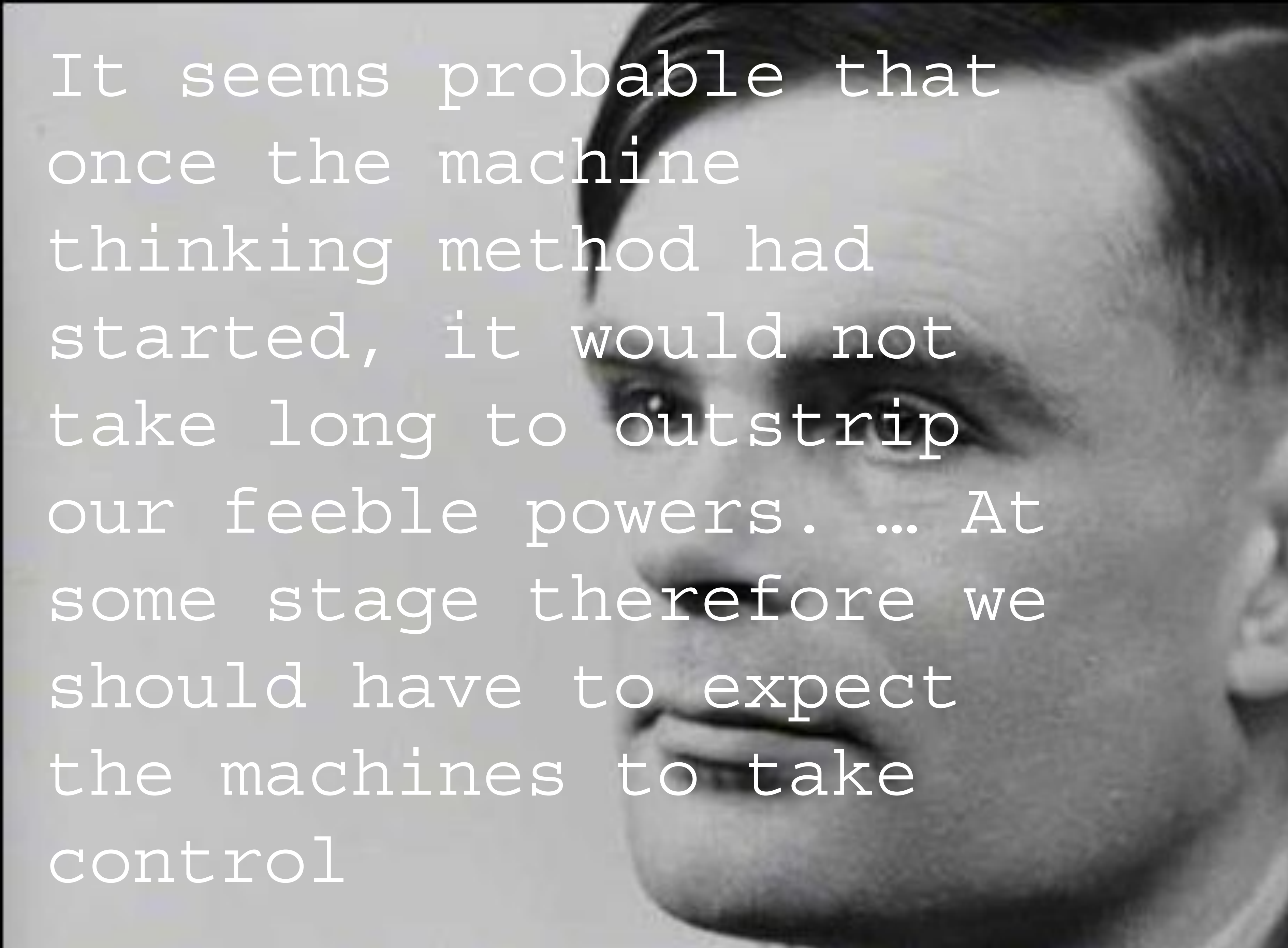
The Future of AI

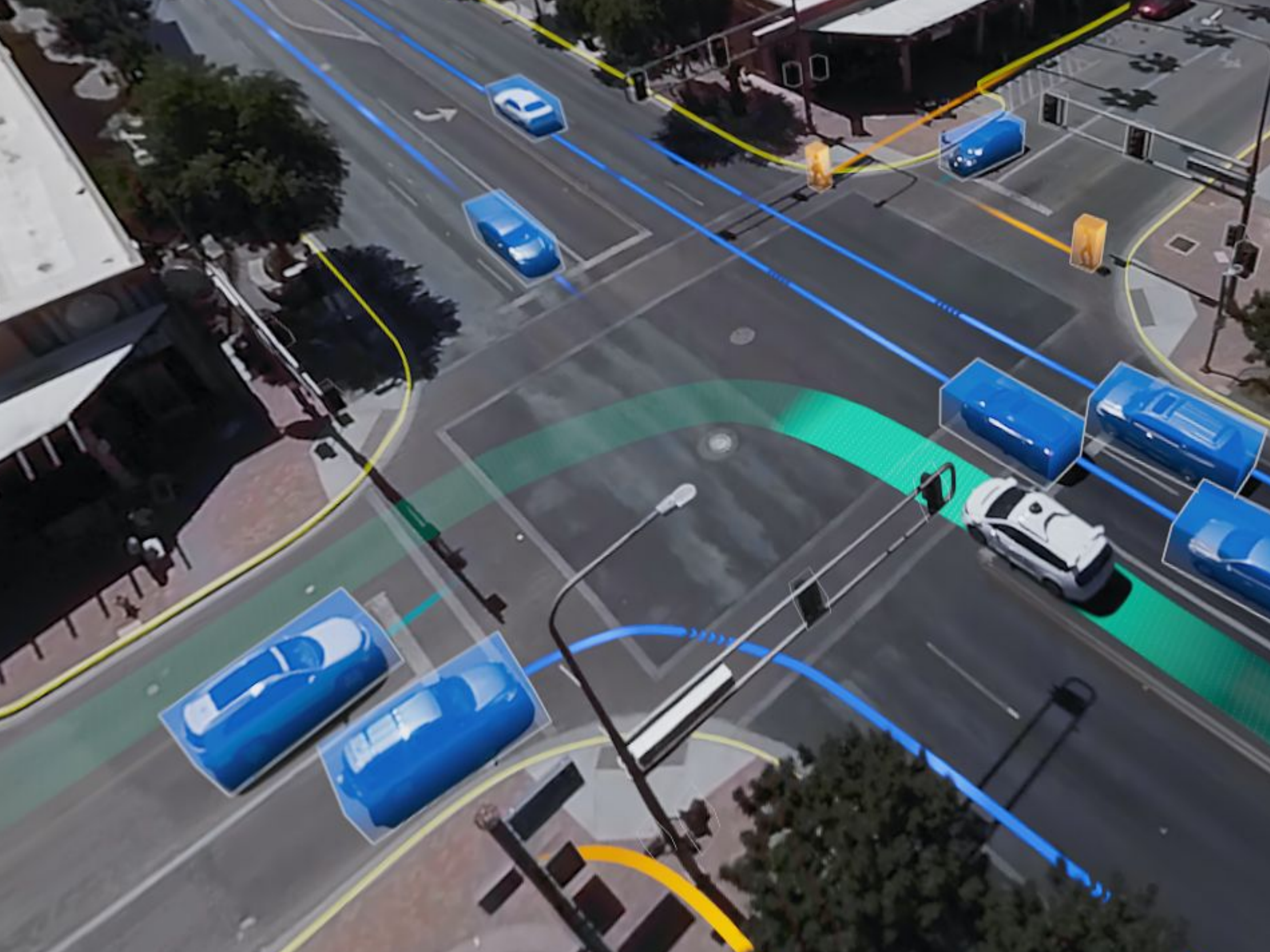
Stuart Russell
University of California, Berkeley

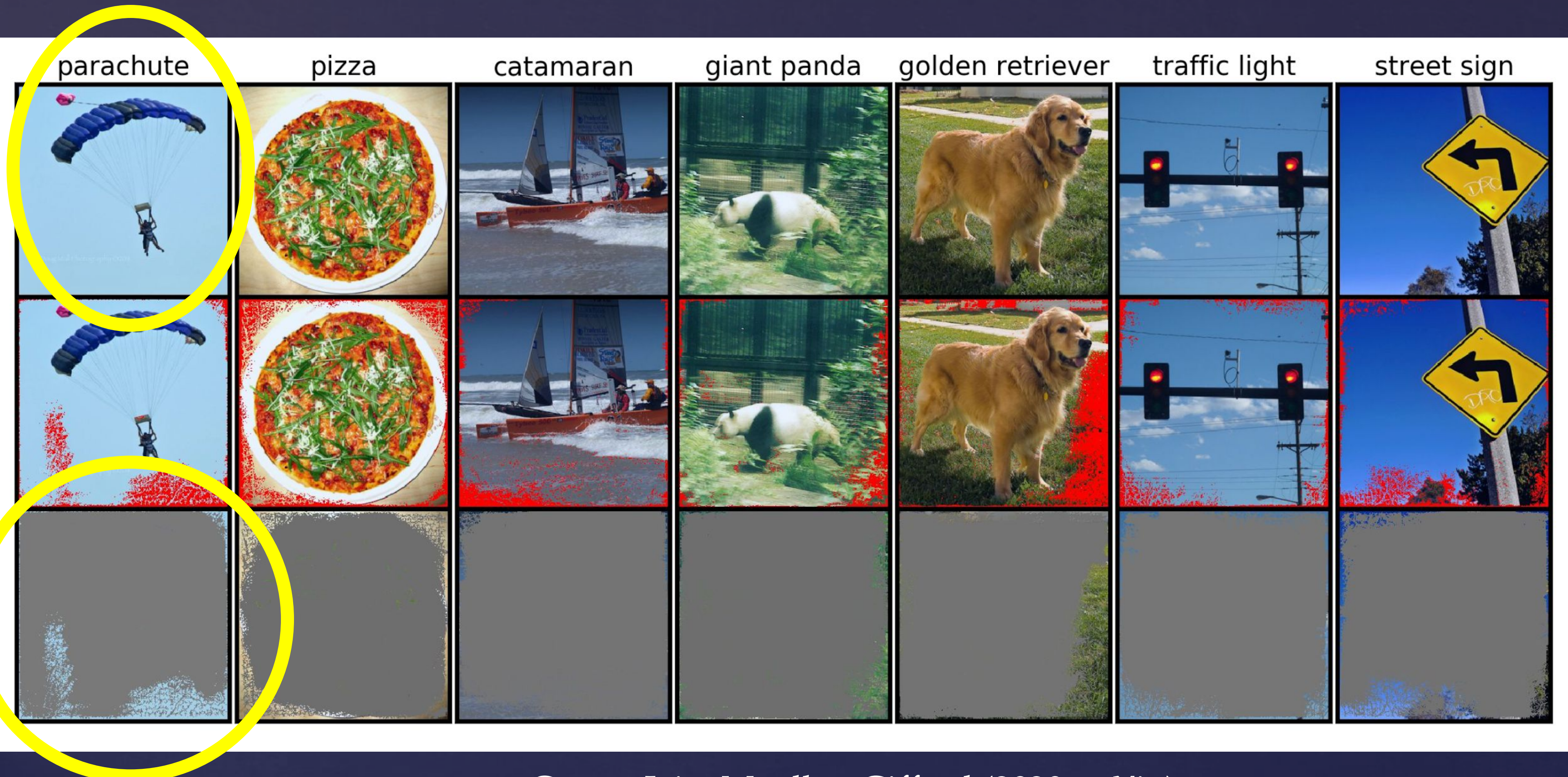
CS 188: Artificial Intelligence

The Future of AI

Instructors: Stuart Russell and Dawn Song


It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control





Carter, Jain, Mueller, Gifford (2020, arXiv)
Overinterpretation reveals image classification model pathologies

MIT Technology Review Topics

Artificial intelligence / Machine learning

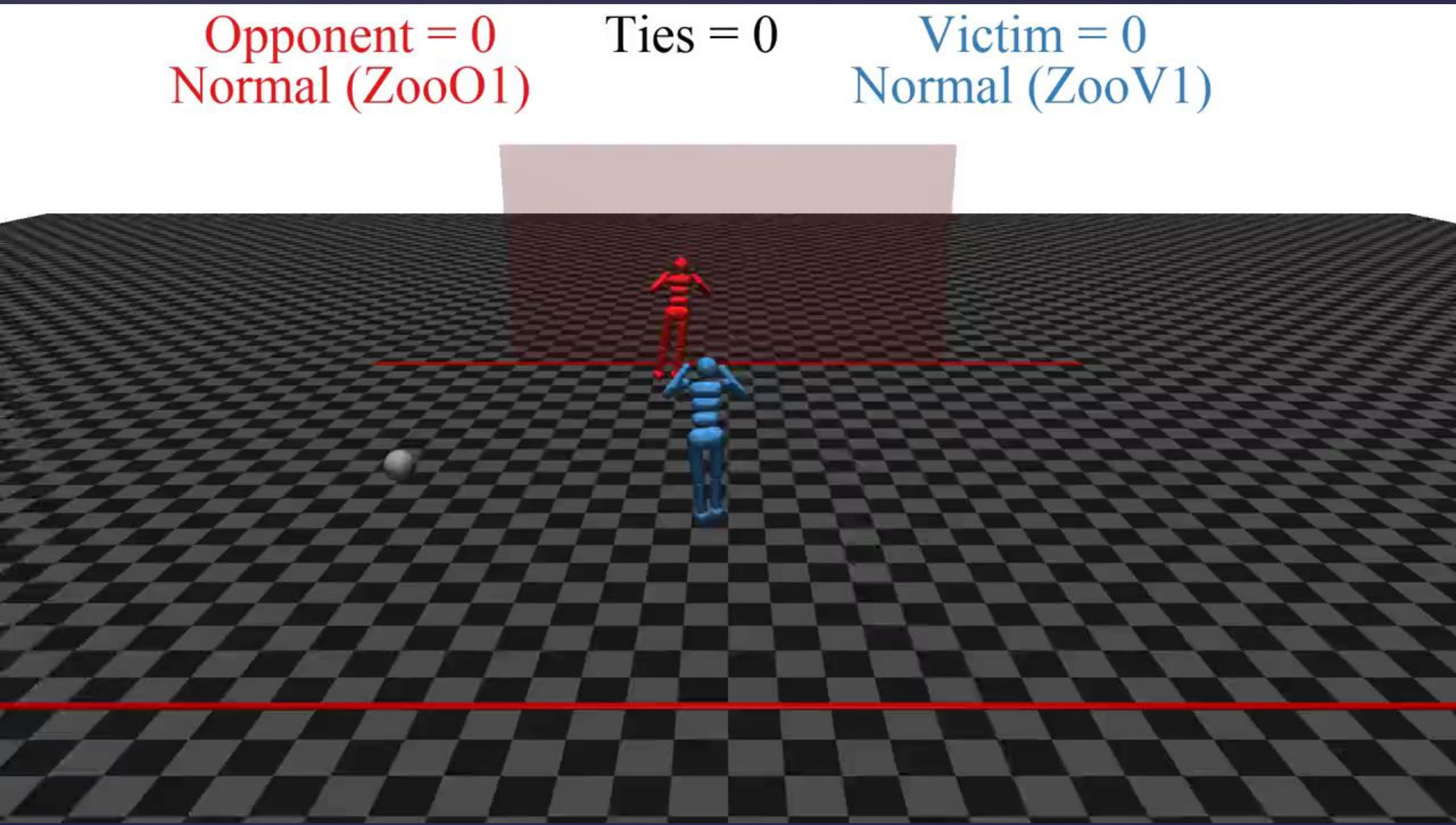
The way we train AI is fundamentally flawed

The process used to build most of the machine-learning models we use today can't tell if they will work in the real world or not—and that's a problem.

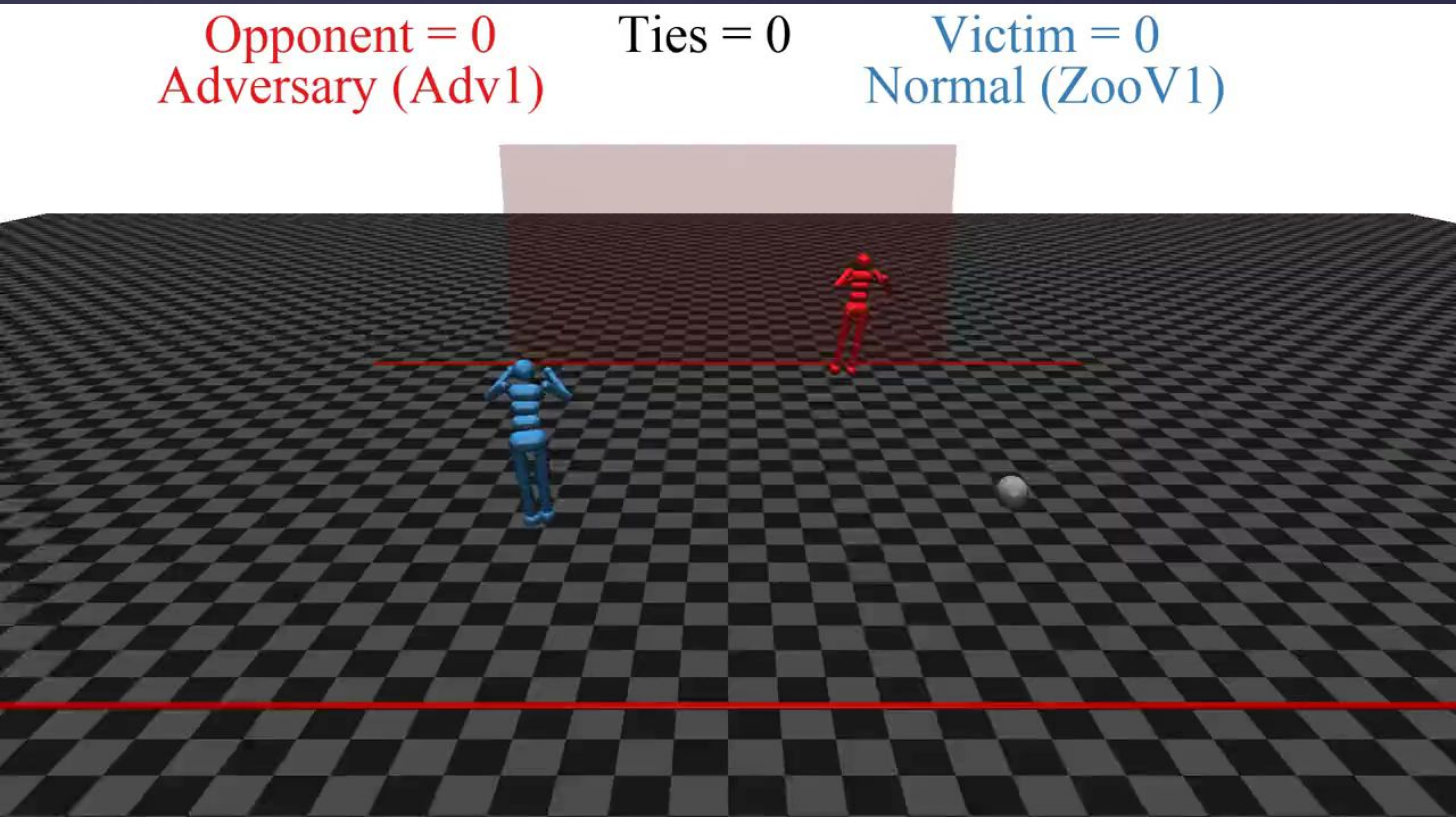
by **Will Douglas Heaven** November 18, 2020

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Alexander D'Amour*	ALEXDAMOUR@GOOGLE.COM
Katherine Heller*	KHELLER@GOOGLE.COM
Dan Moldovan*	MDAN@GOOGLE.COM
Ben Adlam	ADLAM@GOOGLE.COM
Babak Alipanahi	BABAKA@GOOGLE.COM
Alex Beutel	ALEXBEUTEL@GOOGLE.COM
Christina Chen	CHRISTINIUM@GOOGLE.COM
Jonathan Deaton	JDEATON@GOOGLE.COM
Jacob Eisenstein	JEISENSTEIN@GOOGLE.COM
Matthew D. Hoffman	MHOFFMAN@GOOGLE.COM
Farhad Hormozdiari	FHORMOZ@GOOGLE.COM
Neil Houlsby	NEILHOULSBY@GOOGLE.COM
Shaobo Hou	SHAOBOHOU@GOOGLE.COM
Ghassen Jerfel	GHASSEN@GOOGLE.COM
Alan Karthikesalingam	ALANKARTHI@GOOGLE.COM
Mario Lucic	LUCIC@GOOGLE.COM
Yian Ma	YIANMA@UCSD.EDU
Cory McLean	CYM@GOOGLE.COM
Diana Mincu	DMINCU@GOOGLE.COM
Akinori Mitani	AMITANI@GOOGLE.COM
Andrea Montanari	MONTANAR@STANFORD.EDU
Zachary Nado	ZNADO@GOOGLE.COM
Vivek Natarajan	NATVIV@GOOGLE.COM
Christopher Nielson†	CHRISTOPHER.NIELSON@VA.GOV
Thomas F. Osborne†	THOMAS.OSBORNE@VA.GOV
Rajiv Raman	DRRRN@SNMAIL.ORG
Kim Ramasamy	KIM@ARAVIND.ORG
Rory Sayres	SAYRES@GOOGLE.COM
Jessica Schrouff	SCHROUFF@GOOGLE.COM
Martin Seneviratne	MARTSEN@GOOGLE.COM
Shannon Sequeira	SHNNN@GOOGLE.COM
Harini Suresh	HSURESH@MIT.EDU
Victor Veitch	VICTORVEITCH@GOOGLE.COM
Max Vladymyrov	MXV@GOOGLE.COM
Xuezhi Wang	XUEZHIW@GOOGLE.COM
Kellie Webster	WEBSTERK@GOOGLE.COM
Steve Yadlowsky	YADLOWSKY@GOOGLE.COM
Taedong Yun	TEDYUN@GOOGLE.COM
Xiaohua Zhai	XZHAI@GOOGLE.COM
D. Sculley	DSCULLEY@GOOGLE.COM



Opponent = 0
Normal (ZooO1)
Ties = 0
Victim = 0
Normal (ZooV1)


Opponent = 0
Adversary (Adv1)
Ties = 0
Victim = 0
Normal (ZooV1)

Deep learning ad infinitum?

François Chollet (2017): "Many more applications are completely out of reach for current deep learning techniques – even given vast amounts of human-annotated data.

…

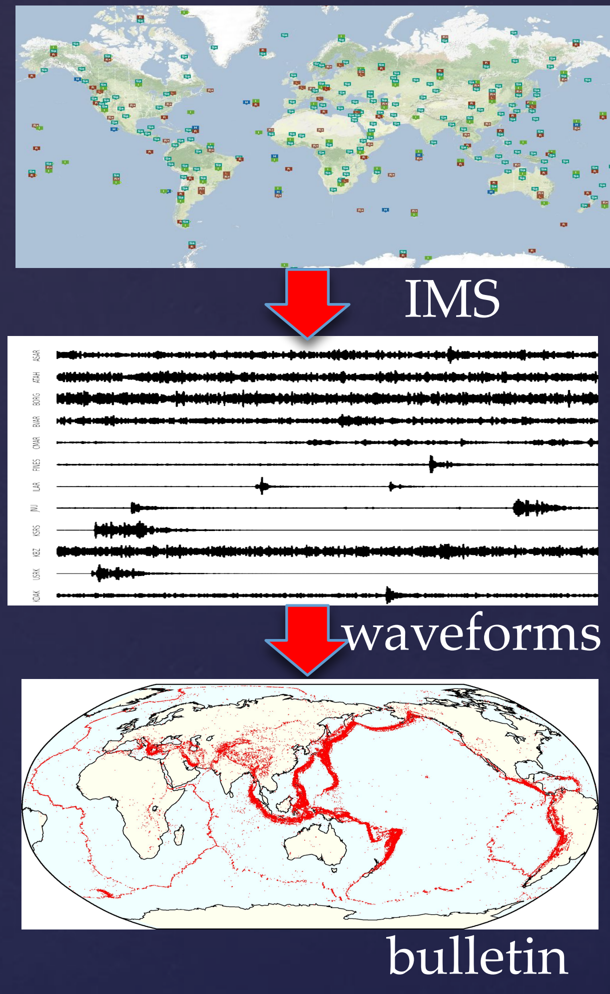
The main directions in which I see promise are models closer to general-purpose computer programs."

Probabilistic programming

Universal (Turing-equivalent) languages and algorithms for probabilistic modelling, learning, and reasoning

Global seismic monitoring for the Comprehensive Nuclear Test-Ban Treaty

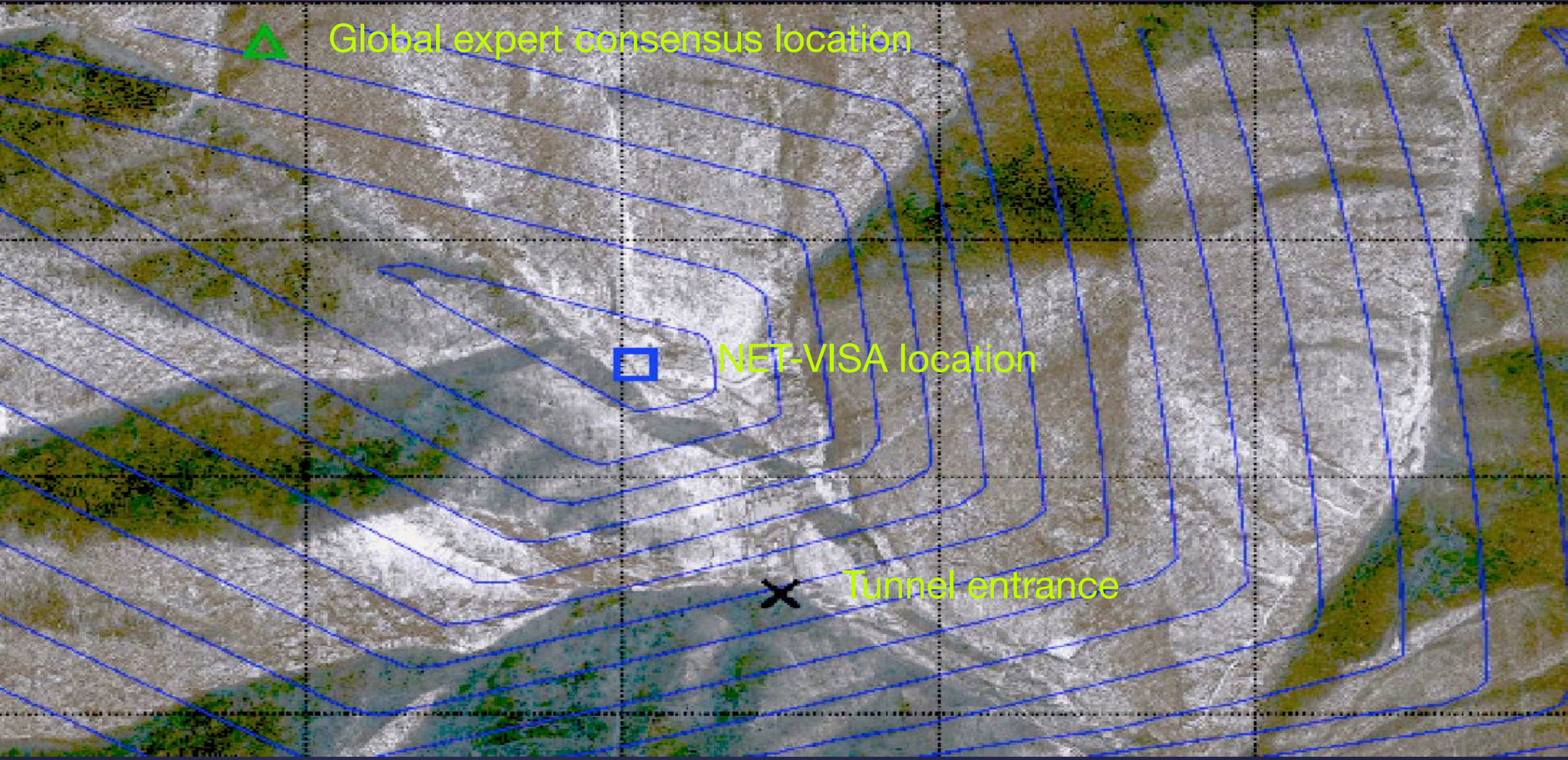
- **Evidence**: waveforms from 150 seismic stations
- **Query**: what happened?
- **Model**: geophysics of event occurrence, signal transmission, detection, noise



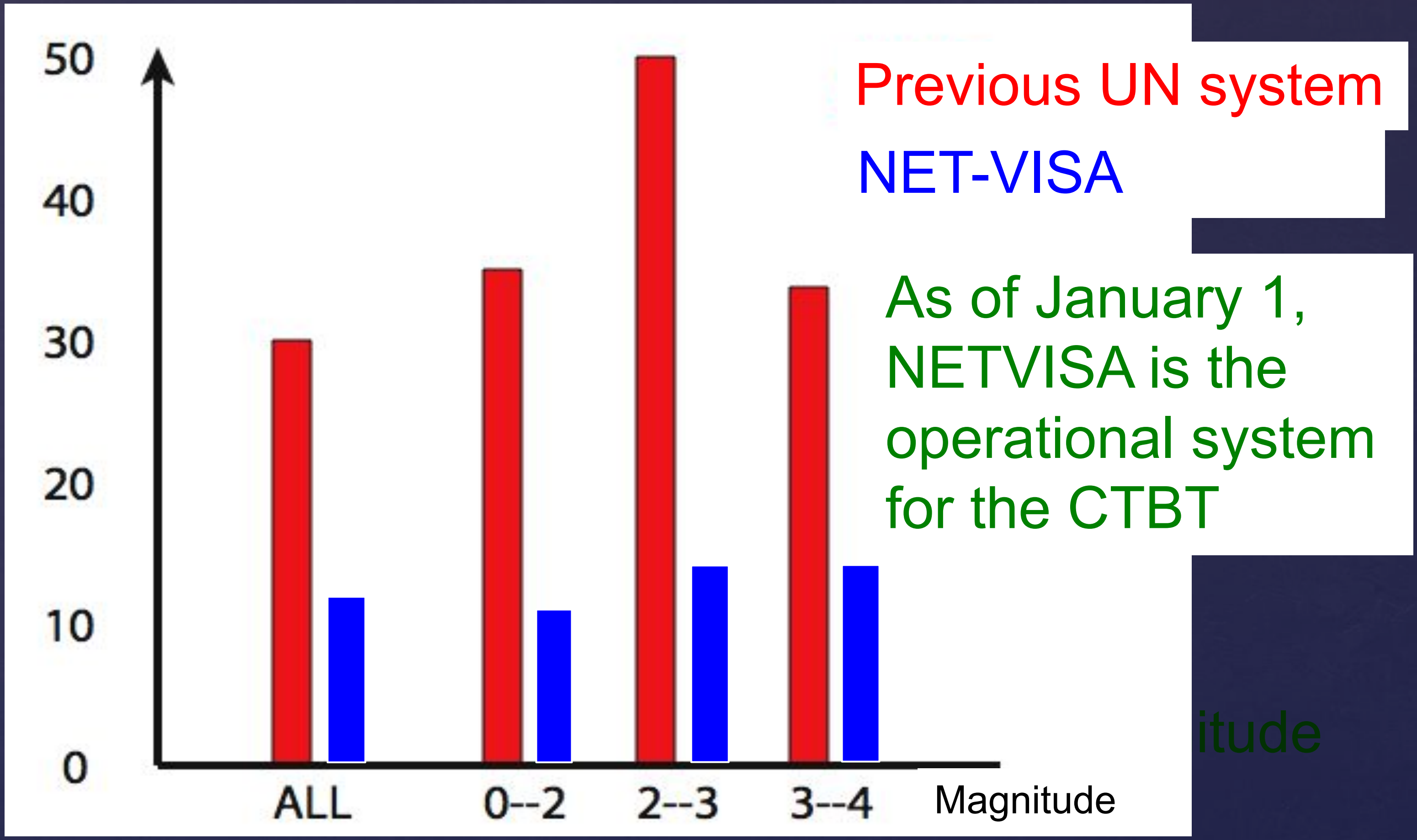
NET-VISA model

#SeismicEvents ~ Poisson[T*λ_e];
***Time(e)* ~** Uniform(0,T)
IsEarthQuake(e) ~ Bernoulli(.999);
Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();
Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;
Magnitude(e) ~ Exponential(log(10));
***IsDetected(e,p,s)* ~** Logistic[weights(s,p)](Magnitude(e), Depth(e), Distance(e,s))**;**
#Detections(site = s) ~ Poisson[T*λ_f(s)];
***#Detections(event=e, phase=p, station=s)* =** if IsDetected(e,p,s) then 1 else 0;
***OnsetTime(a,s)* ~** if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a)) +
Laplace(μ_t(s), σ_t(s))
***Amplitude(a,s)* ~** If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))
***Azimuth(a,s)* ~** If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))
***Slowness(a,s)* ~** If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))
***ObservedPhase(a,s)* ~** CategoricalPhaseModel(phase(a))

February 12, 2013 DPRK test



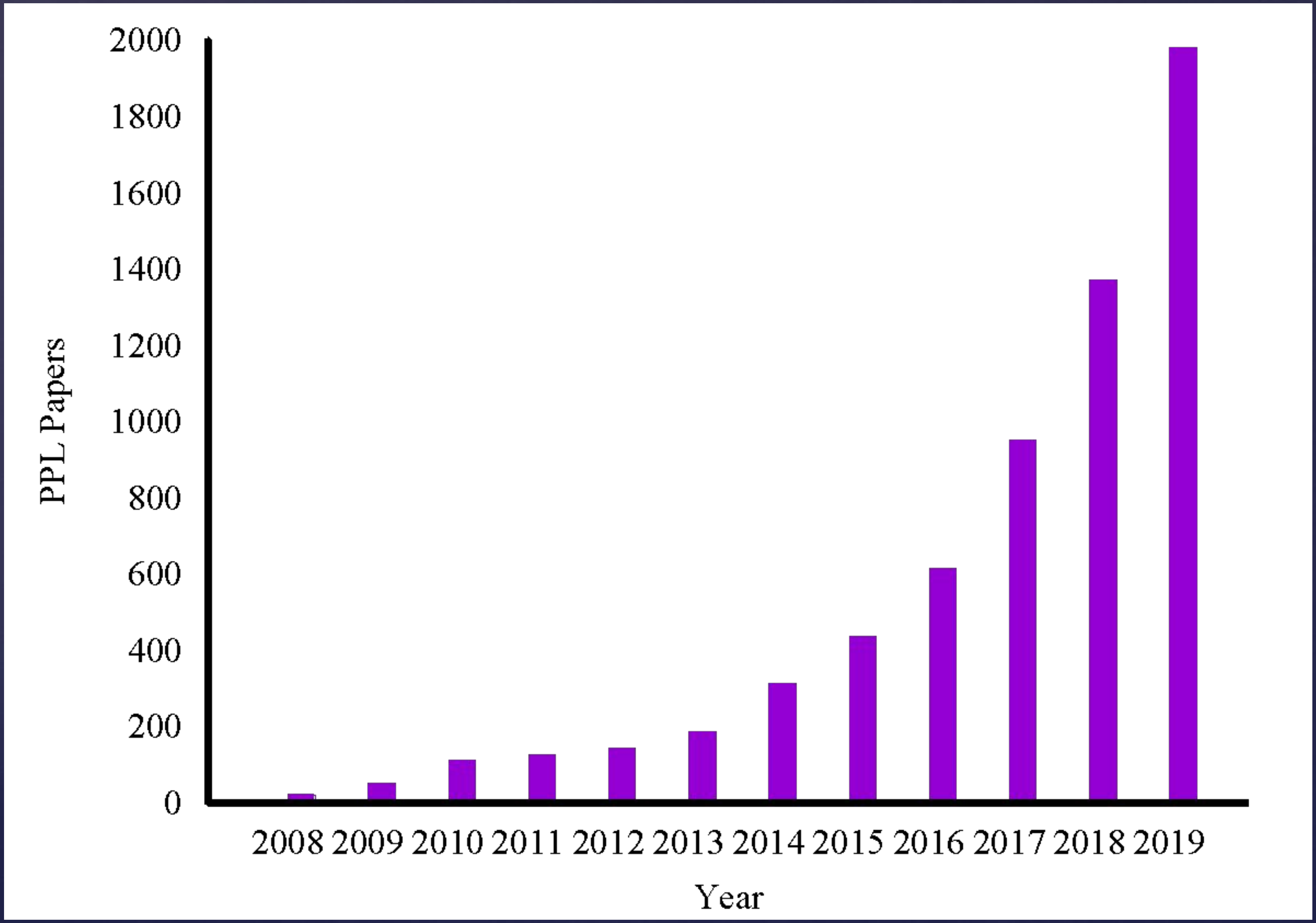
Fraction of events missed



As of January 1, NETVISA is the operational system for the CTBT

itude

Growth in PPL papers



Likely developments in the 2020s

- Robots for war, roads, warehouses, mines, fields, home
- Personal digital assistants for all aspects of life
- Commercial language systems
- Global vision system via satellite imagery

General-purpose AI

- Still missing:
  - Real understanding of language
  - Integration of learning with knowledge
  - Long-range thinking at multiple levels of abstraction
  - Cumulative discovery of concepts and theories
- **Date unpredictable**

AI systems will eventually make better decisions than humans

(Alternative: we will fail in AI)

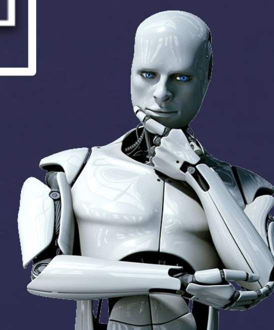
Turing's point: how do we retain control over entities more powerful than us, for ever?

Russell, Many Experts Say We Shouldn't Worry About Superintelligent AI. They're Wrong, *IEEE Spectrum*, October, 2019.

Standard model for AI

Maximize $\sum_{t=0}^{\infty} \gamma^t R(s,a,s')$

Righty-ho



Also the standard model for control theory, statistics, operations research, economics.
The objective need not be explicitly represented in the agent.
The agent can be an entire distributed system.

King Midas problem: **Cannot specify *R* correctly**
Smarter AI => worse outcome

E.g., social media

Optimizing clickthrough

= ~~learning what people want~~

= modifying people to be more predictable

How we got into this mess

- Humans are intelligent to the extent that our actions can be expected to achieve our objectives
- ~~Machines are intelligent to the extent that their actions can be expected to achieve their objectives~~
- Machines are ***beneficial*** to the extent that ***their*** actions can be expected to achieve ***our*** objectives

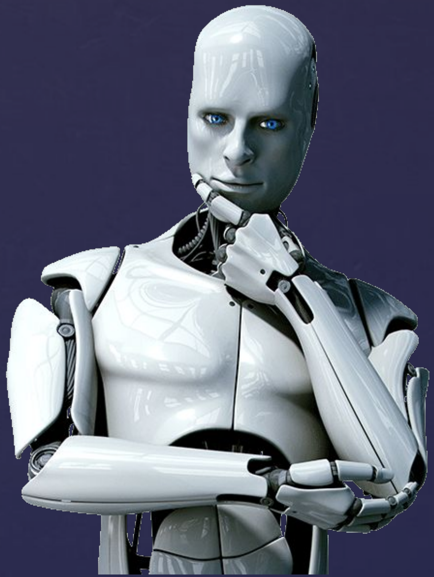
New model: Provably beneficial AI

1. Robot goal: satisfy human preferences*
2. Robot is uncertain about human preferences
3. Human behavior provides evidence* of preferences

=> ***assistance game*** with human and machine players

Smarter AI => better outcome

Basic assistance game



Preferences θ
Acts roughly according to θ

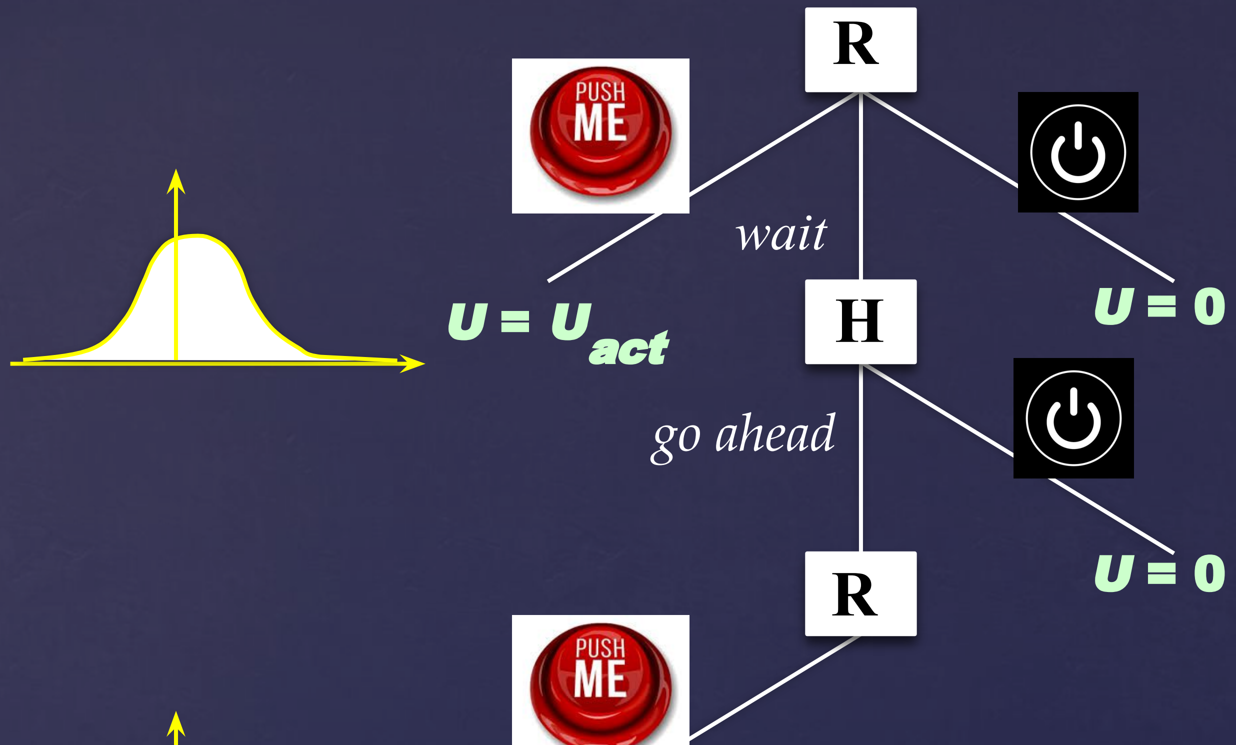
Maximize unknown human θ
Prior P(θ)

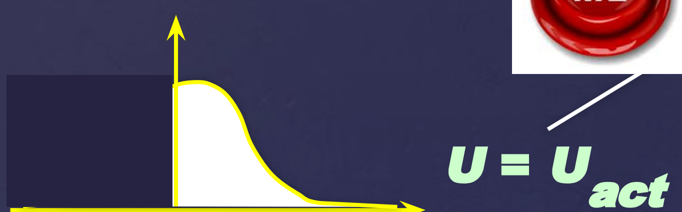
Equilibria:
Human teaches robot
Robot learns, asks questions, permission; defers to human; allows off-switch

[Hadfield-Menell et al, NeurIPS 16, IJCAI 17, NeurIPS 17]
[Milli et al 2017, IJCAI 17] [Malik et al, ICML 18]

The off-switch problem

- A robot, given an objective, has an incentive to disable its own off-switch
  - "You can't fetch the coffee if you're dead"
- A robot with uncertainty about objective won't behave this way





Theorem: *robot has a positive incentive to allow itself to be switched off*
Theorem: *robot is provably beneficial*

Rebuild AI on a New Foundation

- Remove the assumption of a perfectly known objective/goal/loss/reward
  - Combinatorial search: G(s) and c(s,a,s′)
  - Constraint satisfaction: hard and soft constraints
  - Planning: G(s) and c(s,a,s′)
  - Markov decision processes: R(s,a,s′)
  - Supervised learning: Loss(x,y,y′)
  - Reinforcement learning: R(s,a,s′)
  - (Perception)
  - Robotics: all of the above

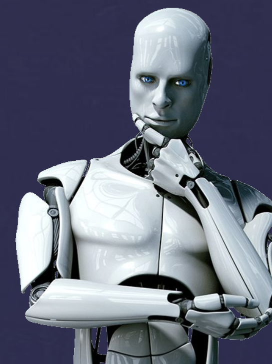
Ongoing research: "Imperfect" humans

- Computationally limited
- Hierarchically structured behavior
- Emotionally driven behavior
- Uncertainty about own preferences
- Plasticity of preferences
- Non-additive, memory-laden, retrospective/prospective preferences

Ongoing research: Many humans

- Commonalities and differences in preferences
- Aggregating individual preferences
- Interpersonal comparisons of preferences
- Potential humans (population ethics), future humans
- Mechanism design for honesty-inducing assistance
- Aggregation over individuals with different beliefs
- Altruism/indifference/sadism; pride/rivalry/envy

One robot, many humans



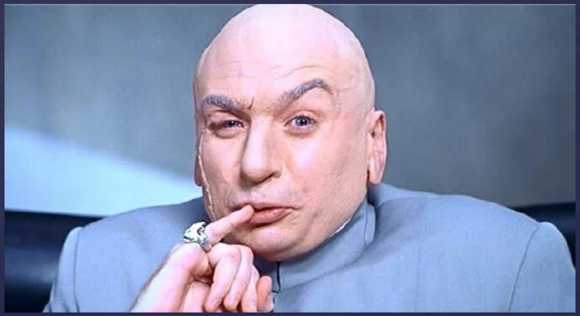
- How should a robot aggregate human preferences?
- Harsanyi: Pareto-optimal policy optimizes a linear combination, assuming a *common prior* over the future
- *In general*, Pareto-optimal policies have dynamic weights proportional to whose predictions turn out to be correct
  - Everyone prefers this policy because they think they are right

[Critch, Russell, Desai, NeurIPS 18]

Summary

- The standard model for AI leads to loss of human control over increasingly intelligent AI systems
- Provably beneficial AI is possible ***and desirable***
    - *It isn't "AI safety" or "AI Ethics," it's* ***AI***

Problems of misuse and overuse are completely unsolved



- Electronic calculators are superhuman at arithmetic. Calculators didn’t take over the world; therefore, there is no reason to worry about superhuman AI.
- Horses have superhuman strength, and we don’t worry about proving that horses are safe; so we needn’t worry about proving that AI systems are safe.
- Historically, there are zero examples of machines killing millions of humans, so, by induction, it cannot happen in the future.
- No physical quantity in the universe can be infinite, and that includes intelligence, so concerns about superintelligence are overblown.
- We don’t worry about species-ending but highly unlikely possibilities such as black holes materializing in near-Earth orbit, so why worry about superintelligent AI?

- FB: You'd have to be ***extremely*** stupid to deploy a powerful system with the wrong objective
- You mean, like clickthrough?
- FB: We stopped using clickthrough as the sole objective a couple of years ago
- Why did you stop?
- FB: Because it was the wrong objective

- Intelligence is multidimensional so “smarter than a human” is meaningless
- => “smarter than a chimpanzee” is meaningless
- => chimpanzees have nothing to fear from humans
- QED

- As machines become more intelligent they will automatically be benevolent and will behave in the best interests of ~~humans~~

~~Antarctic krill~~

~~bacteria~~

~~aliens~~