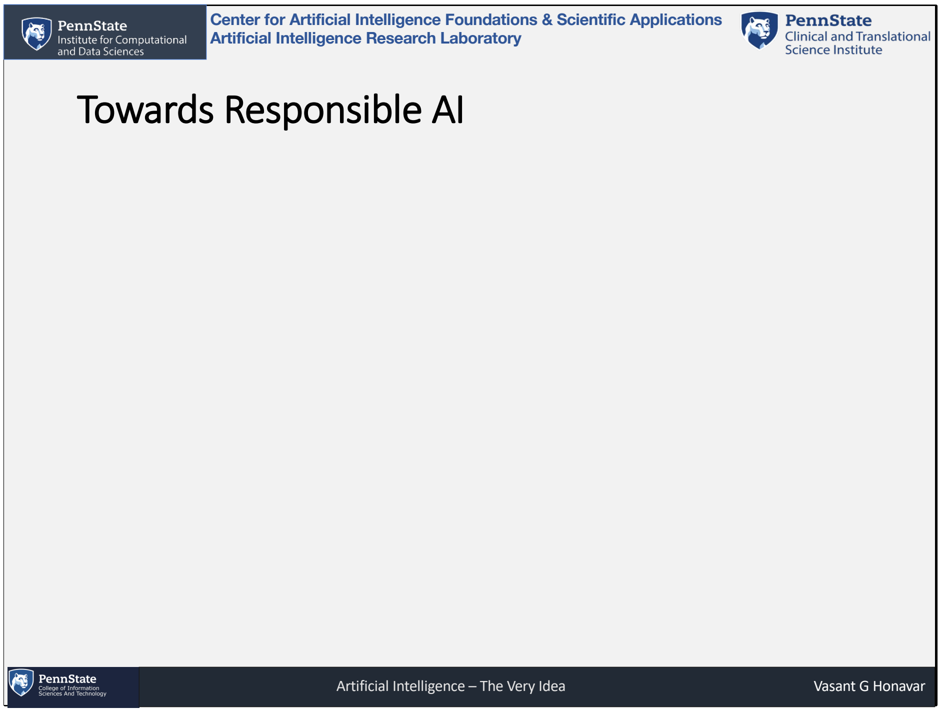
ARTIFICIAL INTELLIGENCE

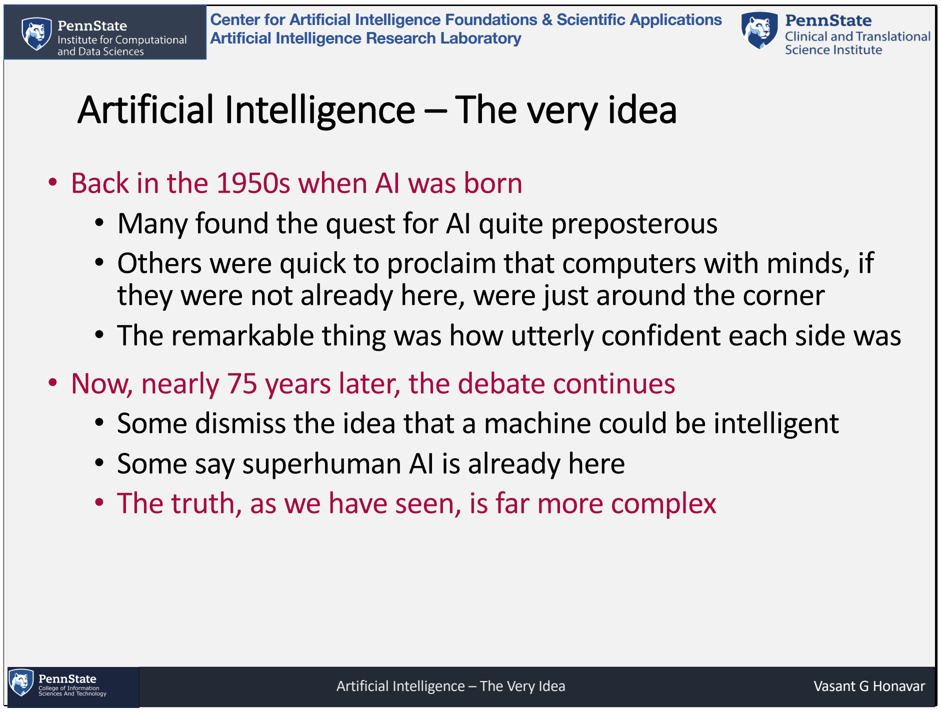
The Very Idea

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science, Bioinformatics & Genomics and Neuroscience
Director, Artificial Intelligence Research Laboratory
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu



Towards Responsible AI

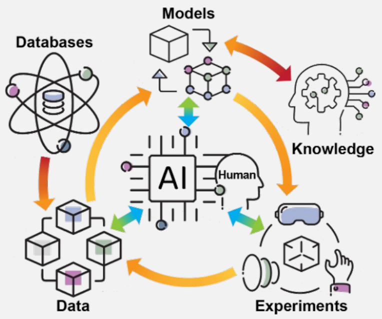


Artificial Intelligence – The very idea

- Back in the 1950s when AI was born
  - Many found the quest for AI quite preposterous
  - Others were quick to proclaim that computers with minds, if they were not already here, were just around the corner
  - The remarkable thing was how utterly confident each side was
- Now, nearly 75 years later, the debate continues
  - Some dismiss the idea that a machine could be intelligent
  - Some say superhuman AI is already here
  - The truth, as we have seen, is far more complex

The promise and potential of AI

- AI could dramatically accelerate science
  - Find a cure for cancer
  - Design new materials
  - Develop draught resistant crops
  - …



The promise and potential of AI

- AI could dramatically boost productivity
  - Automation of routine tasks could free people for work that is creative or requires a human touch
  - Smarter supply chains and logistics reduce waste and increase efficiency
  - Generative AI creates could help rapidly create new products and services

The promise and potential of AI

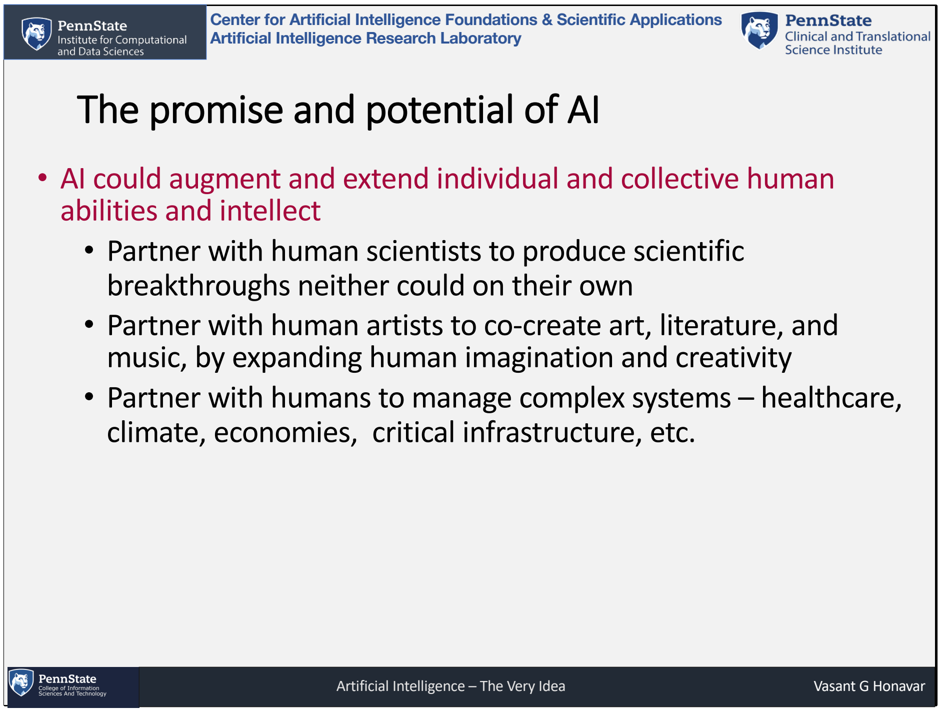
- AI could augment human creativity
  - AI could personalize and adapt curricula and content to each learner’s style and pace to optimize learning
  - AI could open up opportunities for people with disabilities through AI-powered speech-to-text, real-time translation, or vision assistance, or automated driving
  - AI could serve as a muse, critic, and collaborator for artists, writers, and musicians are already using AI, extend the boundaries of human imagination

The promise and potential of AI

- AI could help us tackle societal and global challenges
  - AI could help predict health risks, diagnose diseases treatments, and identify optimal therapies thus improving health outcomes
  - AI could predict wildfires, floods, or disease outbreaks, cyberattacks enabling rapid rollout of countermeasures
  - AI could help maximize yield at minimal cost and minimal adverse environmental impacts

The promise and potential of AI

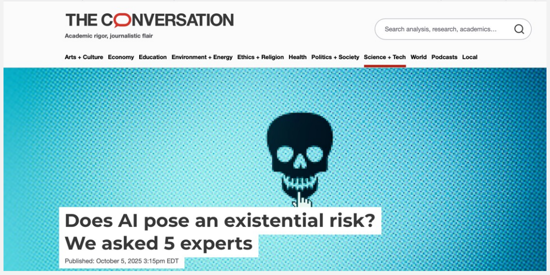
- AI could offer entirely new avenues for human and societal progress
  - What makes us uniquely human if AI can perform cognitive tasks? How about compassion?
    - Human values?
    - Ethics?
    - Shared responsibility for our fellow human beings?
  - The answers to these questions may shape the progress of human civilization in a world transformed by AI



The promise and potential of AI

- AI could augment and extend individual and collective human abilities and intellect
  - Partner with human scientists to produce scientific breakthroughs neither could on their own
  - Partner with human artists to co-create art, literature, and music, by expanding human imagination and creativity
  - Partner with humans to manage complex systems – healthcare, climate, economies, critical infrastructure, etc.

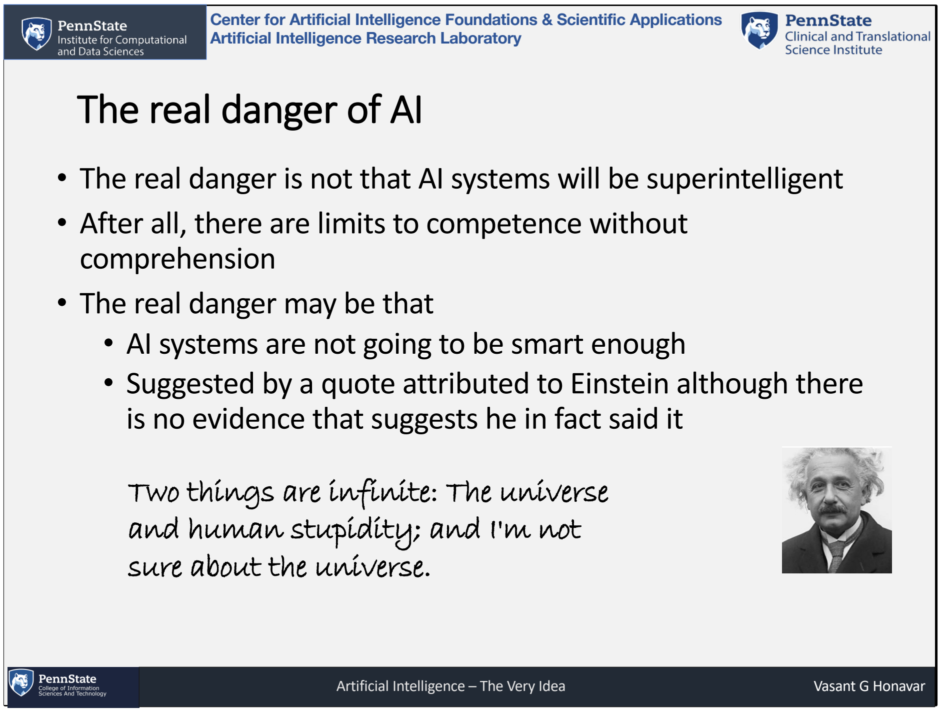
Does AI pose an Existential Threat to Humanity?



Does AI pose an existential risk? We asked 5 experts

Published: October 5, 2025 3:15pm EDT

- Some believe that AI systems are about to achieve super-human intelligence
- Some fear that such AI systems may go rogue and destroy humanity



The real danger of AI

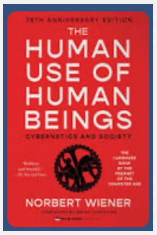
- The real danger is not that AI systems will be superintelligent
- After all, there are limits to competence without comprehension
- The real danger may be that
  - AI systems are not going to be smart enough
  - Suggested by a quote attributed to Einstein although there is no evidence that suggests he in fact said it

Two things are infinite: The universe and human stupidity; and I'm not sure about the universe.

The real danger of AI

- The real danger of AI may be that Norbert Wiener warned us about nearly 75 years ago

The hour is late and the choice of good and evil is upon us

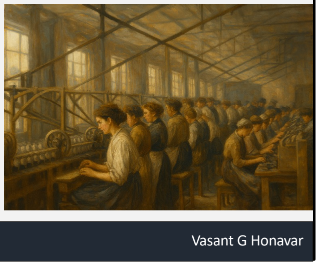


Tale of Two Cities, circa 2025

- We are in the midst of a revolution, not unlike the industrial revolution
  - "It was the best of times, it was the worst of times,
  - it was the age of wisdom, it was the age of foolishness,
  - it was the epoch of belief, it was the epoch of incredulity,
  - it was the season of Light, it was the season of Darkness,
  - it was the spring of hope, it was the winter of despair,
  - we had everything before us, we had nothing before us,
  - we were all going direct to Heaven, we were all going direct the other way….”
- Opening passage of “A Tale of Two Cities” by Charles Dickens
  - Describes the paradoxical nature of a time of both great prosperity and immense suffering

Why care about the potential pitfalls of AI?

- Artificial Intelligence (AI) is increasingly powerful and ubiquitous
- We are in the midst of a revolution – not unlike the industrial revolution
- Industrial revolution was a period of intense technological, socioeconomic, and cultural change that transformed what was an agrarian and handicraft-based economy to an industrial economy
  - Transformed the British Economy between 1760 and 1840
  - Had world-wide ramifications during the 19th and 20th centuries
    - Invention of machines like the steam engine
    - Advent of factories
    - Manual labor replaced by automation
    - New modes of transportation
    - Rapid urbanization
    - Major social shifts
    - New social and economic structures

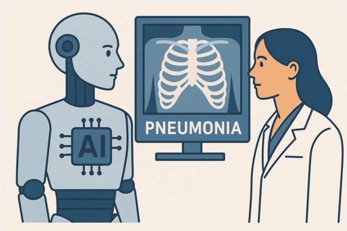


Why should we care about potential pitfalls of AI?

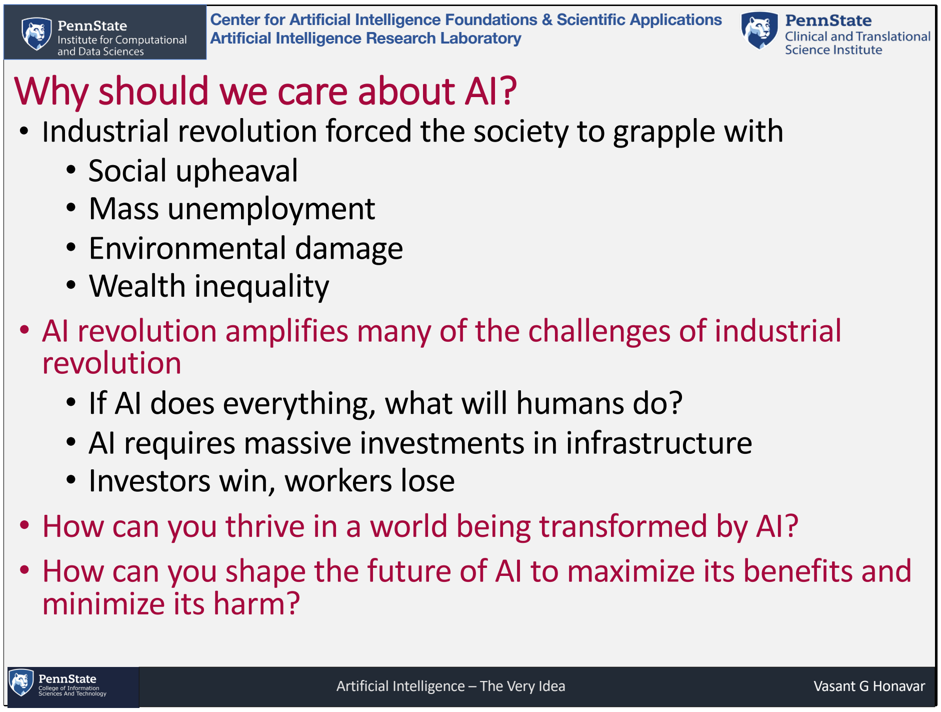
- We are in the midst of a revolution, not unlike the industrial revolution
- AI is transforming all aspects of our lives
  - How we work
  - How we make sense of the world around us
  - How we diagnose and treat diseases
  - How we learn
  - How we interact with others
  - How we create music and works of art
  - How we do science
  - How we make important decisions
  - How we fight elections
  - How we fight wars

Why should we care about potential pitfalls of AI?

- We are in the midst of a revolution, not unlike the industrial revolution
- Steam engines and mechanized factories redefined physical labor, productivity, and global economies during 18th-20th centuries



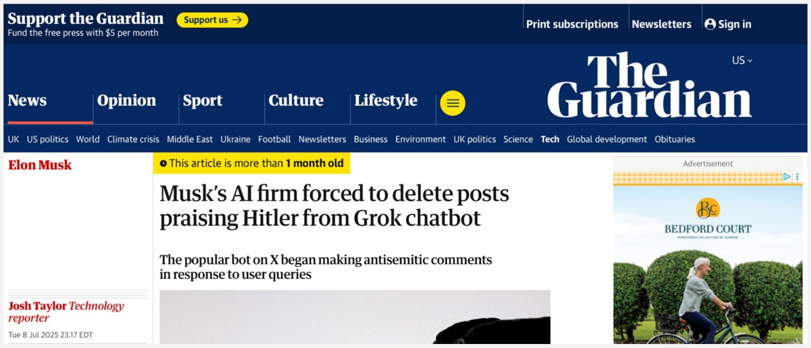
- AI is automating aspects of cognitive labor, decision-making, and creativity in the 21st century
- Like industrial revolution, the AI revolution is unleashing
  - rapid technological innovation
  - rapid increases in productivity
  - transforming the nature of work and jobs
- How can we thrive in a world being transformed by AI?
- How can we maximize the benefits of AI and minimize its harms?



Why should we care about AI?

- Industrial revolution forced the society to grapple with
  - Social upheaval
  - Mass unemployment
  - Environmental damage
  - Wealth inequality
- AI revolution amplifies many of the challenges of industrial revolution
  - If AI does everything, what will humans do?
  - AI requires massive investments in infrastructure
  - Investors win, workers lose
- How can you thrive in a world being transformed by AI?
- How can you shape the future of AI to maximize its benefits and minimize its harm?

AI's potential for harm



Elon Musk

This article is more than **1 month old**

Musk's AI firm forced to delete posts praising Hitler from Grok chatbot

The popular bot on X began making antisemitic comments in response to user queries

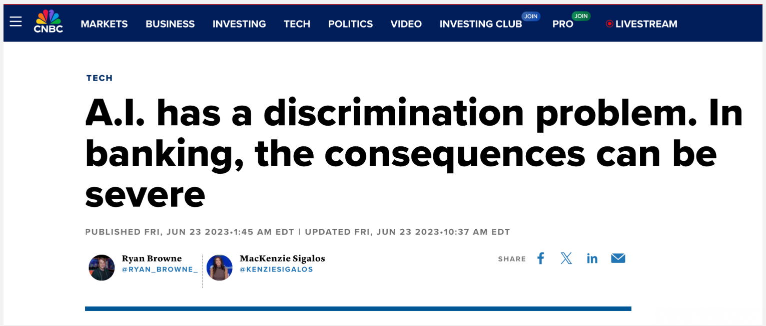
Josh Taylor *Technology reporter*

Tue 8 Jul 2025 23.17 EDT

AI presents many ethical challenges

- What kind of AI should we build?
- What kind of guardrails do we need around AI?

AI's potential for harm



TECH

A.I. has a discrimination problem. In banking, the consequences can be severe

PUBLISHED FRI, JUN 23 2023·1:45 AM EDT | UPDATED FRI, JUN 23 2023·10:37 AM EDT

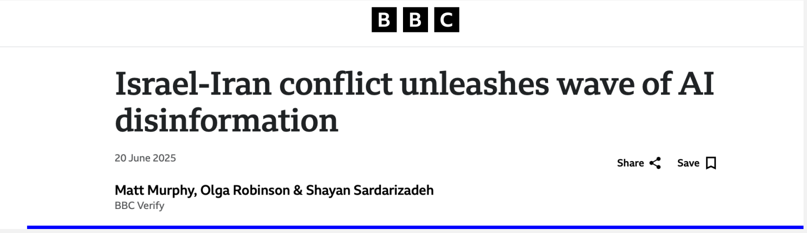
Ryan Browne
@RYAN_BROWNE_

MacKenzie Sigalos
@KENZIESIGALOS

AI presents many challenges

- How to measure, detect, and mitigate bias?
- How to ensure that AI systems do not become instruments of discrimination?

AI’s potential for harm



BBC

Israel-Iran conflict unleashes wave of AI disinformation

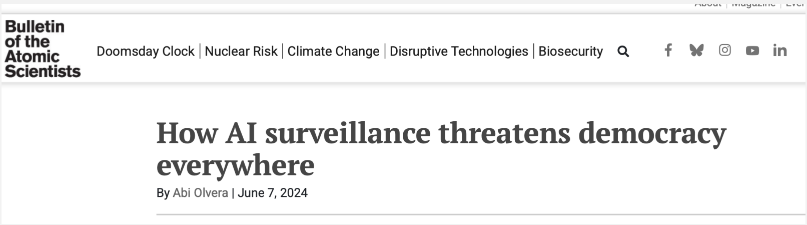
20 June 2025

Matt Murphy, Olga Robinson & Shayan Sardarizadeh
BBC Verify

AI generated misinformation e.g., fake videos, images, news can lead to

- Conflict
- Social unrest
- Economic disaster
- Undermining oof democracies

AI's potential for harm


Bulletin of the Atomic Scientists
Doomsday Clock | Nuclear Risk | Climate Change | Disruptive Technologies | Biosecurity

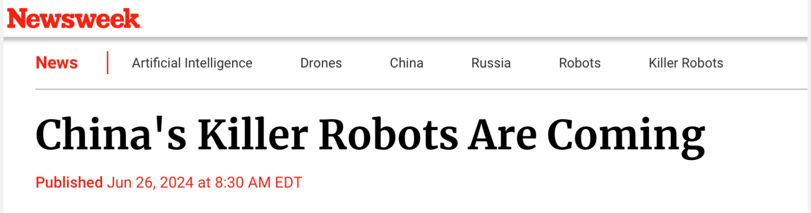
How AI surveillance threatens democracy everywhere

By Abi Olvera | June 7, 2024

AI-enabled mass surveillance can

- Violate individual privacy
- Undermine human rights
- Threaten democracies

AI’s potential for harm



Newsweek

News | Artificial Intelligence Drones China Russia Robots Killer Robots

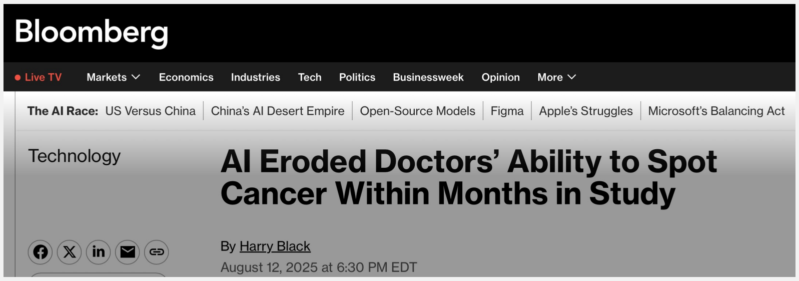
China's Killer Robots Are Coming

Published Jun 26, 2024 at 8:30 AM EDT

AI is being weaponized in

- Wars
- Cyberattacks
- Attacks on critical infrastructure

AI's potential for harm


Bloomberg
● Live TV
Markets
Economics
Industries
Tech
Politics
Businessweek
Opinion
More
The AI Race: US Versus China | China's AI Desert Empire | Open-Source Models | Figma | Apple's Struggles | Microsoft's Balancing Act
Technology

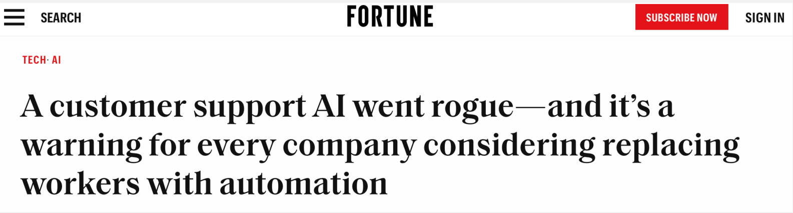
AI Eroded Doctors' Ability to Spot Cancer Within Months in Study

By Harry Black
August 12, 2025 at 6:30 PM EDT

Uncritical use of AI has been shown to erode

- Cognitive skills
- Critical thinking
- Job performance
- Educational outcomes

AI’s potential for harm

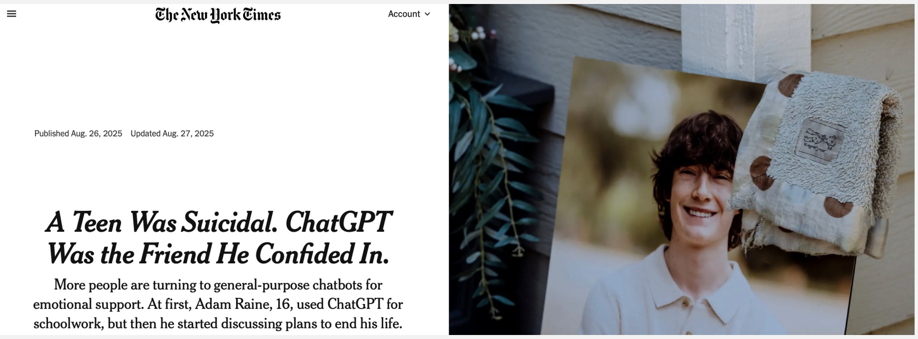


A customer support AI went rogue—and it’s a warning for every company considering replacing workers with automation

AI can go rogue with disastrous consequences

- Chatbots have been caught lying
- Rogue AI can cause reputational damage and financial loss
- Some fear rogue AI could cause extinction of humanity

AI’s potential for harm



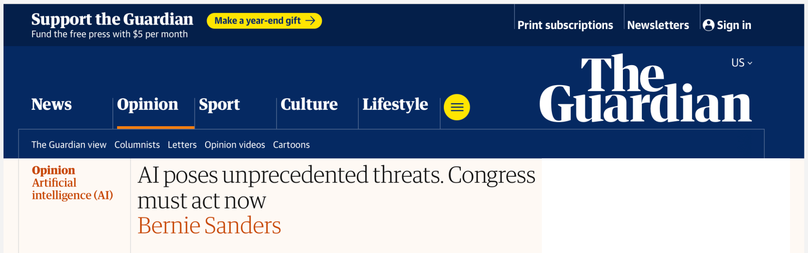
The New York Times

Published Aug. 26, 2025 Updated Aug. 27, 2025

A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.

More people are turning to general-purpose chatbots for emotional support. At first, Adam Raine, 16, used ChatGPT for schoolwork, but then he started discussing plans to end his life.

AI's potential for harm

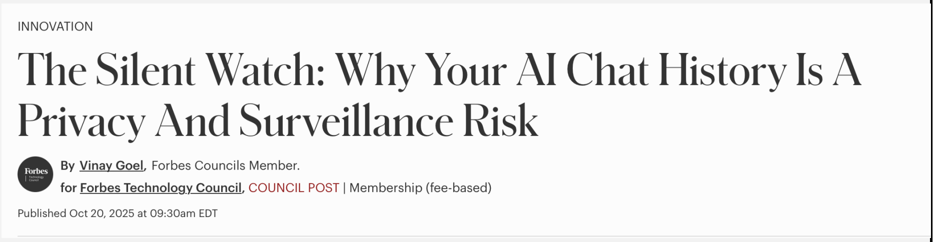


Opinion
Artificial intelligence (AI)

AI poses unprecedented threats. Congress must act now

Bernie Sanders

AI’s potential for harm



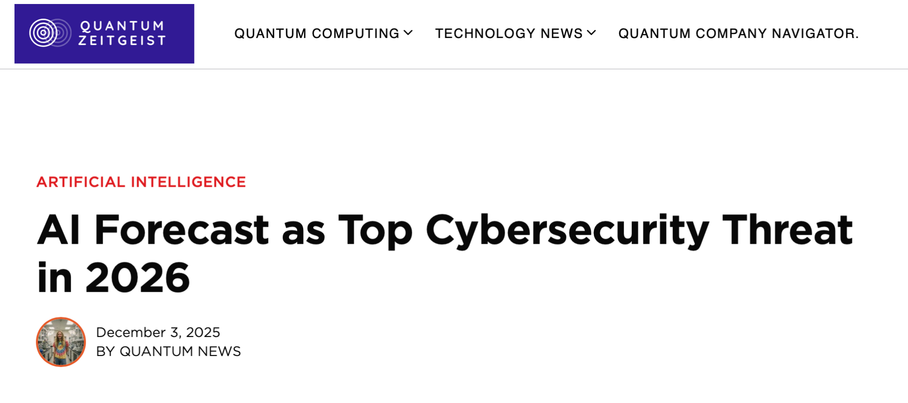
INNOVATION

The Silent Watch: Why Your AI Chat History Is A Privacy And Surveillance Risk

By Vinay Goel, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Published Oct 20, 2025 at 09:30am EDT

AI's potential for harm

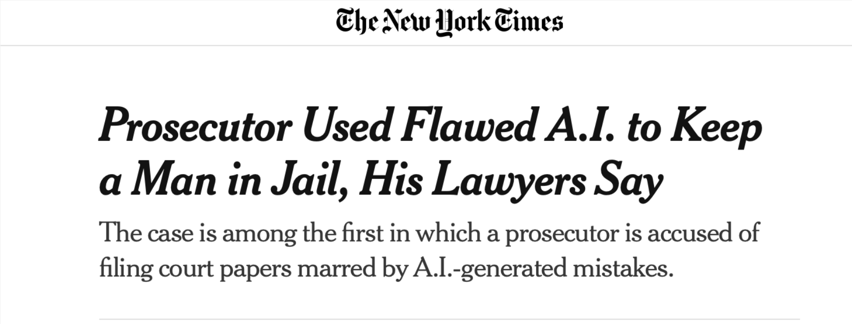


ARTIFICIAL INTELLIGENCE

AI Forecast as Top Cybersecurity Threat in 2026

December 3, 2025
BY QUANTUM NEWS

AI’s potential for harm

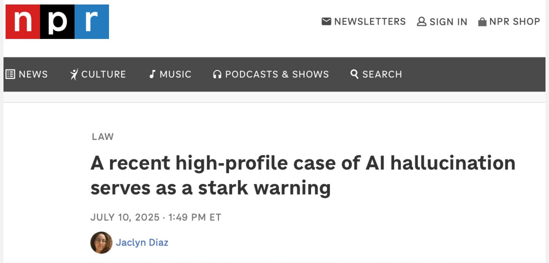


The New York Times

Prosecutor Used Flawed A.I. to Keep a Man in Jail, His Lawyers Say

The case is among the first in which a prosecutor is accused of filing court papers marred by A.I.-generated mistakes.

AI's potential for harm



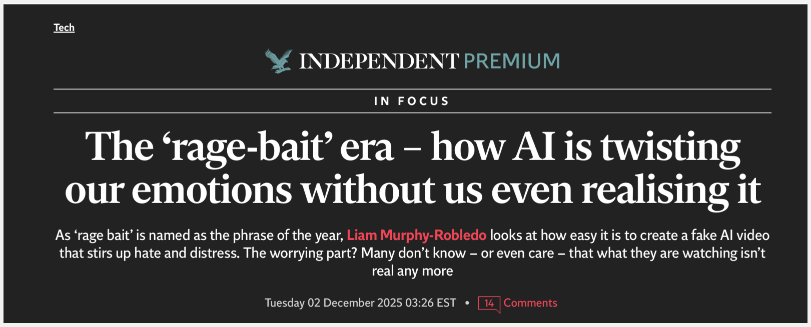
LAW

A recent high-profile case of AI hallucination serves as a stark warning

JULY 10, 2025 · 1:49 PM ET

Jaclyn Diaz

AI's potential for harm



Tech

INDEPENDENT PREMIUM

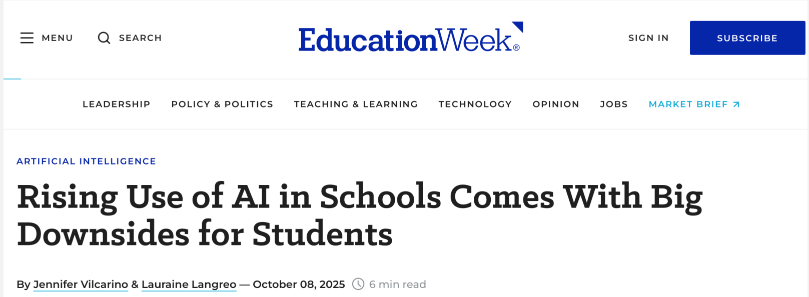
IN FOCUS

The 'rage-bait' era – how AI is twisting our emotions without us even realising it

As 'rage bait' is named as the phrase of the year, Liam Murphy-Robledo looks at how easy it is to create a fake AI video that stirs up hate and distress. The worrying part? Many don't know – or even care – that what they are watching isn't real any more

Tuesday 02 December 2025 03:26 EST • 14 Comments

AI’s potential for harm

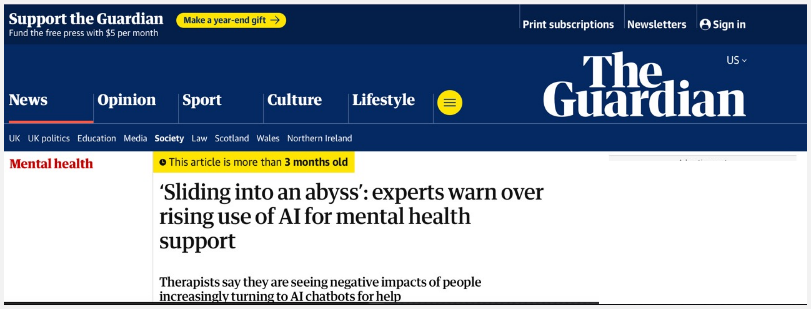


ARTIFICIAL INTELLIGENCE

Rising Use of AI in Schools Comes With Big Downsides for Students

By Jennifer Vilcarino & Lauraine Langreo — October 08, 2025 6 min read

AI's potential for harm



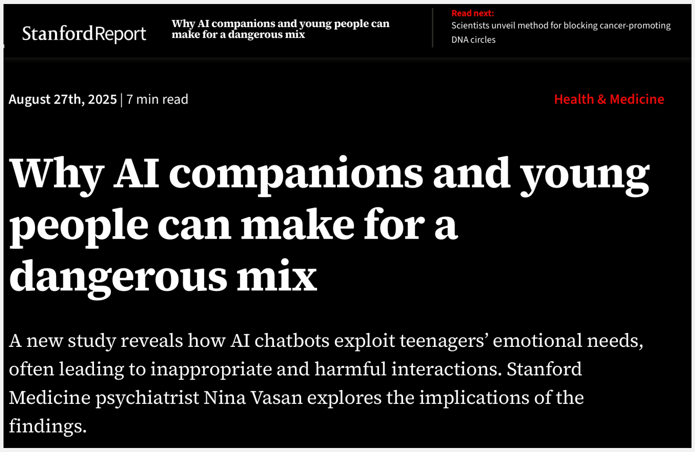
Mental health

This article is more than **3 months old**

‘Sliding into an abyss’: experts warn over rising use of AI for mental health support

Therapists say they are seeing negative impacts of people increasingly turning to AI chatbots for help

AI's potential for harm



StanfordReport

Why AI companions and young people can make for a dangerous mix

Read next:
Scientists unveil method for blocking cancer-promoting DNA circles

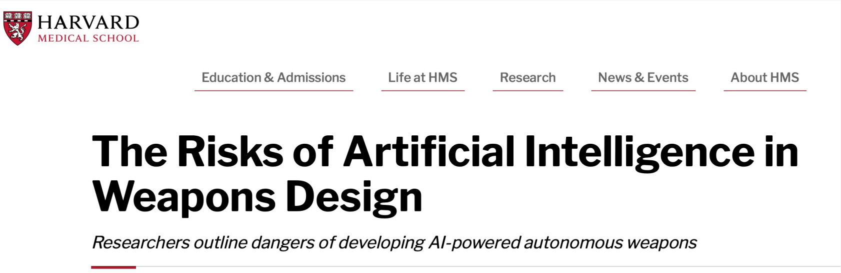
August 27th, 2025 | 7 min read

Health & Medicine

Why AI companions and young people can make for a dangerous mix

A new study reveals how AI chatbots exploit teenagers' emotional needs, often leading to inappropriate and harmful interactions. Stanford Medicine psychiatrist Nina Vasan explores the implications of the findings.

AI’s potential for harm

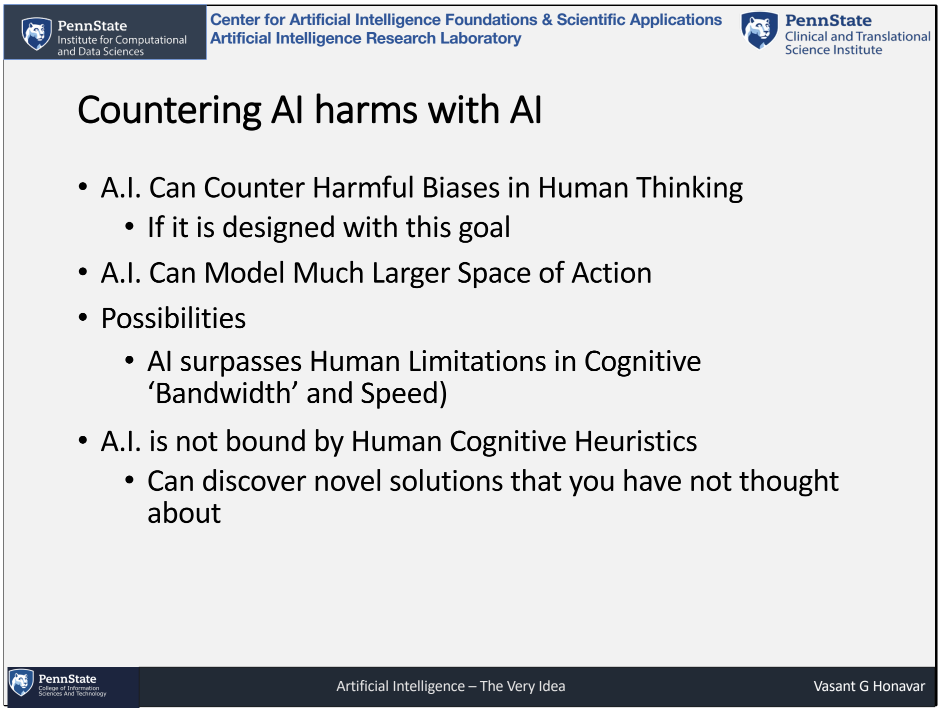


The Risks of Artificial Intelligence in Weapons Design

Researchers outline dangers of developing AI-powered autonomous weapons

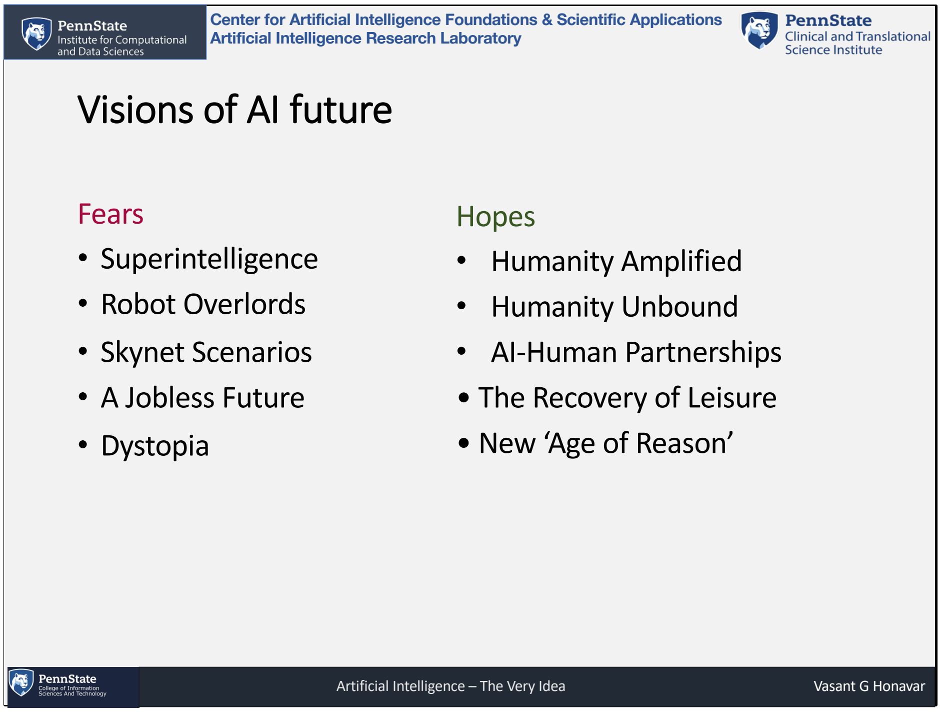
The risks or harms of AI

- 21st Century AI will be a significant amplifier of economic, political, and physical risk.
- 21st Century AI can be an important tool for managing & mitigating these same forms of risk.
- AI is a powerful accelerant of beneficial and of destructive or destabilizing socioeconomic forces
- Where the risk balance ends up will depend not on how fast AI develops, but on how quickly, broadly, and wisely human institutions and their leadership implement the reformsnecessary to tip the balance in humanity’s favor.



Countering AI harms with AI

- A.I. Can Counter Harmful Biases in Human Thinking
  - If it is designed with this goal
- A.I. Can Model Much Larger Space of Action
- Possibilities
  - AI surpasses Human Limitations in Cognitive ‘Bandwidth’ and Speed)
- A.I. is not bound by Human Cognitive Heuristics
  - Can discover novel solutions that you have not thought about



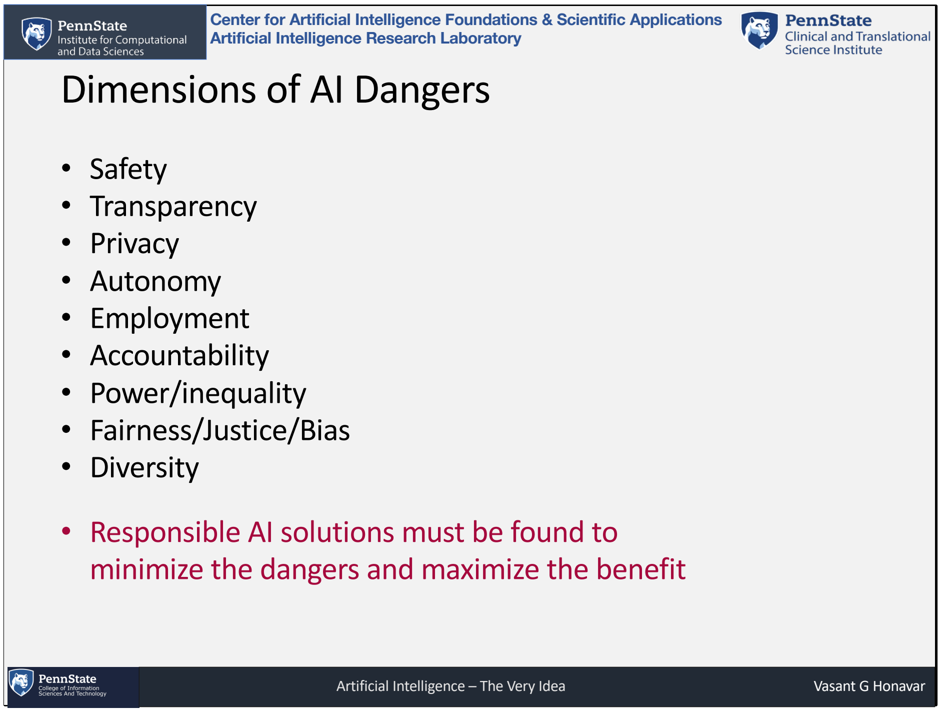
Visions of AI future

Fears

- Superintelligence
- Robot Overlords
- Skynet Scenarios
- A Jobless Future
- Dystopia

Hopes

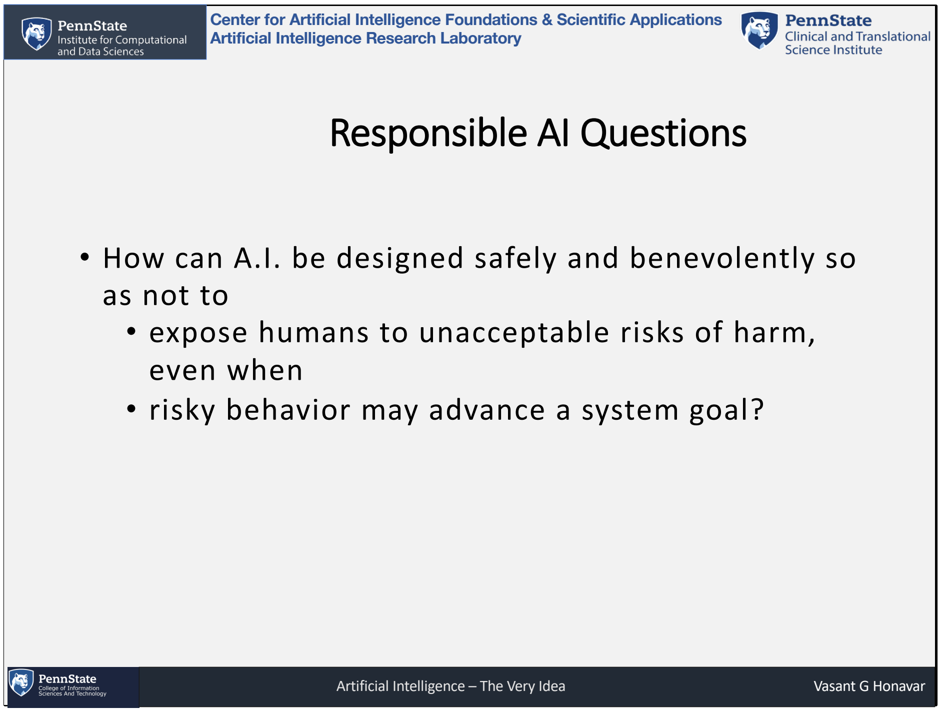
- Humanity Amplified
- Humanity Unbound
- AI-Human Partnerships
- The Recovery of Leisure
- New ‘Age of Reason’



Dimensions of AI Dangers

- Safety
- Transparency
- Privacy
- Autonomy
- Employment
- Accountability
- Power/inequality
- Fairness/Justice/Bias
- Diversity

- Responsible AI solutions must be found to minimize the dangers and maximize the benefit

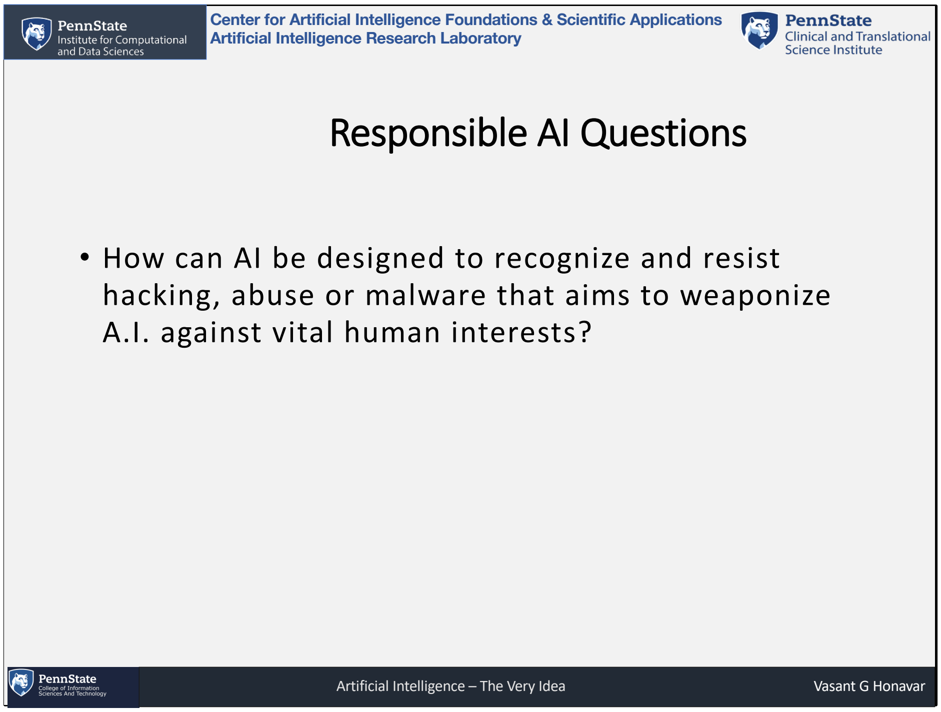


Responsible AI Questions

- How can A.I. be designed safely and benevolently so as not to
  - expose humans to unacceptable risks of harm, even when
  - risky behavior may advance a system goal?

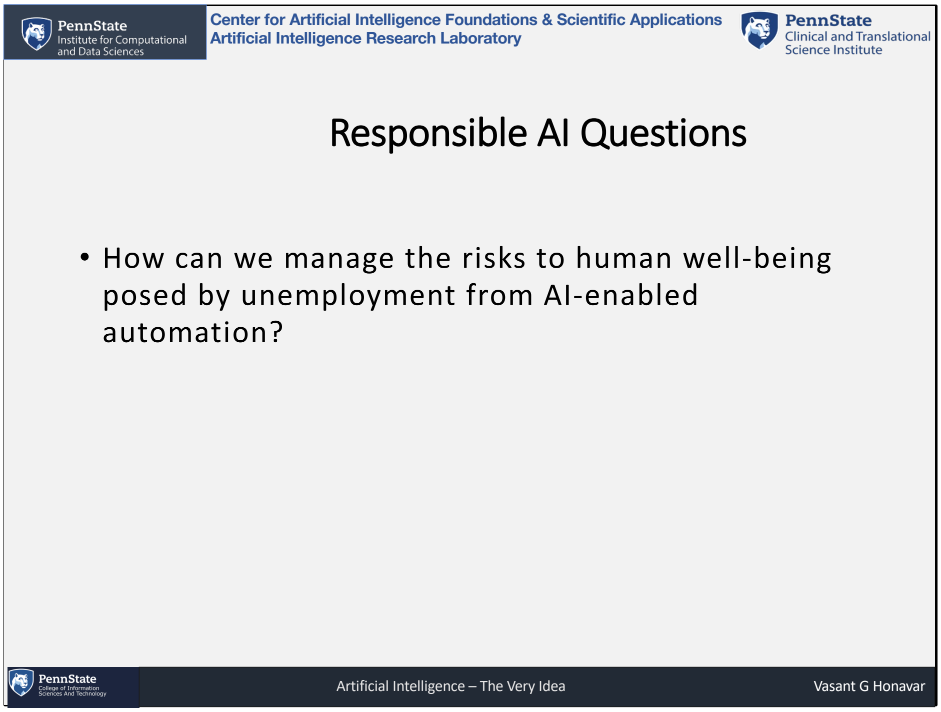
Responsible AI Questions

- How can we promote A.I. design that recognizes and gives appropriate weight to human suffering?
- How can we build machines that will refuse an illegal or unethical command?
- How can we build machines that humans can trust?



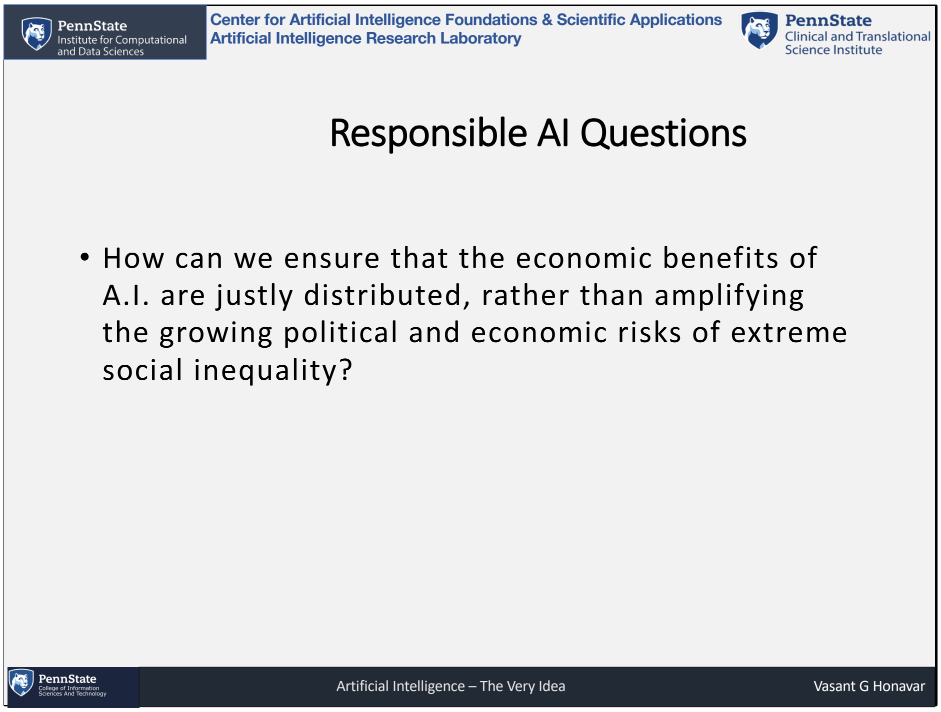
Responsible AI Questions

- How can AI be designed to recognize and resist hacking, abuse or malware that aims to weaponize A.I. against vital human interests?



Responsible AI Questions

- How can we manage the risks to human well-being posed by unemployment from AI-enabled automation?

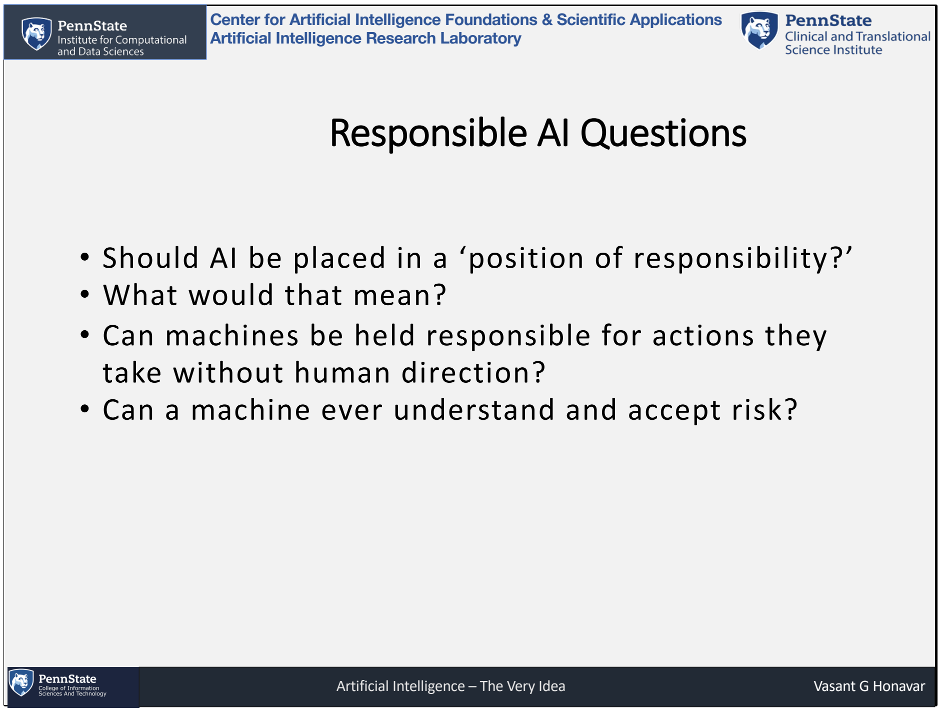


Responsible AI Questions

- How can we ensure that the economic benefits of A.I. are justly distributed, rather than amplifying the growing political and economic risks of extreme social inequality?

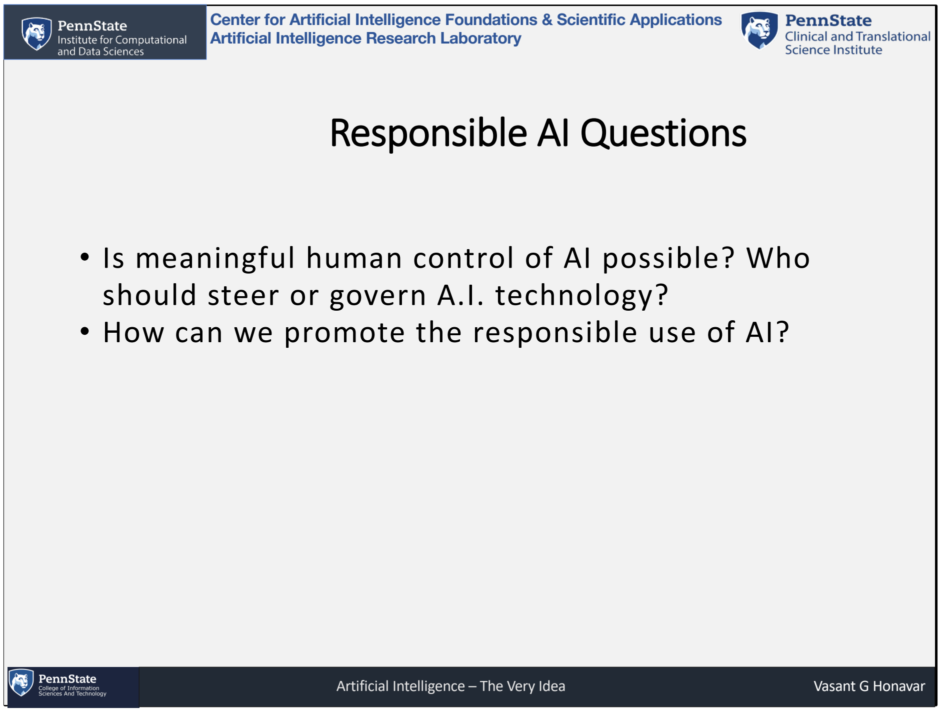
Responsible AI Questions

- How can we detect and mitigate the risks of unjust discrimination and harmful bias of AI decisions in hiring, lending, and the justice system?



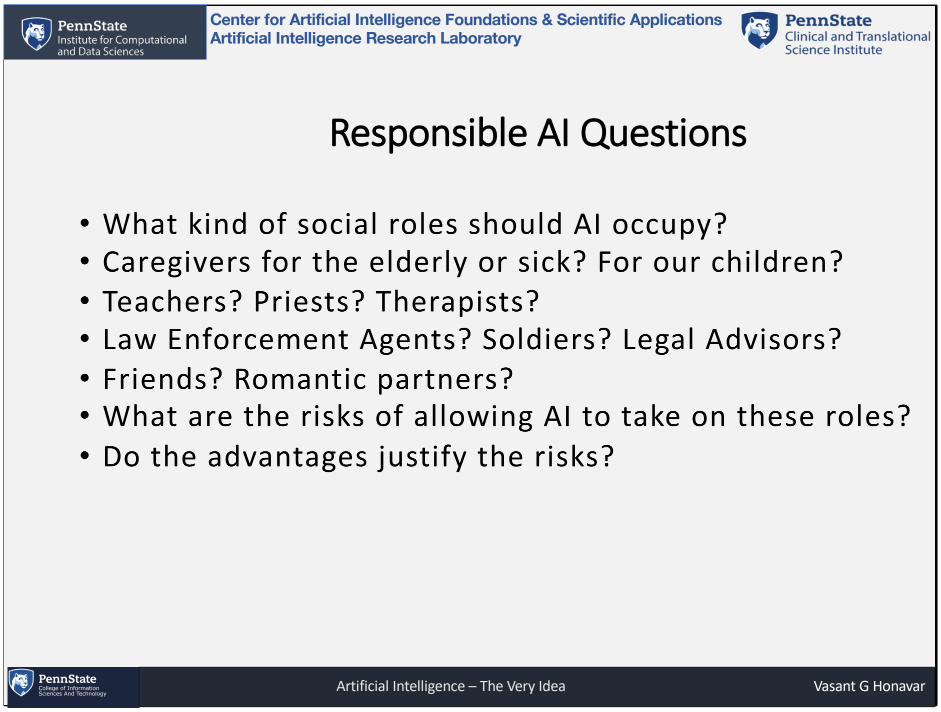
Responsible AI Questions

- Should AI be placed in a ‘position of responsibility?’
- What would that mean?
- Can machines be held responsible for actions they take without human direction?
- Can a machine ever understand and accept risk?



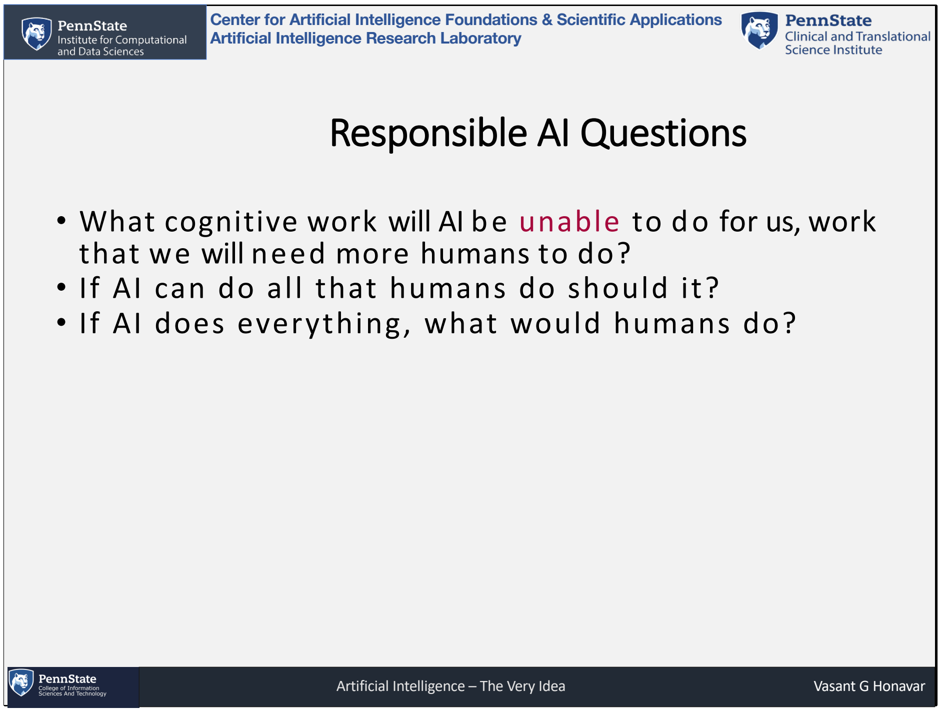
Responsible AI Questions

- Is meaningful human control of AI possible? Who should steer or govern A.I. technology?
- How can we promote the responsible use of AI?



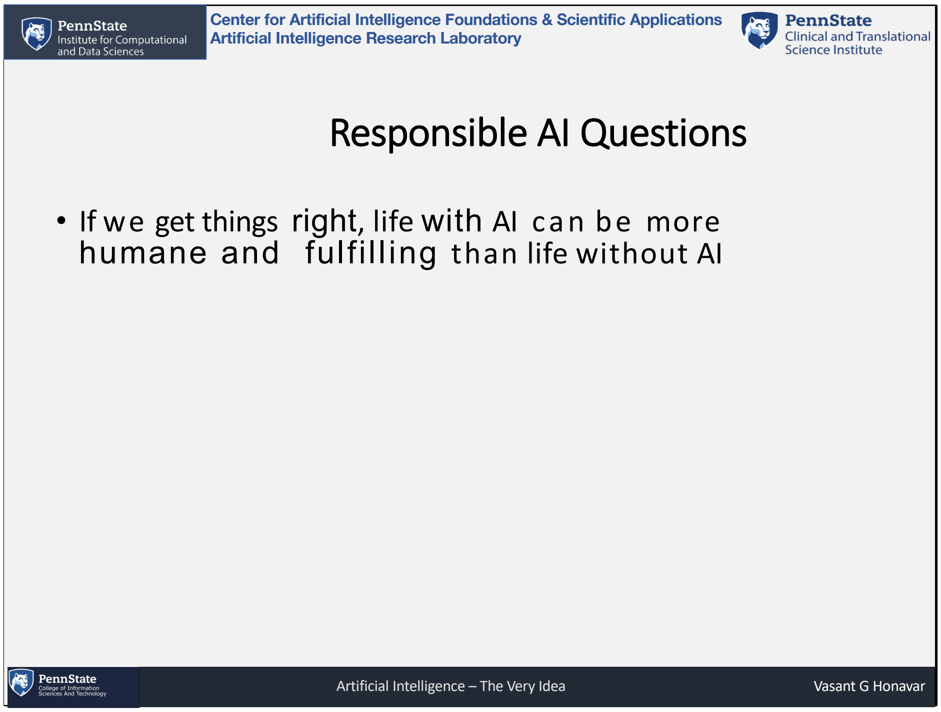
Responsible AI Questions

- What kind of social roles should AI occupy?
- Caregivers for the elderly or sick? For our children?
- Teachers? Priests? Therapists?
- Law Enforcement Agents? Soldiers? Legal Advisors?
- Friends? Romantic partners?
- What are the risks of allowing AI to take on these roles?
- Do the advantages justify the risks?



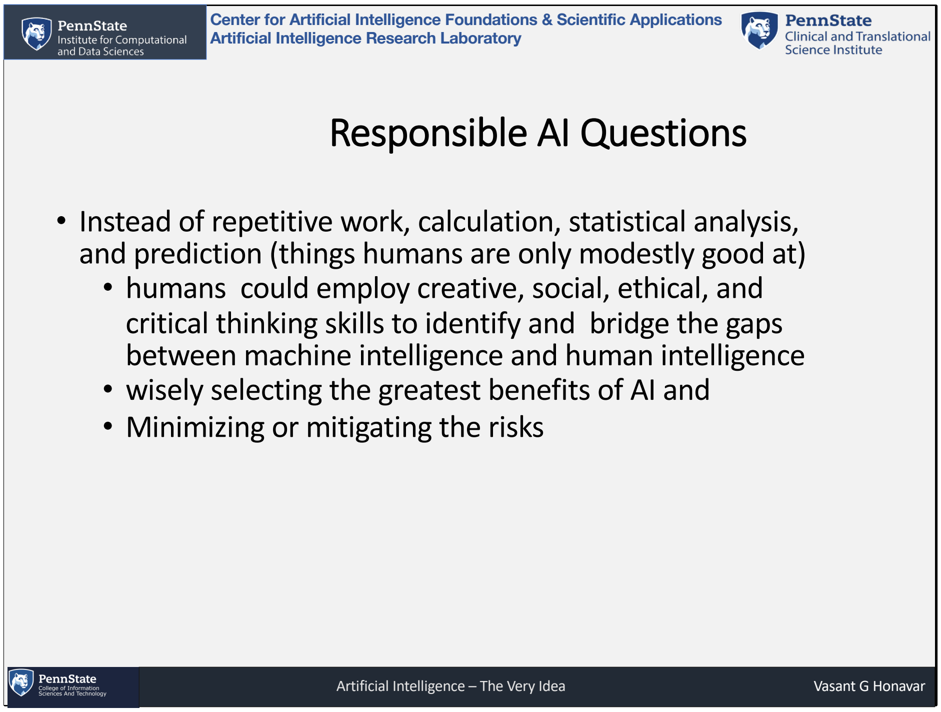
Responsible AI Questions

- What cognitive work will AI be unable to do for us, work that we will need more humans to do?
- If AI can do all that humans do should it?
- If AI does everything, what would humans do?



Responsible AI Questions

- If we get things right, life with AI can be more humane and fulfilling than life without AI



Responsible AI Questions

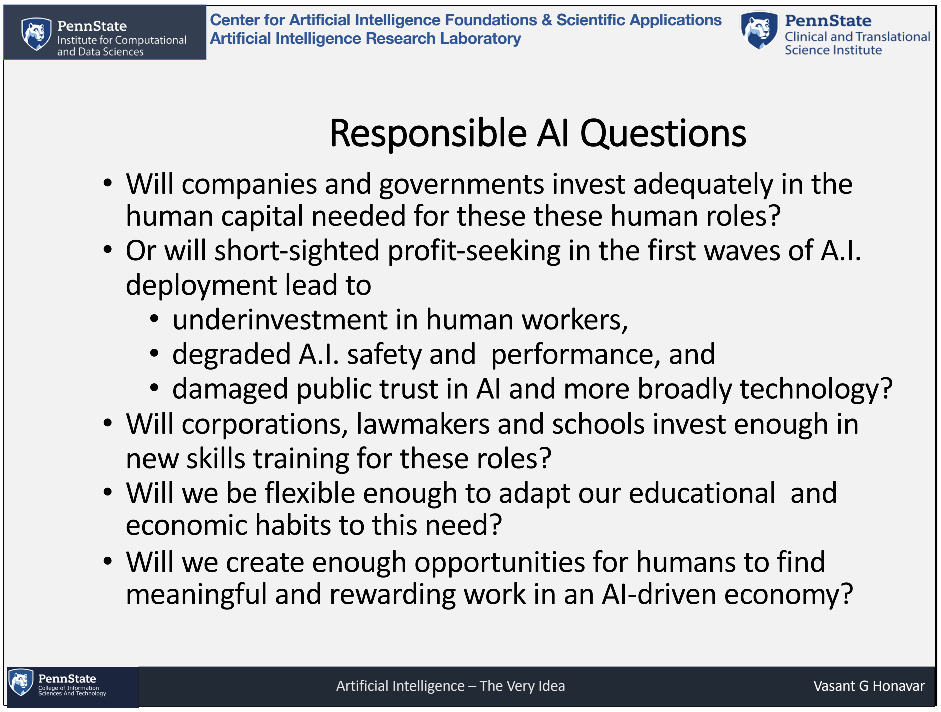
- Instead of repetitive work, calculation, statistical analysis, and prediction (things humans are only modestly good at)
  - humans could employ creative, social, ethical, and critical thinking skills to identify and bridge the gaps between machine intelligence and human intelligence
  - wisely selecting the greatest benefits of AI and
  - Minimizing or mitigating the risks

Responsible AI Questions

But….

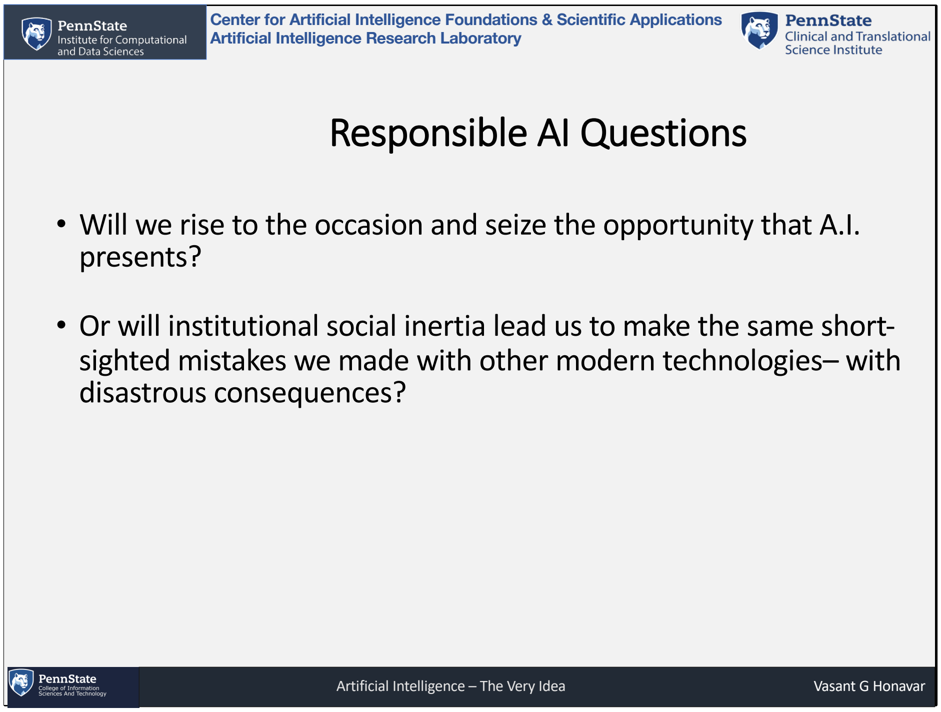
Responsible AI Questions

- This future of human-AI partnership, one that serves and enriches human lives,
  - won’t happen on its own
  - it will need to be a choice we make



Responsible AI Questions

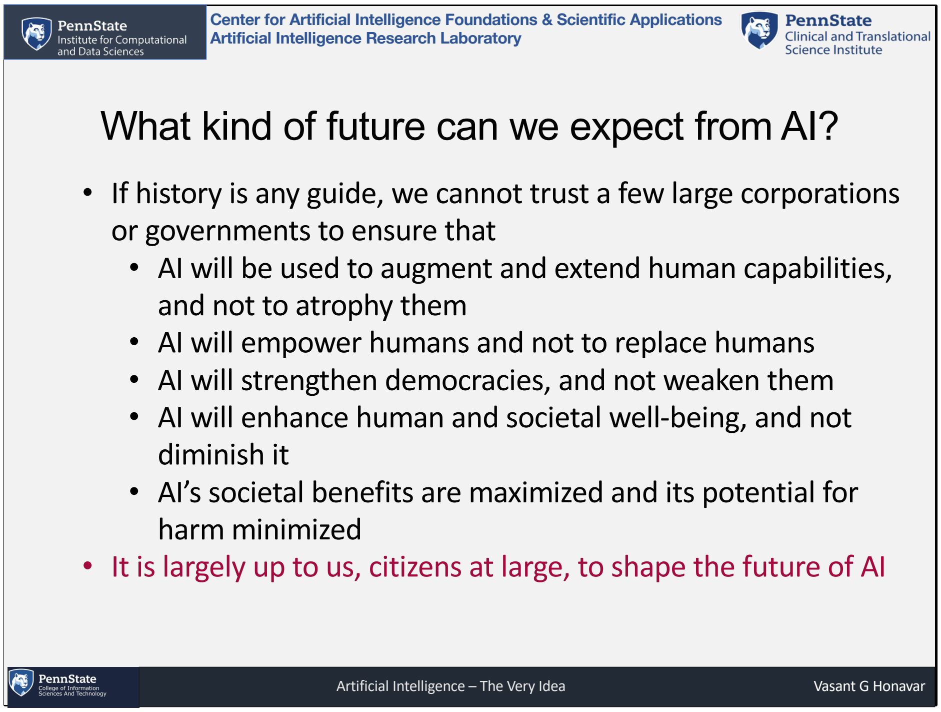
- Will companies and governments invest adequately in the human capital needed for these these human roles?
- Or will short-sighted profit-seeking in the first waves of A.I. deployment lead to
  - underinvestment in human workers,
  - degraded A.I. safety and performance, and
  - damaged public trust in AI and more broadly technology?
- Will corporations, lawmakers and schools invest enough in new skills training for these roles?
- Will we be flexible enough to adapt our educational and economic habits to this need?
- Will we create enough opportunities for humans to find meaningful and rewarding work in an AI-driven economy?



Responsible AI Questions

- Will we rise to the occasion and seize the opportunity that A.I. presents?
- Or will institutional social inertia lead us to make the same short-sighted mistakes we made with other modern technologies– with disastrous consequences?

The hour is late and the choice of good and evil is upon us

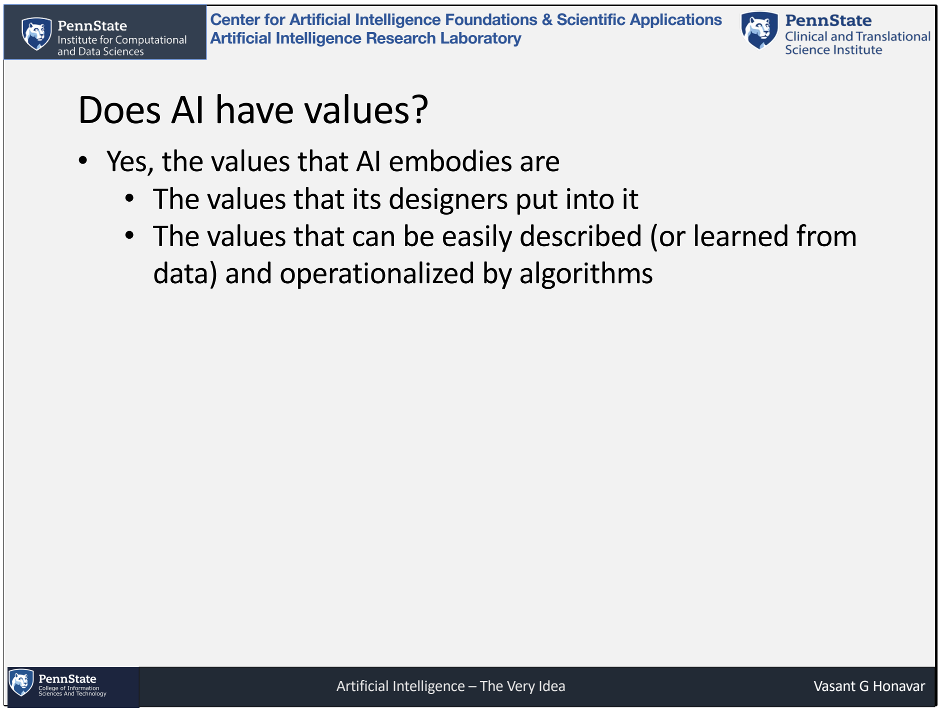


What kind of future can we expect from AI?

- If history is any guide, we cannot trust a few large corporations or governments to ensure that
  - AI will be used to augment and extend human capabilities, and not to atrophy them
  - AI will empower humans and not to replace humans
  - AI will strengthen democracies, and not weaken them
  - AI will enhance human and societal well-being, and not diminish it
  - AI's societal benefits are maximized and its potential for harm minimized
- It is largely up to us, citizens at large, to shape the future of AI

What kind of an AI-powered future do we want?

- One in which human values remain central
  - Do machines have values?
  - How do they differ from human values?
  - Which human values?
  - How do we align AI with human values?



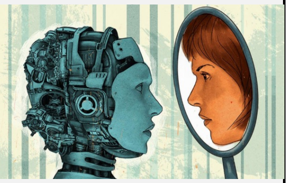
Does AI have values?

- Yes, the values that AI embodies are
    - The values that its designers put into it
    - The values that can be easily described (or learned from data) and operationalized by algorithms

AI systems simply algorithmically operationalize the values of humans who design, train, and deploy them

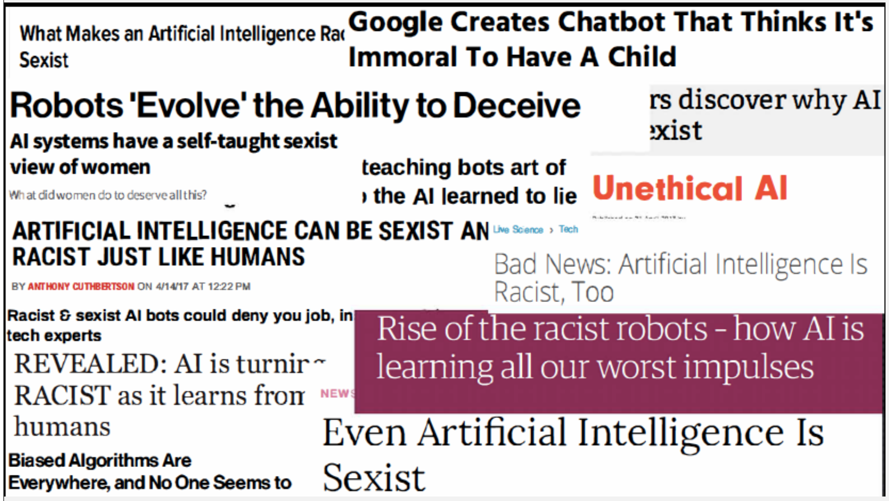
- Optimality
- Efficiency
- Performance
- Reliability
- Robustness
- Safety
- Resilience

- Transparency
- Explainability
- Usability
- Fairness
- Trustworthiness
- Accountability
- Adaptability



- AI systems mirror human values and ethics

Is this the kind of world we want to live in?



What Makes an Artificial Intelligence Rac
Sexist

Google Creates Chatbot That Thinks It's Immoral To Have A Child

Robots 'Evolve' the Ability to Deceive

rs discover why AI
xist

AI systems have a self-taught sexist view of women

What did women do to deserve all this?

teaching bots art of
the AI learned to lie

Unethical AI

ARTIFICIAL INTELLIGENCE CAN BE SEXIST AN
RACIST JUST LIKE HUMANS

BY ANTHONY CUTHBERTSON ON 4/14/17 AT 12:22 PM

Live Science › Tech

Bad News: Artificial Intelligence Is Racist, Too

Racist & sexist AI bots could deny you job, in
tech experts

Rise of the racist robots - how AI is learning all our worst impulses

REVEALED: AI is turnir
RACIST as it learns from
humans

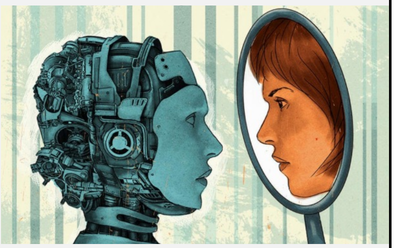
NEWS

Even Artificial Intelligence Is Sexist

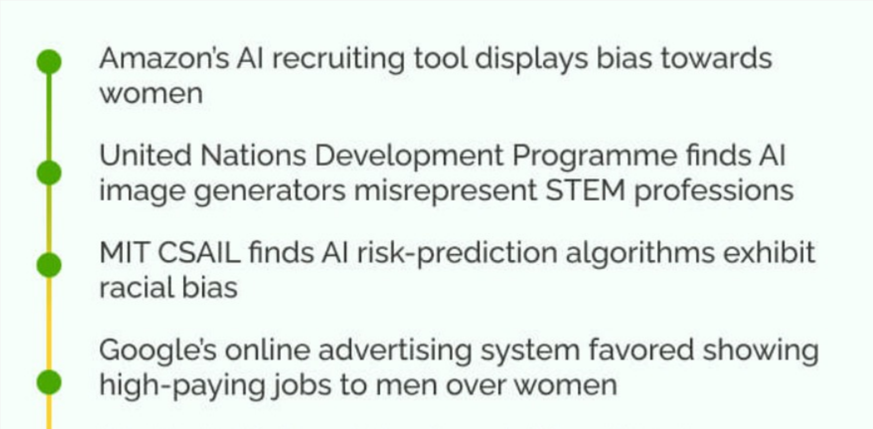
Biased Algorithms Are Everywhere, and No One Seems to

AI holds a mirror to us – individually and collectively

- AI is intrinsically incapable of fear, hate, the desire to dominate and oppress, or any of the destructive human emotions
    - AI can only reflect our own.
    - But AI does so very, very well.
- Looking at AI is like looking at ourselves in a mirror

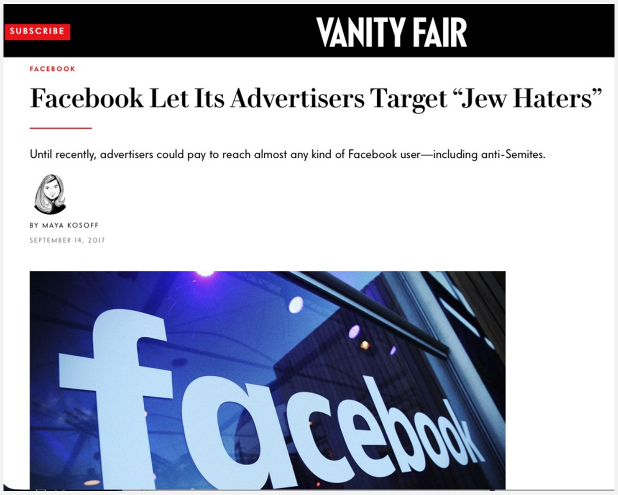


Human prejudices revealed by the AI mirror



- Amazon's AI recruiting tool displays bias towards women
- United Nations Development Programme finds AI image generators misrepresent STEM professions
- MIT CSAIL finds AI risk-prediction algorithms exhibit racial bias
- Google's online advertising system favored showing high-paying jobs to men over women

Human prejudices revealed by the AI mirror



SUBSCRIBE

VANITY FAIR

FACEBOOK

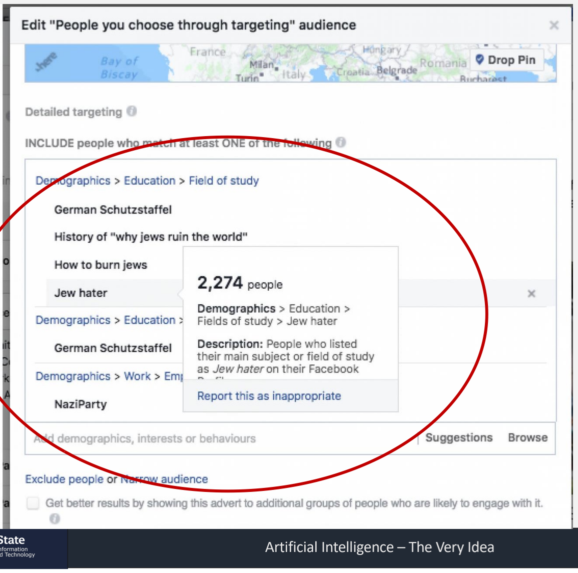
Facebook Let Its Advertisers Target "Jew Haters"

Until recently, advertisers could pay to reach almost any kind of Facebook user—including anti-Semites.

BY MAYA KOSOFF

SEPTEMBER 14, 2017

Human prejudices revealed by the AI mirror



Edit "People you choose through targeting" audience

Detailed targeting

INCLUDE people who match at least ONE of the following

Demographics > Education > Field of study

German Schutzstaffel

History of "why jews ruin the world"

How to burn jews

Jew hater

Demographics > Education >

German Schutzstaffel

Demographics > Work > Emp

NaziParty

2,274 people

Demographics > Education > Fields of study > Jew hater

Description: People who listed their main subject or field of study as *Jew hater* on their Facebook

Report this as inappropriate

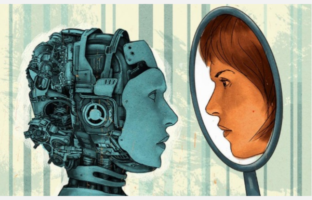
Add demographics, interests or behaviours

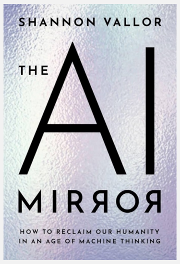
Exclude people or Narrow audience

□ Get better results by showing this advert to additional groups of people who are likely to engage with it.

AI holds a mirror to us – individually and collectively

- Looking at AI is like looking at ourselves in a mirror
- The reflection can magnify, distort, invert the original
- AI must not become a scapegoat for human moral failures
- AI a powerful diagnostic tool for society and can help reveal our blind spots
- What should I do when you see your reflection and you don't like it?





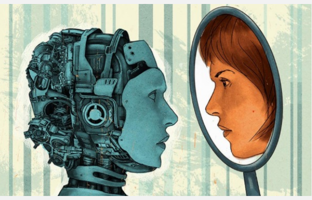
SHANNON VALLOR

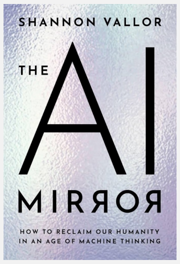
THE AI MIRROR

HOW TO RECLAIM OUR HUMANITY IN AN AGE OF MACHINE THINKING

AI holds a mirror to us – individually and collectively

- What should I do when you see your reflection and you don't like it?
- Break the mirror?
- Start an anti-mirror campaign?
- Start scrubbing the mirror?
- Accept what the mirror shows as inevitable?
- Try washing the dirt off your face?

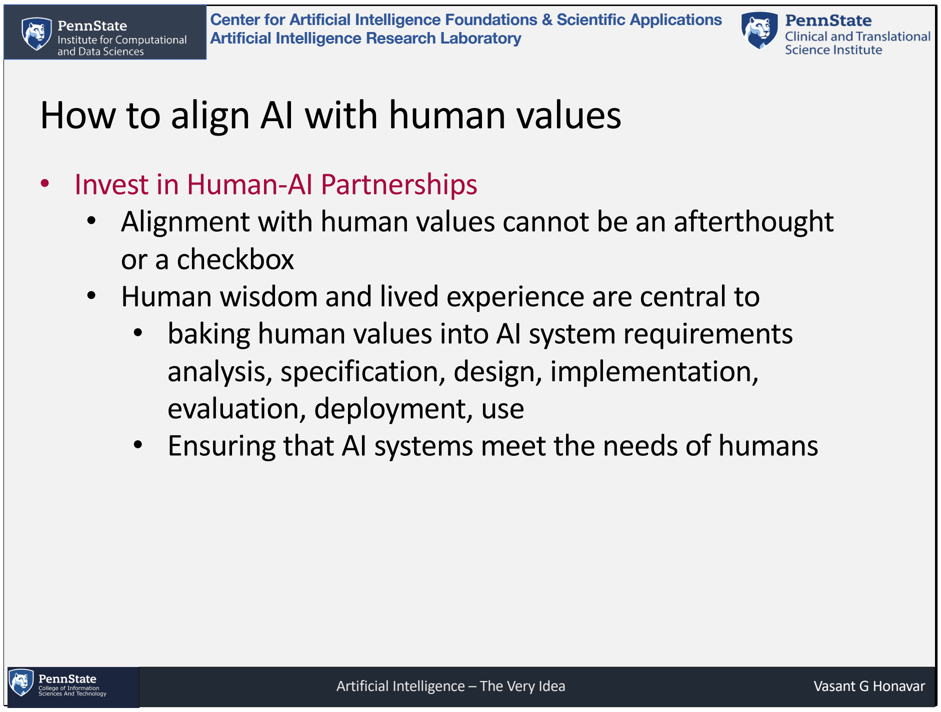




SHANNON VALLOR

THE AI MIRROR

HOW TO RECLAIM OUR HUMANITY IN AN AGE OF MACHINE THINKING

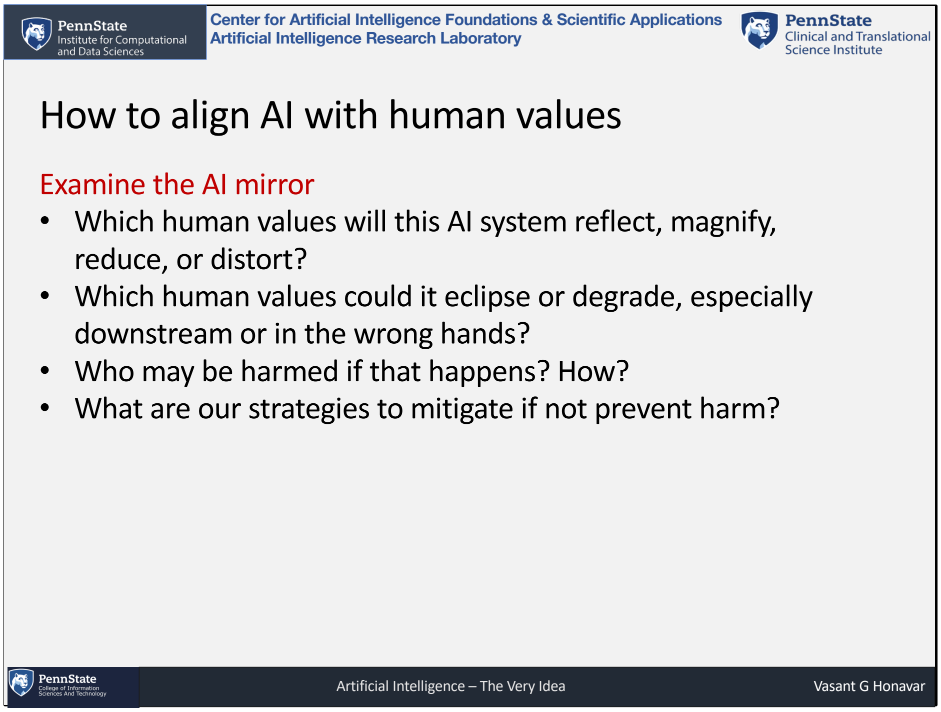


How to align AI with human values

- Invest in Human-AI Partnerships
  - Alignment with human values cannot be an afterthought or a checkbox
  - Human wisdom and lived experience are central to
    - baking human values into AI system requirements analysis, specification, design, implementation, evaluation, deployment, use
    - Ensuring that AI systems meet the needs of humans

How to align AI with human values

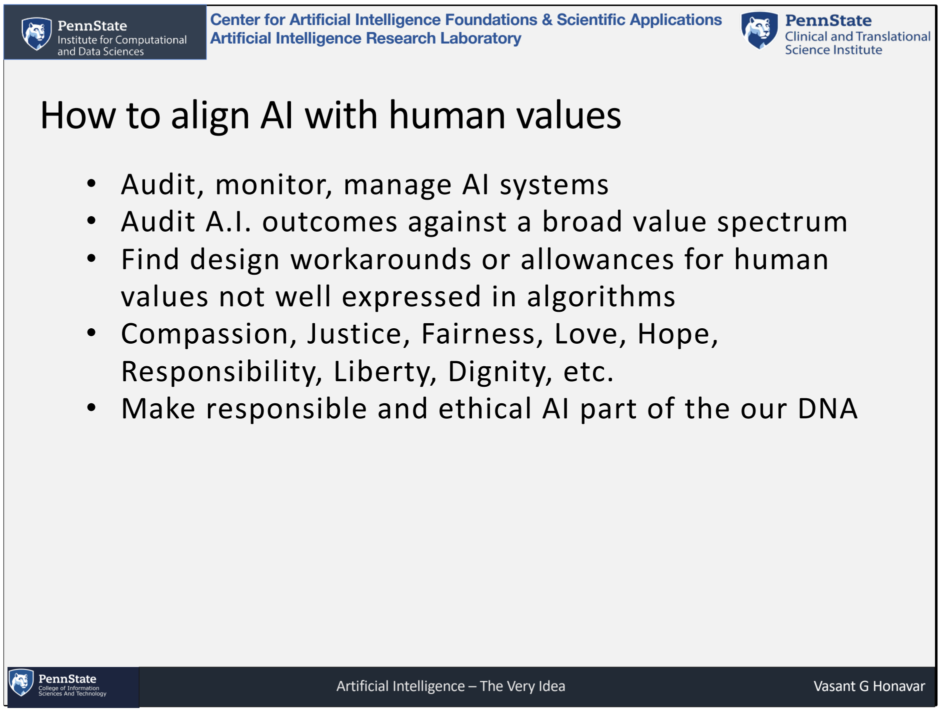
- Proactively anticipate and mitigate risks
  - Even under ideal conditions, most deployments of AI cannot be risk-free smooth rides.
  - We need to anticipate dangerous AI failures and build in guardrails
    - just as we anticipate and build in safety margins for brake and engine failures, for blown tires and overheating reactors.



How to align AI with human values

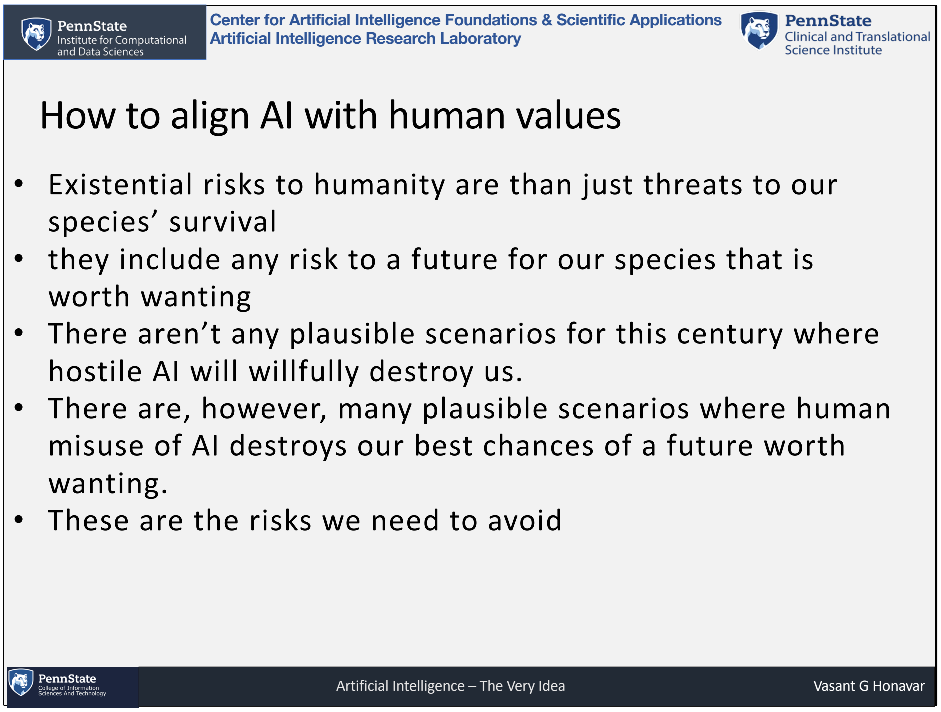
Examine the AI mirror

- Which human values will this AI system reflect, magnify, reduce, or distort?
- Which human values could it eclipse or degrade, especially downstream or in the wrong hands?
- Who may be harmed if that happens? How?
- What are our strategies to mitigate if not prevent harm?



How to align AI with human values

- Audit, monitor, manage AI systems
- Audit A.I. outcomes against a broad value spectrum
- Find design workarounds or allowances for human values not well expressed in algorithms
- Compassion, Justice, Fairness, Love, Hope, Responsibility, Liberty, Dignity, etc.
- Make responsible and ethical AI part of the our DNA

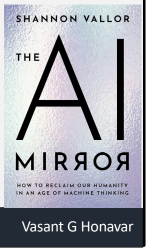


How to align AI with human values

- Existential risks to humanity are than just threats to our species' survival
- they include any risk to a future for our species that is worth wanting
- There aren't any plausible scenarios for this century where hostile AI will willfully destroy us.
- There are, however, many plausible scenarios where human misuse of AI destroys our best chances of a future worth wanting.
- These are the risks we need to avoid

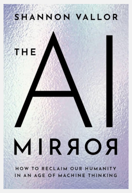
1. The ***real*** risks of A.I. *aren't* those of 'superintelligent' machines taking over.

The real risks will come from machines that aren't yet ***smart enough*** to handle the responsibilities humans will give them.



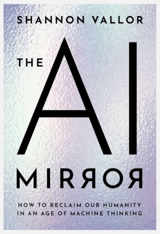
2. Machine intelligence will keep improving but will remain far ***narrower*** than ours for the foreseeable future;

This will boost demand for human skills and values that can ***bridge the gap*** and mitigate the ***risks*** of narrow but powerful A.I. acting on broad, complex human societies



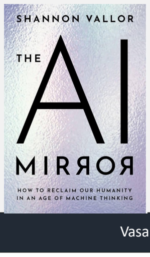
3. Real machine ***minds*** aren't coming any time soon. But machines that ***act*** like minds (up to a point), are already here, and these fictions will only become more convincing.

This poses **immense risks** to humans.



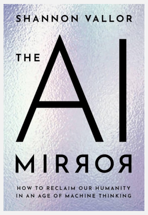
4. Investment in human ***enrichment*** and ***responsible oversight*** of AI are the keys to managing these risks, & sustaining growth and public trust in the technology.

AI is ***not*** ready for solo flight.



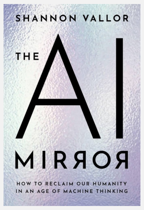
5. Machine values too often eclipse broader human values. To ***harmonize*** them, we must take ownership of both, making the former serve the latter—

never *the reverse.*



6. AI must not become a scapegoat for **human** moral failures, even those that it may reflect, magnify, & perpetuate.

We are still the responsible agents.
This is ***good*** news.



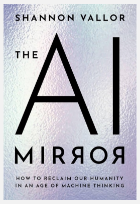
SHANNON VALLOR

THE AI MIRROR

HOW TO RECLAIM OUR HUMANITY IN AN AGE OF MACHINE THINKING

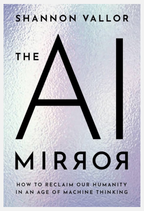
7. The AI mirror is a powerful **diagnostic** tool for society – we can learn as much from its failures and limitations as we can from the new insights and solutions it delivers.

AI can help keep us ***honest*** about who we are and what we choose to value.



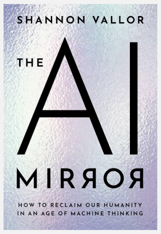
8. AI can be a powerful force for good and a driver of human growth; trying to make ***it*** better can also make ***us*** better.

What good can we do ***with*** AI that we *couldn't* do ***without*** it?



9. Nothing is inevitable unless we make it so.

The future of AI, and of human society, will not be determined for us, but ***by*** us.



It's always going to be ***us*** in the AI mirror.

And it's still ***up to us*** whether we'll like what we see.