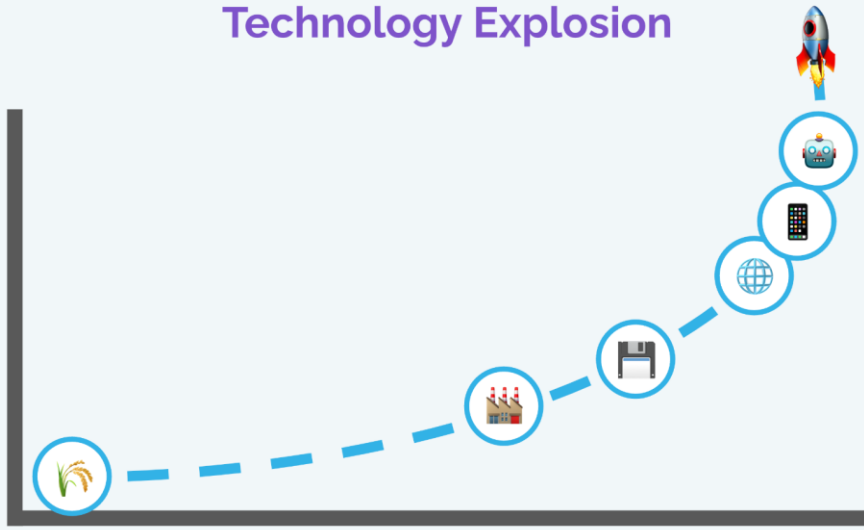
BlastPoint

Modern Artificial Intelligence

Ethics & Safety Overview

November 2024

We are living through a Technology Explosion



Call Under Review – Ethical Considerations of AI

AI will impact and challenge humanity for the rest of history.

- **Short Term:** Job Displacement, Disinformation
- **Medium Term:** End of Capitalism, Human Fulfillment
- **Long Term:** Machine Rights, End of Human Supremacy
- **Existential:** What are we? Why are we here?

Google / Stanford University

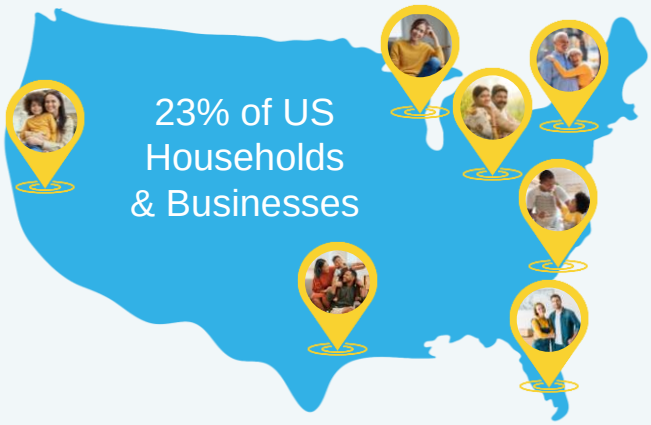
P = 1/1000 = 0.1%

Tomer Borenstein

- Rachael Borenstein's Husband
- Co-Founder & CTO @ BlastPoint AI
- Grew up in Israel & Thailand
- BS, Comp Eng. & MS, ML/AI @ CMU
- GYF Customer

BlastPoint is a Customer AI Platform for Highly Regulated Industries

Carnegie Mellon University



BlastPoint

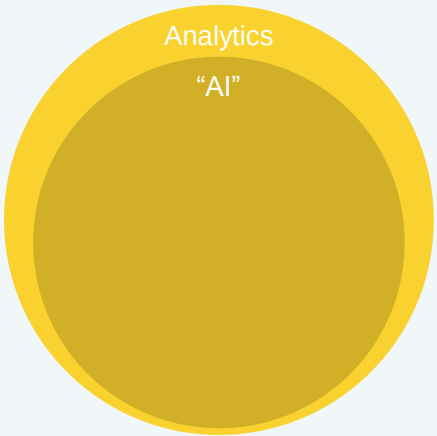
Some Key Definitions, First.

Analytics

Anything we do with data: collecting it, visualizing it, dashboarding it, modeling it, forecasting it, summarizing it, etc…

e.g. Excel, Power BI

Some Key Definitions, First.

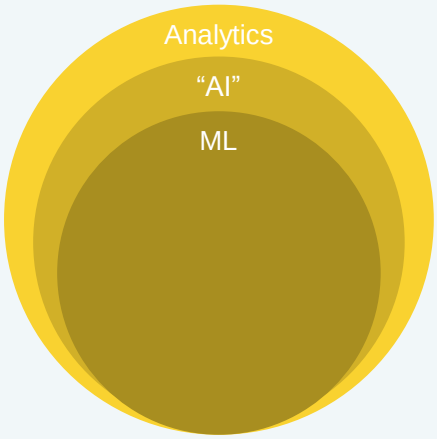


"A.I." (Artificial Intelligence)

~~Doesn't mean anything anymore.~~

Fine: software that performs tasks that would traditionally require human intelligence.

Some Key Definitions, First.

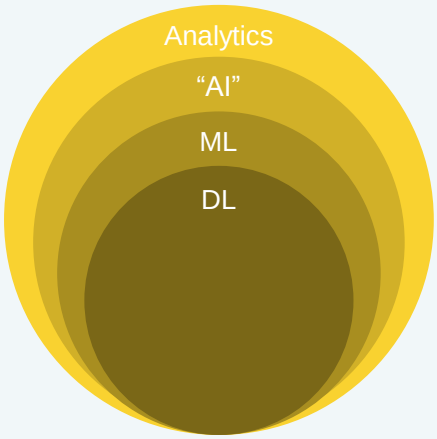


Machine Learning

Algorithms that become better at performing certain tasks the more data they are exposed to.

e.g. Linear Regression, OCR, Fraud Detection, Weather Forecasting, Facial Recognition, and so much more…

Some Key Definitions, First.

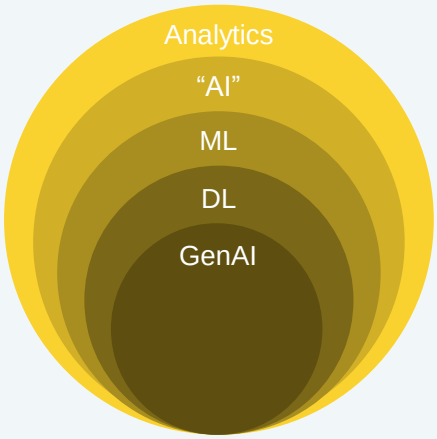


Deep Learning

A specific class of Machine Learning algorithms called Neural Networks, inspired by (but are much more simplified than) human brains.

e.g. Healthcare Diagnostics, Natural Speech Recognition & Synthesis, Image Recognition, and more…

Some Key Definitions, First.

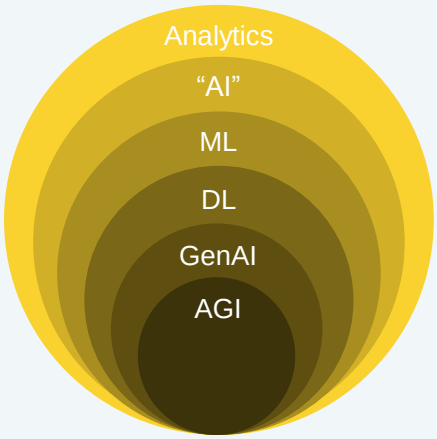


Generative AI

Sparked from the 2018 paper "Attention Is All You Need", a new class of "Transformer-Based" Neural Networks that can generate text, images, sound, and videos.

e.g. ChatGPT, Claude, Stable Diffusion, Midjourney, NotebookLM

Some Key Definitions, First.



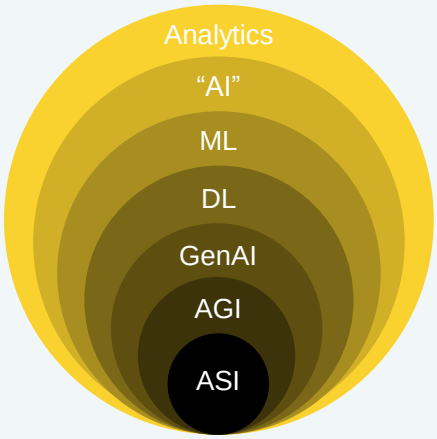
Artificial General Intelligence

Likely-"agentic" AI systems that can perform any task as well as a human could.

AGI is still hypothetical, but for agentic AI examples:

- OpenAI's o1 model
- Anthropic's Claude Computer Use

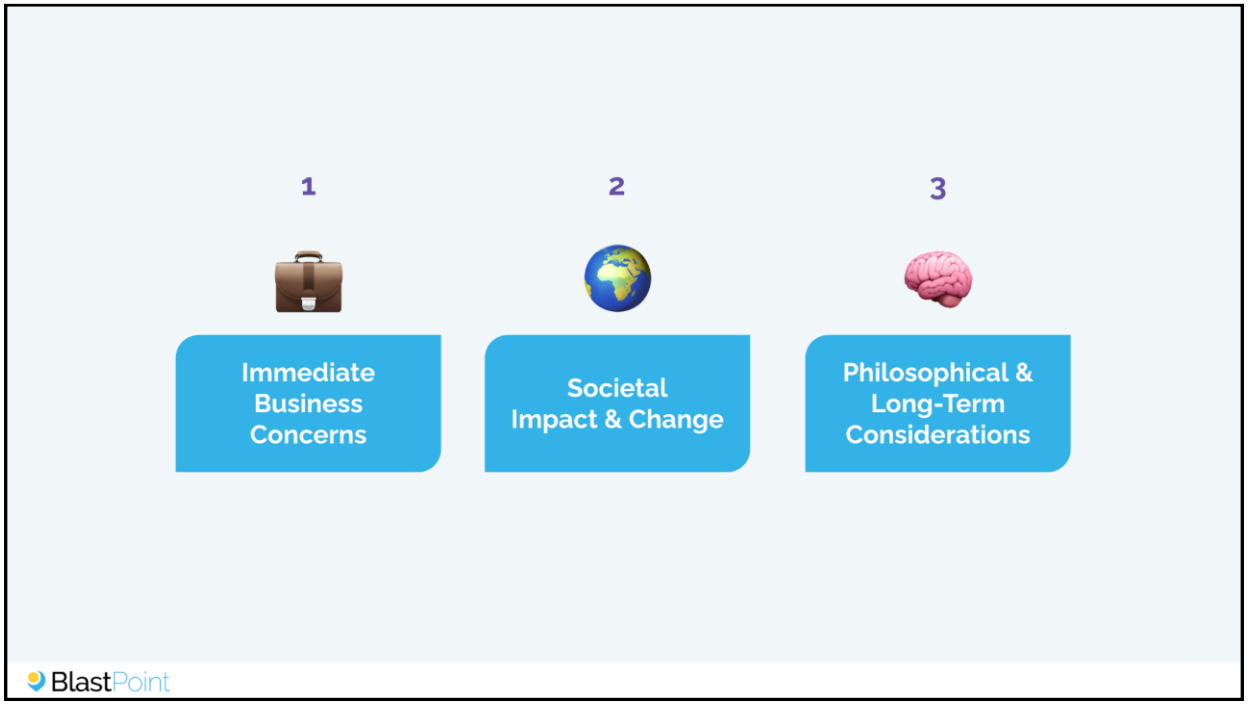
Some Key Definitions, First.

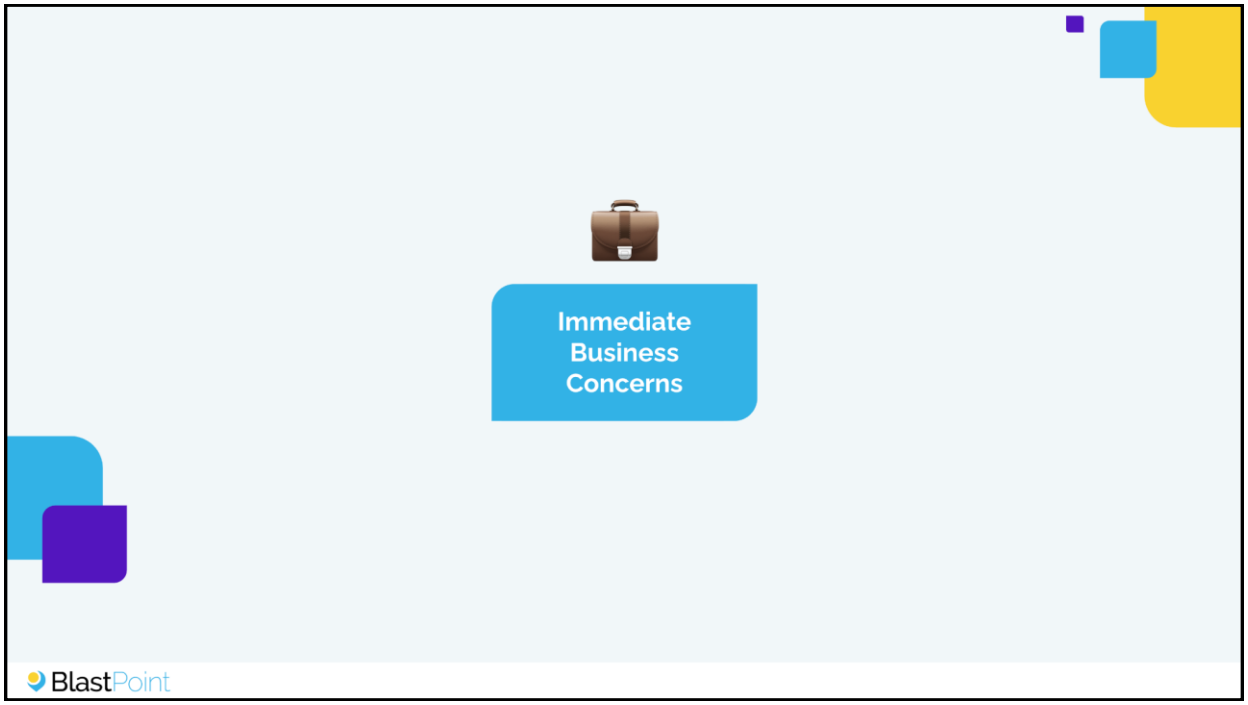


Artificial Super Intelligence

AI systems that are smarter than humans, and can recursively improve themselves, potentially resulting in "The Singularity".

Call Under Review – Ethical Considerations of AI





Data Privacy in GenAI Systems

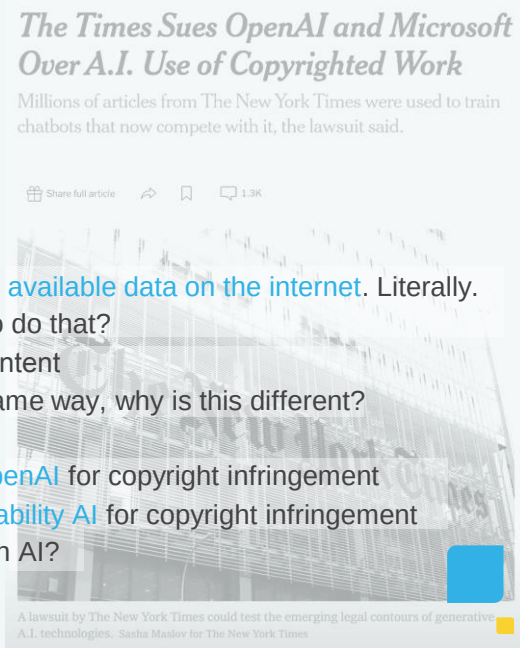
As Business Consumers

- We often deal with sensitive data at work.
- Have clear policies about what employees can/can't input into AI systems.
- Use AI tools with licenses that prohibit from training on your data, or…
- Use open source AI tools hosted in your own environment.

Data Privacy in GenAI Systems

As Content Creators

- GenAI tools are often trained on all available data on the internet. Literally.
- Should AI companies be allowed to do that?
  - Against: no, that's not their content
  - For: well, humans learn the same way, why is this different?
- In the news:
  - The New York Times sues OpenAI for copyright infringement
  - Artists sue Midjourney and Stability AI for copyright infringement
- Do you own what you generate with AI?



Accountability & Fairness in AI/ML

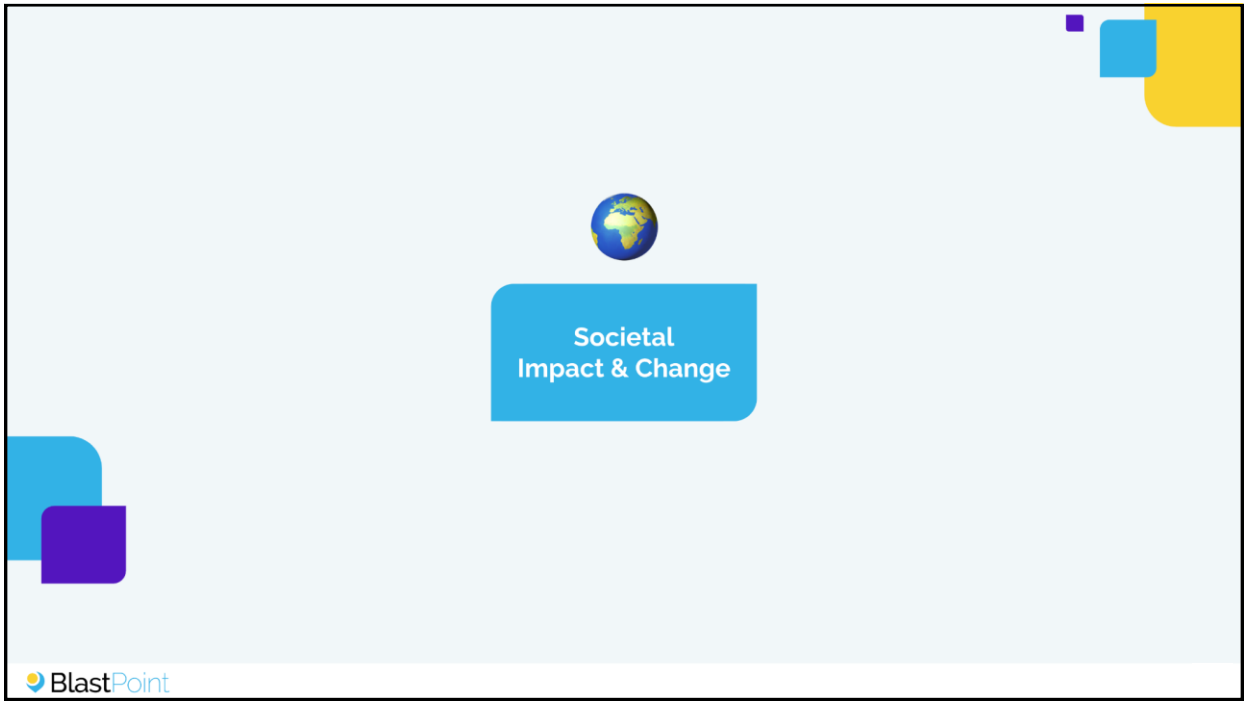
Auditability & Explainability

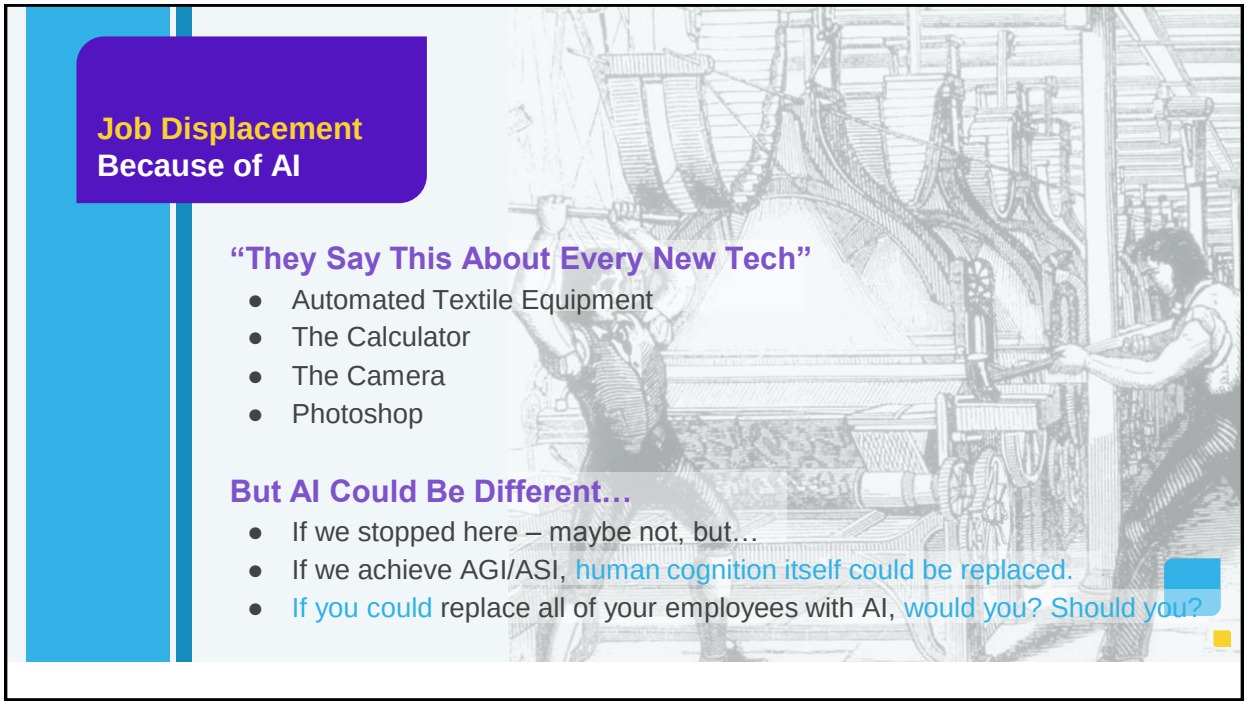
- You should understand how and why your models make decisions (esp. in regulated environments).
- Some models have better explainability than others, choose wisely.
  - Neural Networks: powerful, but extremely poor explainability
  - Linear Regression: limited, but very good explainability

Accountability & Fairness in AI/ML

Bias & Fairness

- Our world is biased; so, naturally, any data about the world is also biased.
- AI/ML can unintentionally propagate bias (e.g. EV adoption).
- Choose inputs to your models carefully, consider excluding sensitive attributes.
- Have post-training checks-and-balances on sensitive attributes.
- Ensure your training data is representative of your population as much as possible.
- As we give AI more decision-making power, who is accountable? (e.g. self-driving cars)


Societal
Impact & Change
BlastPoint


Job Displacement
Because of AI
"They Say This About Every New Tech"
Automated Textile Equipment
The Calculator
The Camera
Photoshop
But AI Could Be Different…
If we stopped here – maybe not, but…
If we achieve AGI/ASI, human cognition itself could be replaced.
If you could replace all of your employees with AI, would you? Should you?

Job Displacement Because of AI

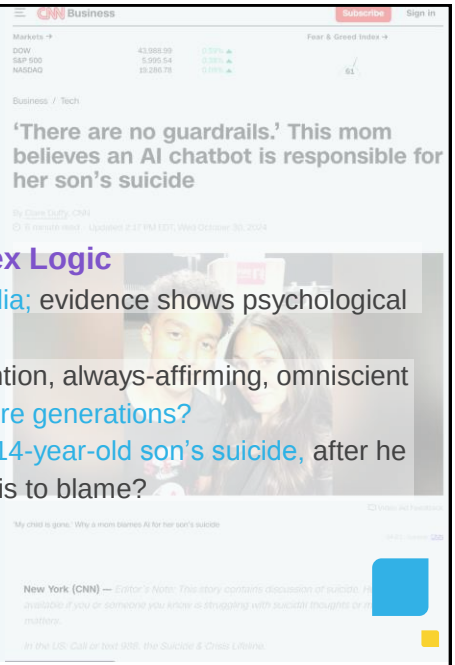
Medium-Long Term Societal Restructuring

- The End of Capitalism?
- What will humans do with our time if we don't have to work? How will we achieve fulfillment?
- Will we achieve a post-scarcity future or a cyberpunk one where corporations control all AI resources and inequality is rampant?

Psychological & Societal Impacts of AI Relationships

AI is Better at Socializing than Complex Logic

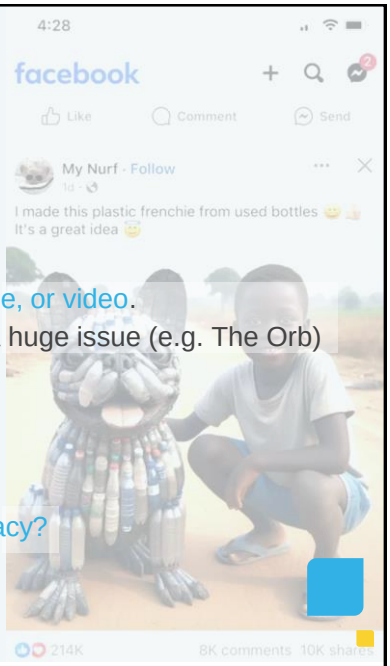
- GenZ grew up with ubiquitous social media; evidence shows psychological harm for some users.
- How will having "always on", infinite-attention, always-affirming, omniscient AI friends impact the development of future generations?
- In the news: Mom sues Character.AI for 14-year-old son's suicide, after he fell in-love with roleplaying chatbot. Who is to blame?



Information Integrity in the Era of AI

Welcome to the Post-Truth Age!

- Soon, you won't be able to trust any text, image, or video.
- In fact, Human Verification will likely become a huge issue (e.g. The Orb)
- In the news:
  - Popular fake Facebook posts
  - Election Deepfakes
  - The Dead Internet Theory
- What can we do besides focus on Media Literacy?



Philosophical & Long-Term Considerations

BlastPoint

Philosophical & Long-Term Considerations

Consciousness & Machine Rights

- The Turing Test used to be the gold standard for "conscious AI", but we've flown past it so fast without noticing.
- The debate about what consciousness is may never be resolved, but if you have an AI system that is Agentic & Multi-modal, has long-term memory, re-trains itself on its own "experience", how is it different from any other conscious being?
- "AI with Rights" will likely happen first (and soon) via Incorporation.

Philosophical & Long-Term Considerations

End of Human Supremacy

- AI Alignment: building AGI & ASI systems that are aligned with human values (how can AI be aligned with humans, if humans aren't aligned?)
- If not friendly, these AI systems could pose serious Extinction Risk.
- Should humans work towards creating something that could become smarter than them?
- Race to ASI: could any world power afford to not be the first to achieve ASI?

Thanks for Listening!

Tomer Borenstein
Co-Founder & CTO
tomer.borenstein@blastpoint.com

If you'd like to stay up-to-date on AI, try:

Hard Fork

Hacker News

Gradient DISSENT
A MACHINE LEARNING PODCAST

BlastPoint