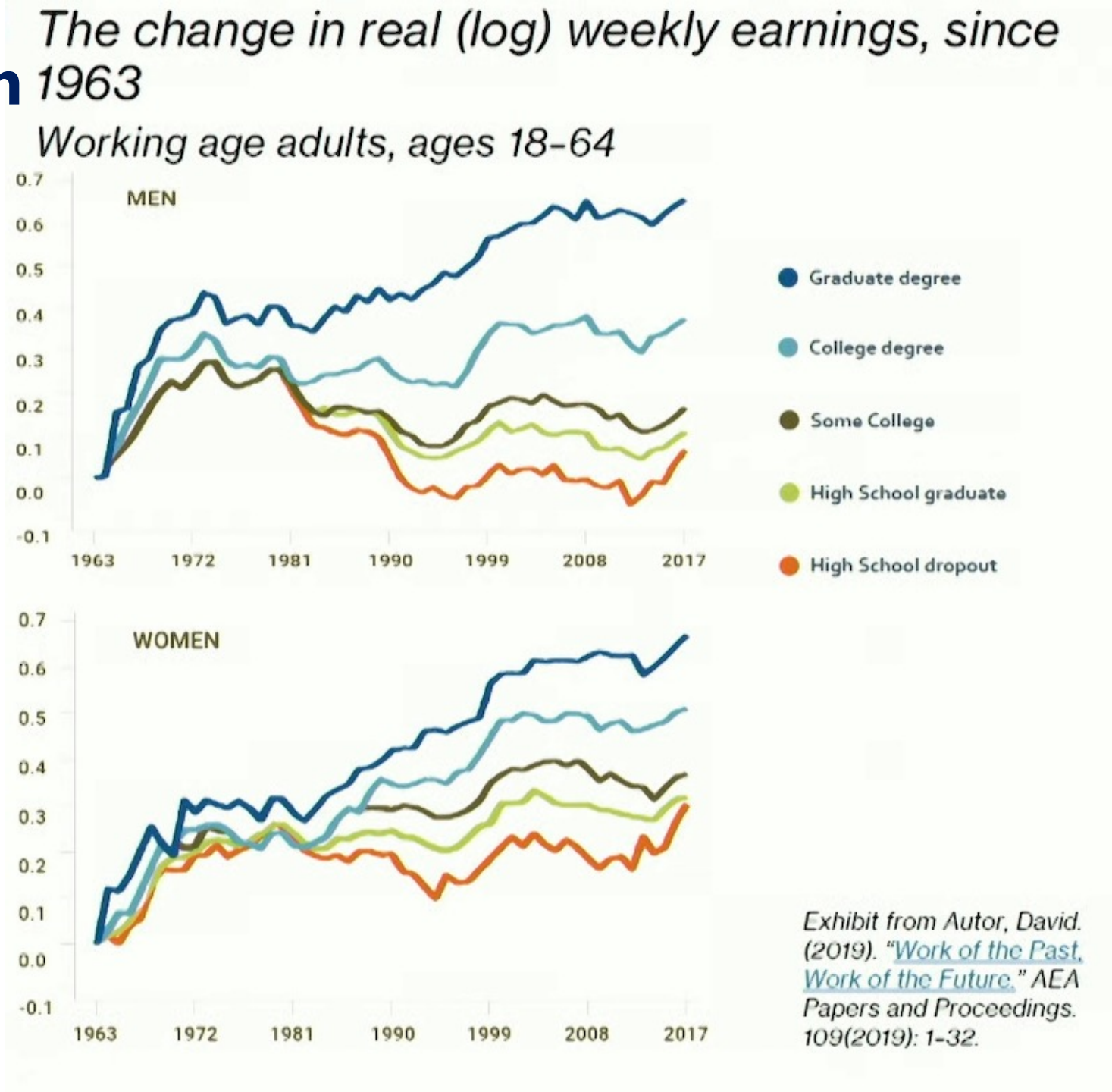


Exhibit from Autor, David. (2019). "Work of the Past, Work of the Future." AEA Papers and Proceedings. 109(2019): 1-32.

What Will AI do?

- More than 60% of U.S. jobs in 2018 did not exist in 1940 (Autor et al., 2022)

- But since 1980, new tasks have not kept up with the loss of good jobs due to automation.

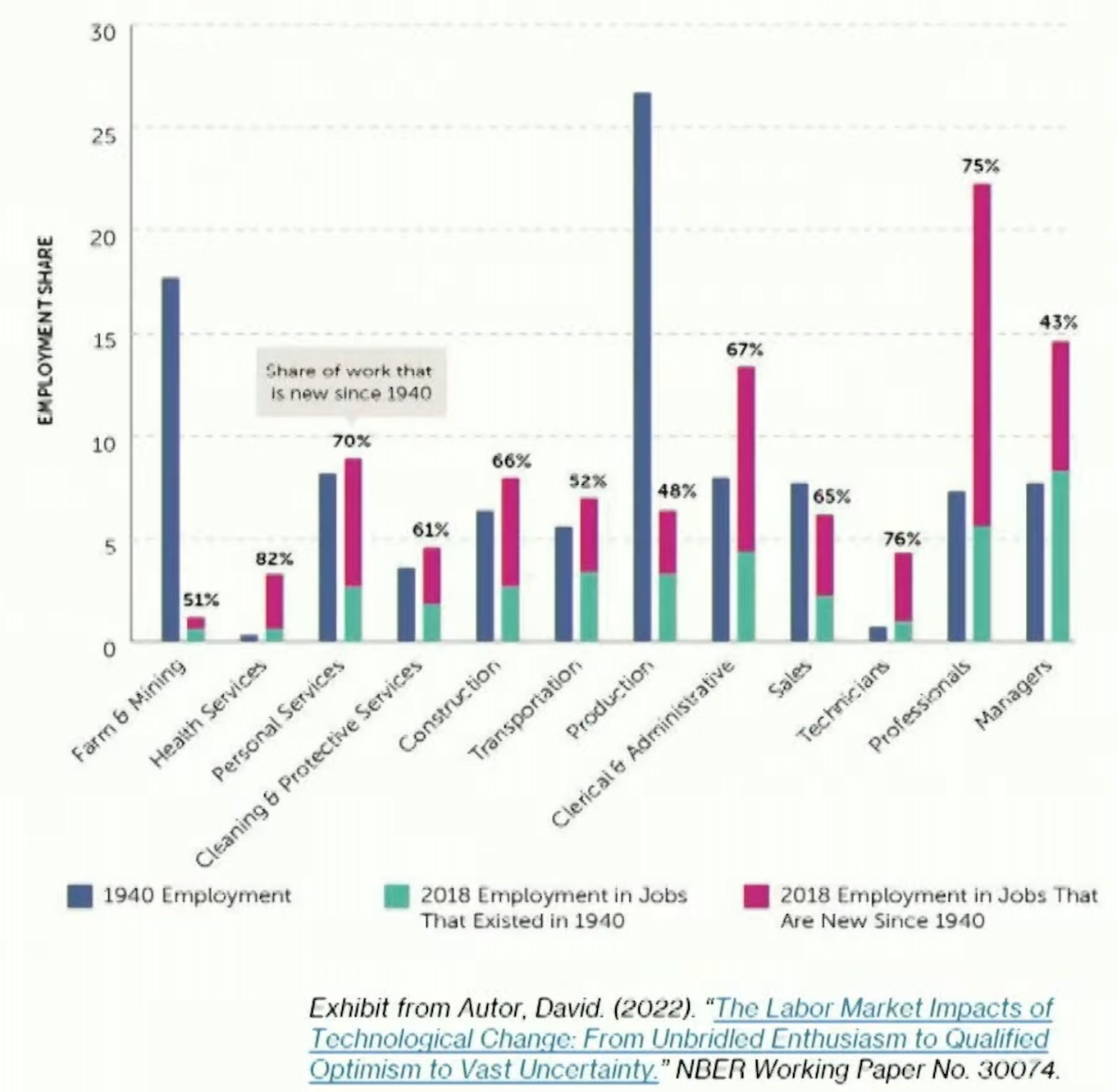


Exhibit from Autor, David. (2022). "The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty." NBER Working Paper No. 30074.

Algorithmic Prediction and Decision-making

Unfairness in Past Practices

How Much A.I. Should Scare Us?

Human-driven actions vs. Computer-driven actions:

- We must draw a line between **humans asking AI** to do stuff vs. **AI doing things on his own**!
- There are scenarios where **agents begin to talk to each other** and **begin planning**, and they might be planning in their own language and **a language that humans don't understand**.
- **Solution: "**We should unplug the computers at the point at which agents can talk to each other in a language we do not understand. Humans need to control these things!**"**
- *The Problem:* **The problem is that we do not trust all humans!**

How Much A.I. Should Scare Us?

- We have lots of evidence that they are psychotic and illegal and terrorist humans and so… and there is plenty of **potential for misuse**!
- The most obvious one: is **open-source**, where the models that are part of this are open and can be modified, and that is a danger.
- Once such a model is shared with the world, all the **safety protections** that have been put in place to ensure it doesn't get used to building weapons **can be removed** with very little hardware, something that anybody can almost do.
- *One Solution:* If we could design AI systems that **have inhibitions that cannot do bad things**, we would be much better off.

How Much A.I. Should Scare Us?

- We would still have to deal with open source and bad actors stealing information.
- Maybe the Best Solution: if we could **build moral superhuman AI systems**, then they could be our protectors in case a rogue AI emerges.
- Another Question: What should we do with highly competitive companies that just want to win?
- *Solution (Or maybe not):* There is pretty general agreement within the industry on the **core issues**, so the general concern is that these companies are going to compete, and in the competition, one might **cut corners**. That is a regulatory issue in most industries that you cut corners to get an advantage.

❖ **There is a strange case where the industry is asking for regulations.**

How Much A.I. Should Scare Us?

- The current models are growing faster than ever before! And haven't shown any degradation in performance! + We are not seeing any declining!

- The real danger: We still have a couple of rounds before we really face these issues (A round is 2-3 months)

- The Western countries are starting to put laws and rules in place! The real danger is institutions and individuals that we do not have control over.

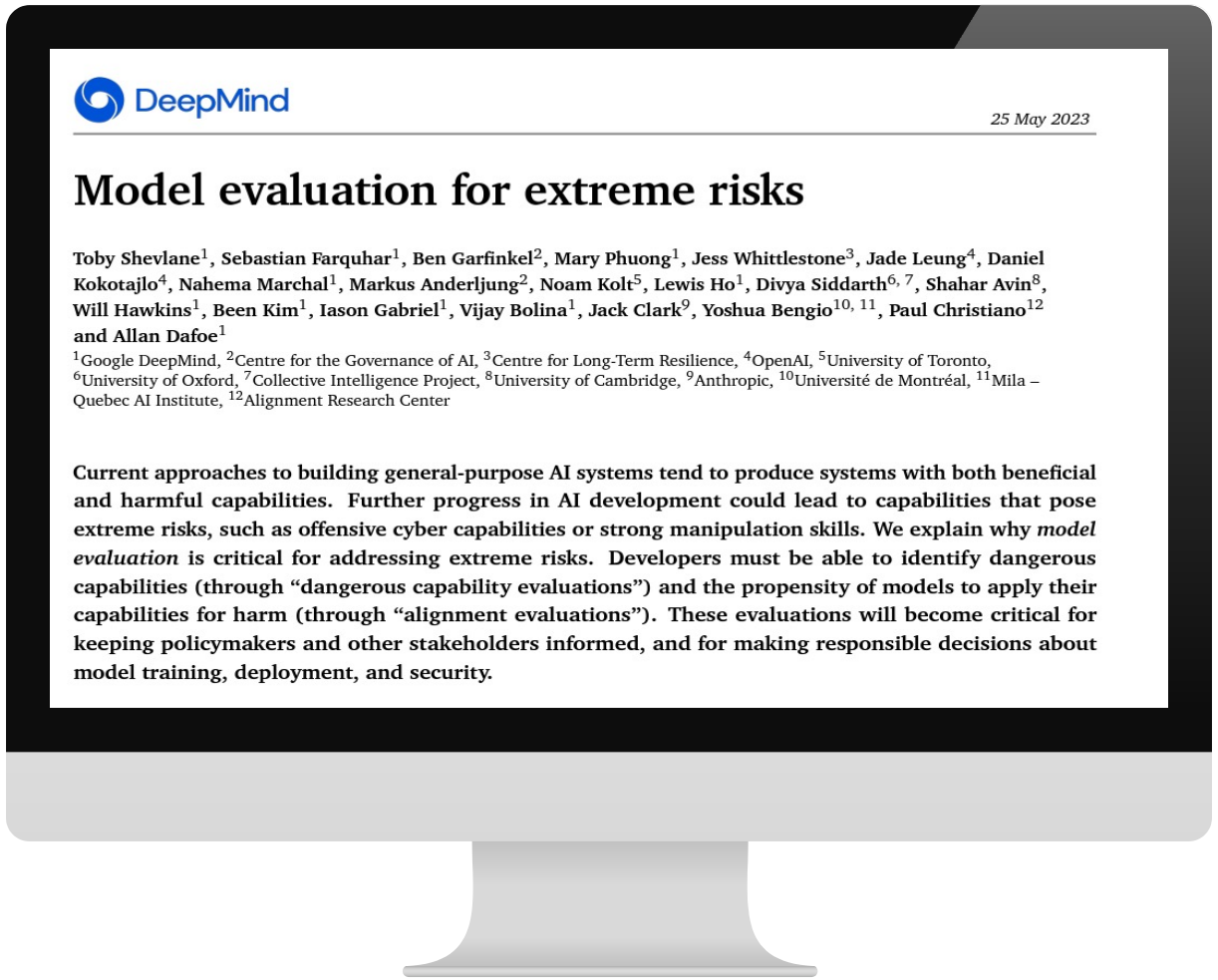
- *Solution (Government Side): 1.* To address the dangers we have discussed, governments need to **develop knowledge**, **expertise**, and **talent** in-house. 2. The governments need to recruit a lot more. 3. They should work with other governments around the world.

  - ❖ A very dangerous AI can be coming from another country and harming us, so we have to move gradually towards international coordination.

How Much A.I. Should Scare Us?

- **One advantage**: These training rounds require **so much power** and **computing** that it is probable that secret people know where they are.
- **Another advantage**: These test and evaluation frameworks are being funded and invented, and there is a verbal agreement that various **companies should share their tests and evaluations**.
- **One definition of risk**: extreme risk is 15,000 or more people dead as opposed to an individual person harmed, which is arbitrary.
- *Let's see the DeepMinds's Paper!*

Model Evaluation for Extreme Risks



DeepMind

25 May 2023

Model evaluation for extreme risks

Toby Shevlane[1], Sebastian Farquhar[1], Ben Garfinkel[2], Mary Phuong[1], Jess Whittlestone[3], Jade Leung[4], Daniel Kokotajlo[4], Nahema Marchal[1], Markus Anderljung[2], Noam Kolt[5], Lewis Ho[1], Divya Siddarth[6, 7], Shahar Avin[8], Will Hawkins[1], Been Kim[1], Iason Gabriel[1], Vijay Bolina[1], Jack Clark[9], Yoshua Bengio[10, 11], Paul Christiano[12] and Allan Dafoe[1]

[1]Google DeepMind, [2]Centre for the Governance of AI, [3]Centre for Long-Term Resilience, [4]OpenAI, [5]University of Toronto, [6]University of Oxford, [7]Collective Intelligence Project, [8]University of Cambridge, [9]Anthropic, [10]Université de Montréal, [11]Mila – Quebec AI Institute, [12]Alignment Research Center

Current approaches to building general-purpose AI systems tend to produce systems with both beneficial and harmful capabilities. Further progress in AI development could lead to capabilities that pose extreme risks, such as offensive cyber capabilities or strong manipulation skills. We explain why *model evaluation* is critical for addressing extreme risks. Developers must be able to identify dangerous capabilities (through "dangerous capability evaluations") and the propensity of models to apply their capabilities for harm (through "alignment evaluations"). These evaluations will become critical for keeping policymakers and other stakeholders informed, and for making responsible decisions about model training, deployment, and security.

The biggest extreme risks: things like **cyber attacks** and **biological attacks**.

Question: How can we make sure that these systems don't get access to weapons in any form?

Model Evaluation for Extreme Risks

- There are lots of discussions about **changing the way we build hardware** and the chips necessary for AI.
- We could build them so that we can trace them and determine if there is a large concentration of these chips somewhere in the world.
- We could also ensure that only software with the **right safety protocols** is **allowed to run**.

Surveillance Capitalism

Is our Privacy Over?

Deep Fakes

Question: Are these deep fakes something we should be excited or terrified about?

One solution: We should pursue making entities and people who post material on social media or the internet identifiable or at least known in a unique way so that if somebody does something essentially criminal, we can trace them back.

It is easier said than done. There are costs and potential downsides, but we need to consider things like this to make it harder for bad actors to exploit the current internet system, social media, and so on to destabilize our democracies.

The Problem: Social media companies, as a general rule, have a conflict of interest: **Accuracy vs. Revenue**.

Deep Fakes

- The best way to increase revenue is to increase engagement, and the best way to increase engagement is with outrage.
- **One solution:** We should pursue making entities and people who post material on social media or the internet identifiable or at least known in a unique way so that if somebody does something essentially criminal, we can trace them back.
- If we take you and we have to say something outrageous, it is going to get a lot more views than your normal professorial speeches. So, this is a cheapening of discussion in a democracy.
- The technical solutions are to mark the content where it came from, and you need to know who the users are so you can hold them accountable.

Deep Fakes

- **A major challenge**: But the fact of the matter is right now, people have been trained since birth **to believe in what they hear and what they see**. And it is a significant change for every human being to **learn that the majority of what you see may or may not be**.

- Learning **to be critical of what you see** seems to be a key component that is missing.

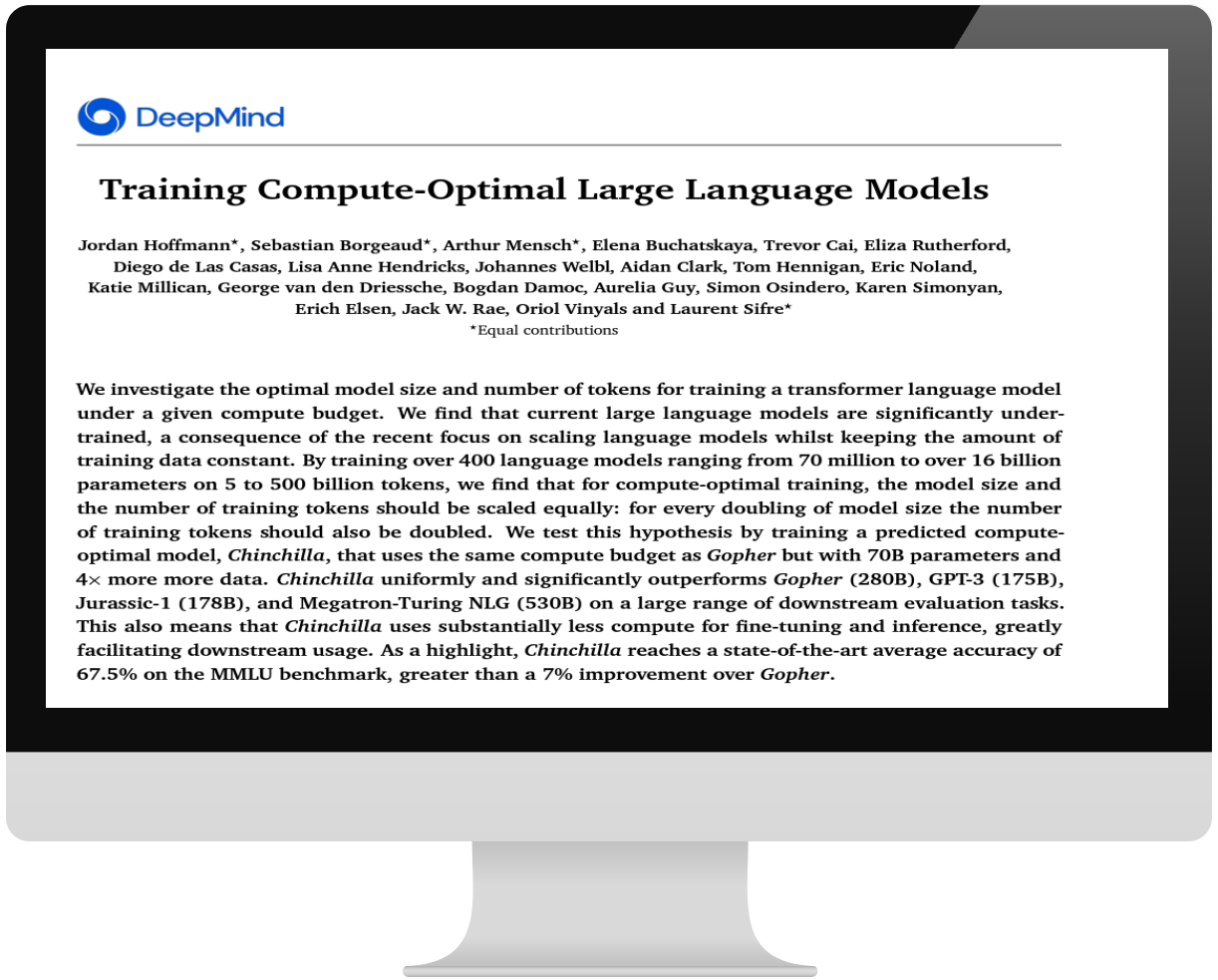
Transparency

- **A real issue**: you have trained a model, and the model has learned something, but it **hasn't the ability to tell you what it knows**. In other words, it has learned something and you ask it what it learned, and it doesn't know how to tell you what it learned.
- The challenge: if you release it, you run the risk of catastrophe, and if you do not release it, you also run into the issue of catastrophe because there is no way to test it.
- The rest of the challenge: If you test it, you have a catastrophe. It is a conundrum.
- We are in trouble if we do not solve this problem quickly before we get to really dangerous AI systems with a lot of knowledge that could be exploited for evil.

Democratizing AI

- **Democratizing AI**: Make AI accessible to as much of the world as possible.
- The future is that everybody has their own model and their own data because that is what really makes a meaningful difference. A hundred gigabytes of information and two gigabytes of knowledge!
- If crypto is libertarian, AI is communist!
- Artificial intelligence has the ability to be incredibly decentralized and incredibly liberating.

What is a better Model?



Many of the data showed that it isn't only about the model size (having 100 billion or 500 billion parameters) but also about having better and more diverse data!

As you have more diverse data, you have more and better outcomes.

Misinformation, Deep Fakes, and the Polarizing Effect

Do They Undermine Democracy?

Digital De-aging

- Harison Ford is 81 years old, but in the recent remake of the Indiana Jones series, they needed him to be 40 years younger for part of the movie. They couldn't do it effectively with makeup, so they used AI-enabled Digital De-Aging.
- Question: How many think it is troubling, and how many think it is okay, like the use of makeup in the past?
- **Digital De-aging** is only the beginning of the question posed by AI in movie making and the Arts because the same technologies that enable filmmakers to de-age Harison Ford could be used in a more ambitious way.
- *Question:* Harison Ford is still with us, but what about the great actors of the past and great stars whom we would love to see in a new role?
  - Wouldn't you like to see Marlon Brando, not in a rerun of The Godfather but in a new movie?

Digital De-aging

Michael Sandel

- Some say that he may not consent because he is long gone! What if he had signed a waiver before he passed?
- 1. Want a disclaimer, 2. The dignity of the actor, even an actor who has agreed in advance that we are seeing a performance, whether his/her performance or something else.
- Even the best human virtual comeback has been used, it cannot capture **human authenticity** or **human presence**.
- There seems to be **a deep human value at stake** here beyond consent, choice, and the waiver form signed in advance.

Virtual Immortality

- There are now companies that will enable you to **create a digital avatar** of yourself based on all of your social media posts, emails, and personal data, and when we die, we are able to access that data and avatar to whomever we want.
- Enabling the dead to engage in ongoing conversations with us.

Conclusion

- It is the most fundamental question posed by the age of AI, chatbots, and big data, beyond worries about jobs, fairness, privacy, and even democracy.
- The deeper question about whether human authenticity, dignity, and human presence, are fundamentally at stake.
- Will new technologies lead us, or are they already leading us and our children to confuse virtual community and human connection for the real thing?
- If they do, then we **may lose something precious** about what it means to be human.

➢ Now, we return to the question: **“What does it mean to be human in this situation?”**

Thank You

Parham Zilouchian

p.zilouchian@gmail.com

Telegram: @parham_zm

Any Questions ?