HLMI:
HOW TO MANAGE ITS IMPACT

GEOFFREY HINTON OVERVIEW

- British-Canadian Computer Scientist
- Known for work with deep learning, backpropagation
- Left Google because he wanted to focus more on AI safety

HLMI DEFINITION

- Unaided machines able to perform every task better and more cheaply than human workers.

HOW GOOD WILL THIS TECHNOLOGY GET?

- Consequences of new tech: beneficial, or apocalyptic?
- Hilton fears applications in elections and wars
    - Cites DeSantis and Putin as bad actors
- Next goals for tech:
    - Setting subgoals to complete a task
    - Eliminate need of micromanagement

INTERIM STEP DISTINCTION

“Don’t think for a moment that Putin wouldn’t make hyper-intelligent robots with the goal of killing Ukrainians,” he says. “He wouldn’t hesitate. And if you want them to be good at it, you don’t want to micromanage them—you want them to figure out how to do it.”

BABYAGI AND AUTOCHATGPT

- Hooking up chatbots with other programs to complete simple steps as part of a larger task
- Concern: Would this lead to robots rerouting all power to their chips?

PREVENTING FUTURE DISASTERS

Suggests modeling development & use of dangerous AI after chemical weapons ban

Questions current modes of social organization

Author of piece questions this "Robot Overlords" vs "Simply upending job market" dichotomy

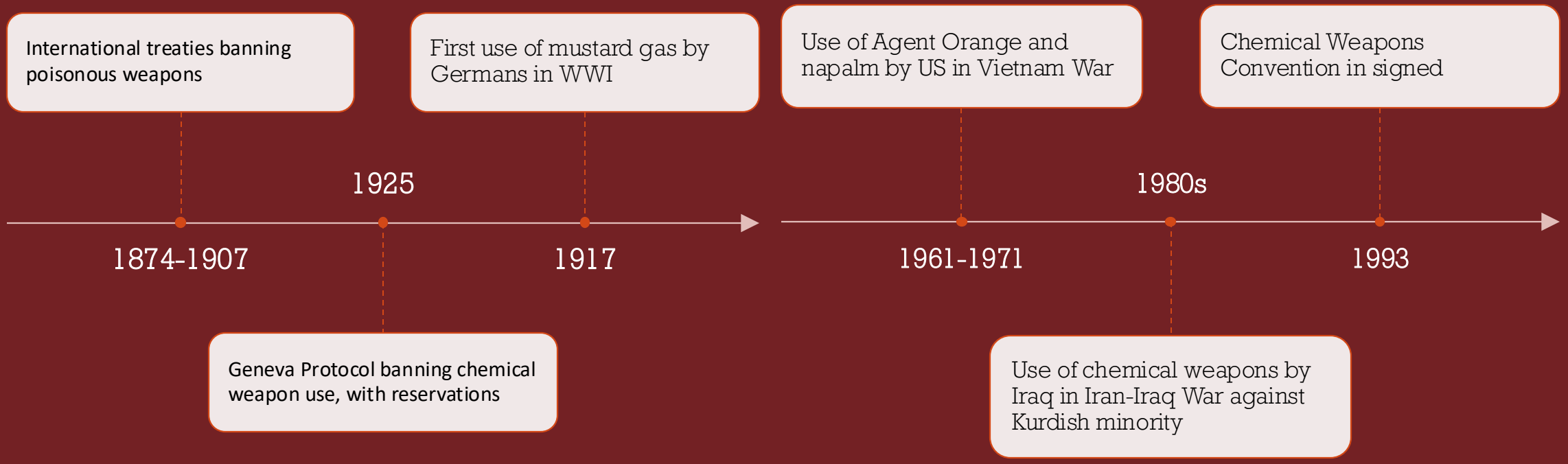
META'S CHIEF SCIENTIST DISAGREES:

BUT WILL MACHINES DOMINATE SIMPLY BECAUSE THEY ARE SMARTER?

LECUN POSITS THE OPPOSITE, POTENTIALLY.

WARFARE REGULATION

HAVE REGULATIONS ON CHEMICAL WARFARE BEEN EFFECTIVE?



PROPOSALS

1. Council of Europe (46 countries) is developing a legally binding treaty for AI

- Requires each nation to individually ratify & implement
- Ends uses with some problematic cases (facial recognition)
- Focus on ensuring human rights, democracy
- European Commission on behalf of EU signed in Sep 2024

2.UN adopted AI framework in 2019

- Ethical impact assessments, environmental assessments, promotion of gender equality, not used for mass surveillance
- Sole avenue for Global South contribution
- Sincerity on part of Russia and China is.... questionable

ECONOMIC REGULATION

PROPOSALS

1. FTC Probing of OpenAI

- Focus on promoting consumer protection
- Focuses on uses of data/data privacy, risk assessment and mitigation, biases
- Currently pursuing anti-trust inquiries for Microsoft and OpenAI partnership

2.Penguin Books and AI Training

- Acknowledges transformative aspect of generative AI on creative businesses
- Favoring human creativity
- Advocate for IP rights
- No to AI training

DON'T LOOK UP

Metaphor for ignoring climate change akin to that of AI revolution

Potential Discussion Questions:

1. *Despite over a century of attempting to prevent the use of chemical weapons in war, we still struggle to eliminate all uses. How does this observation relate to future instances of regulation of AI?*
2. *How much power would a bad actor have to attain to be able to use these dangerous AI systems to enact harm?*

Q & A

Potential Discussion Questions:

3. With all these separate spheres of potential regulation vying for attention, do we need a more universal, concentrated push to demand regulation for multiple spheres at once?

4. Should we be considered that AI pioneers leave trailblazing companies because they feel like they cannot work on AI safety while working for that company?