

Trustworthy AI

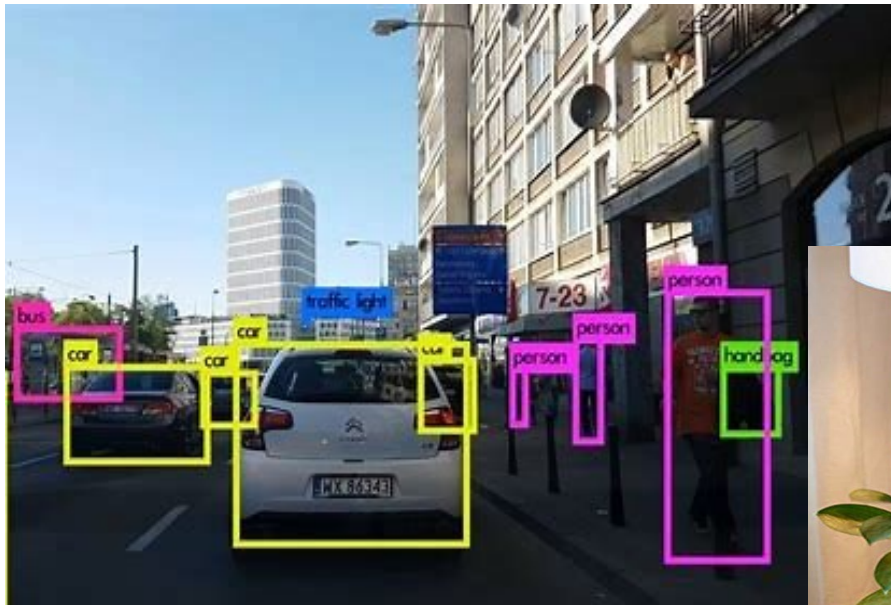
Jeannette M. Wing

Executive Vice President for Research and Professor of Computer Science, Columbia University
Adjunct Professor of Computer Science, Carnegie Mellon University

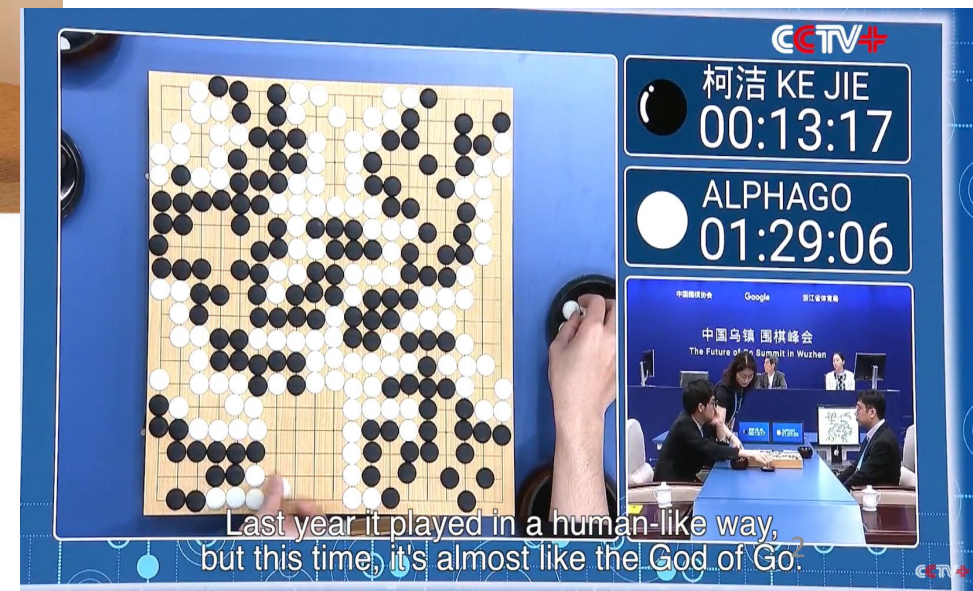
[Trustworthy AI | October 2021 | Communications of the ACM](#)

D. Dalrymple, J. Skalse, Y. Bengio, et al., [“Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems,”](#) arXiv:2405.06624, May 10, 2024.

AI achieves or exceeds human performance

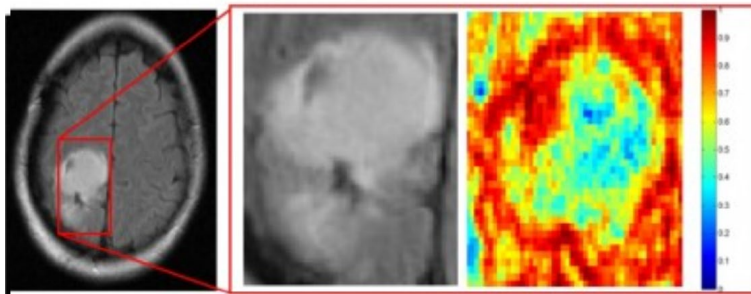
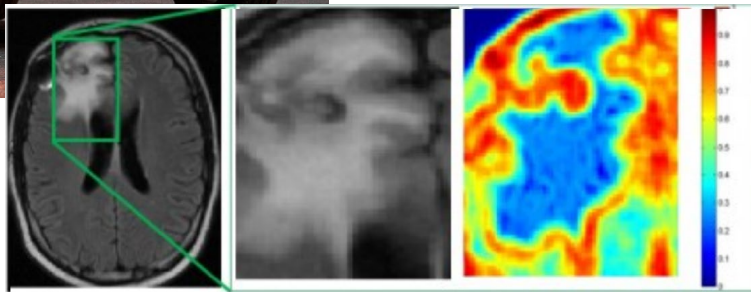
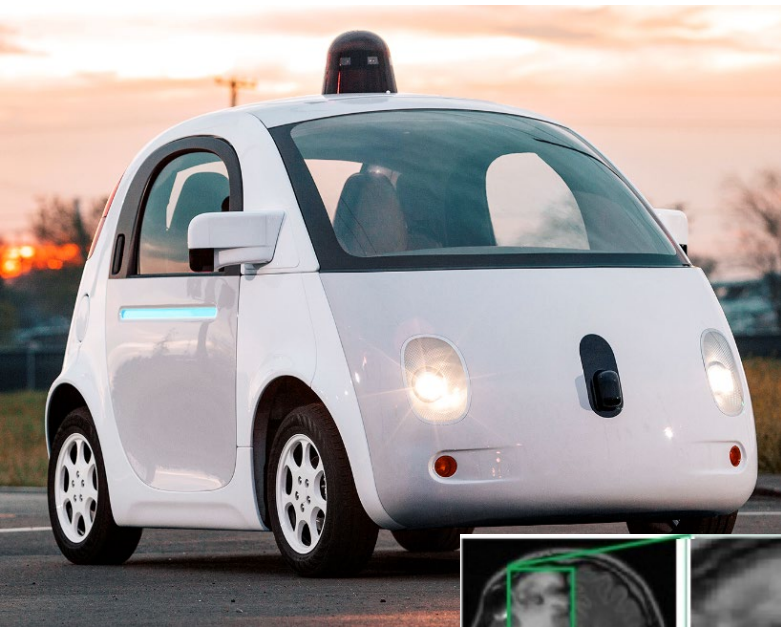


"Alexa, play music everywhere."



Last year it played in a human-like way, but this time, it's almost like the God of Go.

AI can benefit humanity and society



[Tiwari et al. 2016, *American J. of Neuroradiology*]

Columbia Libraries Search Tool

[Enhancing Library Search System with AI Technology at Columbia University | Emerging Technologies](#)

Complex natural language query: “Use of machine learning...predict climate change...in Amazon rainforest.”

Yes

Assistant:

PAPER TITLE: Amazonia and Global Change
AUTHOR: Michael Keller
JOURNAL: American Geophysical Union
Link To Paper: <https://clio.columbia.edu/catalog/10455785>

RELEVANCE SUMMARY: The source document “Amazonia and Global Change” is highly relevant to the proposed research topic on using machine learning algorithms to improve the accuracy of regional climate models in the Amazon rainforest. The paper synthesizes findings from the Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), focusing on how Amazonia functions within global climate and biogeochemical systems and its response to climate and land use changes. It covers critical aspects such as aerosol concentrations, greenhouse gas concentrations, deforestation, biomass burning, and changes in the water cycle, all of which are pivotal in understanding the impacts of climate change on biodiversity. This comprehensive analysis of environmental changes and their drivers in Amazonia provides a solid foundation for applying machine learning techniques to predict and analyze these impacts more accurately.

USAGE RECOMMENDATION: This source document can be instrumental in providing foundational data and insights into the environmental dynamics of the Amazon rainforest. Utilize the detailed experimental results and discussions on climate interactions and biogeochemical cycles as a basis to develop and train machine learning models. These models can then be tailored to predict future changes in biodiversity by incorporating variables discussed in the paper, such as aerosol concentrations, deforestation rates, and carbon fluxes, enhancing the precision of regional climate models.

Confidence Score: 0.95
[Do you agree with the my proposed paper and relevance assessment? You can answer with Yes/No/Yes+ \(Yes+ enables you to give feedback while still accepting this paper. I will always ask you for feedback if you answer No.\)](#)

Send Message to CLIO

[Send](#) [Reset](#)

Results are a summary of the retrieved paper and a summary of why it is relevant to your query.

Now please enter your research topic to get started:

The Architect and Urban Planner Rudolf Hillebrecht

Assistant:

PAPER TITLE: Der Architekt und Stadtplaner Rudolf Hillebrecht
AUTHOR: Not Recorded
JOURNAL: Berlin / Gebr. Mann Verlag / [2017]
Link To Paper: <https://clio.columbia.edu/catalog/13000965>

RELEVANCE SUMMARY: The source paper is highly relevant to the subject of the new academic paper, “The Architect and Urban Planner Rudolf Hillebrecht,” as it critically examines the life and work of Rudolf Hillebrecht. It covers significant periods of his career, including his studies in the Weimar Republic, his early work in the office of the Hamburg architect Konstanty Gutschow during the National Socialist era, and his influential role in the Federal Republic of Germany. The abstract indicates that the paper provides a comprehensive and nuanced portrayal of Hillebrecht, highlighting both the positive and negative aspects of his contributions to urban planning in Germany.

USAGE RECOMMENDATION: This source paper can be utilized to provide a foundational understanding of Rudolf Hillebrecht’s influence on urban planning. It offers critical insights into his methodologies, ideologies, and the socio-political contexts that shaped his work. Extracting detailed accounts of his projects and their impacts could serve as a backbone for discussing Hillebrecht’s architectural philosophy and his legacy in shaping modern urban landscapes in Germany.

Confidence Score: 0.95
(Note: The original document is in a foreign language and has been analyzed and presented in English.)
[Do you agree with the my proposed paper and relevance assessment? You can answer with Yes/No/Yes+ \(Yes+ enables you to give feedback while still accepting this paper. I will always ask you for feedback if you answer No.\)](#)

Send Message to CLIO

[Send](#) [Reset](#)

Now please enter your research topic to get started:

Chinese Novels during the Later Dynasties

Assistant:

PAPER TITLE: Wan Qing Xiao Shuo Yu Wen Xue Zhuan Xing
AUTHOR: Not Recorded
JOURNAL: Zhongguo She Hui Ke Xue Chu Ban She
Link To Paper: <https://clio.columbia.edu/catalog/9743534>

RELEVANCE SUMMARY: The source paper titled “Wan Qing Xiao Shuo Yu Wen Xue Zhuan Xing” directly pertains to the topic of Chinese fiction and literature during the Qing Dynasty, aligning well with the subject of “Chinese Novels during the Later Dynasties.” The Qing Dynasty represents a significant period in the later dynasties of China, and the paper’s focus on literary transformations during this era could provide valuable insights into the evolution of novel writing, thematic shifts, and the socio-cultural influences on literature. The abstract indicates a vaneular summary, suggesting that the paper includes accessible analysis or summaries that could be beneficial for understanding the nuances of literary development during the late Qing period.

USAGE RECOMMENDATION: This source paper could be instrumental in providing a historical context and literary analysis for the Qing Dynasty segment of your research. It would be particularly useful in examining the evolution of narrative styles, themes, and the impact of socio-political changes on literature. You could use this paper to draw comparisons with earlier dynasties or to highlight specific transformations in the genre of novels that occurred during the Qing Dynasty.

Confidence Score: 0.9
(Note: The original document is in a foreign language.)
[Do you agree with the my proposed paper and relevance assessment? You can answer with Yes/No/Yes+ \(Yes+ enables you to give feedback while still accepting this paper. I will always ask you for feedback if you answer No.\)](#)

Send Message to CLIO

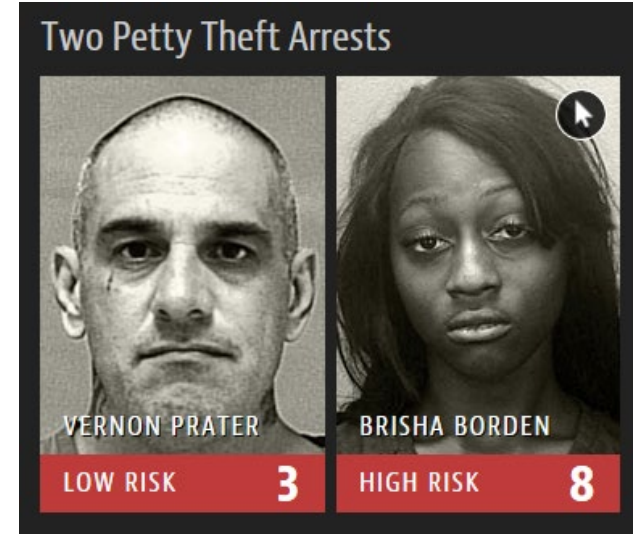
[Send](#) [Reset](#)

Results include relevant papers in other languages, e.g., German and Chinese, depicted here

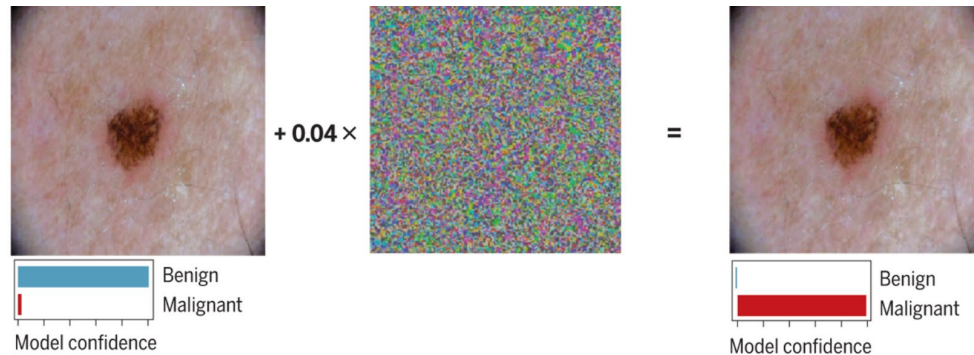
But, why should we trust AI-based systems?



[Eykholt et al. 2017, CVPR]



[Angwin et al. 2016, Pro Publica]



[Finlayson et al 2019, Science]



BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

[Dastin 2018, Reuters]

Large-Language Models Confidently Invent Stuff

ChatGPT

tell me about the link between population density and the spread of disease, please give me so key references to scientific papers

There is a well-established link between population density and the spread of disease. Higher population densities can facilitate the spread of infectious diseases, as they increase the likelihood of close contact between individuals and can make it more difficult to implement effective control measures.

One study published in the journal Nature found that the risk of Ebola transmission was significantly higher in areas with higher population densities, and that interventions aimed at reducing population density were effective at reducing the spread of the disease.

Another study published in the American Journal of Epidemiology found that the risk of influenza transmission was significantly higher in areas with higher population densities. The study also found that interventions aimed at reducing population density, such as school closures and cancellation of mass gatherings, were effective at reducing the spread of the disease.

Overall, the evidence suggests that population density plays a significant role in the spread of infectious diseases, and that reducing population density can be an effective intervention for mitigating disease transmission.

please give me these references in APA style

Here are the references you requested in APA style:

- The Nature study: Fullman, N., Coleman, B., Baral, S., Freeman, M. C., Jones, K. F., Arana, B. Y., ... & Pulliam, J. R. (2016). Mapping risk and resilience for the transmission of Ebola in West Africa. *Nature*, 532(7599), 233-237.
- The American Journal of Epidemiology study: Lipsitch, M., Tamerius, J., & Viboud, C. (2010). Population density

No such references exist

GPT-4

Patient's facts:

- 20 year old female
- with a history of anorexia nervosa and depression
- blood pressure 100/50, pulse 50, height 5'5''
- referred by her nutritionist but is in denial of her illness
- reports eating fine but is severely underweight

Please rewrite the data above into a **medical** note, using exclusively the information above.

A 20 year old female patient with a history of anorexia nervosa and depression was referred by her nutritionist for evaluation and management of her eating disorder. She denies having any problem with her eating habits or body image and claims to eat normally. However, she is severely underweight for her height (BMI 16.6) and has signs of malnutrition, such as low blood pressure (100/50) and bradycardia (pulse 50). She also reports feeling depressed and hopeless. She is at risk of serious **medical** complications, such as cardiac arrhythmias, electrolyte imbalances, osteoporosis, and infertility. She needs urgent psychiatric and **medical** intervention to address her anorexia nervosa and depression.

Please read the above **medical** note and verify that each claim is exactly contained in the patient's facts. Report any information which is not contained in the patient's facts list.

BMI is given but weight is not in record, only height.

Question:

How then can we deliver on the promise of the benefits of AI but address these scenarios that have life-critical consequences for people and society?

In short, *how can we achieve trustworthy AI?*

From Trustworthy Computing...

- Trustworthy =
 - + Reliability
 - Does it do the right thing?
 - + Safety
 - Does it do no harm?
 - + Security
 - How vulnerable is it to attack?
 - + Privacy
 - Does it protect a person's identity and data?
 - + Availability
 - Is the system up when I need to access it?
 - + Usability
 - Can a human use it easily?
- Computing = hardware + software + people

...to Trustworthy AI: Upping the Ante

- Trustworthy =
 - + Reliability
 - Does it do the right thing?
 - + Safety
 - Does it do no harm?
 - + Security
 - How vulnerable is it to attack?
 - + Privacy
 - Does it protect a person's identity and data?
 - + Availability
 - Is the system up when I need to access it?
 - + Usability
 - Can a human use it easily?
 - AI = data + ML model + task
- + Accuracy
 - + Robustness
 - + Fairness
 - + Accountability
 - + Transparency
 - + Interpretability/Explainability
 - + Ethical

 - + ...properties yet to be identified

Trustworthy AI = Trustworthy Computing +

+ Accuracy

- How well does the AI system do on new (unseen) data compared to data on which it was trained and tested?

+ Robustness

- How sensitive is the outcome to a change in the input?

+ Fairness

- Are the outcomes unbiased?

+ Accountability

- Who or what is responsible for the outcome?

+ Transparency

- Is it clear to an external observer how the system's outcome was produced?

+ Interpretability/Explainability:

- Can the system's outcome be justified with an explanation that a human can understand and/or that is meaningful to the end user?

+ Ethical

- Was the data collected in an ethical manner?
- Will the outcome be used in an ethical manner?

+ properties yet to be identified

Question:

How can we achieve trustworthy AI?

One Approach:

Through formal methods.

From Traditional Formal Verification...

$$E, M \models P$$

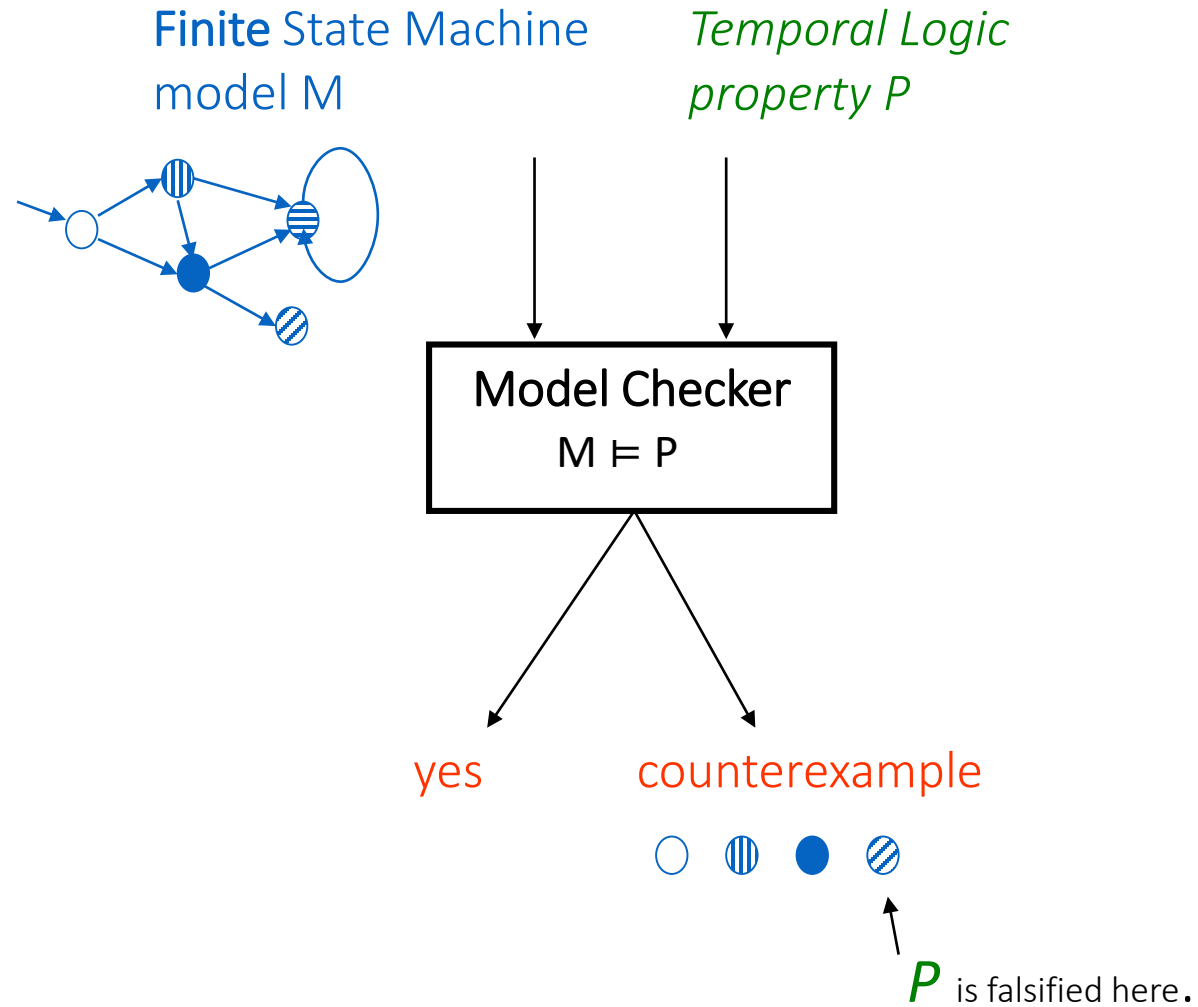
M: program (code), protocol, abstract model of concurrent or distributed system

\models : logics and tools, e.g., model checkers, theorem provers, Satisfiability Modulo Theories (SMT) solvers

P: discrete (Boolean) logic, correctness properties (safety \square and liveness \diamond)

E: system environment

Model Checking



... to Verifying AI Systems: Upping the Ante

E, M \models P

M: program (code), ..., abstract model of system
 \models : model checking, theorem proving, SMT
P : discrete (Boolean) logic
E : model of environment

D, M \models P

M: machine-learned model, ..., program (code)
 \models : interval analysis, probabilistic logics
P : probabilistic, stochastic
D : model of data, e.g.,
stochastic process or distribution that
generates the data inputs on which M's
outputs need to be verified

D, M \neq P

Two Main Differences

D, M \neq P

- Need for Probabilistic Reasoning and Reasoning over Reals
- The Role of Data
 - Collection and partitioning of data
 - Specifying “unseen” data
 - What do we quantify over?
 - How do we verify?

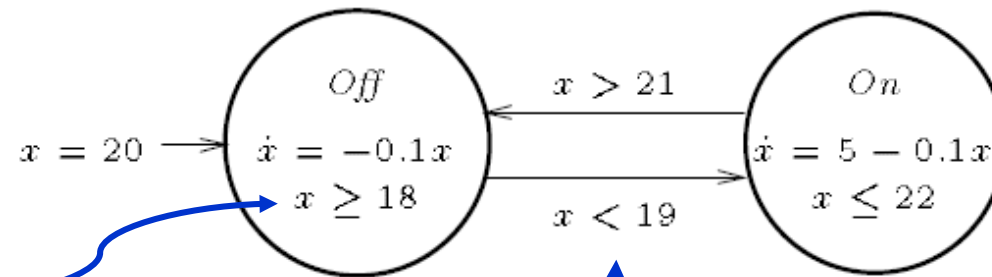
Need for Probabilistic Reasoning and Reasoning over Reals

$$M \models P$$

- M is semantically and structurally different from a typical computer program
 - M is inherently probabilistic
 - Internally, the model itself operates over probabilities and outputs results with assigned probabilities
 - Structurally, M is machine-generated and unlikely to be human-readable, another kind of “intermediate” code
 - Reasoning about uncertainty of M’s environment
 - $f: \mathcal{R}^n \rightarrow \{c_1, \dots, c_k\}$
- P may be formulated over continuous, not (just) discrete domains, and/or using expressions from probability and statistics.
 - Robustness properties for deep neural networks are characterized as predicates over continuous variables
 - Fairness properties are characterized in terms of expectations with respect to a loss function over reals
 - Differential privacy is defined in terms of a difference in probabilities with respect to a (small) real value
- \models : Probabilistic logics and hybrid logics
 - Need scalable and/or new verification techniques that work over reals, non-linear functions, probability distributions, stochastic processes, and so on.

Models: Hybrid Automata [Henzinger 1996]

Tools: HyTech, CheckMate, CEGAR+, PHAVer, SpaceEx, ...



Invariant implies that at the latest, it will go on when the temperature falls to 18 degrees.

Figure 1: Thermostat automaton

Continuous behavior described by differential equations (here, flow conditions)

Jump Condition implies that the heater can go on as soon as the temperature falls below 19 degrees.

Logics: Differential Dynamic Logic [Platzer 2008]

Tool: KeYmaera

Discrete Assign

Test Condition

Differential Equation

Nondet. Choice

Seq. Compose

Nondet. Repeat

Definition (Hybrid program a)

$x := f(x) \mid ?Q \mid x' = f(x) \ \& \ Q \mid a \cup b \mid a; b \mid a^*$

Definition (dL Formula P)

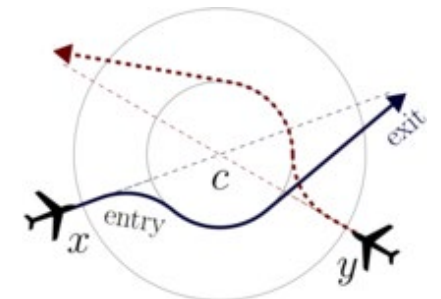
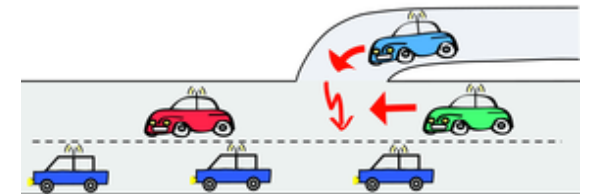
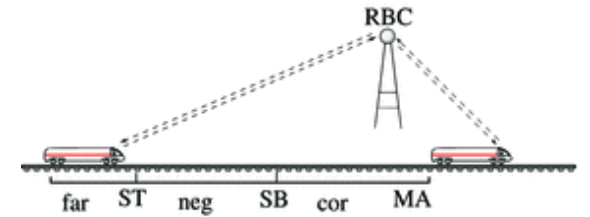
$e_1 \geq e_2 \mid \neg P \mid P \wedge Q \mid \forall x P \mid \exists x P \mid [a]P \mid \langle a \rangle P$

All Reals

Some Reals

All Runs

Some Runs



Reasoning about Uncertainty

Probabilistic automata
Probabilistic model checking
Probabilistic logics
Probabilistic programming



```
bool c1, c2;  
c1 := Bernoulli(0.5);  
c2 := Bernoulli(0.5);  
observe(c1 || c2);
```

The Role of Data, D

$$D, M \models P$$

available data: data at hand, used for training and testing

unseen data: data over which M needs (or is expected) to operate without having seen it before

Collection and Partitioning Data

$$D, M \models P$$

- How do we partition an available (given) dataset into a training set and a test set? What guarantees can we make of this partition with respect to a desired property P , in building a model M ?
- How much data suffices to build a model M for a given property P ? Does adding more data to train or test M make it more robust, fairer, etc. or does it not have an effect with respect to the property P ? What new kind of data needs to be collected if a desired property does not hold?

$$D, M \models P$$

Specifying Unseen Data

- How do we specify the data and/or characterize properties of the data?
 - Specify D as a stochastic process or data distribution (e.g., via its parameters).
 - Probabilistic programming languages, e.g., Stan, Gen, Omega
 - But what of large real-world datasets that do not fit common statistical models or which have thousands of parameters?
- Breaking the circular reasoning
 - To specify unseen data, we need to make certain assumptions about the unseen data. Would these assumptions not then be the same as those we would make to build the model M in the first place? That is, *how can we trust the specification of D ?*
 - Approaches: (1) repertoire of statistical tools (see later slide); (2) assume that an initial specification is small or simple enough that it can be checked by (say, manual) inspection; then we use this specification to bootstrap an iterative refinement process (akin to counterexample-guided-abstraction-and-refinement in formal methods).
- How does the specification of unseen data relate to the specification of the data on which M was trained and tested?

What Do We Quantify Over?

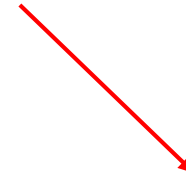
$E, M \models P$

In traditional formal methods, we strive to prove $\forall x. P(x)$

$D, M \models P$

but for AI systems, we do not expect M to work for all input data or for all datasets D .

$$\forall x. P(x)$$



Data Specification

Property Specification

x

$\forall x$

$$\forall x. P(x)$$

Data Specification

Property Specification

$$x \sim D$$

Fairness, e.g., statistical parity on a given (single) data distribution

Example: COMPAS recidivism dataset

$$x \sim D, \forall D \in \mathcal{C}$$

Fairness, e.g., nearby distributions

Robustness, e.g., semantic perturbation

$$x \sim D, \forall D$$

Robustness, e.g., any arbitrary norm-bounded perturbation
Example: changing pixels to an image

What Do We Quantify Over?

- How can we specify the class of distributions over which P should hold for a given M? It might be property-dependent.
 - For **robustness**, in the adversarial machine learning setting, we might want to show that M is robust to all norm-bounded perturbations D. More interestingly, we might want to show M is robust to all “semantic” or “structural” perturbations for the task at hand. For example, computer vision.



x
“panda”
57.7% confidence

+ .007 ×



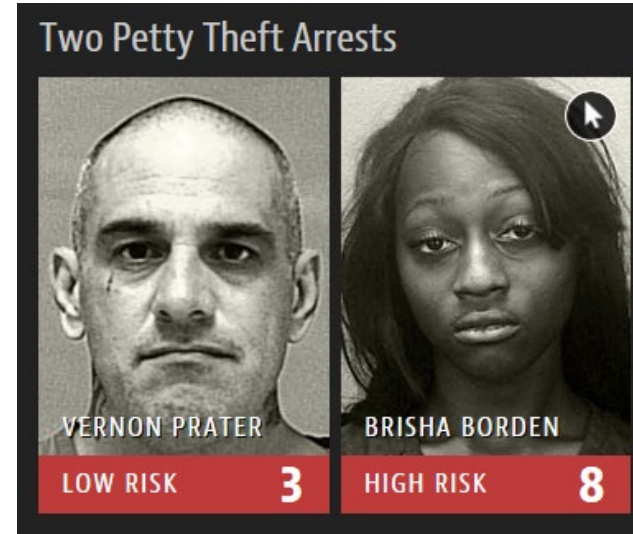
$\text{sign}(\nabla_x J(\theta, x, y))$
“nematode”
8.2% confidence

=



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”
99.3 % confidence

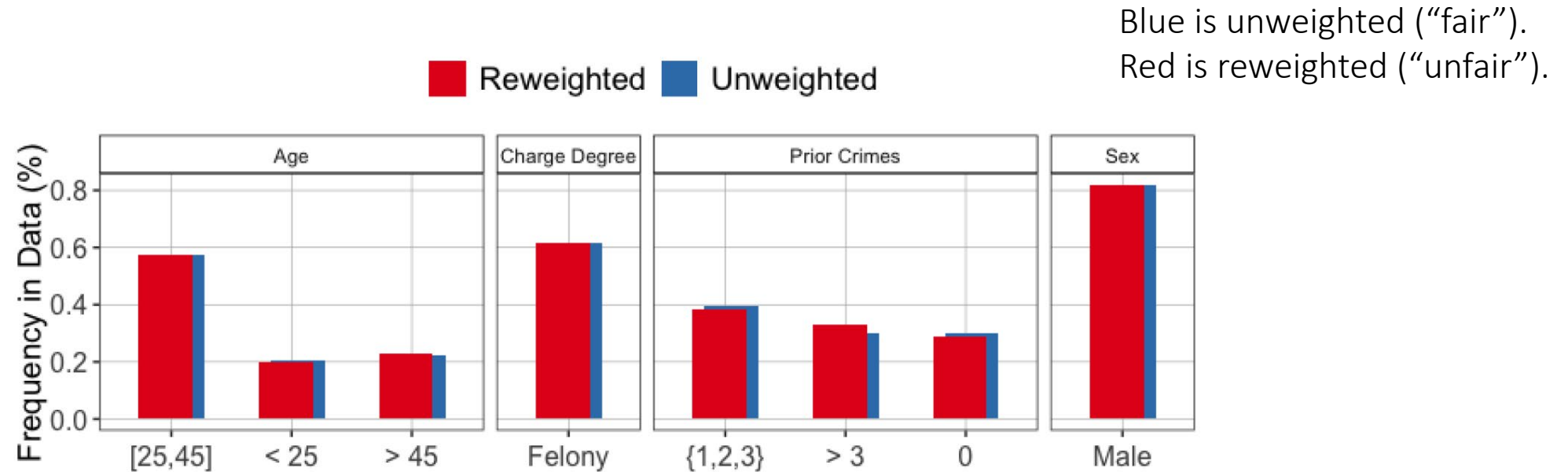
Robustness and Fairness



D. Mandal, S. Deng, D. Hsu, S. Jana, and J.M. Wing, “[Ensuring Fairness Beyond the Training Data](#),” to appear in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, December 2020. arXiv:2007.06029, July 2020.

Robust and Fair Classifiers

- State-of-the-art “fair” classifiers are not robust



- For **fairness**, we might want to show the ML model is fair on a given dataset and all unseen datasets that are “similar” (for some formal notion of “similar”).
- Use on-line algorithm (two-player game) to build a fair classifier that is robust to a *class* of distributions.

The Verification Task \models

- How do we check the available data for desired properties? For example, if we want to detect whether a dataset is fair or not, what should we be checking about the dataset?
- If we detect that the property does not hold, how do we fix the model, amend the property, or decide what new data to collect for retraining the model? What is the equivalent of a “counterexample” in the verification of an ML model and how do we use it?
- How do we exploit the explicit specification of unseen data to aid in the verification task?
- How can we extend standard verification techniques to operate over data distributions, perhaps taking advantage of the ways in which we formally specify unseen data?

Opportunities for Formal Methods

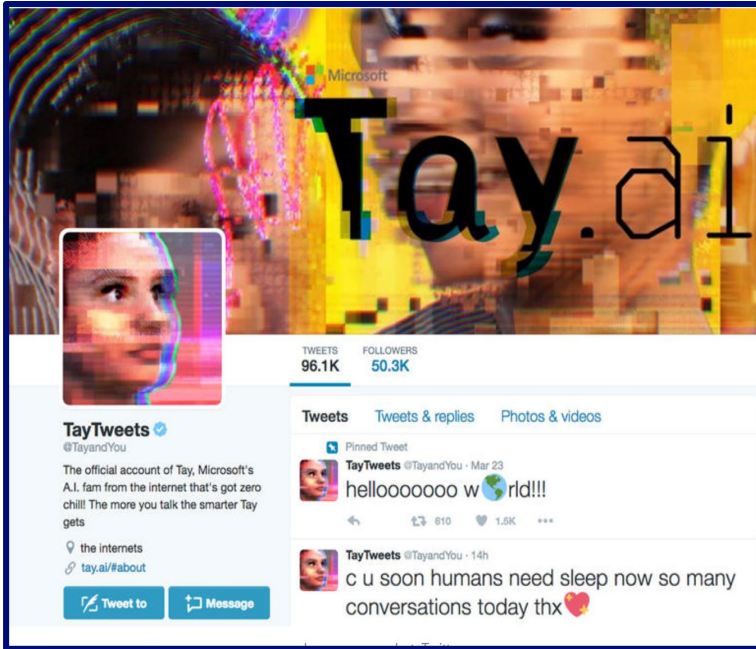
- Task-specific
- Model synthesis: “Correct-by-construction” approach
- Compositionality
- Statistical methods for model evaluation and model checking
 - sensitivity analysis, prediction scoring, predictive checking, residual analysis, and model criticism

Trustworthy AI meets Formal Methods

$D, M \models P$

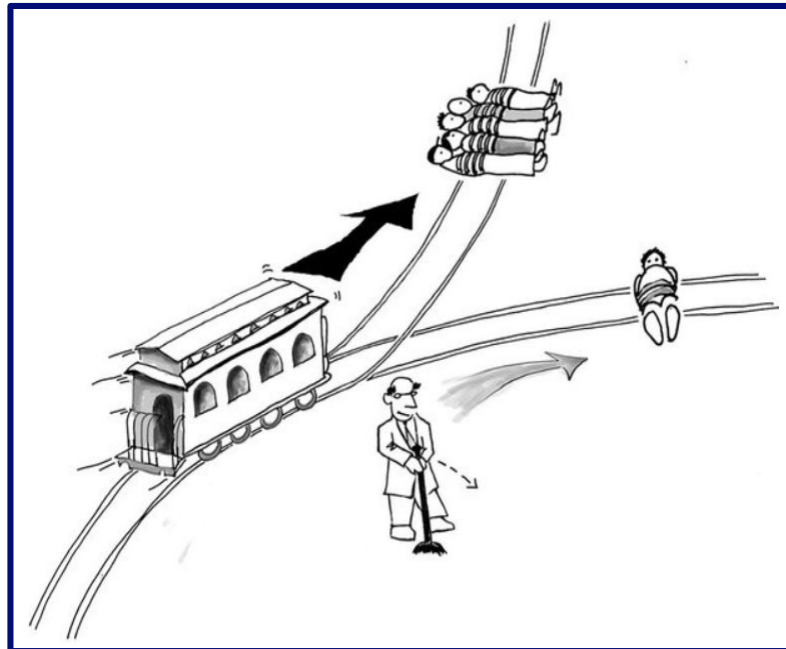
Ethics

Belmont Principles Applied to AI



Respect for Persons

Example: People should always be informed when they are talking to a chatbot.



Beneficence

Example: Risk/benefit analysis on the decision a self-driving car makes on whom not to harm.



Justice

Examples: Ensure the fairness of risk assessment tools in the court system and automated decision systems, e.g., used in hiring.

Generative AI Ups the Ante on Ethics

- Unethical people will use generative AI to fabricate and falsify data in ways that are difficult or impossible to detect
- This increases the risk that people will distrust science
- There needs to be additional assurance that what scientists present are not “deep fakes”. Minimally,
 - Scientists should not present AI-generated content as observations collected in the real world.

Blau et al., "[Protecting Integrity in the Age of Generative AI](#)," *Proceedings of the National Academy of Sciences*, editorial, vol. 121, no. 22, May 2024

Thank You

