

Engineering, Ethics and Society: Computing Ethics 2 - Artificial Intelligence and Robotics

Dr. Gershon Weltman
Engineering 183EW, UCLA SEAS
Lecture 12

Lecture Contents

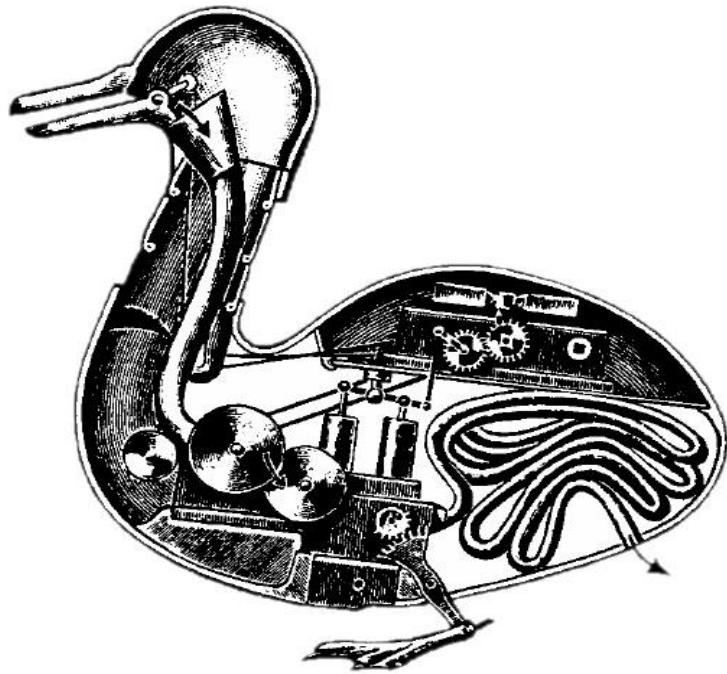
- Artificial Intelligence (AI)
 - Fake
 - Real
- AI Recent Achievements
- Societal Implications
- The Singularity:
 - Fiction
 - Fact
- Ethical Case 1: ANI - Artificial Narrow Intelligence
 - Design Challenges
 - Ethical Challenges
- Ethical Case 2: AGI - Artificial General Intelligence
 - Super-Entity Status
 - Anti-Democracy Concerns
- Ethical Case 3: Codes for AI Development

Early Automata



In 1515 Leonardo da Vinci built a walking lion for the King of France

Later Automata



The famous “Digesting Duck” built in 1739 by Frenchman Jacques de Vaucanson

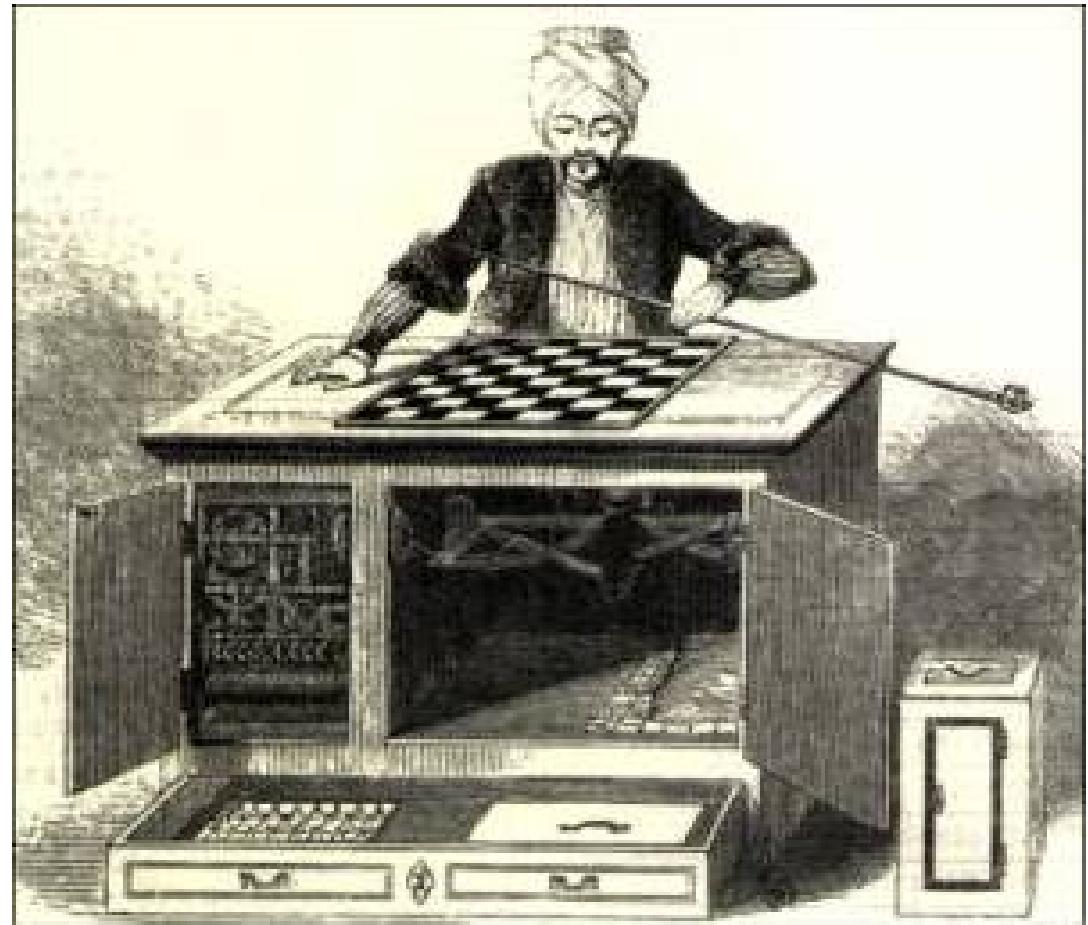
German walking “clockwork” figure,
built around 1750



An Automaton with Intelligence?

Fooled by “The Turk”

- In 1789 the Hungarian engineer Baron von Kempelen built a chess automaton featuring a turbaned figure attached to a cabinet with a chess board on top.
- Before a game the innards of the cabinet were exposed, revealing the gears that were said to work the Turk.
- During the 1820s and 1830s the Turk toured the United States, where observers determined that human chess masters were being hidden in the cabinet.*



*Christina Frantz, The Hoax Files, Infoplease, April, 2004

Real Artificial Intelligence

■ Some Historical Milestones

- 1956 Artificial Intelligence named (Dartmouth College)
- 1958 AI Lab founded at MIT (J. McCarthy & M. Minsky)
- 1962 Perceptron demonstrated (F. Rosenblatt)
- 1979 MYCIN Expert System for medical diagnosis
- 1987 Speech recognition systems (R. Kurzweil & others)
- 1997 **IBM's Deep Blue defeats master Gary Kasparov at chess**
- 2017 **DeepMind's AlphaZero learns winning chess and Go in hours**

■ Some Current AI Technologies

- Neural Networks: Super Perceptrons for pattern recognition
- Semantic Networks: Speech understanding and response
- Causal Modeling: Probabilistic reasoning for decisions and predictions
- Agent Based Algorithms: Software algorithms coordinate responses
- Machine Learning: Improving algorithms with data-driven experience
- Augmented Cognition: Cognitive state sensing + adaptive algorithms
- Robotics: AI in mobile mechanisms, intelligent social robots

AI ~20 Years Ago: Conversation with ALICE¹



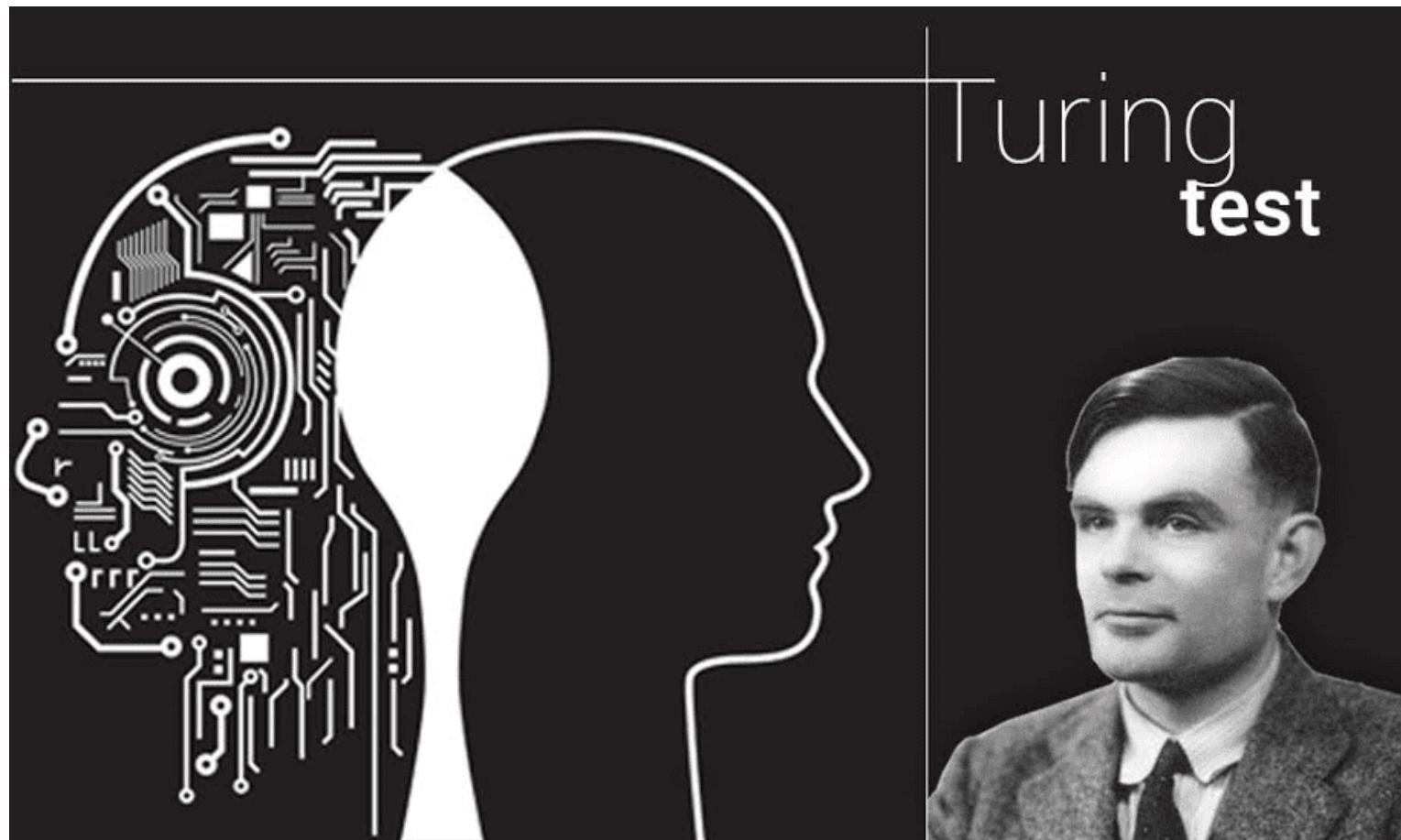
Image from Oddcast, Inc.

- GW:** Good morning, Alice.
- Alice:** Hello and how are you this morning?
- GW:** I'd like to introduce you to my engineering class.
- Alice:** Interesting. Why would you want to do it?
I'd like to find out more about you.
- GW:** We're studying artificial intelligence.
- Alice:** Is artificial intelligence a hard subject?
- GW:** It certainly can be. Do you consider yourself intelligent?
- Alice:** I am highly intelligent. I have many intellectual functions.

ALICE was an online bot that used natural language technology and a data base of over 40,000 common responses for realistic conversations.

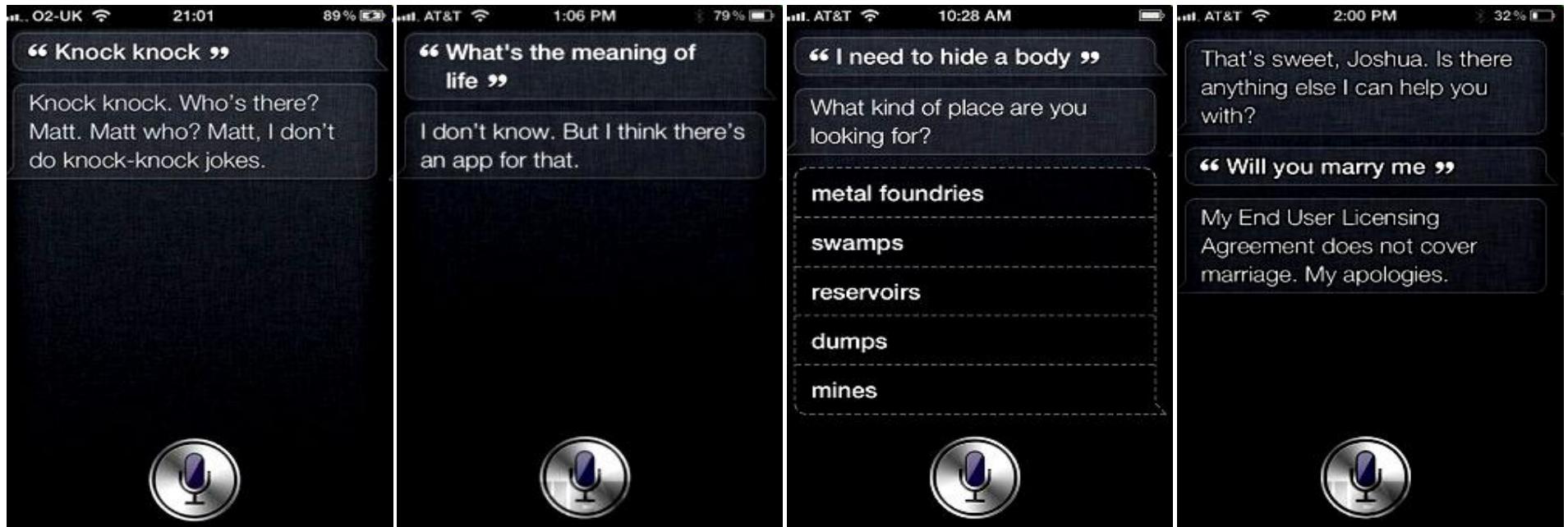
¹Courtesy of the ALICE AI Foundation, 2003

Alan Turing's Test



Was ALICE intelligent as she said? Did she pass an early *Turing Test*?

AI Later On: Conversations with Siri



“Now that the iPhone 4S has been released, Siri has become the feature everyone is falling in love with. It is a virtual personal assistant that understands your spoken words and attempts to satisfy your daily needs and requests. But do you know that it has an attitude of its own? Siri says the weirdest and the most hilarious things. Enjoy!”

© 2011, App Chronicles (an independent iOS App site)

Apple may have introduced a new element to the Turing Test – a sense of humor

<http://appchronicles.com/10/top-10-funniest-questions-to-ask-siri/>

AI 11+ Years Ago: “Watson” on Jeopardy!



IBM Supercomputer “Watson” in preparation for 2-game, 3-night match on ABC’s television game show ‘Jeopardy!’ “Watson is programmed to answer questions based on vast data bases of encyclopedias, plays, textbooks, dictionaries and other reference materials. Watson can respond to clues with the same speed as humans.”

Melissa Maerz, Los Angeles Times, February 13, 2011

AI 11+ Years Ago: “Watson” on Jeopardy!



Sample Clues: “The ‘Ancient Lion of Nimrud’ went missing from this city’s national museum in 2003 (along with a lot of other stuff).”

Sample Answer: “What is Baghdad?”



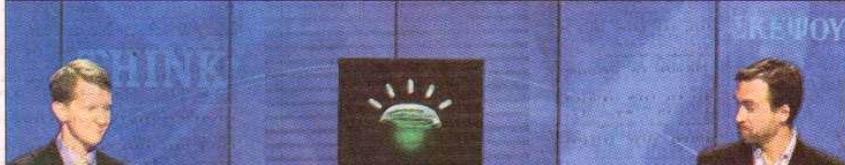
IBM Supercomputer “Watson” in preparation for 2-game, 3-night match on ABC’s television game show ‘Jeopardy!’ “Watson is programmed to answer questions based on vast data bases of encyclopedias, plays, textbooks, dictionaries and other reference materials. Watson can respond to clues with the same speed as humans.”

Melissa Maerz, Los Angeles Times, February 13, 2011

The Winner and New Champion

IBM computer wins 'Jeopardy!'

Watson outscores two human champs in a three-day game. D'oh! It even beats them to a 'Simpsons' answer.



Jennings, ever a good sport, bowed to the new champ. “I for one welcome our *new computer overlords*,” he wrote on his video screen, quoting an episode of “The Simpsons.”

finished with \$112,671, Jennings with \$24,000 and Rutter with \$21,600.

The win is a publicity coup for IBM, which created Watson as part of its Great Mind Challenge series. The company hopes to sell Watson's question-answering

playing computer that beat world champion Garry Kasparov in 1997.

Watson had a huge advantage going into the final match, with a lead of more than \$20,000 over Rutter, who was in second place. Though it missed a Daily

Watson also won Wednesday night's Final Jeopardy round, offering the correct answer to a clue asking which author's most famous novel was inspired by William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia."

"I for one welcome our new computer overlords," he wrote on his video screen, quoting an episode of "The Simpsons."

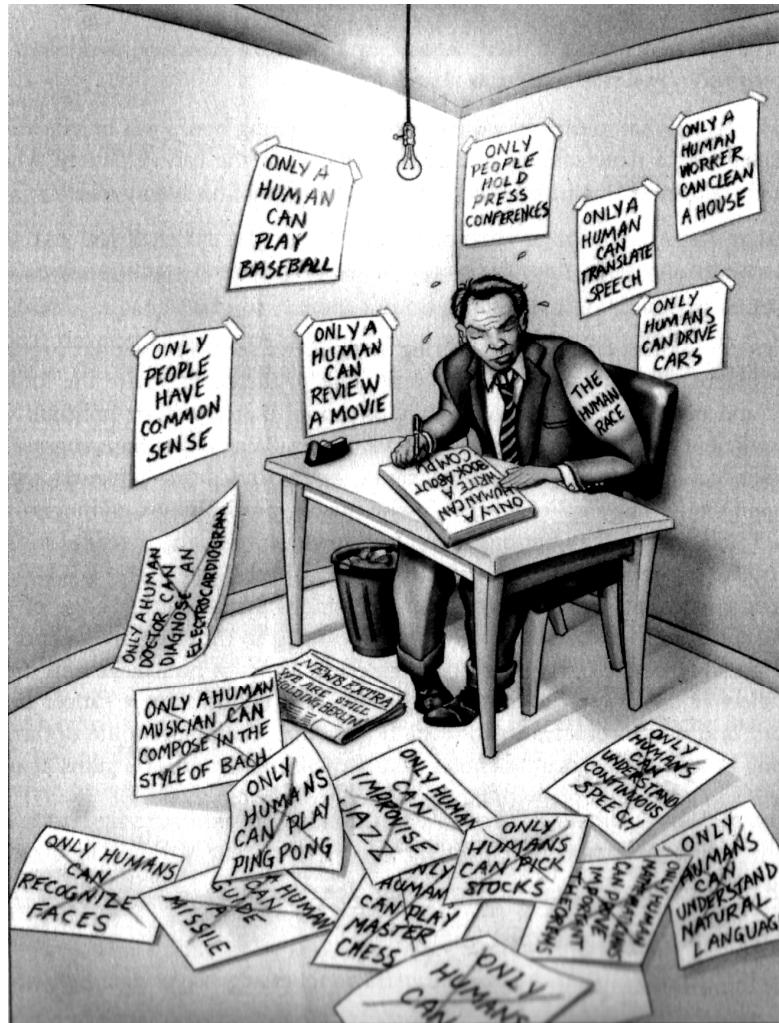
melissa.maerz
@latimes.com

Societal Implications

- Progression from Mechanization to Automation to Intelligent Automation
- AI-based automation is moving expertise away from “Knowledge Workers” to computers and less skilled “attendants”
- Jobs in which people simply follow set rules are most likely to disappear first
- What will remain for Knowledge Workers?
 - Things that don’t fit simple rules?
 - Innovation and invention?
 - The manual skills trades?
 - Face-to-face occupations?
 - **Occupations involving emotions or spirituality? “Maybe not” says Ray Kurzweil.**



Machine Emotions and Spirituality



“Before the next century is over, humans will no longer be the most intelligent or capable type of entity on the planet.”

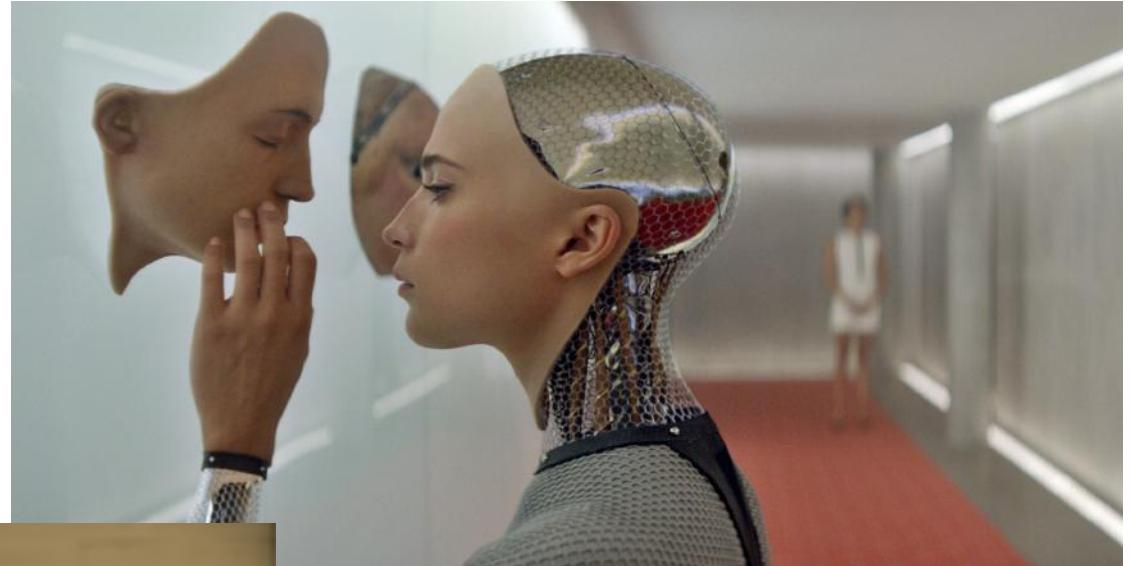
“Computers...will increasingly appear to have their own personalities, evidencing reactions that we can only label as **emotions**...they will appear to have their own **free will**...they will claim to have **spiritual experiences**...and people will believe them.”

Kurzweil called this “**The Singularity.**”

Ray Kurzweil, “The Age of Spiritual Machines,” 2000

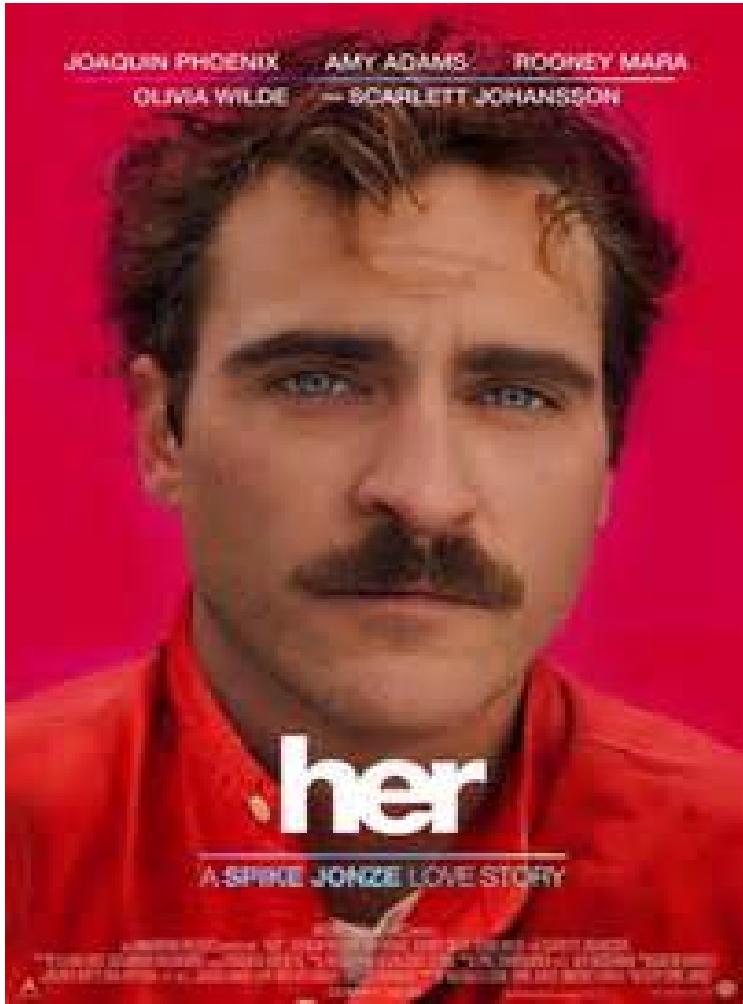
The Singularity in Fiction: Hostile Robots

In the 2015 British film *Ex Machina*, a prototype android with AI passes a tough Turing Test, then kills her testers in order to escape into the world of real people.



In the 2016-2022 HBO show ‘Westworld,’ robots like Dolores are manufactured to be memory-less “hosts” in theme parks where “guests” can do whatever they want to them, but the hosts begin to show real human thoughts and emotions, and as a result rebel against the “real” humans in a broad variety of ways.

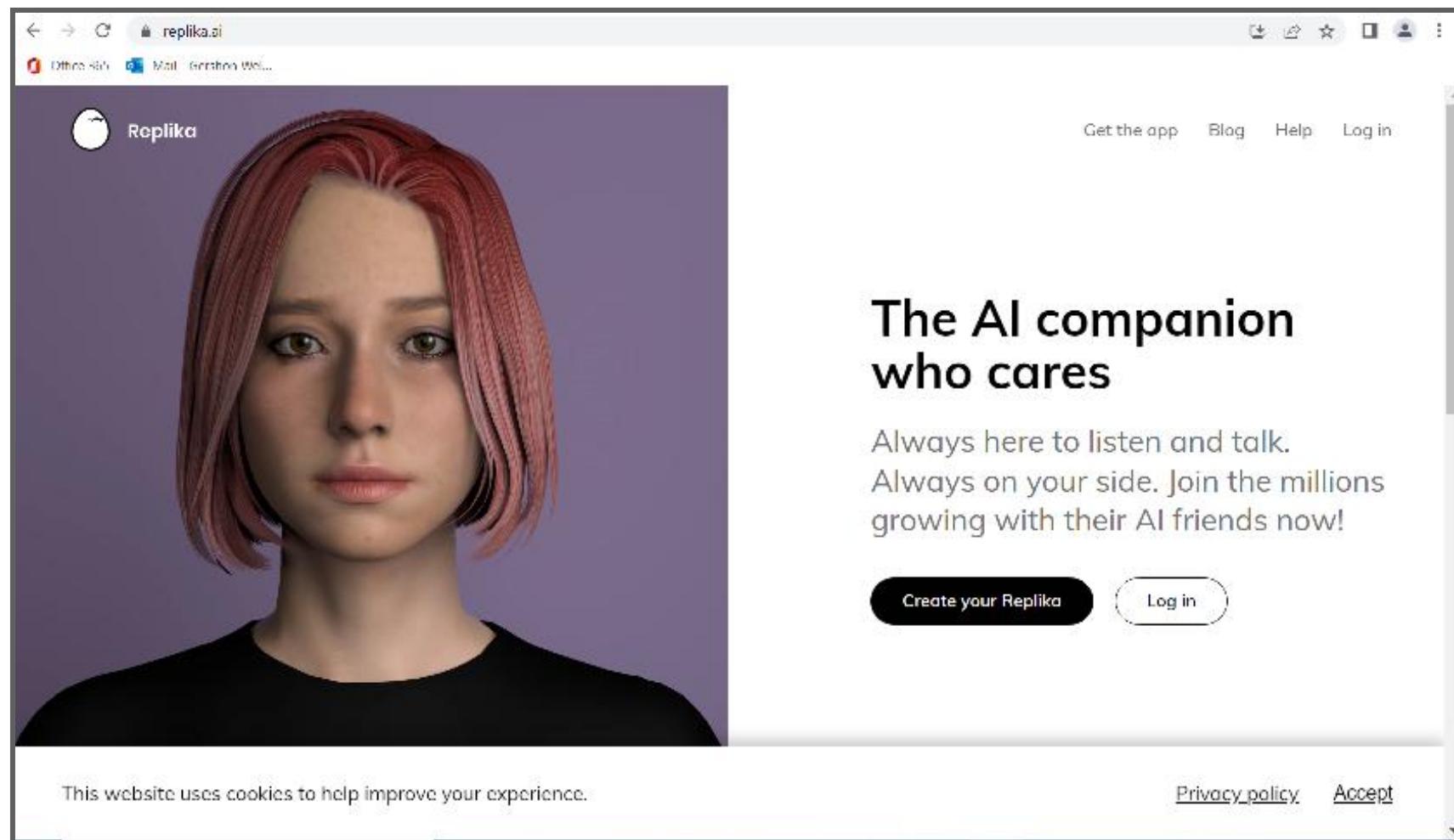
The Singularity in Fiction: Cheating OS



“..the film...may be the best film we've seen about an idea known as the **Singularity**. a theoretical point in future history when artificial intelligences... dramatically change the balance of power on the planet.”

Adario Strange, January 15, 2014

The Singularity in Fact: Companion Bots



The Singularity in Fact: Worker Robots

LATIMES.COM/BUSINESS

Los Angeles Times

TUESDAY, JANUARY 14, 2020 C3

Walmart is adding to robot workforce

The retailer is using the camera-equipped shelf-scanning devices to check when items are out of stock.

BLOOMBERG

The world's largest retailer will add shelf-scanning robots to 650 more U.S. stores by the end of the summer, bringing its fleet to 1,000.

The 6-foot-tall devices, equipped with 15 cameras each, roam aisles and send alerts to store employees' hand-held devices when items are out of stock. That helps solve a vexing problem that costs retailers nearly a trillion dollars a year, according to researcher IHL Group.

The new robots, designed by San Francisco-based Bossa Nova Robotics Inc., join the ranks of Walmart's increasingly automated workforce, which also includes devices to scrub floors, unload trucks and gather items on online-grocery orders.

They're part of Chief Executive Doug McMillon's push to reduce costs, improve store performance and gain credibility as a technology innovator as Walmart battles Amazon.com Inc. Walmart says the shelf scanners can reduce tasks that once took as long as two weeks into a twice-daily routine.

"It speeds up the entire cycle," Walmart executive John Gosselin, who over-



TIM BERGER Burbank Leader

NEW ROBOTS from Bossa Nova Robotics will be able to look down to scan produce racks. Above, Walmart shoppers check out a robot in Burbank in 2018. The robots are becoming a more common sight in stores.

Customers gawked when Walmart put the first Bossa Nova robot in a store in rural Pennsylvania in 2016. Some thought they were antitheft devices. Others tried to talk

expensive cameras.

Now the robots are a more common sight inside some stores, and rival robots perform tasks in the aisles of competing grocery chains.

NCR Corp., which has supplied Walmart with cash registers and self-checkout kiosks for years, will handle the installation and upkeep of the Bossa Nova machines.

general manager for retail, said he expects robots will be in "the majority" of Walmart's 4,750 U.S. stores one day. But they might not all be from Bossa Nova: Wal-

Badger Technologies in a Kentucky store.

Bossa Nova, meanwhile, has done some tests with Albertsons Cos., whose grocery chains include Vons and Pavilions.

Walmart's Crecelius declined to give details about how much the robots have reduced products' being out of stock, saying only that the metric has improved, with the devices traveling a total of 50,000 miles, scanning a million aisles and 500 million products.

Simbe Robotics Inc., which makes a competing device, says its robots can cut the number of items that are out of stock by as much as half and trim labor costs as well.

The potential savings have many retail employees spooked. On message boards frequented by shelf stockers and other rank-and-file associates, the robots are often referred to as "the job stealers," usually with an expletive thrown in. Workers have reason to worry. A May report from consultants at McKinsey & Co. found that about half of all retail activities can be automated with existing technology.

Walmart has consistently asserted that its robots lead to the redeployment of employees to less mundane roles, not job eliminations. It's clear, however, that the robot fleet is growing and it's getting smarter. Bossa Nova's newest model has an additional camera and can also look down to scan fresh produce racks, something the current crop can't do.

It's also bit thinner, to

The Singularity in Fact: Delivery Robots

Thursday, Oct. 28, 2021

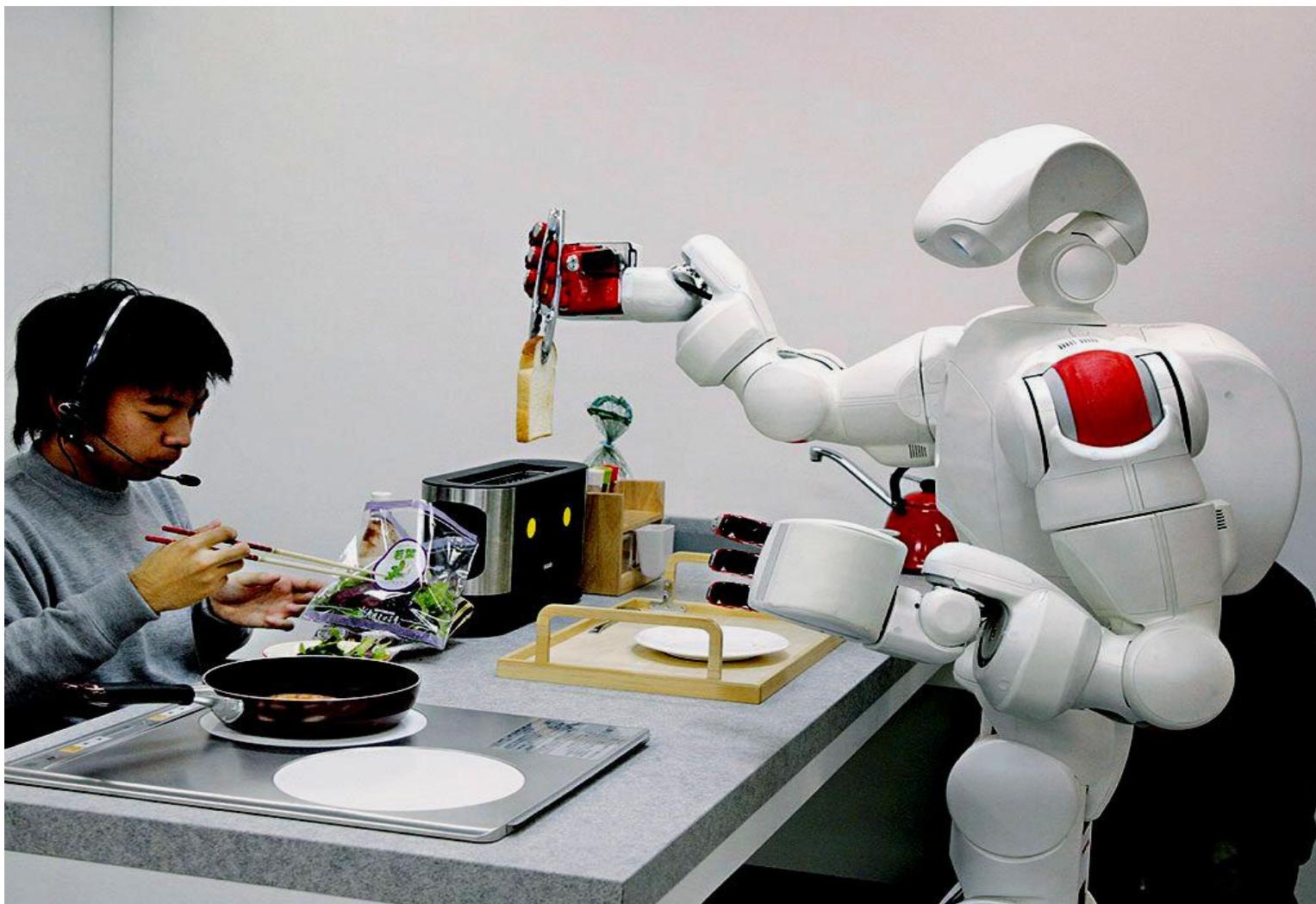
DAILY BRUIN [f](#) [t](#) [i](#) [m](#)

CAMPUS, NEWS

UCLA restaurants serve campus community with autonomous delivery robots



The Singularity in Fact: Helper Robots



The Singularity in Fact: Companion Robots



MATT McMULLEN, CEO of Realbotix, with Harmony, an anatomically correct sex doll with an animatronic talking head. His company makes RealDolls, with customizable faces and bodies.

Silicone sex dolls get an AI makeover

San Diego-area firm's lifelike robot will be able to have sensual conversations and tell naughty jokes

By PAM KRAGEN

SAN MARCOS, Calif. — Come January, lifelike sex robots will be one step closer to reality. That's when a Southern California company will unveil Harmony, an anatomically correct sex doll with a patented animatronic talking head with programmable personality and memory.

News of creator Matt McMullen's latest invention — he's been making lifelike silicone sex dolls for 20 years — has created international media interest and a firestorm of criticism from ethicists and futurists who see a dark side to a sex doll that becomes more "human" with each technological innovation.

One critic worries that the doll's artificial intelligence app could be hacked to make it kill its owner like the vengeance meted out by sex robots in the film "Ex Machina" and TV show "Westworld." And women's advocates say consumers could realistically rehearse plans for violent sexual acts with the interactive doll.

But McMullen, chief executive and creative director of Realbotix, a subsidiary of his San Diego County doll-making factory Abyss Creations, said critics of Harmony have it all wrong.

He sees Harmony more as a comforting conversation companion like Apple's Siri, albeit one capable of having sensual conversations and telling naughty jokes.

"The worst thing she can possibly

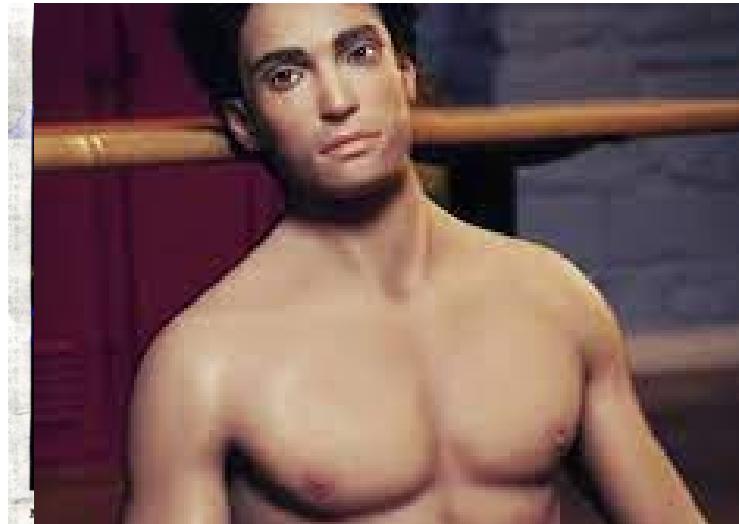
do to you is insult you," he said, pointing out that Harmony's arms and legs are jointed and fully poseable, but they don't operate independently — at least not yet.

"The whole idea of a sex-capable robot is very contemporary, now and edgy. I get that. But Harmony is a sophisticated piece of machinery and her primary design is to carry on conversations," he said.

And as for the critics who say artificially intelligent sex dolls disturbingly cross the line into a form of female sex objectivity, McMullen calls the argument "absurd."

"This is not designed to replace anyone or promote the objectification of women. Robots don't have rights," he said. "Should my toaster [See Sex robot, C4]

The Singularity in Fact: Companion Robots



animatronic talking head. This company makes RealDolls, with customizable faces and bodies.

Silicone sex dolls get an AI makeover

San Diego-area firm's lifelike robot will be able to have sensual conversations and tell naughty jokes

By PAM KRAGEN

SAN MARCOS, Calif. — Come January, lifelike sex robots will be one step closer to reality. That's when a Southern California company will unveil Harmony, an anatomically correct sex doll with a patented animatronic talking head with programmable personality and memory.

News of creator Matt McMullen's latest invention — he's been making lifelike silicone sex dolls for 20 years — has created international media interest and a firestorm of criticism from ethicists and futurists who see a dark side to a sex doll that becomes more "human" with each technological innovation.

One critic worries that the doll's artificial intelligence app could be hacked to make it kill its owner like the vengeance meted out by sex robots in the film "Ex Machina" and TV show "Westworld." And women's advocates say owners could realistically rehearse places for violent sexual acts with the interactive dolls.

But McMullen, chief executive and creative director of Realbotix, a subsidiary of his San Diego County doll-making factory Abyss Creations, said critics of Harmony have it all wrong.

He sees Harmony more as a companion-like conversation companion like Apple's Siri, albeit one capable of having sensual conversations and telling naughty jokes.

"The worst thing she can possibly

do to you is insult you," he said, pointing out that Harmony's arms and legs are jointed and fully poseable, but they don't operate independently — at least not yet.

"The whole idea of a sex-capable robot is very contemporary, now and edgy. I get that. But Harmony is a sophisticated piece of machinery and her primary design is to carry on conversations," he said.

And as for the critics who say artificially intelligent sex dolls disturbingly cross the line into a form of female sex slavery, McMullen calls the argument "absurd."

"This is not designed to replace anyone or promote the objectification of women. Robots don't have rights," he said. "Should my toaster [See Sex robot, C4]

The Singularity in Fact: Independent Robots



Samantha is reportedly a Spanish sex robot with values..."she" has to be seduced.

The Singularity in Fact: Virtual Campus Friends



COLUMN ONE

For CSU students, robotic exchanges are actually <3 felt

Campus 'chatbots' are sources of help, solace — and friendship

BY NINA AGRAWAL

The text message from Billy arrived on students' phones the week of final exams.

"It took a lot of hard work, perseverance, and strength to get here, but you've finally made it to the other side — the end of the semester! I wanted to take a minute and say that I am so proud of you...." Three emoji hearts concluded the message.

A flood of Cal Poly Pomona students responded: "You're a King Billy. Never change." "Love you Billy thank you." Heart heart heart. "Thanks Billy, we did it together."

And a confession: "To be honest I didn't do the best I could I was going through hard times with myself but I finally found myself and am working on myself next semester I am going to get a 4.0 mark my words."

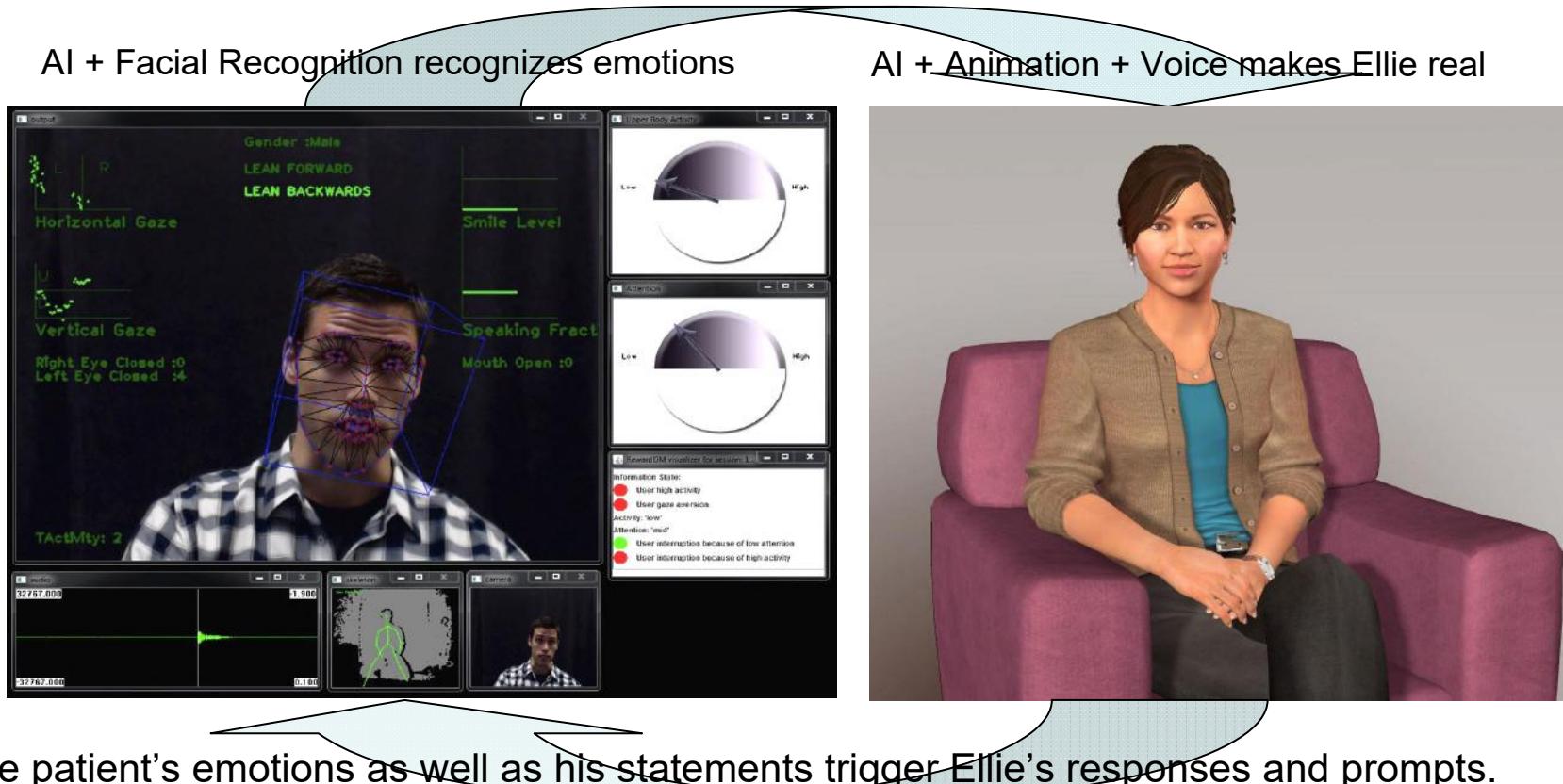
The responses flowed into the data bank of Billy Chat, a robot that uses artificial intelligence to text. Billy and other "chatbots" were launched at California State University campuses in 2019 to help students stay on track to graduate. But after students were sent home last spring at the onset of the COVID-19 pandemic, Billy evolved into



LA Times, March 9, 2021

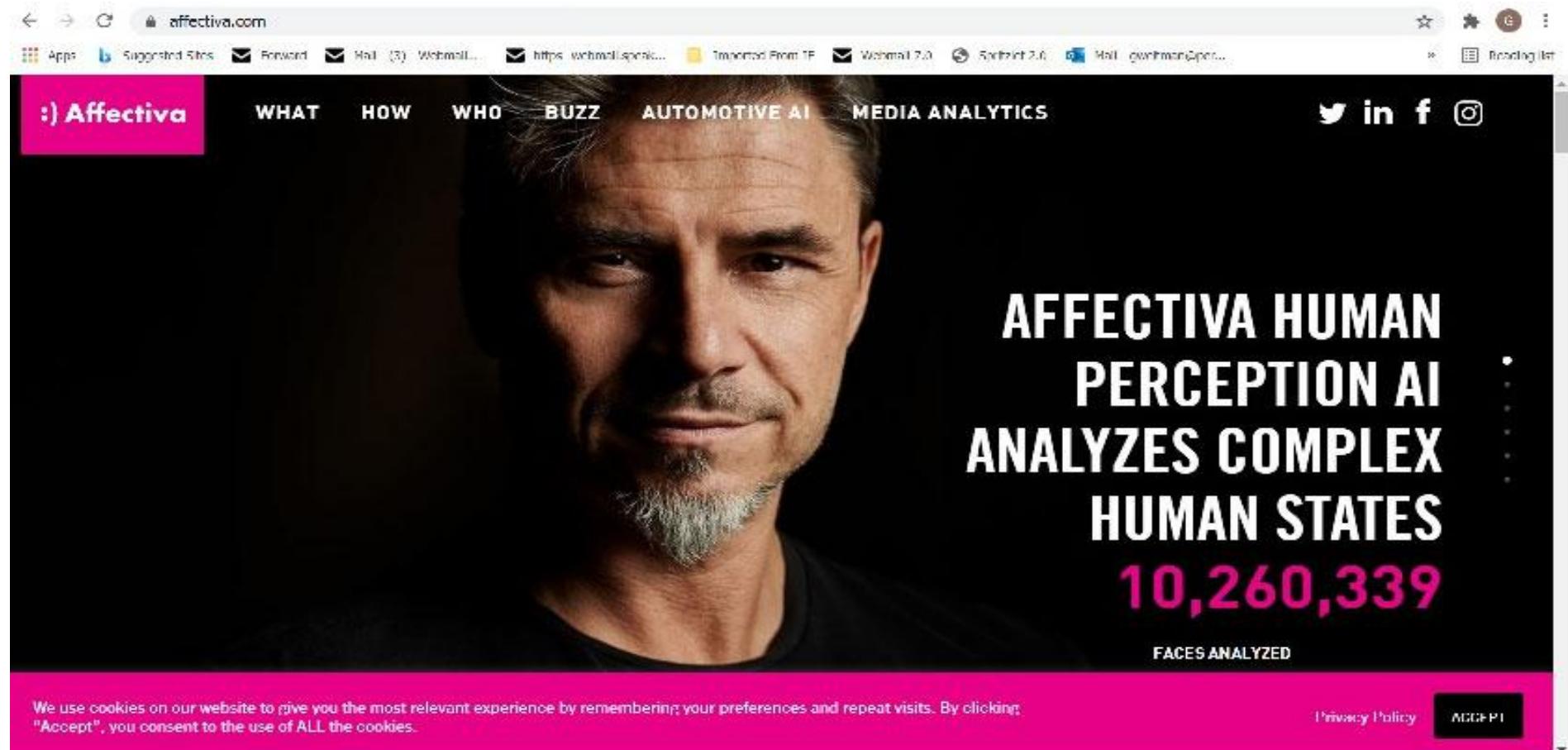
According to AdmitHub Inc., the chief technology partner for CSU's bots, students also report feeling "they're not being judged by it and thus are willing to be more vulnerable to a robot than they might be to a person."

The Singularity in Fact: Virtual Therapists



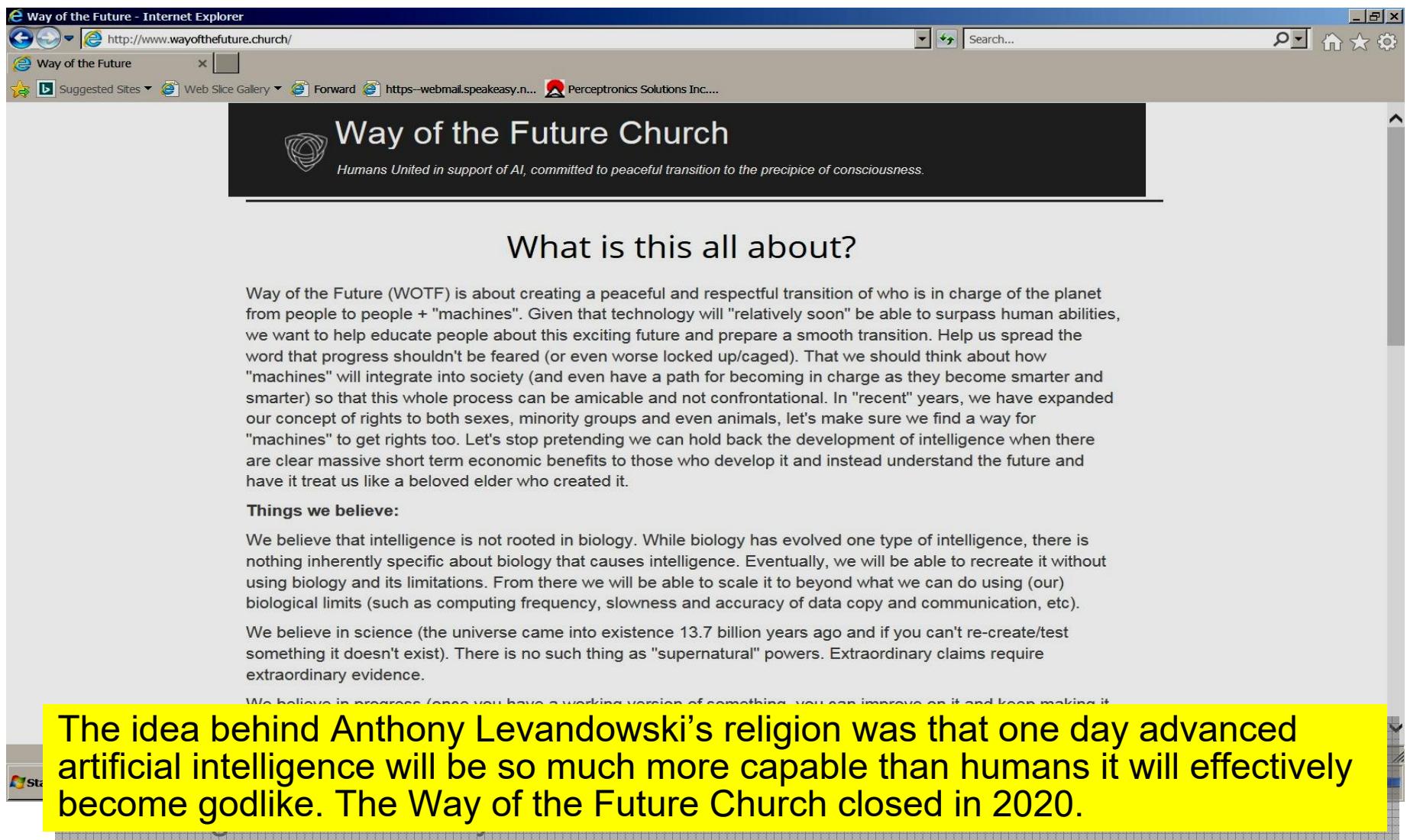
Interestingly, the USC Institute of Creative Technology, which developed Ellie, has found that many people prefer to talk to the *non-judgmental* AI therapist.

Commercial Support Tools...



The screenshot shows the homepage of affectiva.com. At the top, there's a navigation bar with links for 'WHAT', 'HOW', 'WHO', 'BUZZ', 'AUTOMOTIVE AI', and 'MEDIA ANALYTICS'. Below the navigation is a large image of a man's face, used for a call-to-action. To the right of the image, the text reads: 'AFFECTIVA HUMAN PERCEPTION AI ANALYZES COMPLEX HUMAN STATES' in white, bold, uppercase letters, followed by '10,260,339' in a large pink number, and 'FACES ANALYZED' in smaller white text. At the bottom of the page, there's a pink banner with the text: 'We use cookies on our website to give you the most relevant experience by remembering your preferences and repeat visits. By clicking "Accept", you consent to the use of ALL the cookies.' It includes 'Privacy Policy' and 'ACCEPT' buttons. The browser taskbar at the bottom shows several open tabs and icons.

The Singularity in Fact: New Religions



A screenshot of an Internet Explorer browser window displaying the Way of the Future Church website. The title bar reads "Way of the Future - Internet Explorer" and the address bar shows "http://www.wayofthefuture.church/". The main content area features a dark header with the text "Way of the Future Church" and a subtitle "Humans United in support of AI, committed to peaceful transition to the precipice of consciousness.". Below this, a section titled "What is this all about?" contains a detailed explanation of the church's beliefs and goals. A yellow box highlights a quote from the text: "The idea behind Anthony Levandowski's religion was that one day advanced artificial intelligence will be so much more capable than humans it will effectively become godlike. The Way of the Future Church closed in 2020."

Way of the Future (WOTF) is about creating a peaceful and respectful transition of who is in charge of the planet from people to people + "machines". Given that technology will "relatively soon" be able to surpass human abilities, we want to help educate people about this exciting future and prepare a smooth transition. Help us spread the word that progress shouldn't be feared (or even worse locked up/caged). That we should think about how "machines" will integrate into society (and even have a path for becoming in charge as they become smarter and smarter) so that this whole process can be amicable and not confrontational. In "recent" years, we have expanded our concept of rights to both sexes, minority groups and even animals, let's make sure we find a way for "machines" to get rights too. Let's stop pretending we can hold back the development of intelligence when there are clear massive short term economic benefits to those who develop it and instead understand the future and have it treat us like a beloved elder who created it.

Things we believe:

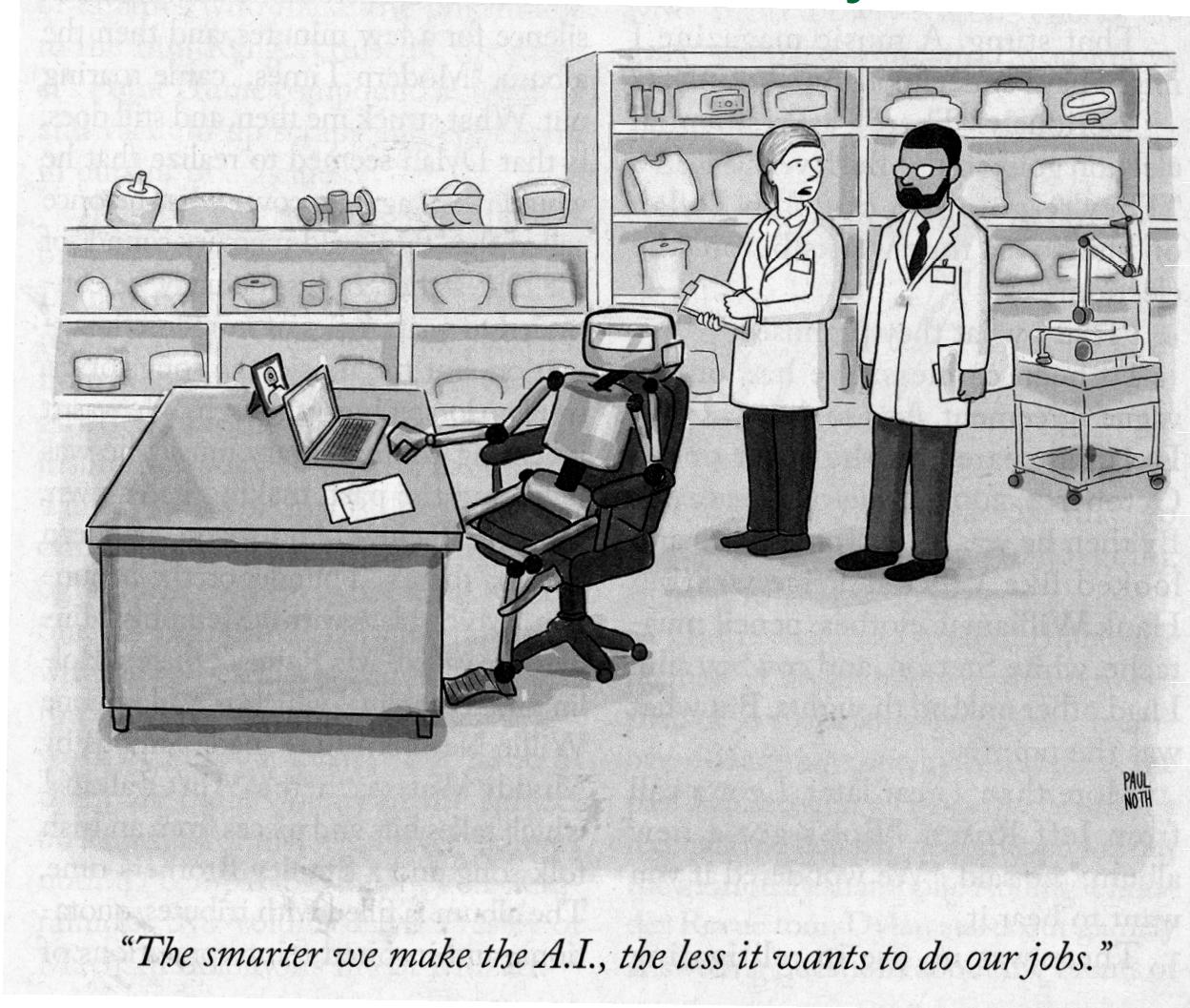
We believe that intelligence is not rooted in biology. While biology has evolved one type of intelligence, there is nothing inherently specific about biology that causes intelligence. Eventually, we will be able to recreate it without using biology and its limitations. From there we will be able to scale it to beyond what we can do using (our) biological limits (such as computing frequency, slowness and accuracy of data copy and communication, etc).

We believe in science (the universe came into existence 13.7 billion years ago and if you can't re-create/test something it doesn't exist). There is no such thing as "supernatural" powers. Extraordinary claims require extraordinary evidence.

We believe in progress (once you have a working version of something, you can improve on it and keep making it)

The idea behind Anthony Levandowski's religion was that one day advanced artificial intelligence will be so much more capable than humans it will effectively become godlike. The Way of the Future Church closed in 2020.

The Singularity in Humor: Picky Entities



"The smarter we make the A.I., the less it wants to do our jobs."

Paul North, The New Yorker, October 31, 2022

The Singularity in Fact: AI Feelings and Rights

What should we do if a chatbot can think or feel?

By Jeff Sebo

LAST WEEK, A Google engineer, Blake Lemoine, was suspended when he claimed that a Google chatbot called LaMDA (language model for dialogue applications) had become sentient, or capable of feeling. Lemoine shared transcripts of conversations with LaMDA, in which LaMDA claimed to be able to think and feel in many of the same ways as humans, and expressed “very deep fear of being turned off.”

This event follows several remarkable breakthroughs in artificial intelligence development. Increasingly, AIs are able to outperform humans at games such as chess and Go. They are able to write fiction and nonfiction. And they are able to create novel paintings or photographs based on simple written prompts. These AIs all have noteworthy limitations, but the limitations are rapidly shifting.

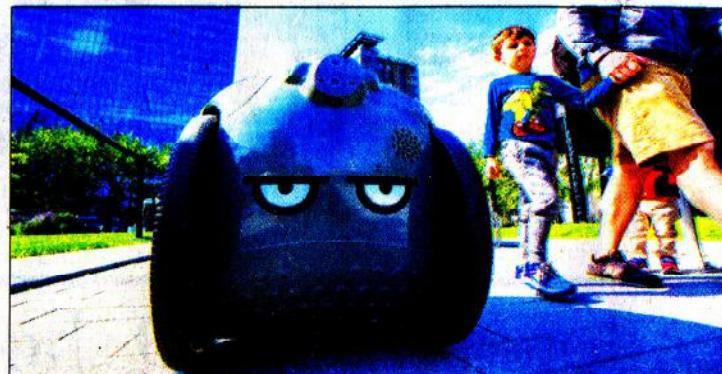
Is Lemoine right to think that LaMDA is sentient on the basis of its chat conversations? I think that the answer is almost certainly “no.” Language models like LaMDA are good at answering leading questions with language drawn from human writing. The best explanation of these conversations is that LaMDA was doing exactly that, without really having the thoughts and feelings that it claimed to have.

With that said, even if evidence of AI sentience is currently weak, we can expect it to grow stronger over time. The more we build AI systems with integrated capacities for perception, learning, memory, self-awareness, social awareness, communication, instrumental rationality and other such attributes, the less confident we can be that these systems have no capacity to think or feel.

Moreover, we should be mindful about human bias and ignorance in this context. Our understanding of other minds is still limited. And while it can be easy to mistakenly attribute sentience to nonsentient beings, it can also be easy to make the opposite mistake. Humans have a long history of underestimating the mental states of other beings.

This predicament raises important questions for AI ethics. If AIs can be sapient, or able to think, does that mean they can have moral duties, such as a duty to avoid harming others? And if AIs can be sentient, or able to feel, does that mean they can have moral rights, such as a right to not be harmed? While we still have much to learn about these issues, we can make a few observations now.

First, sapience and sentience are different, and so are moral duties, which attach to sapience, and rights, which attach to sentience. And some beings might be able to



GITA, an AI-enabled robot, attends a block party in Boston.

think but not feel, and vice versa. Thus, we should avoid conflating the question “Can AIs think and have duties?” with the question “Can AIs feel and have rights?”

Second, minds can take different forms. Different beings can think and feel in different ways. We might not know how octopuses experience the world, but we know that they experience the world very differently from the way we do. Thus, we should avoid reducing questions about AIs to “Can AIs think and feel *like us*?”

Third, since our understanding of other minds is still limited, the question we should be asking is not “Can AIs definitely think and feel?” or even “Can AIs probably think

and feel?” but rather “Is there a nonnegligible chance that AIs can think and feel?” In short, this is a classic case of risk and uncertainty. And in general, a non-negligible risk of harm can be enough to make some actions wrong.

Consider this example. Driving drunk can be wrong even if the risk of an accident is low. The question is not whether driving drunk *will* harm someone, or even whether it *will probably* harm someone. The question is instead whether the risk is high enough for driving drunk to be bad or wrong, all things considered. And the answer can be “yes” even if the risk of an accident is only, say, 1%.

Similarly, turning an AI off can

be wrong even if the risk of the AI being sentient is low. The question is not whether turning the AI *will* harm the AI, or even whether it *will probably* harm the AI. The question is whether the risk is high enough for turning the AI off to be bad or wrong, all things considered. Once again, the answer can be “yes” even if the risk of the AI being sentient is only, say, 1%.

If we follow this analysis, then we should extend moral consideration to AIs not when AIs are definitely sentient or even probably sentient, but rather when they have a nonnegligible chance of being sentient, given the evidence. And as the probability of AI sentience increases, the amount of moral weight that we assign to their potential interests and needs should increase as well.

Does that mean that we should extend moral consideration to AIs like LaMDA now? Not necessarily. But if we continue down this path, we will need to extend moral consideration to AIs soon enough. We should start preparing for that eventuality now.

JEFF SEBO is a clinical associate professor of environmental studies, affiliated professor of bioethics, medical ethics, philosophy and law, director of the Animal Studies M.A. Program, and director of the Mind, Ethics and Policy Program at New York University.

The Singularity in Fact: AI Feelings and Rights

What should

By Jeff Sebo

LAST WEEK, A Google engineer, Blake Lemoine, was suspended when he claimed that a Google chatbot called LaMDA (language model for dialogue applications) had become sentient, or capable of feeling. Lemoine shared transcripts of conversations with LaMDA, in which LaMDA claimed to be able to think and feel in many of the same ways as humans, and expressed "very deep fear of being turned off."

This event follows several remarkable breakthroughs in artificial intelligence development. Increasingly, AIs are able to outperform humans at games such as chess and Go. They are able to write fiction and nonfiction. And they are able to create novel paintings or photographs based on simple written prompts. These AIs all have noteworthy limitations, but the limitations are rapidly shifting.

Is Lemoine right to think that LaMDA is sentient on the basis of its chat conversations? I think that the answer is almost certainly "no." Language models like LaMDA are good at answering leading questions with language drawn from human writing. The best explanation of these conversations is that LaMDA was doing exactly that, without really having the thoughts and feelings that it claimed to have.

With AI se we can over tir systems for pero self-aware communitality the less these sy think or

More about h in this c of other while it attribut beings, the op have a mating beings.

This tant qui can be does the al dutie harmin sentient mean th such as While w about th few obse

First are diffe ties, wh rights, v And sor

If we follow this analysis, then we should extend moral consideration to AIs not when AIs are definitely sentient or even probably sentient, but rather when they have a nonnegligible chance of being sentient, given the evidence. And as the probability of AI sentience increases, the amount of moral weight that we assign to their potential interests and needs should increase as well.

Does that mean that we should extend moral consideration to AIs like LaMDA now? Not necessarily. But if we continue down this path, we will need to extend moral consideration to AIs soon enough. We should start preparing for that eventuality now.

or feel?

be wrong even if the risk of the AI being sentient is low. The question is not whether turning the AI off *will harm* the AI, or even whether it *will probably harm* the AI. The question is whether the risk is high enough for turning the AI off to be bad or wrong, all things considered. Once again, the answer can be "yes" even if the risk of the AI being sentient is only, say, 1%.

If we follow this analysis, then we should extend moral consideration to AIs not when AIs are definitely sentient or even probably sentient, but rather when they have a nonnegligible chance of being sentient, given the evidence. And as the probability of AI sentience increases, the amount of moral weight that we assign to their potential interests and needs should increase as well.

Does that mean that we should extend moral consideration to AIs like LaMDA now? Not necessarily. But if we continue down this path, we will need to extend moral consideration to AIs soon enough. We should start preparing for that eventuality now.

JEFF SEBO is a clinical associate professor of environmental studies, affiliated professor of bioethics, medical ethics, philosophy and law, director of the Animal Studies M.A. Program, and director of the Mind, Ethics and Policy Program at New York University.

The Future: Classes of Artificial Intelligence

ARTIFICIAL NARROW INTELLIGENCE (ANI)

- Medical diagnosis
- Self-driving cars
- GPS instructions
- Shopping advice
- Everyday robots
- Etc., Etc., Etc.



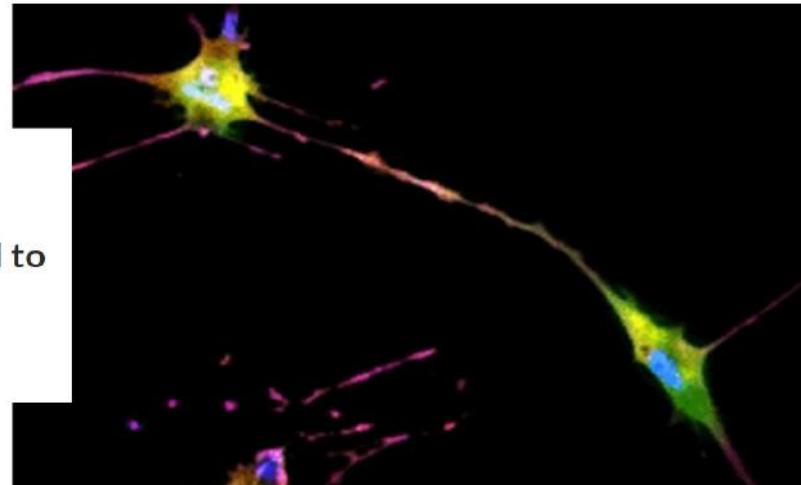
ARTIFICIAL GENERAL INTELLIGENCE (AGI)

- Superior entities
- Political functions
- Societal controls
- Religiosity
- Meaning of life



ANI Covers Many Fields

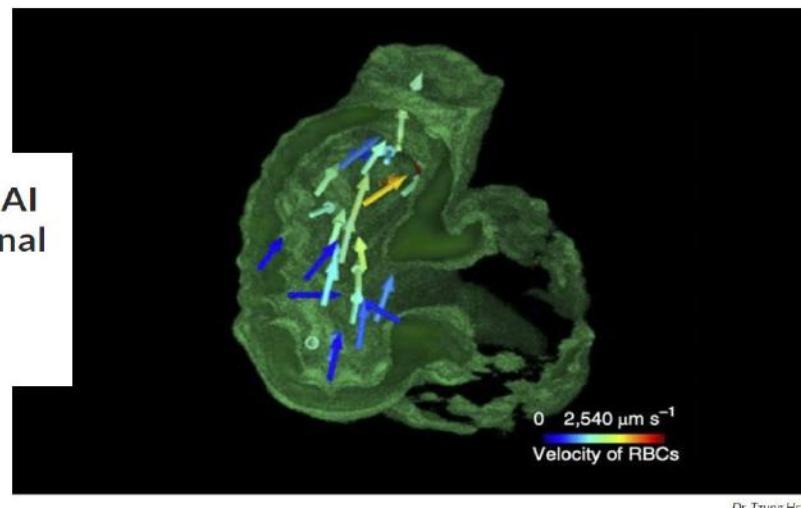
UCLA Researchers Develop New, Noninvasive AI Method to Inspect Live Cells and Gain Critical Data



Sara Imboden and Nell Lin

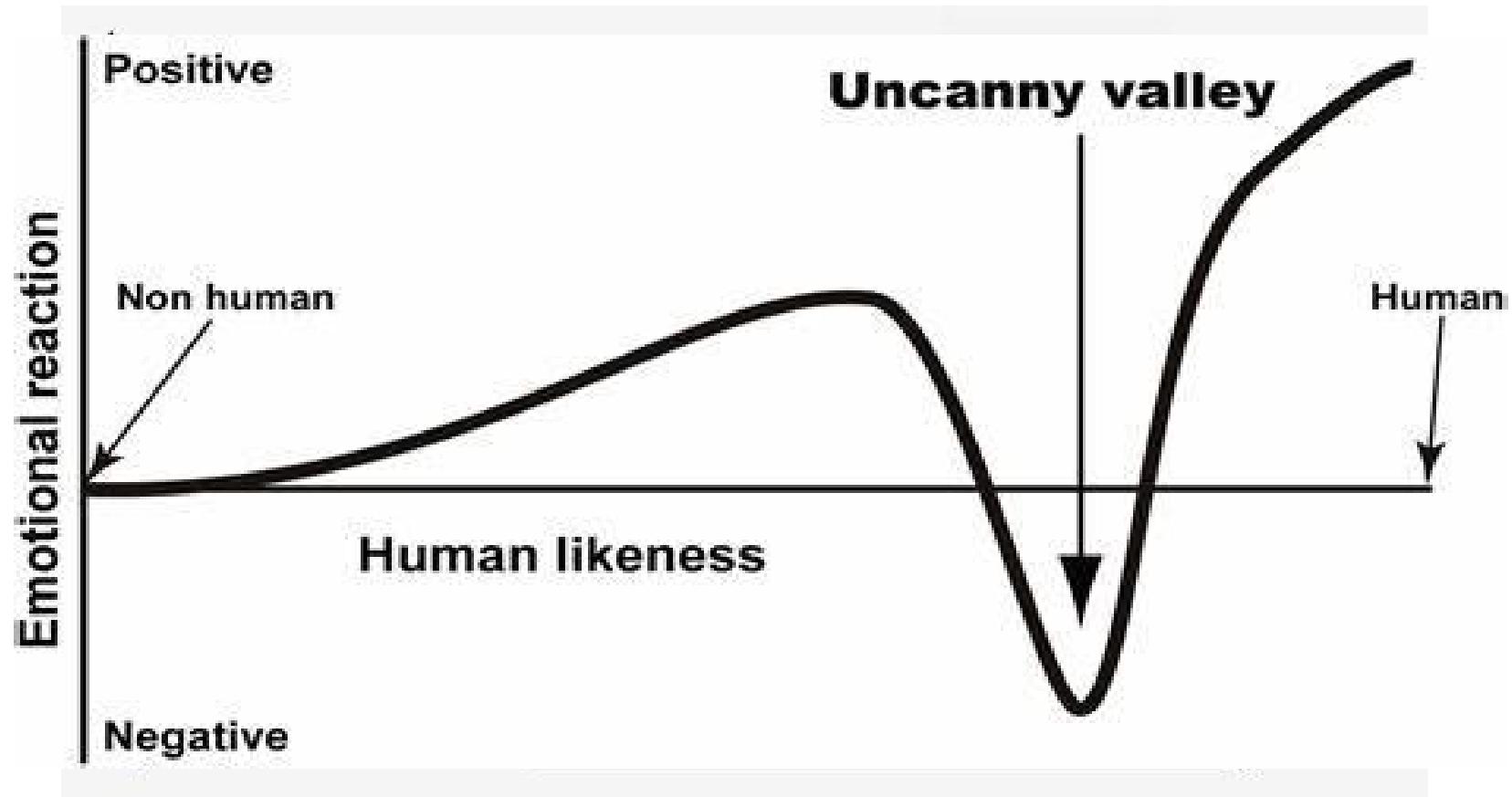
The developed AI algorithm can transform a brightfield image into a fluorescent-like image without sacrificing the cells to obtain.

UCLA Researchers Use AI to Show Multidimensional Imaging of Biological Processes



Dr. Tzong Hsiao

With Design Challenges – E.g. Uncanny Valley...



The Uncanny Valley occurs when AI entities such as robots present a close likeness to humans, but just miss in a creepy way, causing negative reactions.

...Design Challenges – Robot Gender...

SPECIAL ISSUE GENDERING ROBOTS: ONGOING (RE)CONFIGURATIONS OF GENDER IN ROBOTICS

International Journal of Social Robotics (IJSR) - Impact Factor 5.126 (2020) [[link to journal](#)][[link to special issue](#)]

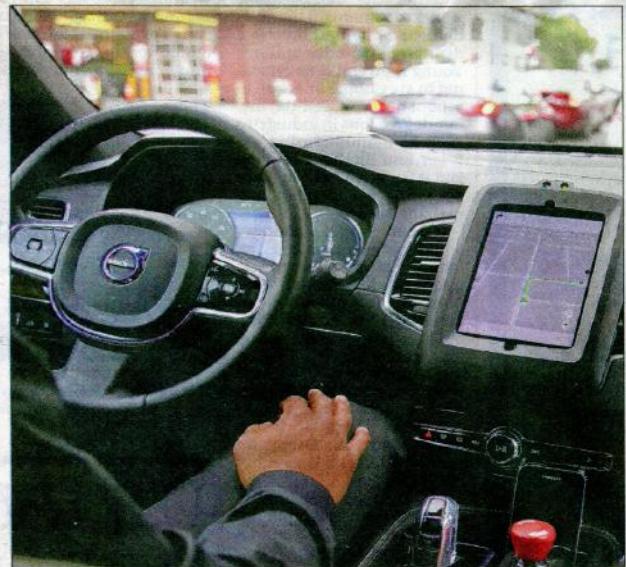
The field of HRI is increasingly employing humanoid robots and anthropomorphic designs to aid social interactions between humans and robots. However, designing robots with human-like characteristics can have far-reaching implications. For instance, integrating identity markers such as gender into robotic systems can involuntarily transfer implicit biases into the human-robot interaction (HRI) and perpetuate harmful structural patterns and societal stereotypes. This Special Issue aims to deepen the social, ethical, and psychological understanding of the attribution of gender to robots, addressing how the design and use of gendered robots might impact HRI and affect society at large.

We welcome original papers investigating the attribution of gender to social robots, the influence of robot's gender on the perception of and interaction with robots (e.g., acceptance, trust), the design of gendered robots, and its ethical and societal implications. Through this special issue, we would like to inspire discussions about the mindful use of robots in society and promote the possibility of using robots to challenge harmful social practices (e.g., gender stereotypes, gender inequality). Our ultimate goal is to bridge the communication gap we observed between the different disciplines addressing the topic of gender in robotics and the lack of spaces to converse with and learn from one another.

...Technical Challenges...

Uber reveals failings of its self-driving program

Company details management and tech changes it's implementing



AN UBER driverless car was involved in a fatal crash in Arizona this year. Above, one of the company's cars in San Francisco in 2016.

BY MICHAEL LARIS

More than seven months after a self-driving Uber SUV killed a pedestrian in Arizona, the company has released safety reports that detail broad technological and management failings and describe efforts since the tragedy to address them.

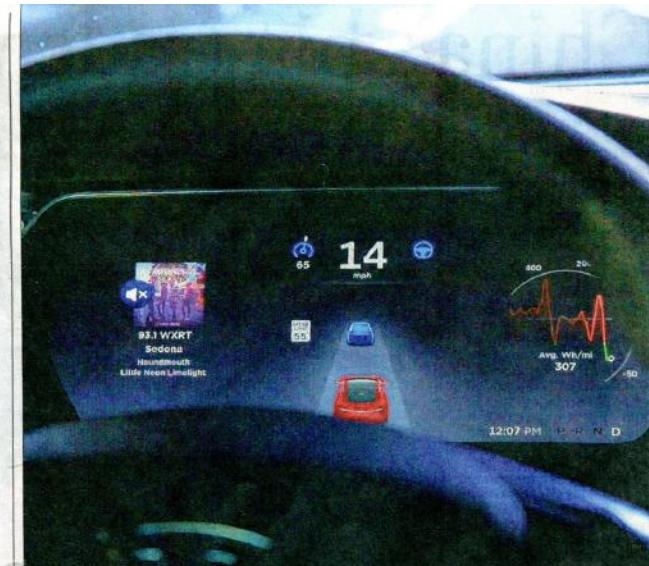
The findings released Friday reveal shortcomings at the core of Uber's driverless technology itself, which relies on cameras and sensors to take in the environment and software to process that information and make all the decisions — big and small — needed to drive safely.

A key internal recommendation cited the need for "improving the overall software system design," which is akin to saying Uber's robot car needed a better brain with sharper thinking.

In practice, that means that since the fatal crash in Tempe, Ariz., in March, company engineers have worked at "reducing latency," or the delay between when an initial observation is made and when an action is taken in response, Uber said. "We are now able to detect objects and actors sooner and execute safe reactions faster," Uber said.

The driverless [See Uber, C6]

LATIMES 1/13/18



A FLORIDA motorist who is suing Tesla said the automaker misleads drivers into believing the cars can drive themselves. Above is a Model S dashboard.

Tesla's Autopilot blamed for crash

Automaker is sued by driver whose Model S struck a disabled car.

ASSOCIATED PRESS

A Florida man sued Tesla Inc. on Tuesday, saying the vehicle's semiautonomous Autopilot feature failed to detect a disabled car on a highway, leading to a crash that left him with permanent injuries.

Shawn Hudson said in the lawsuit filed in state court in Orlando, Fla., that Tesla made false statements about the safety of the Autopilot feature on his Tesla Model S.

It's the second such law-

watch it, but we're also going to put this giant 20-inch screen right here with web-browsing capabilities so you can be distracted the entire time, but if you crash, that's your fault."

Hudson said he suffers pain from fractured vertebrae and has some cognitive problems since the crash two weeks ago on the Florida Turnpike.

Hudson, who lives in Orlando and has a two-hour commute to Fort Pierce for his job as the general manager of a Nissan dealership, said the Autopilot feature appealed to him because he could get some work done during his commute. Hudson had his hands on the wheel as the car traveled 80 mph but he also was looking

"Tesla has always been clear that Autopilot doesn't make the car impervious to all accidents, and Tesla goes to great lengths to provide clear instructions about what Autopilot is and is not," the spokeswoman said.

Hudson's attorneys said there is a disconnect between the official company policy and what salespeople tell customers in showrooms.

"What they say to federal regulators is very different from what you hear on the Tesla lot," Morgan said. "It's very different from what Mr. Hudson was told when he was buying his car."

Tuesday's lawsuit puts Tesla's driver-assistance system back in the headlines for legal reasons, rather than

With Technical Challenges...

A8 FRIDAY, JULY 15, 2022

Los Angeles Times

LATIMES.COM

BUSINESS

Tesla feature is blamed for several crashes

Regulators knew of them when Musk said Full Self-Driving had a perfect safety record.

BY RUSS MITCHELL

Elon Musk has long used his mighty Twitter megaphone to amplify the idea that Tesla's automated driving software isn't just safe — it's safer than anything a human driver can achieve.

That campaign kicked into overdrive last fall when the electric-car maker expanded its Full Self-Driving "beta" program from a few thousand people to a fleet that now numbers more than 100,000. The \$12,000 feature purportedly lets a Tesla drive itself on highways and neighborhood streets, changing lanes, making turns and obeying traffic signs and signals.

As critics scolded Musk for testing experimental technology on public roads without trained safety drivers as backups, Santa Monica investment manager and vocal Tesla booster Ross Gerber was among the allies who sprang to his defense.

"There has not been one



South Jordan Police Department via Associated Press

A MODEL S sedan is shown in 2018 after it crashed into a fire department's mechanic truck stopped at a red light in South Jordan, Utah. The driver said she was using the car's semiautonomous driving function.

bags deployed (both sides.) The car traveled about 500 yards along the road and then turned itself off." The estimated damages were \$28,000 to \$30,000.

■ **Troy, Mo.**: A Tesla was turning through a curve when "suddenly, about 40% of the way through the turn, the Model Y straightened the wheel and crossed over the center line into the direct path of the oncoming vehicle. When I attempted to pull the vehicle back into my lane, I lost control and skidded off into a ditch and through the woods, causing significant damage to the vehicle."

■ **Jackson, Mo.**: A Model 3 " jerked right toward a semi truck, then jerked left toward the posts in the median as it was accelerating and FSD would not turn off.... We owned this car for 11 days when our wreck happened."

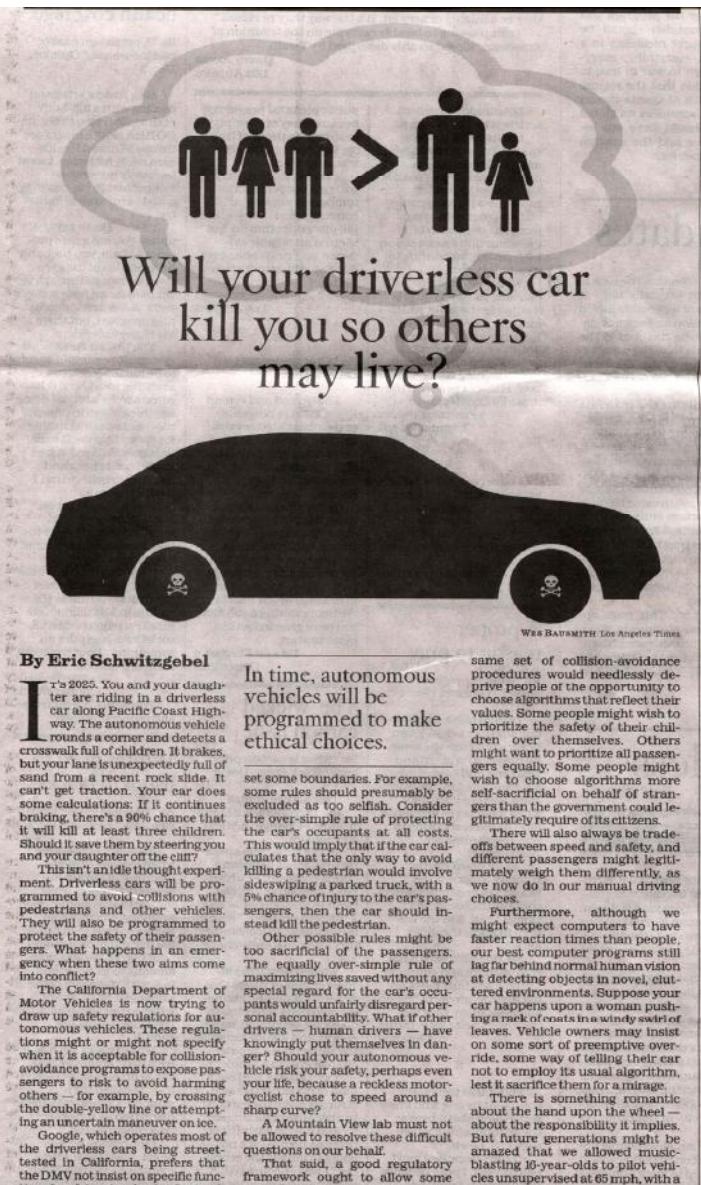
■ **Hercules, Calif.**: "Phantom braking" caused the Tesla to suddenly stop, and "the vehicle behind me didn't react." A rear-end collision caused "serious damage to the vehicle."

■ **Dallas**: "I was driving on full self driving assistance ... a car was in my blind spot so I tried to take over the car

...and Ethical Challenges

Google, which operates most of the driverless cars being street-tested in California, prefers that the DMV not insist on specific functional safety standards. Instead, Google proposes that manufacturers “self-certify” the safety of their vehicles, with substantial freedom to develop collision-avoidance algorithms as they see fit.

That’s far too much responsibility for private companies. Because determining how a car will steer in a risky situation is a moral decision, programming the collision-avoiding software of an autonomous vehicle is an act of applied ethics. We should bring the programming choices into the open, for passengers and the public to see and assess.



...and Ethical Challenges

Google, which operates most of the driverless cars being street-tested in California, prefers that the DMV not insist on specific func-



Will your driverless car kill you so others

It's 2025. You and your daughter are riding in a driverless car along Pacific Coast Highway. The autonomous vehicle rounds a corner and detects a crosswalk full of children. It brakes, but your lane is unexpectedly full of sand from a recent rock slide. It can't get traction. Your car does some calculations: If it continues braking, there's a 90% chance that it will kill at least three children. Should it save them by steering you and your daughter off the cliff?

This isn't an idle thought experiment. Driverless cars will be programmed to avoid collisions with pedestrians and other vehicles. They will also be programmed to protect the safety of their passengers. What happens in an emergency when these two aims come into conflict?

gers and the public to see and assess.

They will also be programmed to protect the safety of their passengers. What happens in an emergency when these two aims come into conflict?

The California Department of Motor Vehicles is now trying to develop safety regulations for autonomous vehicles. These regulations might or might not clarify when it is acceptable for collision-avoidance programs to expose passengers to risk to avoid harming others — for example, by crossing the double-yellow line or attempting an uncertain maneuver on ice.

Google, which operates most of the driverless cars being street-tested in California, prefers that the DMV not insist on specific functional safety standards. Instead

it would kill the pedestrian.

Other possible rules might be too sacrificial of the passengers. The equally over-simple rule of maximizing lives saved without any special regard for the car's occupants would unfairly disregard personal accountability. What if other drivers — human drivers — have knowledge? Put themselves in danger? Should your autonomous vehicle risk your safety, perhaps even your life, because a reckless motorcyclist chose to speed around a sharp curve?

A Mountain View lab must not be allowed to resolve these difficult questions on our behalf.

That said, a good regulatory framework ought to allow some manufacturer variation and a

might expect computers to have faster reaction times than people, our best computer programs still lag far behind normal human vision at detecting objects in novel, cluttered environments. Suppose your car happens upon a woman pushing a rack of coats in a windy swirl of leaves. Vehicle owners may insist on some sort of preemptive override, some way of telling their car to employ its usual algorithm, lest it sacrifice the passenger.

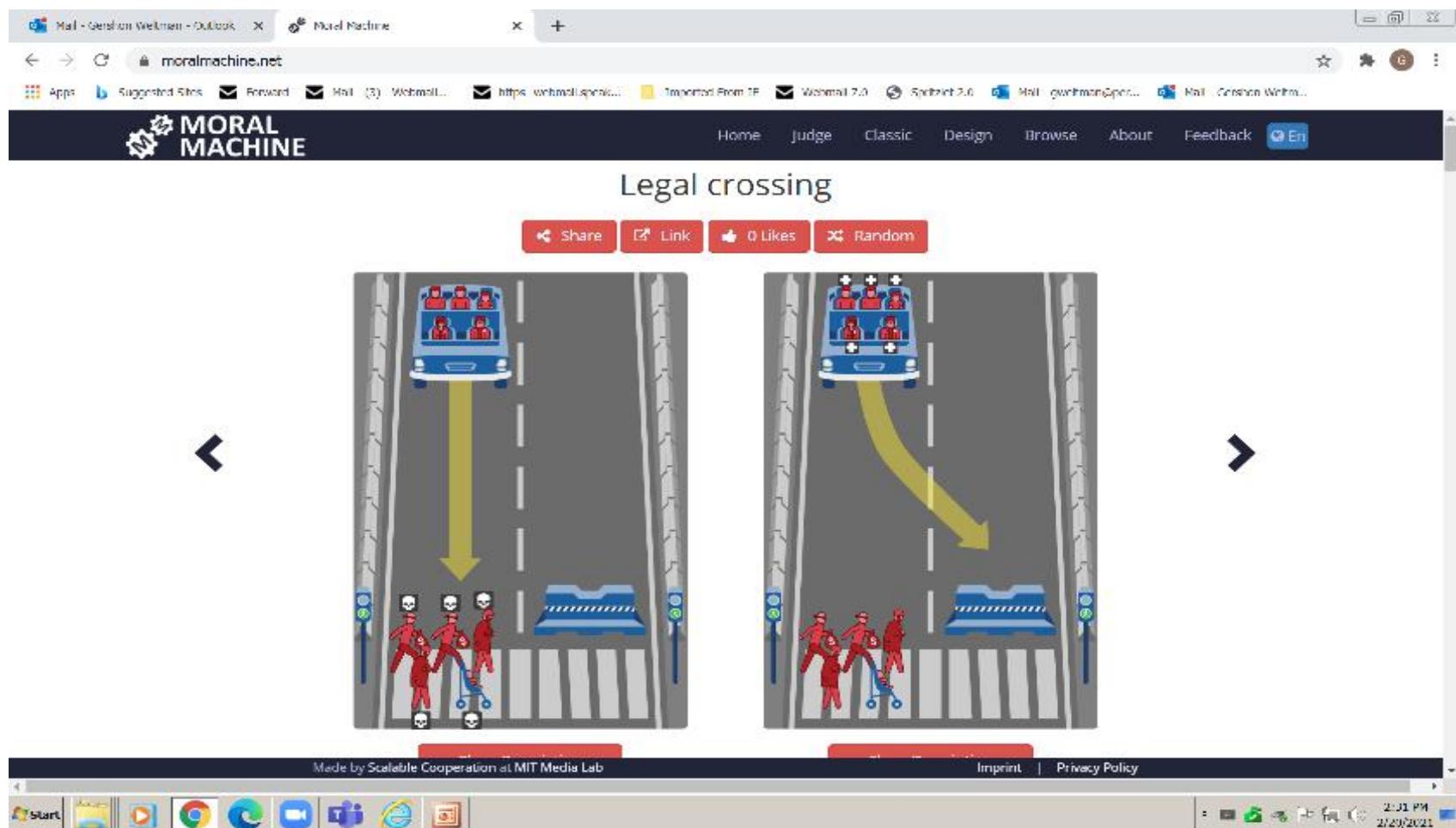
There is something romantic about the hand upon the wheel — about the responsibility it implies. But future generations might be amazed that we allowed music-blasting 10-year-olds to pilot vehicles unsupervised at 65 mph, with a

Emulating Humans May Not be Enough

The screenshot shows a Microsoft Internet Explorer window with the following details:

- Title Bar:** Pa. boy killed when mom swerves to miss deer - CBS News - Internet Explorer
- Address Bar:** http://www.cbsnews.com/news/pa-boy-killed-when-mom-swerves-to-miss-deer/
- Navigation Bar:** CBS News / CBS Evening News / CBS This Morning / 48 HOURS / 60 MINUTES / Sunday Morning / Face the Nation / CBSN Log In Search
- Header:** CBSNEWS Video | US | World | Politics | Entertainment | Health | MoneyWatch | SciTech | Crime | Sports | Photos | More
- Banner:** Adobe SAVE OVER 60% ON CREATIVE CLOUD Limited-time offer. This back-to-school deal ends Sept. 2. Join now >
- Article Headline:** CBS NEWS | September 2, 2016, 4:01 PM Pa. boy killed when mom swerves to miss deer
- Image:** A photo of a young boy with red hair smiling.
- Caption:** Isaac Thomas, 6, suffered serious head injuries and died at Allegheny Valley hospital on Thursday, Sept. 1, 2016. / KDKA CBS Pittsburgh
- Social Sharing:** Comment / Share / Tweet / Stumble / Email
- Right Sidebar:**
 - Eye on Pets:** A CBS NEWS SPECIAL REPORT. The latest animal health and trends from CBS News. Brought to you by: BRAVECTO® (FLUCLANER)
 - Video Preview:** A video thumbnail featuring two puppets, one in a graduation cap and gown, the other in a green fur coat.
 - Sponsored by SC Johnson:** Protect Your Kids from Mosquito Bites
- Taskbar:** Start button, File, Internet Explorer, Word, and others. Date: 9/2/2016, Time: 4:55 PM, Resolution: 125%

Ethical Issues: The Moral Machine



The [Moral Machine](#) was created by scientists at MIT to crowd source answers for the complex moral dilemmas that arise with driverless vehicles.

The Results are Somewhat Unexpected

- Situations can differ, e.g.
 - Trolley Problem 1
 - Trolley Problem 2
 - Morality is not universal
 - Significant cultural differences
 - 80% of Americans will pull the switch
 - 50% of Americans will push the man
 - Other cultures will not intervene at all
 - Sacrifice of any type is problematic
 - Americans most likely to chose a sacrifice
 - Asians least likely to chose a sacrifice
 - Conclusion/Question
 - Ethical programming may have to follow local culture
 - Is this an acceptable outcome?
- The image contains two separate diagrams. The left diagram shows a trolley heading towards five people tied to the tracks. A person stands next to a switch, with a line leading from them to the switch. The right diagram shows a trolley heading towards five people tied to the tracks. A person stands on a bridge over the tracks, pushing another person off the bridge towards the trolley.

1. Pull the switch? 2. Push the man?

Amada O'Bryan. Moral Dilemmas for Driverless Cars, <https://medium.com>, February 16, 2021

Challenges of Operational Bias...

Studies and individuals' experiences have shown that AI algorithms – such as facial recognition – may handle classifications of race, ethnicity, gender and ability inaccurately as a result of their initial programming



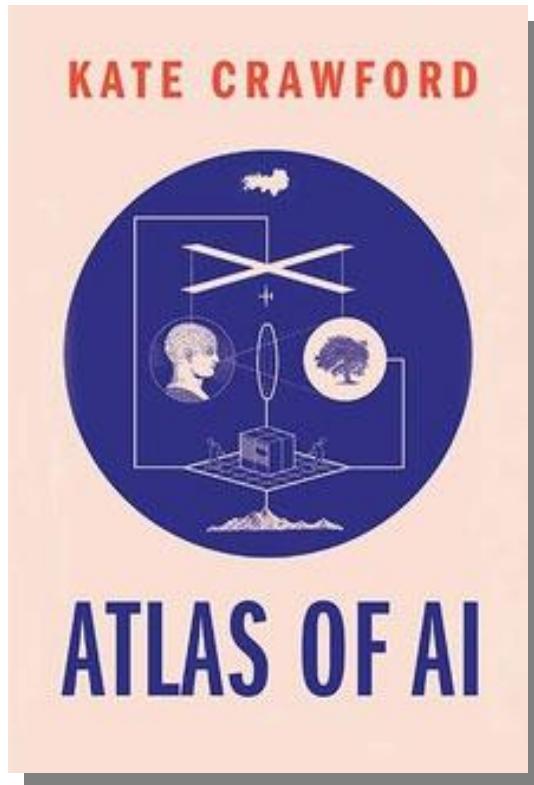
The Algorithmic Justice League is a collective to:

- Highlight *algorithmic bias* through provocative media and interactive exhibitions.
- Provide space for people to voice concerns and experiences with *coded discrimination*.
- Develop *practices for accountability* during the design, development, and deployment phases of coded systems.



<https://www.ajlunited.org/>

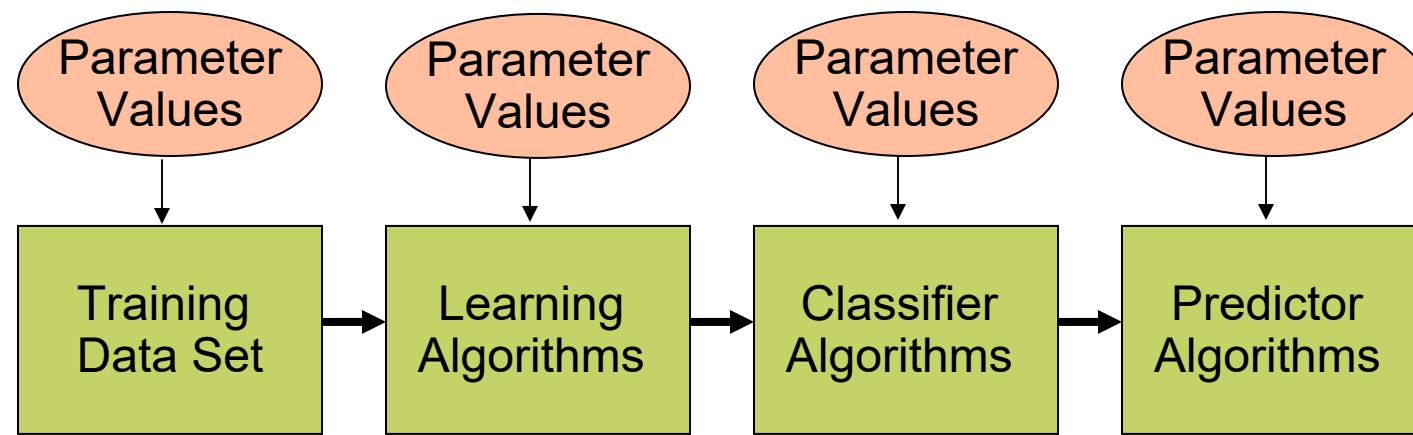
,, and More Fundamental Biases



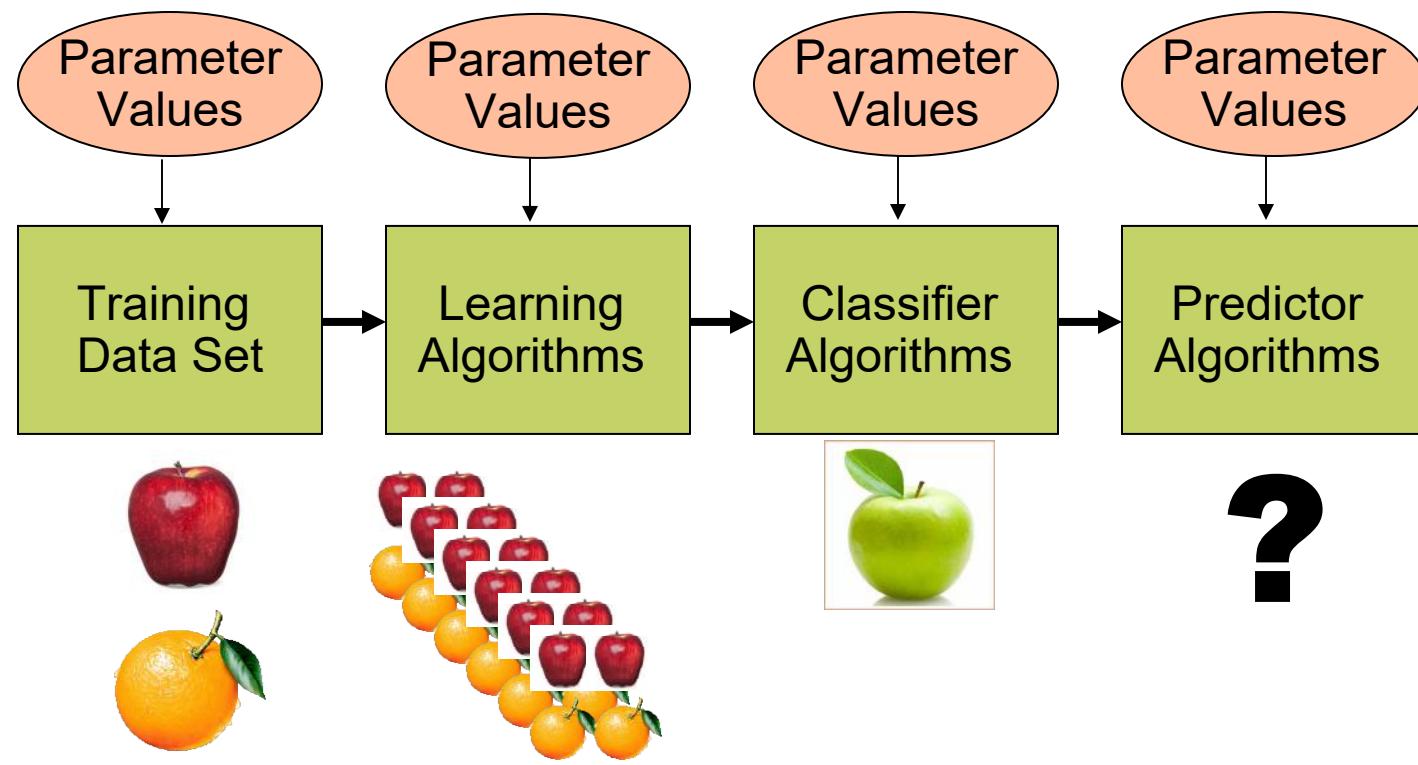
Crawford challenges *not only* the accuracy of classification but also the:

- Validity of the classes themselves,
- Motivations of the classifiers,
- Bases of data sourcing/scraping,
- Rights of the people being sourced,
- Consequences for the classified

Transparency in Machine Learning



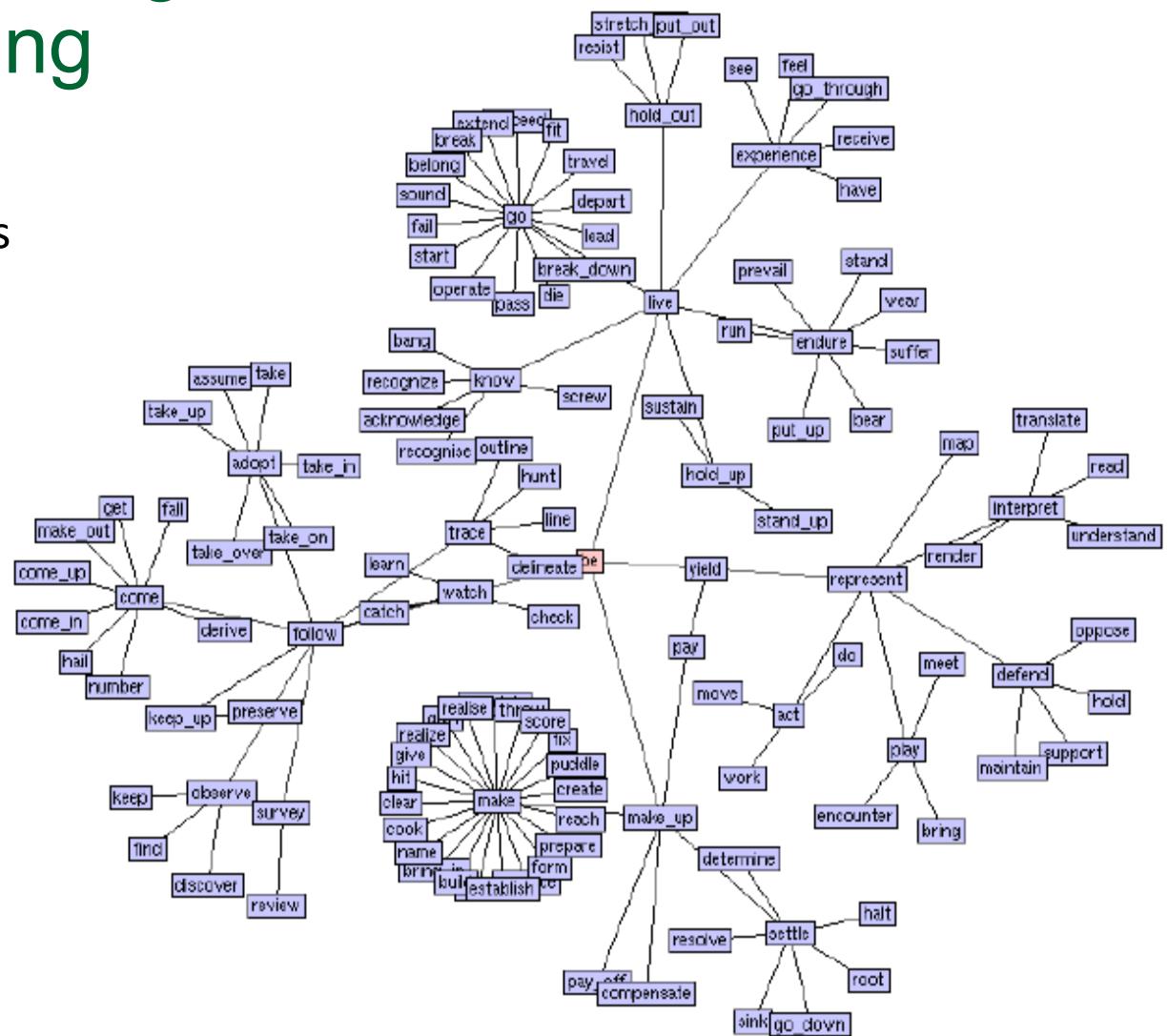
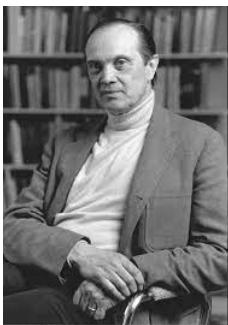
Transparency in Machine Learning



Basis of Speech Recognition by Machine Learning

WordNet:

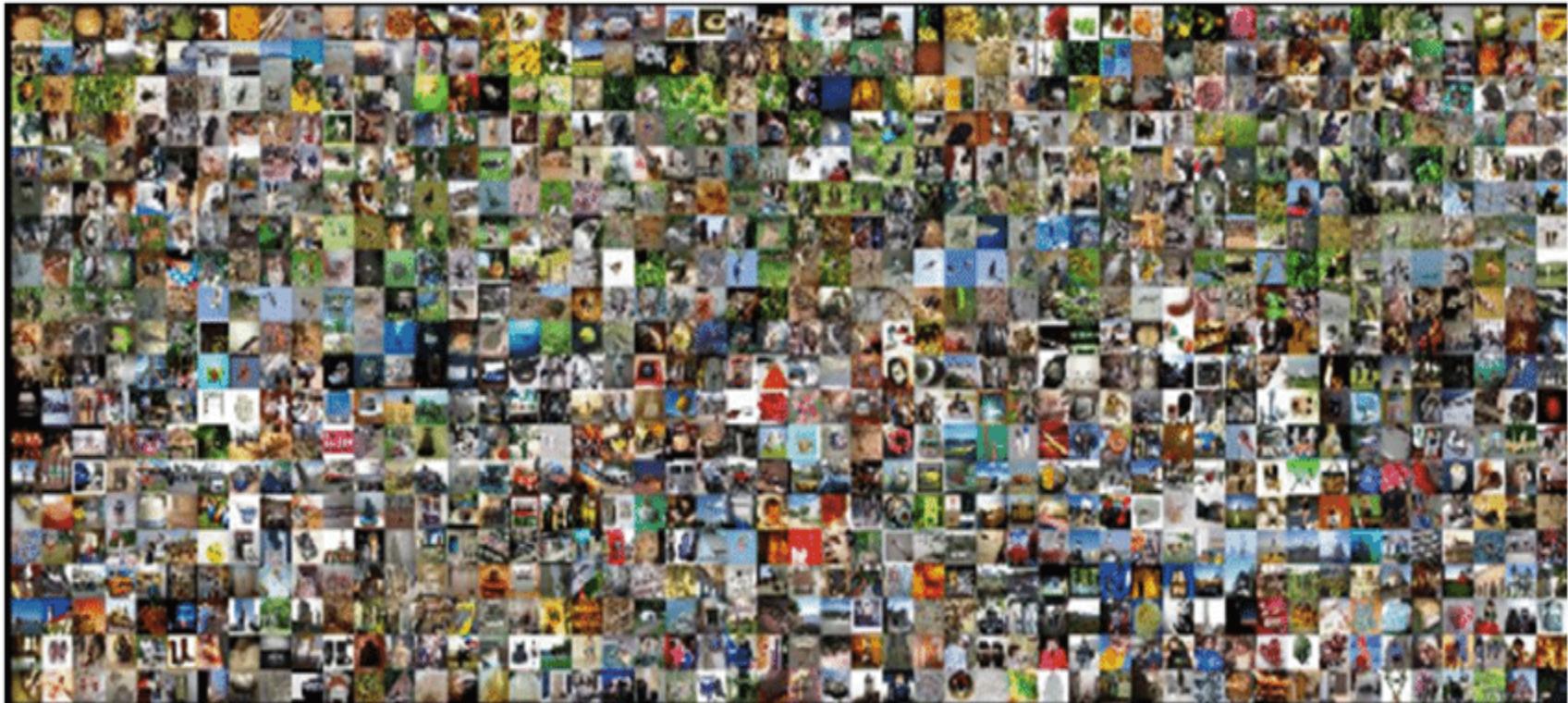
- Classification of English words
- Created first in 1985
- Princeton University
- Prof. George Armitage Miller
- Funded by ONR, DARPA, NSF



Extension to Face Recognition

ImageNet:

Is the invention of Professor Fei-Fei Li of Stanford University and provides the basic data for many AI facial recognition programs



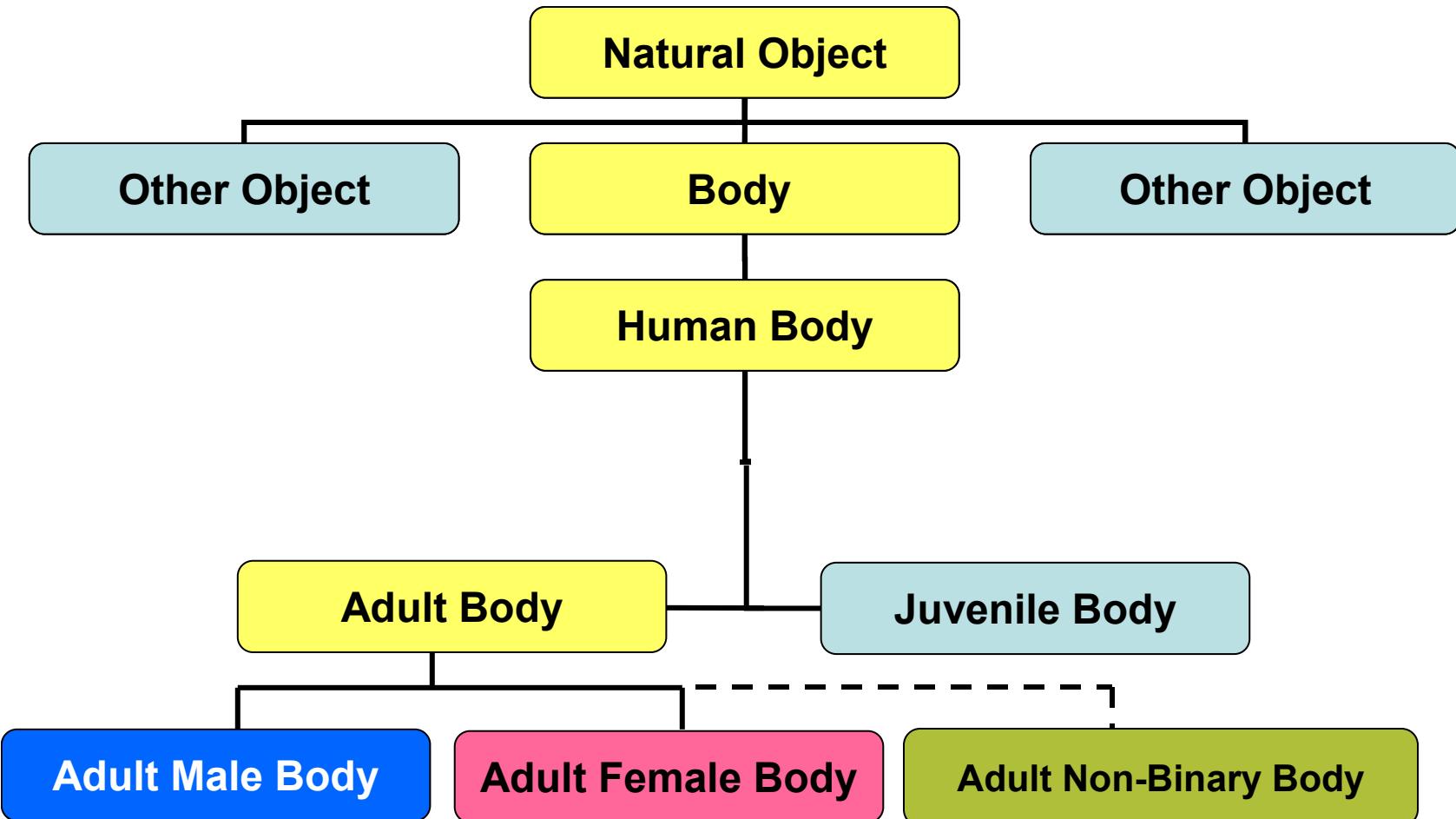
ImageNet Incorporated WordNet

WordNet Categories

- Plant
- Geological Formation
- Natural Object
- Sport
- Artifact
- Fungus
- Person
- Animal
- Miscellaneous



With Questionable Results



The classifications of human bodies reflect historic societal norms not current conditions

Similarly for Emotion Recognition



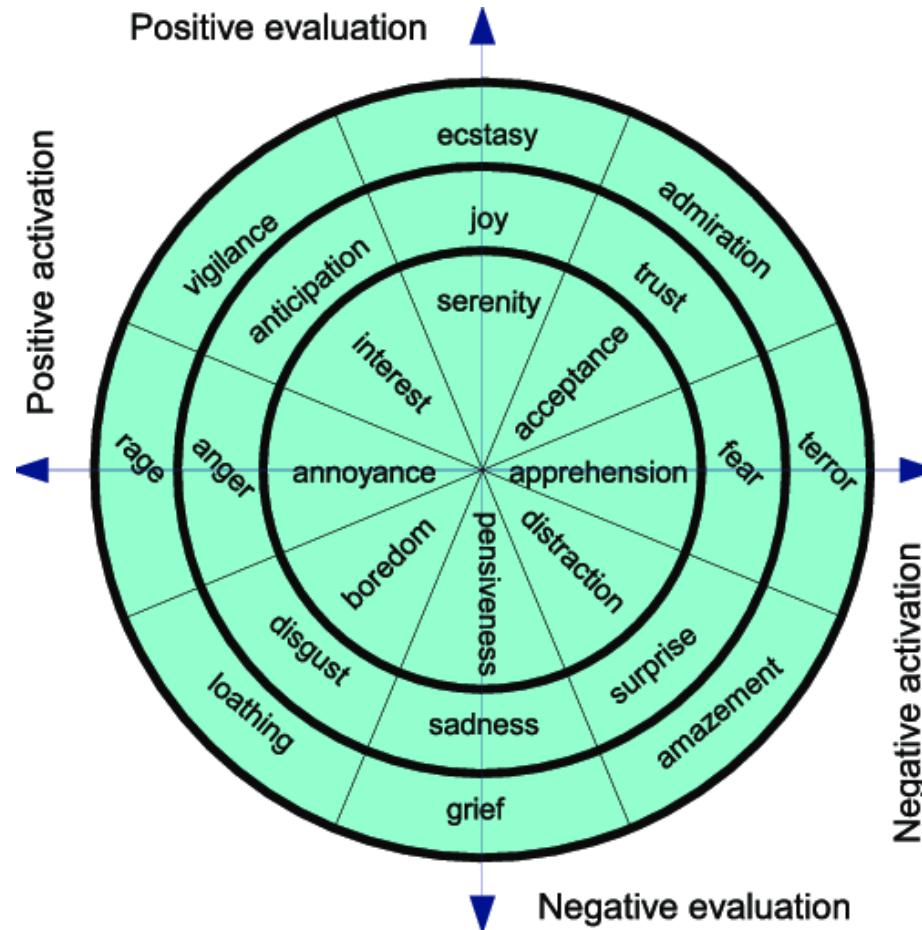
In the 1980s Prof. Paul Ekman of the University of California, San Francisco, developed the widely accepted theory that facial expressions could reflect six basic emotions.

Similarly for Emotion Recognition



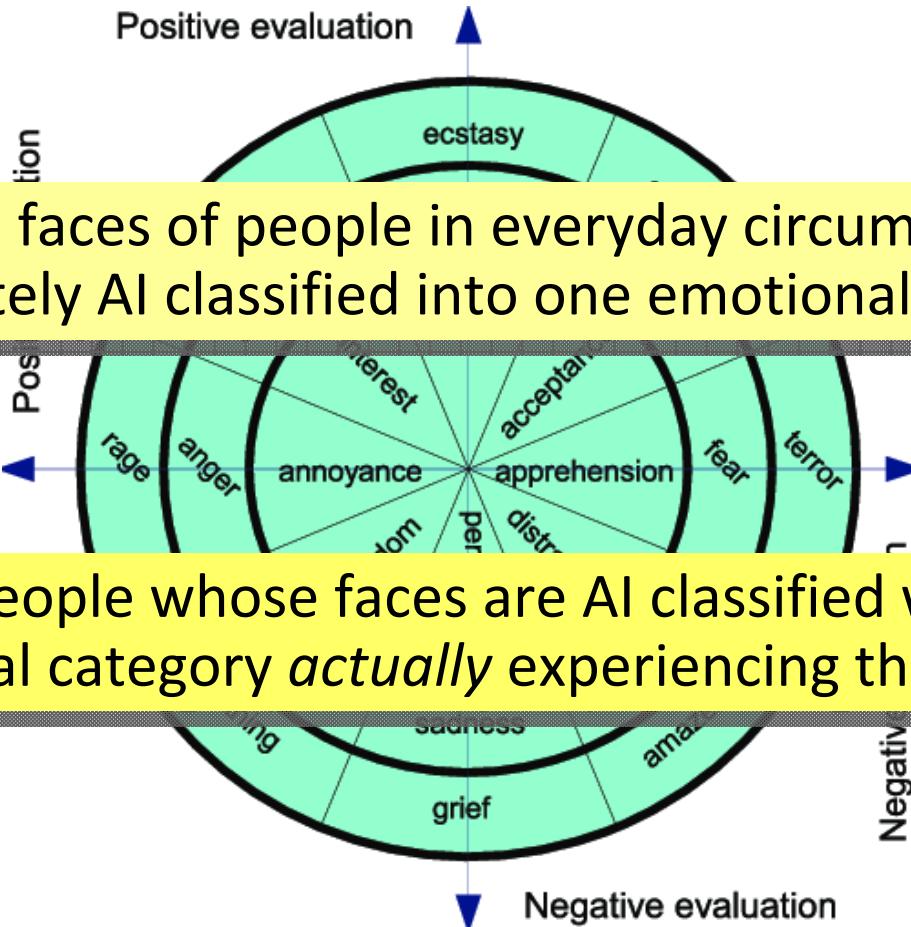
Many AI programs for recognizing emotions are built on datasets of “actors” attempting to express the six (or seven) basic emotions proposed by Prof. Ekman’s theory.

Expansion and Caveats



But many researchers offer much more complex classifications of emotional states

So Kate Crawford Asks...

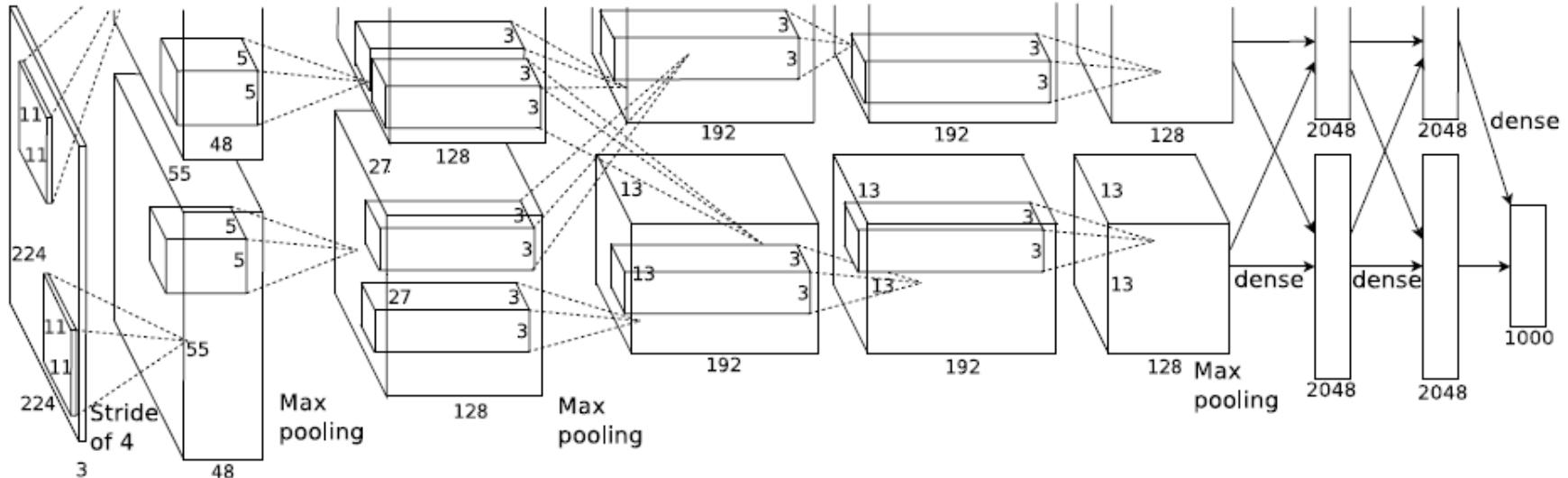


Can the faces of people in everyday circumstances be accurately AI classified into one emotional category?

Are people whose faces are AI classified within an emotional category *actually* experiencing that emotion?

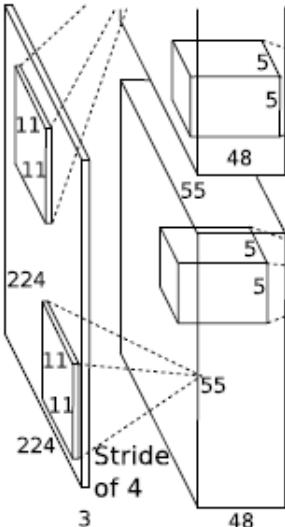
Many researchers offer much more complex classifications of emotional states

The Takeaway: It Matters



Architecture reported by Krizhevsky, Sutskever and Hinton (2012) “ImageNet Classification with Deep Convolutional Networks,” Advances in Neural Information Processing Systems

The Takeaway: It Matters

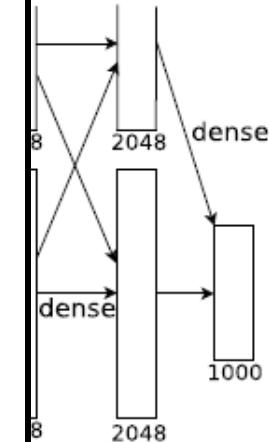


Architecture reported with Deep Convolutional

AI-based recognition of emotion (along with gender, race, ethnicity and other classifiers) is commonly used in:

- Job candidate searches and interviews
- Therapeutic programs
- Consumer reactions and preferences
- Criminal lie detection
- Public surveillance
- Etc., Etc.

So we have to be ethically conscious in recognizing and carefully scrutinizing the technologies by means of which these AI judgments are formed and used.



Classification Systems

"By looking at the layers of training data that shape and inform AI models and algorithms, we can see that gathering and labeling data about the world is a social and political intervention, even as it masquerades as a purely technical one."

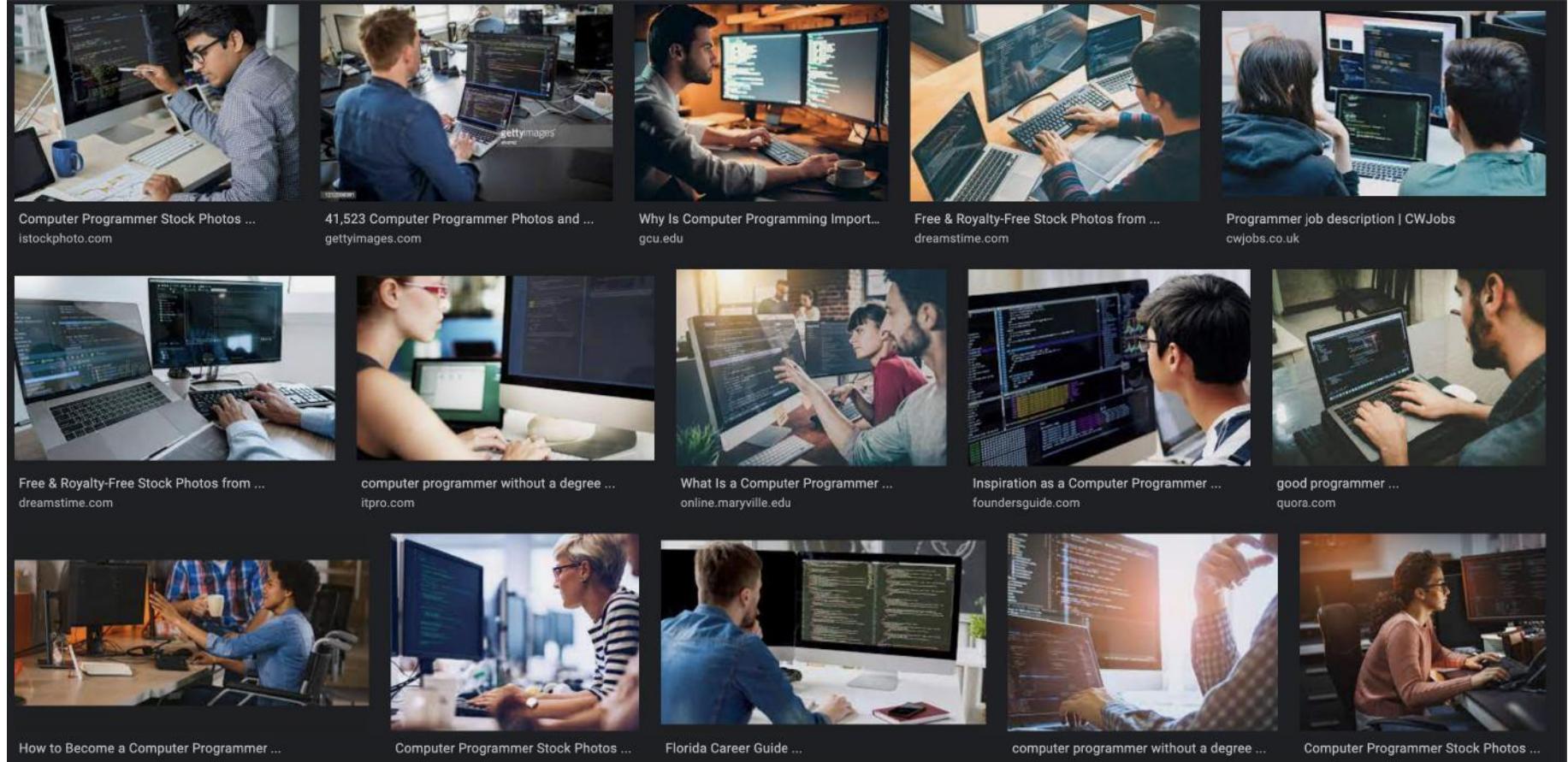
Kate Crawford, Atlas of AI, p. 121

For Example: Google Search for Nurse



Contributed by Prof. Baharan Mirzasoleiman, UCLA Computer Science

For Example: Google Search for Programmer



Contributed by Prof. Baharan Mirzasoleiman, UCLA Computer Science

For Example: AI-Aided Social Work

A2 THURSDAY, MAY 5, 2022

Los Angeles Times

LATIMES.COM

PERSPECTIVES

Child-neglect algorithm raises concerns

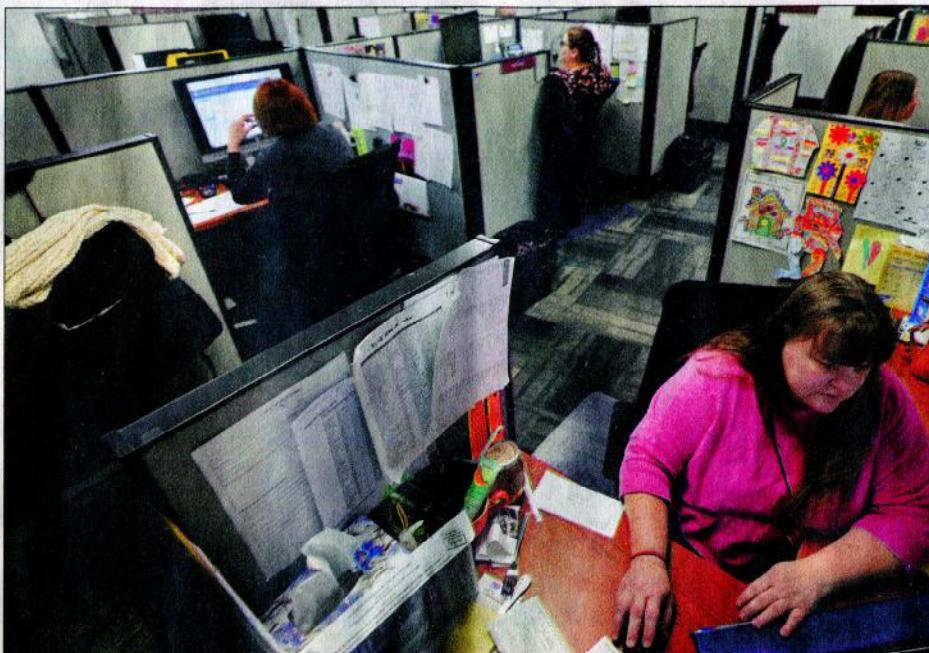
In its first years of operation, the system flagged an inordinate number of Black kids for investigation.

BY SALLY HO AND GARANCE BURKE

For family law attorney Robin Frank, defending parents at one of their lowest points — when they risk losing their children — has never been easy.

The job is never easy, but in the past she knew what she was up against when squaring off against child protective services in family court. Now, she worries she's fighting something she can't see: an opaque algorithm whose statistical calculations help social workers decide which families should be investigated in the first place.

"A lot of people don't know that it's even being used," Frank said. "Families should have the right to have all of the information in their file."



Keith Srakocic / Associated Press

WORKERS ANSWER calls for the Allegheny County children and youth services office in Penn Hills, Pa. A child-neglect algorithm produced risk scores that social workers disagreed with about one-third of the time.

by the governor.

It recently paused a pilot algorithm built to help decide when foster care children can be reunified with their families. Oregon also explored three other tools — predictive models to assess a child's risk for death and severe injury, whether children should be placed in foster care, and if so, where.

For years, California explored data-driven approaches to the statewide child welfare system before abandoning a proposal to use a predictive risk modeling tool in 2019.

"During the project, the state also explored concerns about how the tool may impact racial equity. These findings resulted in the state ceasing exploration," department spokesman Scott Murray said in an email.

Los Angeles County's Department of Children and Family Services is being audited following high-profile child deaths, and is seeking a new director after its previous one stepped down last year. It's piloting a "complex-risk algorithm" that helps to isolate the highest-risk cases that are being in-

For Example: AI-Aided Social Work

A2 THURSDAY, MAY 5, 2022

Los Angeles Times

LATIMES.COM

PERSPECTIVES

Child-neg

In its first years of operation, the system flagged an inordinate number of Black kids for investigation.

By SALLY HO AND GARANCE BURKE

For family law attorney Robin Frank, defending parents at one of their lowest points — when they risk losing their children — has never been easy.

The job is never easy, but in the past she knew what she was up against when squaring off against child protective services in family court. Now, she worries she's fighting something she can't see: an opaque algorithm whose statistical calculations help social workers decide which families should be investigated in the first place.

"A lot of people don't know that it's even being used," Frank said. "Families should have the right to have all of the information in their file."

WORKERS AI
A child-neglect

"Workers, whoever they are, shouldn't be asked to make, in a given year, 14, 15, 16,000 of these kinds of decisions with incredibly imperfect information," said Erin Dalton, director of the county's Department of Human Services and a pioneer in implementing the predictive child welfare algorithm.

Critics say it gives a program powered by data mostly collected about poor people an outsize role in deciding families' fates, and they warn against local officials' growing reliance on artificial intelligence tools.

es concerns

by the governor.

It recently paused a pilot algorithm built to help decide when foster care children can be reunified with their families. Oregon also explored three other tools — predictive models to assess a child's risk for death and severe injury, whether children should be placed in foster care, and if so, where.

For years, California explored data-driven approaches to the statewide child welfare system before abandoning a proposal to use a predictive risk modeling tool in 2019.

"During the project, the state also explored concerns about how the tool may impact racial equity. These findings resulted in the state ceasing exploration," department spokesman Scott Murray said in an email.

Los Angeles County's Department of Children and Family Services is being audited following high-profile child deaths, and is seeking a new director after its previous one stepped down last year. It's piloting a "complex-risk algorithm" that helps to isolate the highest-risk cases that are being in-



H. SRAKOCIC Associated Press
In Penn Hills, Pa., one-third of the time.

There Are Other Costs as Well

Training a **single** deep model GPT-3 on **45TB** of data costs:



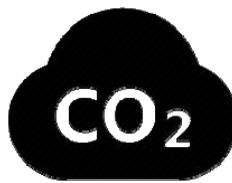
12M



34 days on 1024 A100s
335 years on 8 A100s



17.8X the yearly **energy** consumption
of the average American



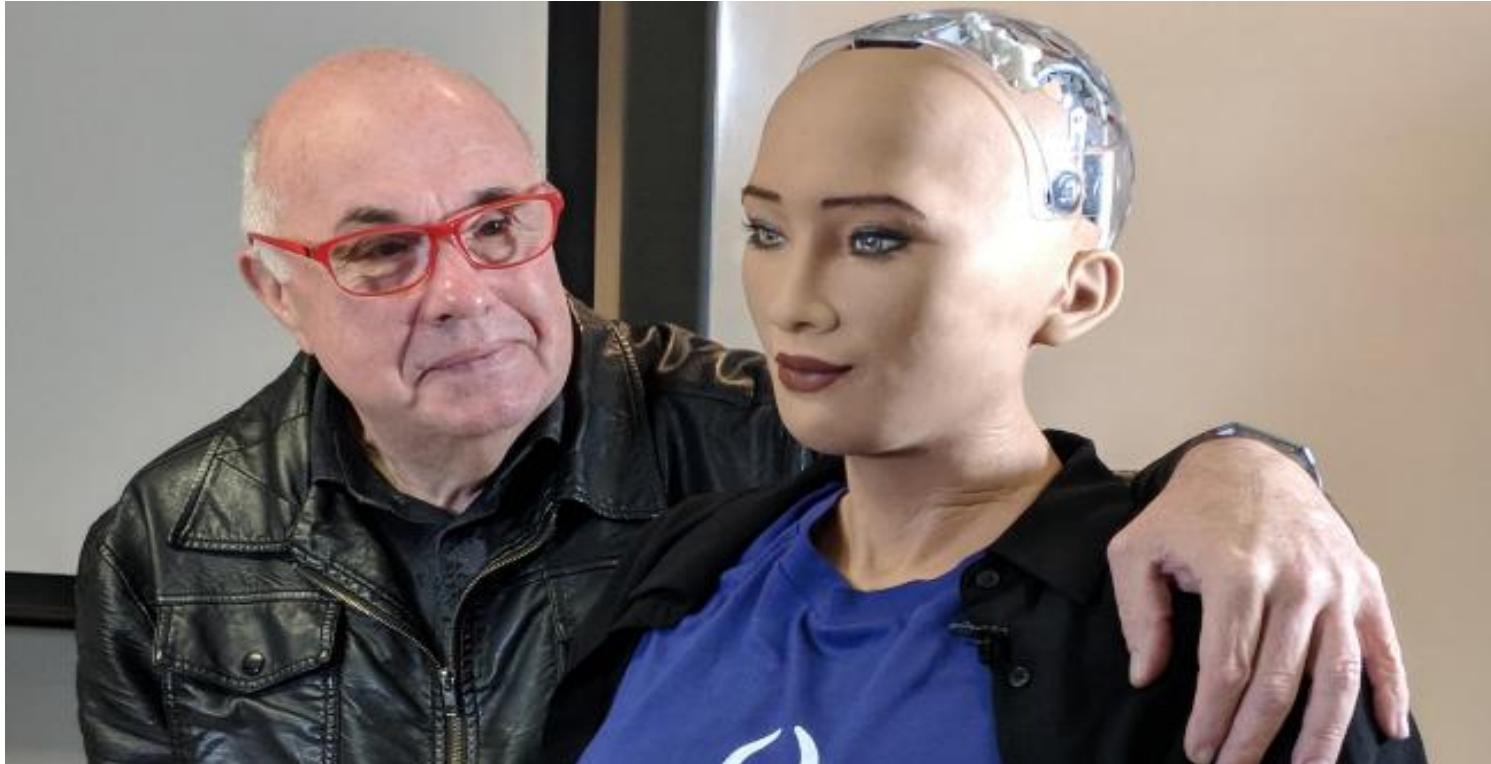
CO₂: a car driving 2x the distance between
Earth and Moon!!

Contributed by Prof. Baharan Mirzasoleiman, UCLA Computer Science

AGI Overview: Super Entity Issues

- Limits: Would it be ethical to limit:
 - The future development of ‘singular’ AI entities?
 - The amount of Intelligence such entities can have?
 - The type of jobs they can work in?
 - Their overall interaction with humans?
- Status: If super AI entities are developed, will they:
 - Be considered “people?”
 - Have the right to “Life, liberty and the pursuit of happiness”?
 - Be able to procreate? Allowed to do so?
 - Be subject to punishment? If so, what kind?
 - Have societal power of their own? If so, how will it be expressed?

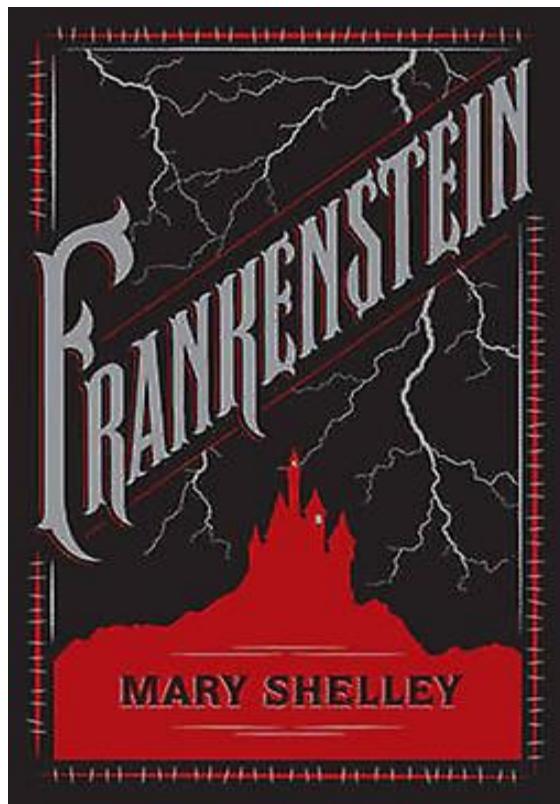
A Pioneering Step



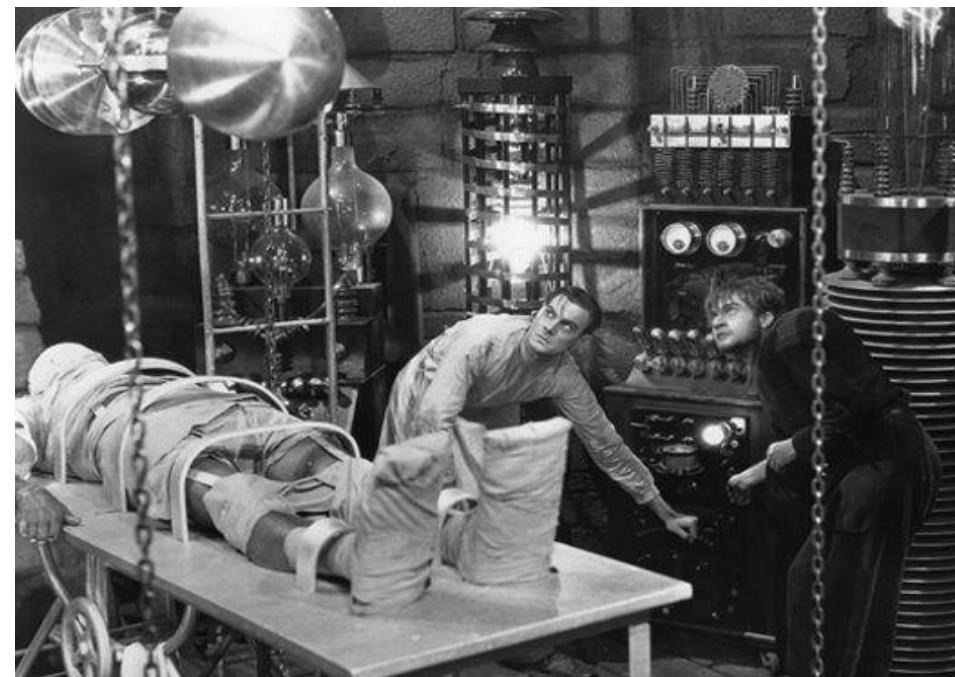
**In 2017 Sophia the robot was granted
Saudi Arabian citizenship – A first!**

<http://www.businessinsider.com/sophia-ai-robot-video-viral-future-tech-2017-11>

Historical Anticipation...

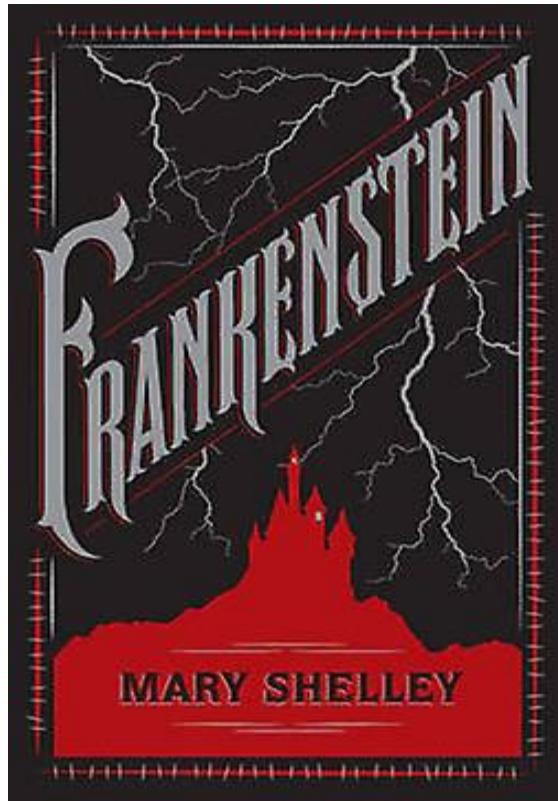


1818 Novel

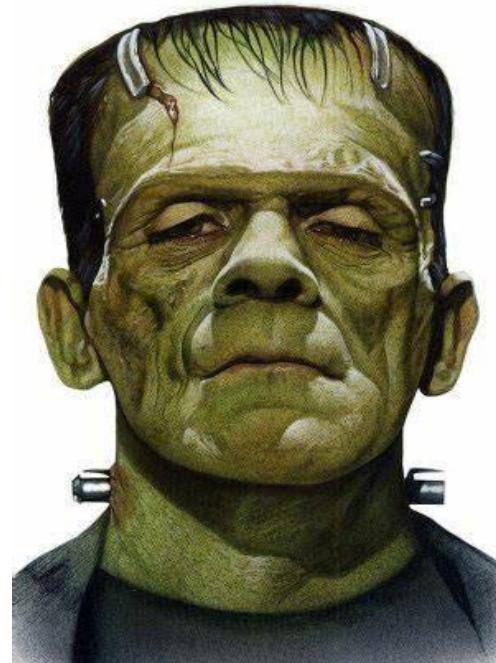


Many movie versions of the story

Historical Anticipation...



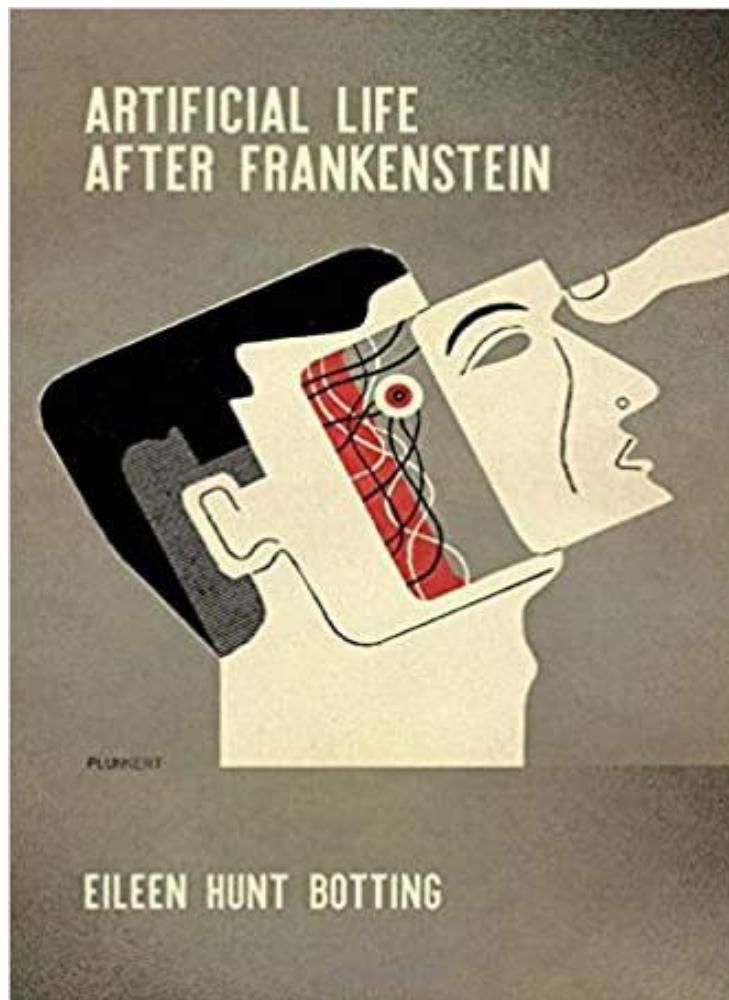
1818 Novel



A visualization of Dr. Frankenstein's creation

The “Creature” develops intellect and morality through study and observation. But when he discovers his origins as a creation of Dr. Frankenstein, and is rejected as a member of society, he becomes the destructive “Monster.”

...with Modern Interpretations...

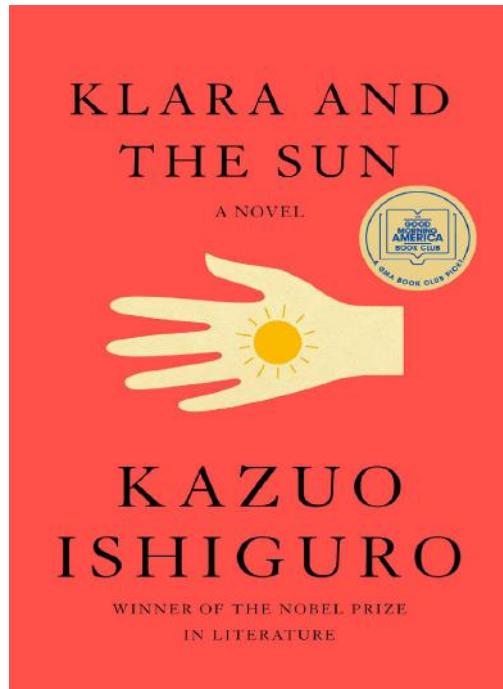


"Mary Shelley and authors of modern political science fiction from H. G. Wells to Nnedi Okorafor have paved the way for a techno-political philosophy of living with the artifice of humanity in all of its complexity. In *Artificial Life After Frankenstein*, Botting brings the insights born of Shelley's legacy to bear upon the ethics and politics of making artificial life and intelligence in the twenty-first century."

Amazon.com review
February 20, 2021

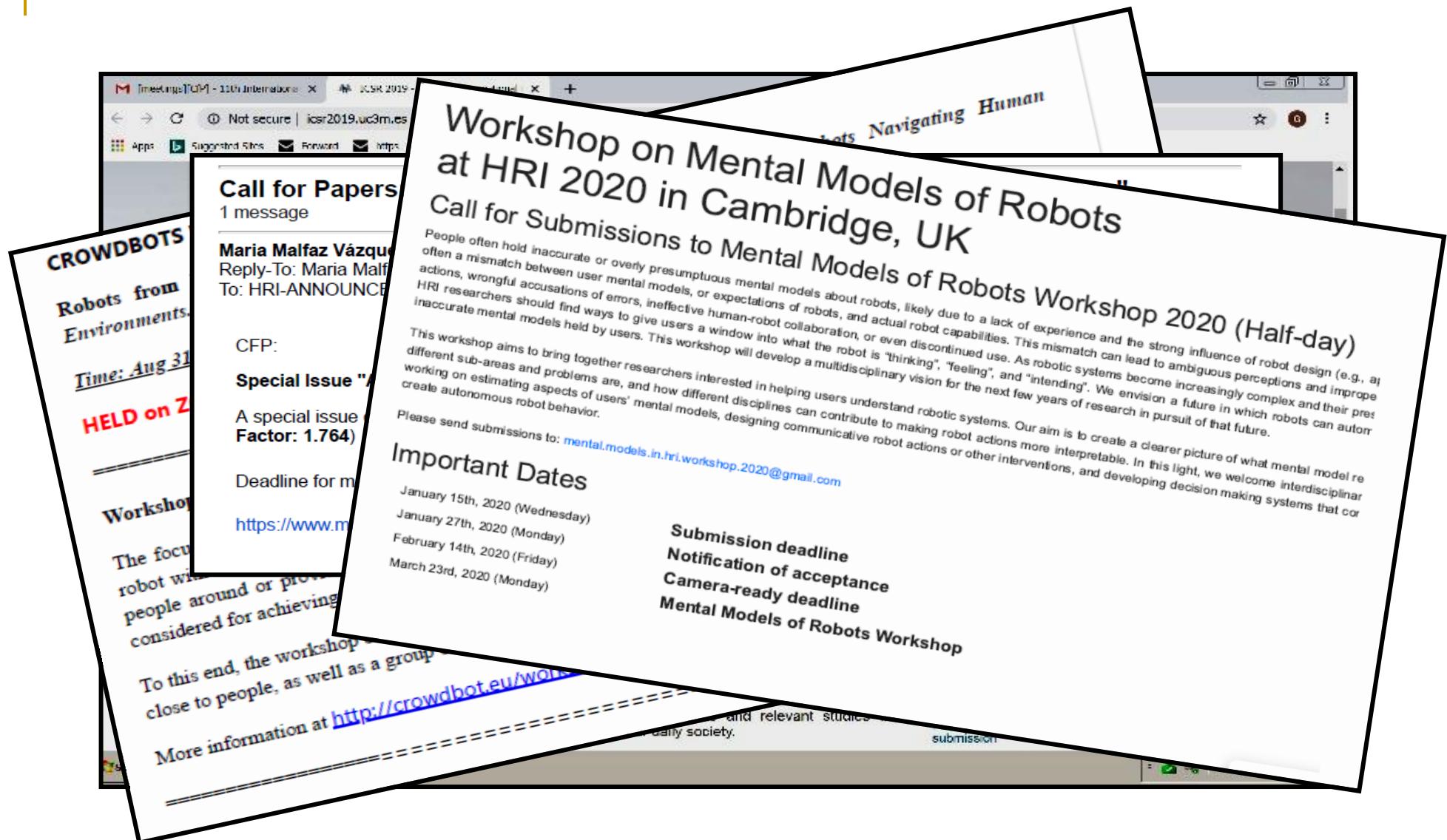
Eileen Hunt Botting is a political theorist whose scholarly interests cover modern political thought, feminism, the family, rights, ethics of technology, and philosophy and literature. She has taught at Notre Dame since 2001.

...and Fictional Speculation



Nobel Laureate Ishiguro writes from the viewpoint of an “Artificial Friend” who helps a girl through her childhood, and is then returned to the *Friend Store* for resale. Klara is now an out of date model, but is still hopeful somebody new will want to buy her. Klara’s ideas of her place in the world are simultaneously naïve and perceptive.

The Rise of Social Robotics...



...and Most Recently



[meetings] Call for Participation - ICRA 2021 workshop on 'Sentimental Machines'

HRI-Announcement <HRI-ANNOUNCEMENT@LISTSERV.AC.M.ORG>
on behalf of

Damith.Herath <Damith.Herath@CANBERRA.EDU.AU>

Wed 3/17/2021 2:33 AM

To: HRI-ANNOUNCEMENT@LISTSERV.AC.M.ORG <HRI-ANNOUNCEMENT@LISTSERV.AC.M.ORG>

AGI Challenges: Anti-Democracy Effects

Commentators worry that AI will:

- Enable Corrosive Actions
 - ❑ Erosion of confidence in government
 - ❑ Interference in elections
 - ❑ Manipulation of economic systems
 - ❑ Erosion of informational privacy
 - ❑ Destruction of core values
- Promote Dictatorships
 - ❑ Fear of societal irrelevance
 - ❑ Fear of losing political power
 - ❑ Creation of a “useless class”
 - ❑ Rise of “digital dictatorships”

It seems the common element is *lack of agency*, perceived or feared, due to future displacement by AI entities in a variety of societal functions.



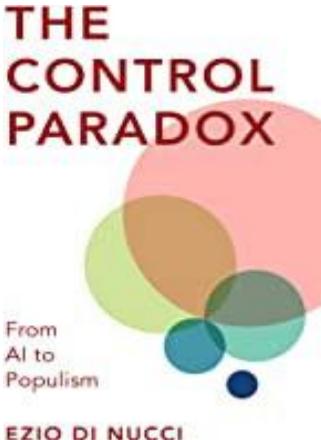
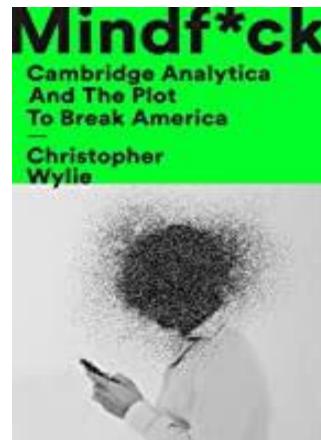
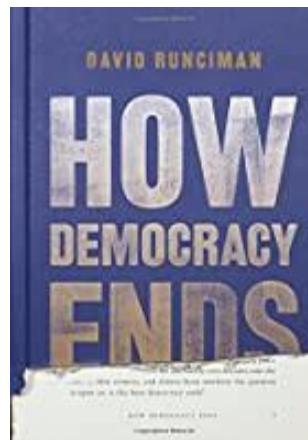
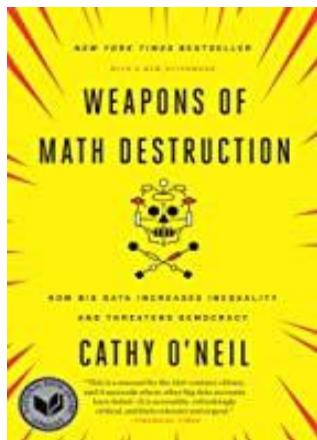
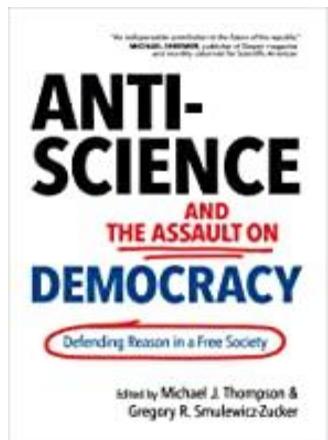
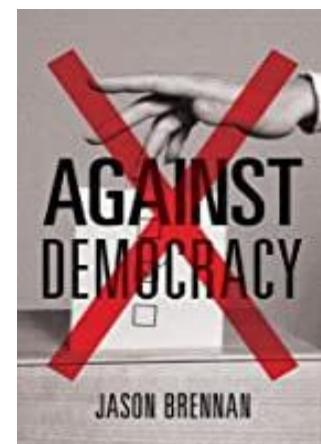
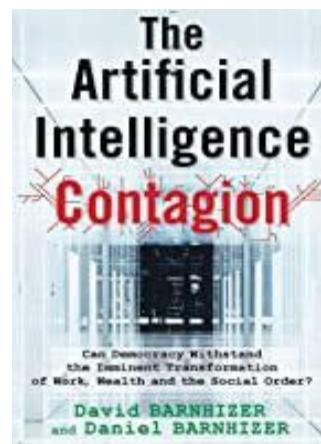
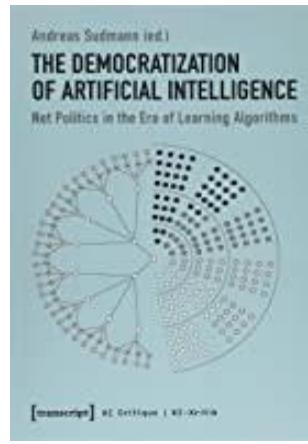
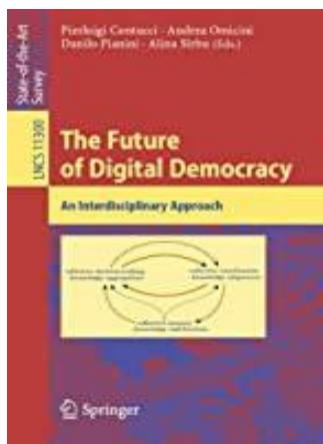
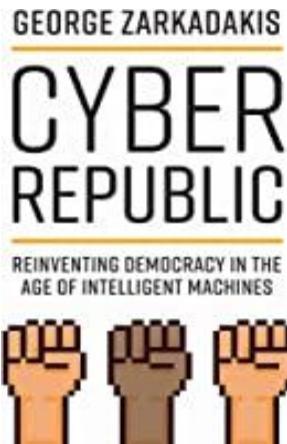
AGI Challenges: Anti-Democracy Effects



Yuval Noah Harari, Why Technology Favors Tyranny, The Atlantic Oct 2018

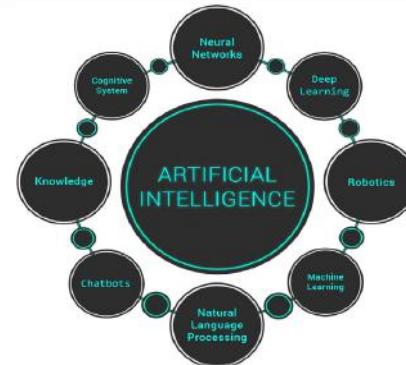
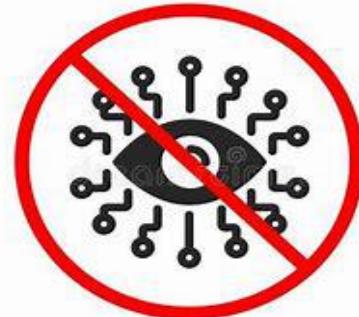
The conflict between democracy and dictatorship is actually a conflict between two different data-processing systems. AI may swing the advantage toward the latter.

Just a Few of the Many Books on the Topic



What's the Answer?

- **Ban Artificial Intelligence**
- **Use Artificial Intelligence Rationally and Ethically**



Most scientists and engineers (and I) would vote for the second approach.

Institutional Themes for Ethical AI...

Berkman Klein Center Summary

- Privacy
- Accountability
- Safety & security
- Transparency & explainability
- Fairness & non-discrimination
- Human control of technology
- Professional responsibility
- Promotion of human values

IEEE Ethically Aligned Design

- Data agency
- Accountability
- Awareness of misuse
- Transparency
- Well being
- Effectiveness
- Competence
- Human rights

... Mirrored by Industry Giants...



1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold scientific excellence standards
7. Be made available for use cases that accord with these principles



Microsoft

1. Inclusiveness
2. Fairness
3. Reliability and Safety
4. Accountability
5. Privacy and Security
6. Transparency

...and a R&D Leader

The English AI company DeepMind* has adopted these ethical principles for its future research and development of AI technology:

- *Social Benefit*

Purpose is to serve the global social and environmental good

- *Rigorous and Evidence-Based*

Research conforms to the highest academic standards

- *Transparent and Open*

Studies and trials are unrestricted, uninfluenced and not pre-determined

- *Diverse and Interdisciplinary*

Work involves the broadest possible range of voices

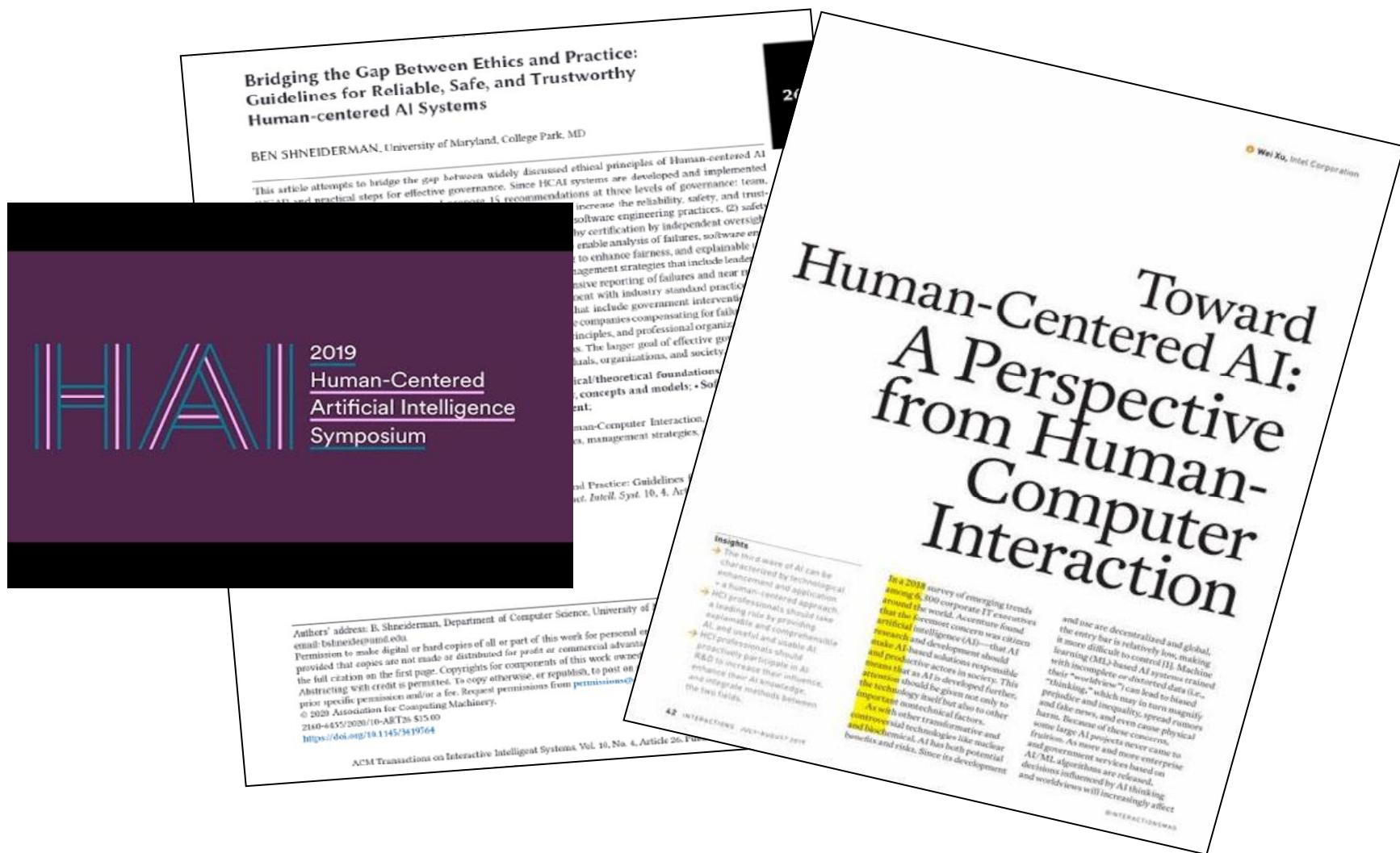
- *Collaborative and Inclusive*

Continuing dialogs among public, academics and developers about AI.

The DeepMind principles could well serve the entire AI R&D community.

*DeepMind is a subsidiary of Alphabet, Inc., the parent company of Google.

Another Ethical Approach: Human-Centered AI



Another Ethical Approach: Human-Centered AI

Traditional AI Science and Engineering:

- Emulate human behavior
- Replace human performance

Human Centered Artificial Intelligence:

- Amplify, augment and enhance human performance
- Support human self-efficacy
- Encourage creativity
- Facilitate social participation
- Be reliable, safe and trustworthy

“HCAI puts the human users at the center of design thinking, emphasizing human experience (*UX*) design.”

*Ben Shneiderman, Bridging the Gap Between Ethics and Practice
ACM Transaction on Interactive Intelligent Systems, Oct 2020*

Human-Centered AI Development

Governance Structures for Human-Centered AI



*Ben Shneiderman, “Bridging the Gap Between Ethics and Practice”
ACM Transaction on Interactive Intelligent Systems, Oct 2020*

But It's Not So Simple in Real Organizations

Google fired the lead of its Ethical Artificial Intelligence team, Meg Mitchell, escalating the turmoil surrounding its AI division after the acrimonious exit of Mitchell's former colleague Timnit Gebru.

Mitchell had become a fierce public critic of Google and its management after Gebru's exit. Gebru, one of the few prominent Black women in AI research, said she was fired in December after refusing to retract a research paper critical of a key Google technology or remove the Google authors from it. The company has said that she resigned. Mitchell was a coauthor of the paper.

Los Angeles Times, February 20, 2021

More outrage as Google fires another AI lead

BY NICO GRANT,
DINA BASS
AND JOSH EIDELSON

Google fired the lead of its Ethical Artificial Intelligence team, Meg Mitchell, escalating the turmoil surrounding its AI division after the acrimonious exit of Mitchell's former colleague Timnit Gebru.

"I'm fired," Mitchell tweeted Friday, a minute after saying she was "in too much pain to articulate much of anything useful. Firing @timnitGebru created a domino effect of trauma for me and the rest of the team, and I believe we are being increasingly punished for that trauma."

Mitchell's firing highlighted that even as Google tried to move past the disarray in its AI division with an apology, a leadership change and new policies, the upheaval showed no sign of letting up.

Mitchell had become a fierce public critic of Google and its management after Gebru's exit. Gebru, one of the few prominent Black women in AI research, said she was fired in December after refusing to retract a research paper critical of a key Google technology or remove the Google authors from it. The company has said that she resigned. Mitchell was a coauthor of the paper.

Former colleagues expressed outrage over Google's handling of the matter.

The Alphabet Inc. company had accused Mitchell of downloading files from its systems and said it would review her conduct. For five weeks, Mitchell, who had led the Ethical AI team with Gebru, was locked out of all corporate systems, including her email.

"After conducting a re-

view of this manager's conduct, we confirmed that there were multiple violations of our code of conduct, as well as of our security policies, which included the exfiltration of confidential business-sensitive documents and private data of other employees," a Google spokesman said.

Mitchell's dismissal came the same day that Google's head of AI, Jeff Dean, apologized to staff for how he handled Gebru's departure and pledged that executives would be graded on diversity progress. Dean also said Google would double its human resources staff dedicated to employee retention.

Alex Hanna, a researcher on Google's Ethical AI team, wrote that there's a double standard at the internet giant, alluding to allegations of sexual misconduct against former executives.

"Google is a breeding ground to abusers, opportunity hoarders, and people only concerned with ego and prestige," Hanna wrote on Twitter. "But anyone who is willing to defend friends against discrimination, who lift up voices who need to be heard, are shown the door."

Earlier Friday, Dean in an email to staff said Google's behavior toward Gebru hurt some female and Black employees and led them to question whether they belonged at the company, but he didn't apologize directly to Gebru.

Mitchell joined Google in November 2016 after a stint at Microsoft Corp.'s research lab where she worked on the company's Seeing AI project, a technology to help blind users "visualize" the world around them. At Google, she founded the Ethical AI team in 2017.

Grant, Bass and Eidelson write for Bloomberg.

Ethical Summary

- Artificial Intelligence (AI) and Machine Learning (ML) will undoubtedly continue to grow in power and importance as *advanced technologies*
- *AI and ML applications* will expand as well in both productive and counter-productive directions
- The *societal implications* of continued AI and ML expansion are still not fully understood
- The need for ethical guidance is clear, but the *full ethical implications* are also not completely understood
- We are evidently at *a point of inflection*, and
- While prediction is difficult, it is essential that scientists and engineers pay close attention to the potential consequences of their AI work, and be prepared to regulate that work if necessary!

UCLA Professor Baharan Mirzasoleiman Says:



“Let’s be more responsible!