Managing the continuity of Human biases in AI SE applications

September 28, 2023

WALTER SCOTT, JR.
COLLEGE OF ENGINEERING
COLORADO STATE UNIVERSITY

Dr. Steven Conrad
Heidi Perry

Trustworthy AI in Systems Engineering

Heidi C. Perry

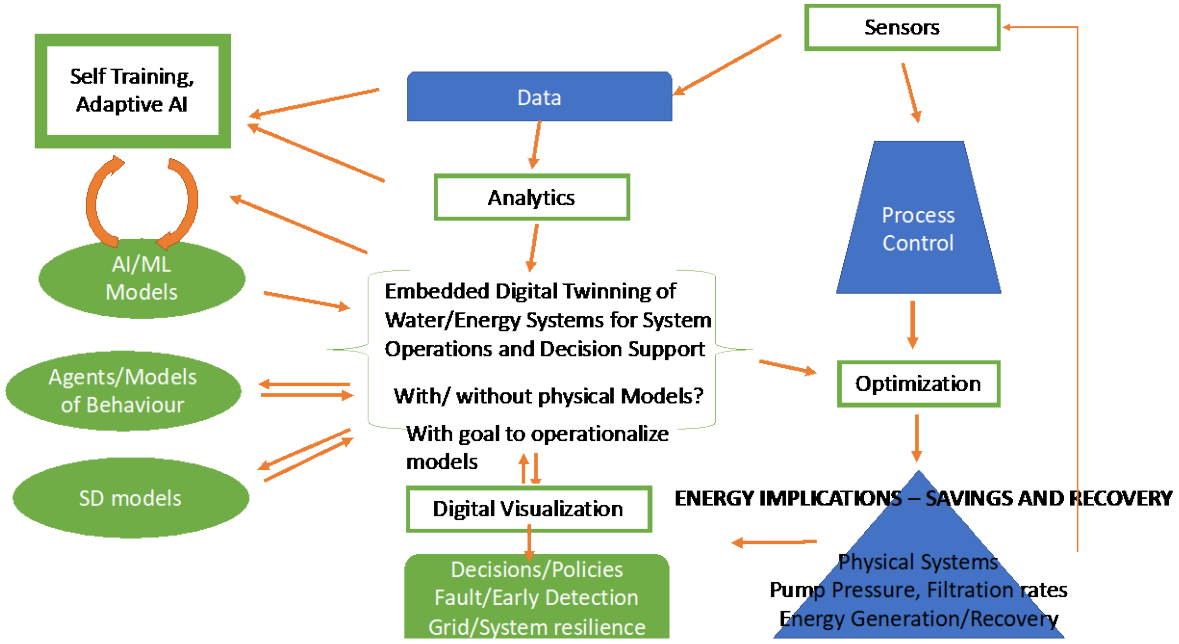
28 September 2023

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

Human Coupled Systems Modelling Lab - integrated test bed for Urban Environments

ENERGY INSTITUTE
COLORADO STATE UNIVERSITY

- Coupling social sciences and engineering to investigate and model the dynamics of human-water-energy-carbon-food systems to inform system optimization and decision making
- Ongoing AI research
  - Water and energy urban system performance
  - Applications of Systems Thinking and Dynamics/Agent modelling
  - Human behavioural responses
    - Choices
    - AI Augmented Decision Making
    - Decision visualization
  - Visual cognition of smart displays - Adaptive human machine interfaces
  - City Sustainability and Resilience
    - Urban metabolism and resource behaviours
    - Urban water/energy design challenge

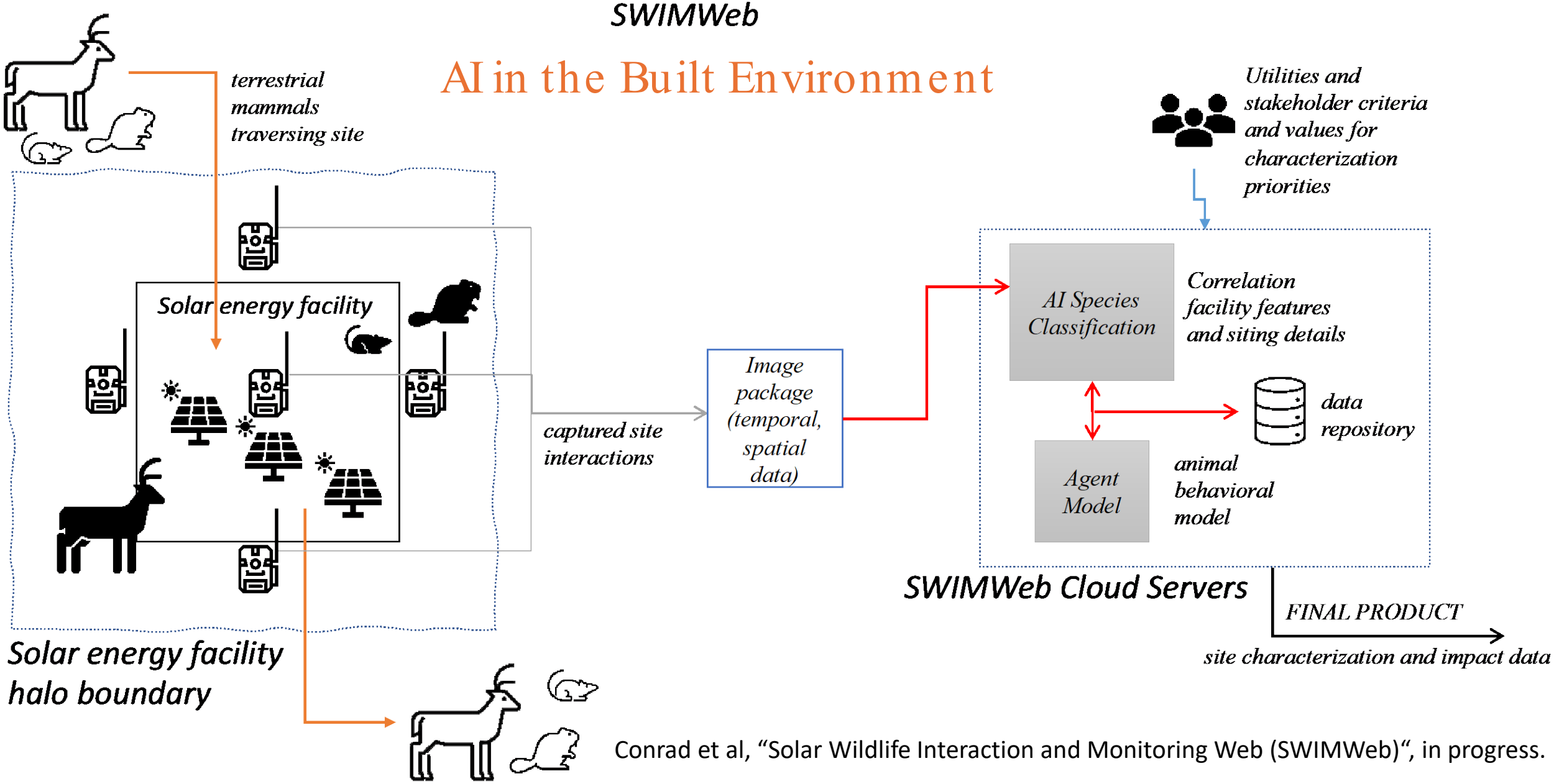


Overview of the modelling efforts in the Human Coupled Systems Modelling Lab

WATER ENERGY URBAN DESIGN CHALLENGE

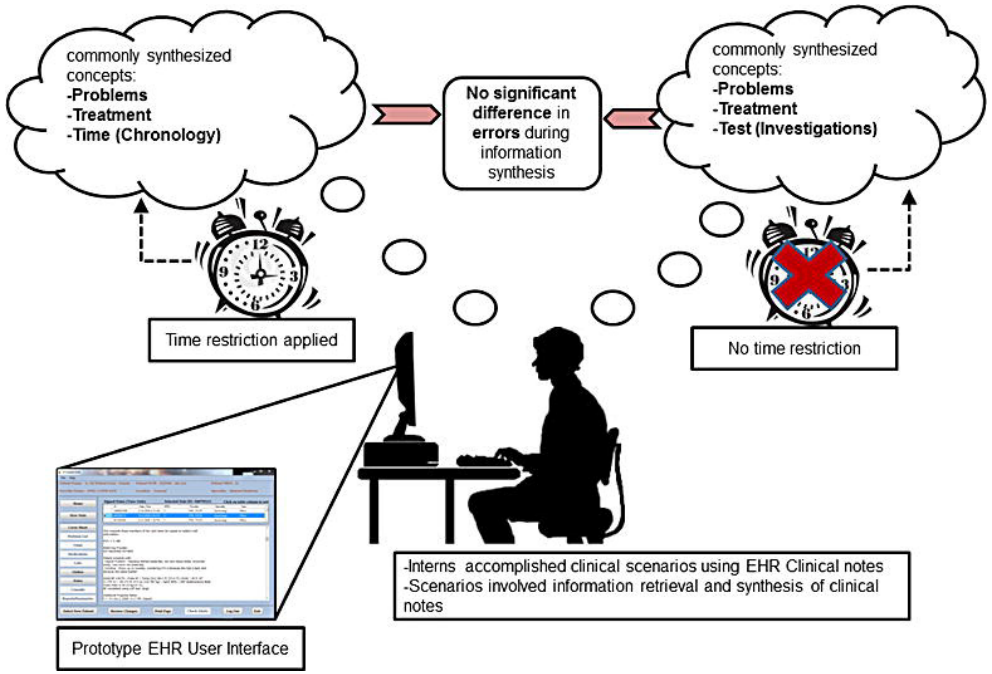
SWIMWeb

AI in the Built Environment

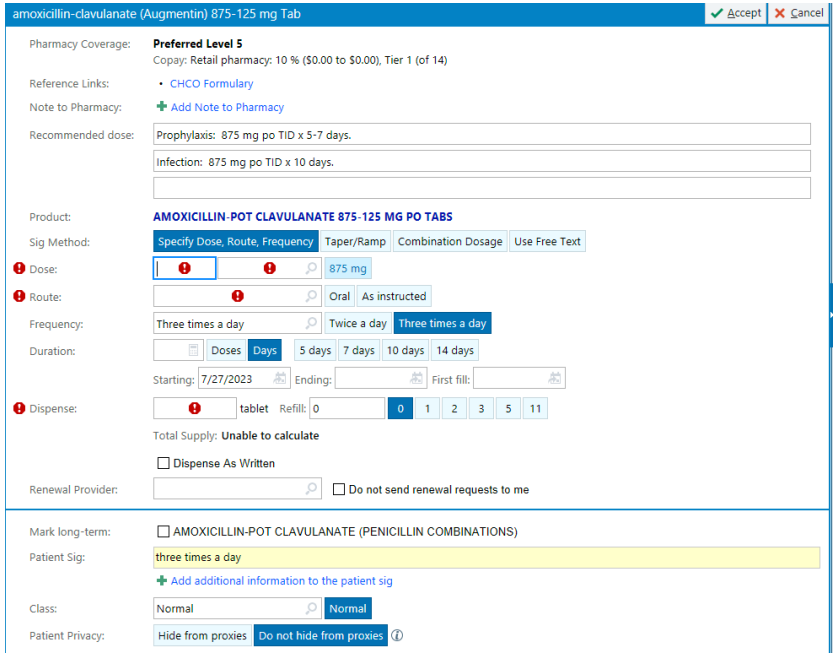


Conrad et al, "Solar Wildlife Interaction and Monitoring Web (SWIMWeb)", in progress.

Designing AI for time constrained decisions (alleviate cognitive burden)



Farri, O., Monsen, K. A., Pakhomov, S. V., Pieczkiewicz, D. S., Speedie, S. M., & Melton, G. B. (December 01, 2013). Effects of time constraints on clinician-computer interaction: A study on information synthesis from EHR clinical notes. *Journal of Biomedical Informatics, 46,* 6, 1136-1144.

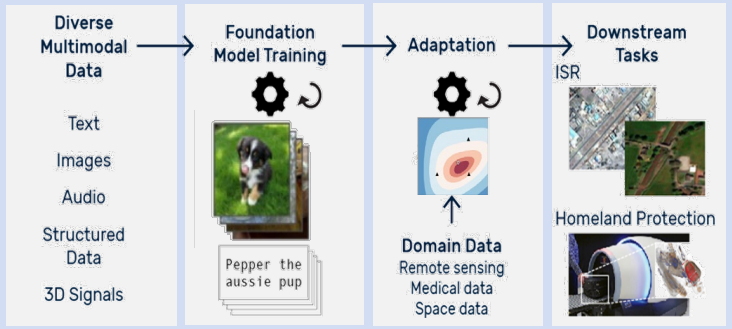


Use of Artificial Intelligence to Facilitate EHR Documentation in Emergency Departments

Conrad, Ozkaynak, Simske, Project in progress

MIT Lincoln Laboratory – A Sampling of AI Research

Foundation Models for DoD Applications

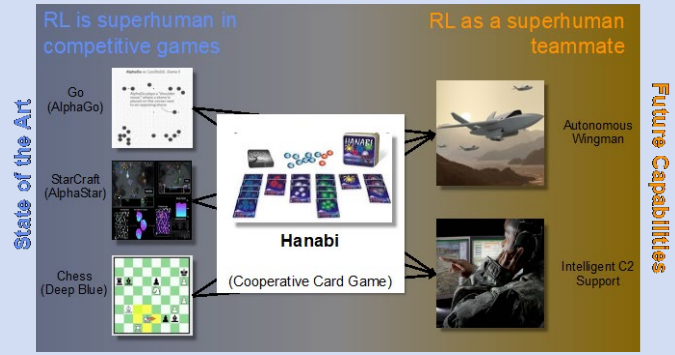


Design foundation models to rapidly adapt to national security requirements and data sets

Trustworthy AI

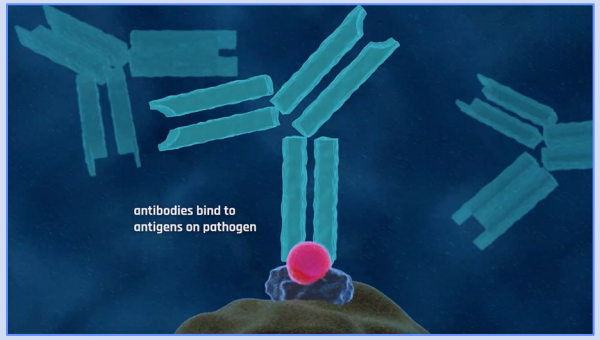
Develop SAFER (safe, fair, resilient, ethical and robust) AI for DoD Applications

Human Machine Teaming



Extend Human-AI interaction to autonomous wingman and intelligent C2 applications

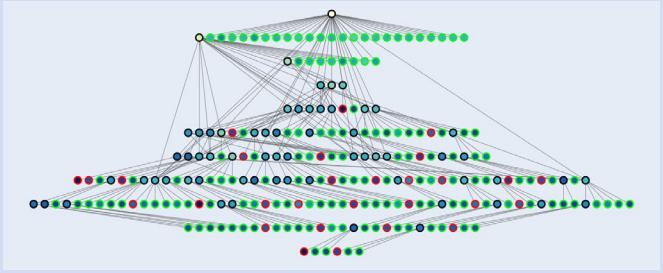
AI Driven Experimental Design



Design End-to-End* ML-driven framework applied to rapid and cost-effective design of antibodies

Lincoln Laboratory supports a broad portfolio of national security AI research applied in Air, Land, Sea, Space, Bio and Cyber domains

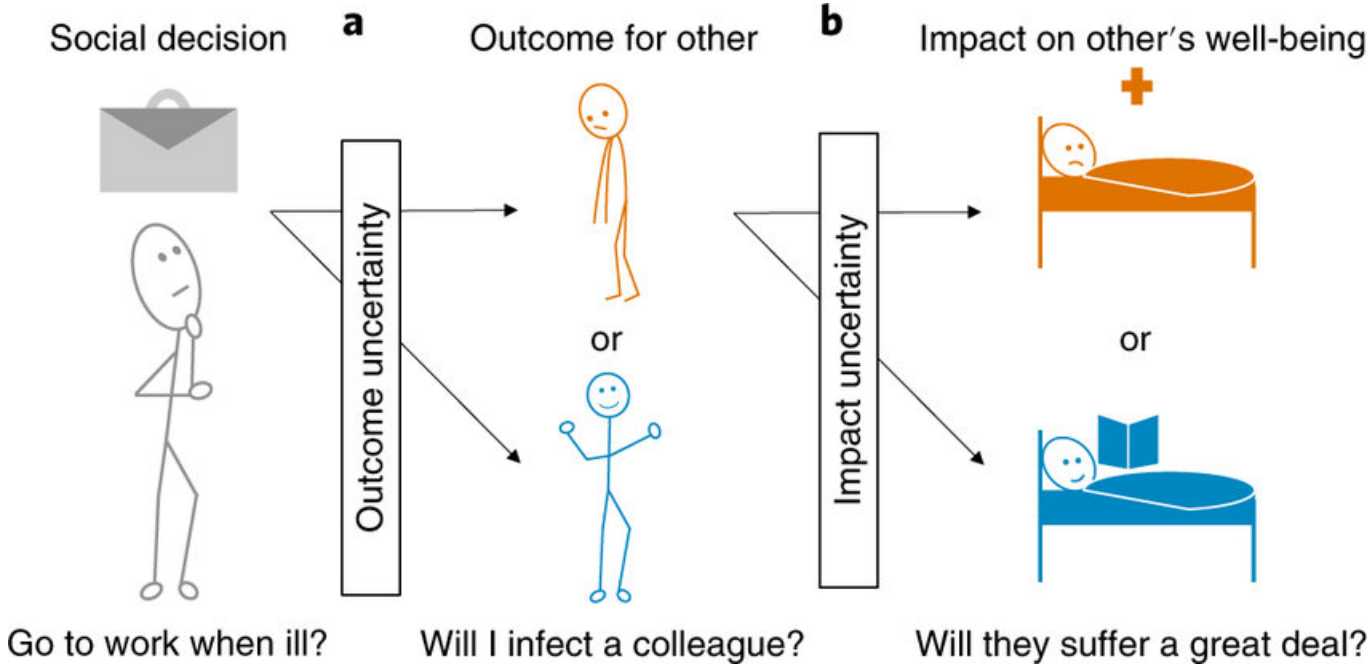
AI for Challenging Environments



Develop AI to classify signatures under sparse data conditions with no initial training data

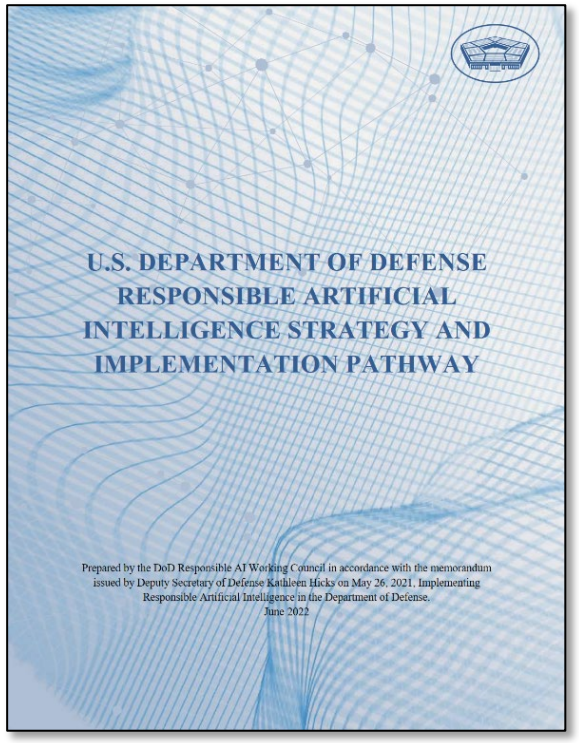
* Li, L., Gupta, E., Spaeth, J. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. ***Nature Communications,* 14**, 3454 (2023). https://doi.org/10.1038/s41467-023-39022-2

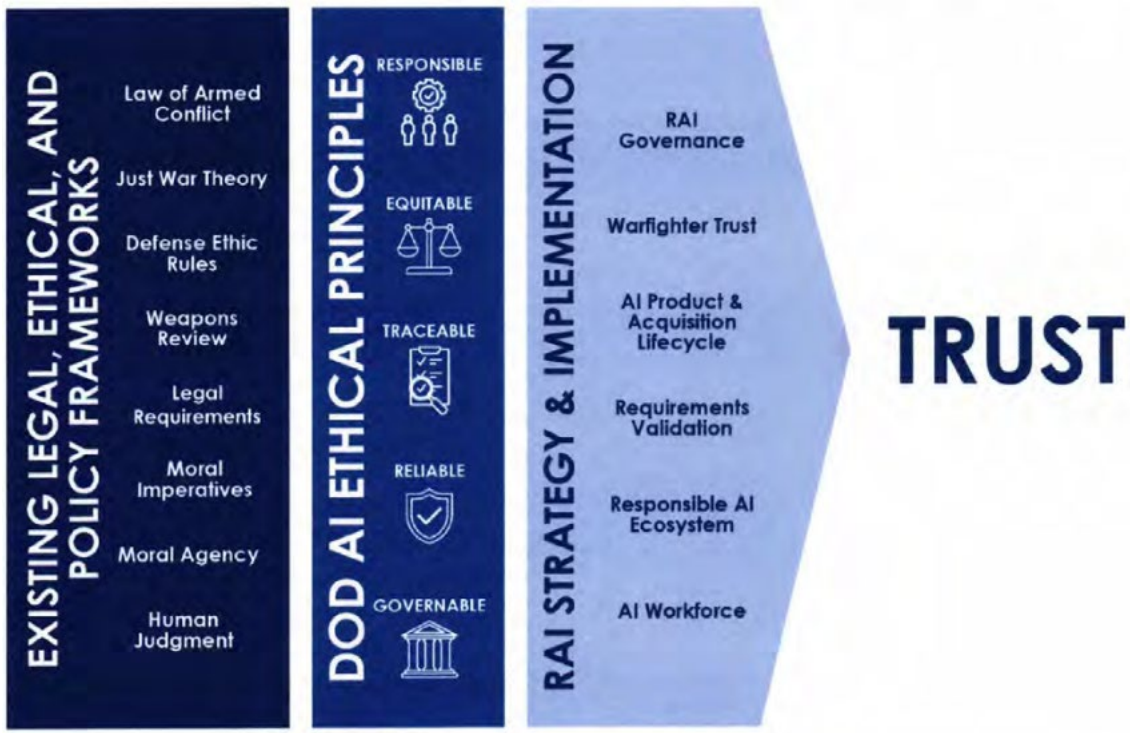
The AI promise of uncertainty reduction



Kappes, Andreas & Nussberger, Anne-Marie & Faber, Nadira & Kahane, Guy & Savulescu, Julian & Crockett, Molly. (2018). Uncertainty about the impact of social decisions increases prosocial behaviour. Nature Human Behaviour. 2. 10.1038/s41562-018-0372-x.

DoD Responsible AI Pathway to Trust

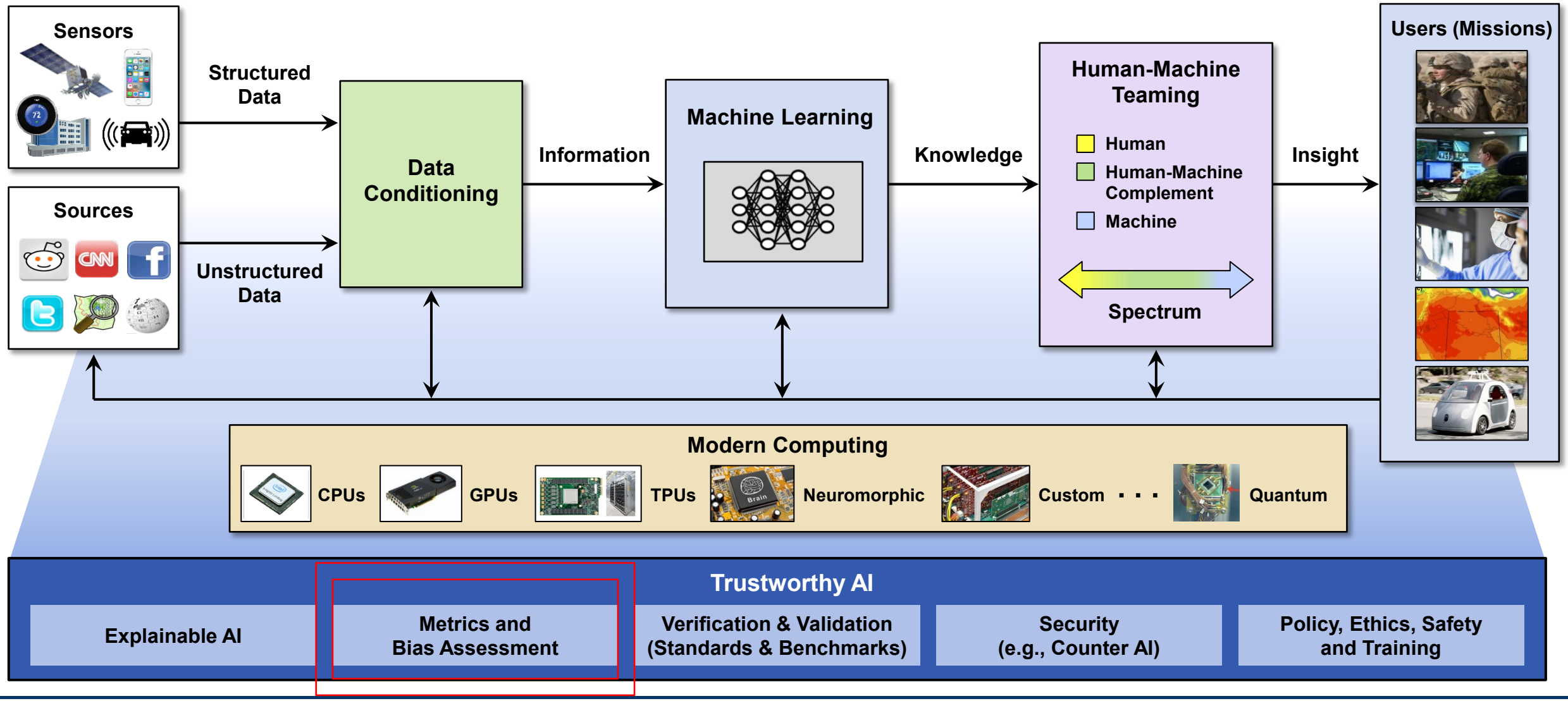




- **DoD strategy and implementation "pathway" for development of responsible AI**
- **BUT, to achieve trust, the underlying AI technology must be trustworthy…**

"DoD Responsible AI Strategy and Implementation Pathway," DoD Responsible AI Working Council, June 2022

AI Canonical Architecture



GPU = Graphics Processing Unit

TPU = Tensor Processing Unit

Human Cognitive Limitations and Perception Blocks

- Retain only limited information in short-term memory
- Display different types and degrees of intelligence
- Closed belief systems restrict information search
- Propensity for risk varies
- Level of aspiration correlated to desire for information

- Difficulty in isolating a problem
- Delimiting the problem space too closely
- Inability to see the problem from various perspectives
- Stereotyping
- Cognitive saturation or overload

Morris, Alan H. "Human cognitive limitations. Broad, consistent, clinical application of physiological principles will require decision support." *Annals of the American Thoracic Society* 15.Supplement 1 (2018): S53-S56.

Clemen, Robert T., and Making Harding Decisions. "An introduction to decision analysis." *R. Clemen, & T. Reilly, Making Hard Decisions with Decision Tools* (2001): 2.

AI as a rational or human thinker?

Lee, C. (2019) The Game of Go: Bounded Rationality and Artificial Intelligence

Table 1: Four Categories of AI Definitions

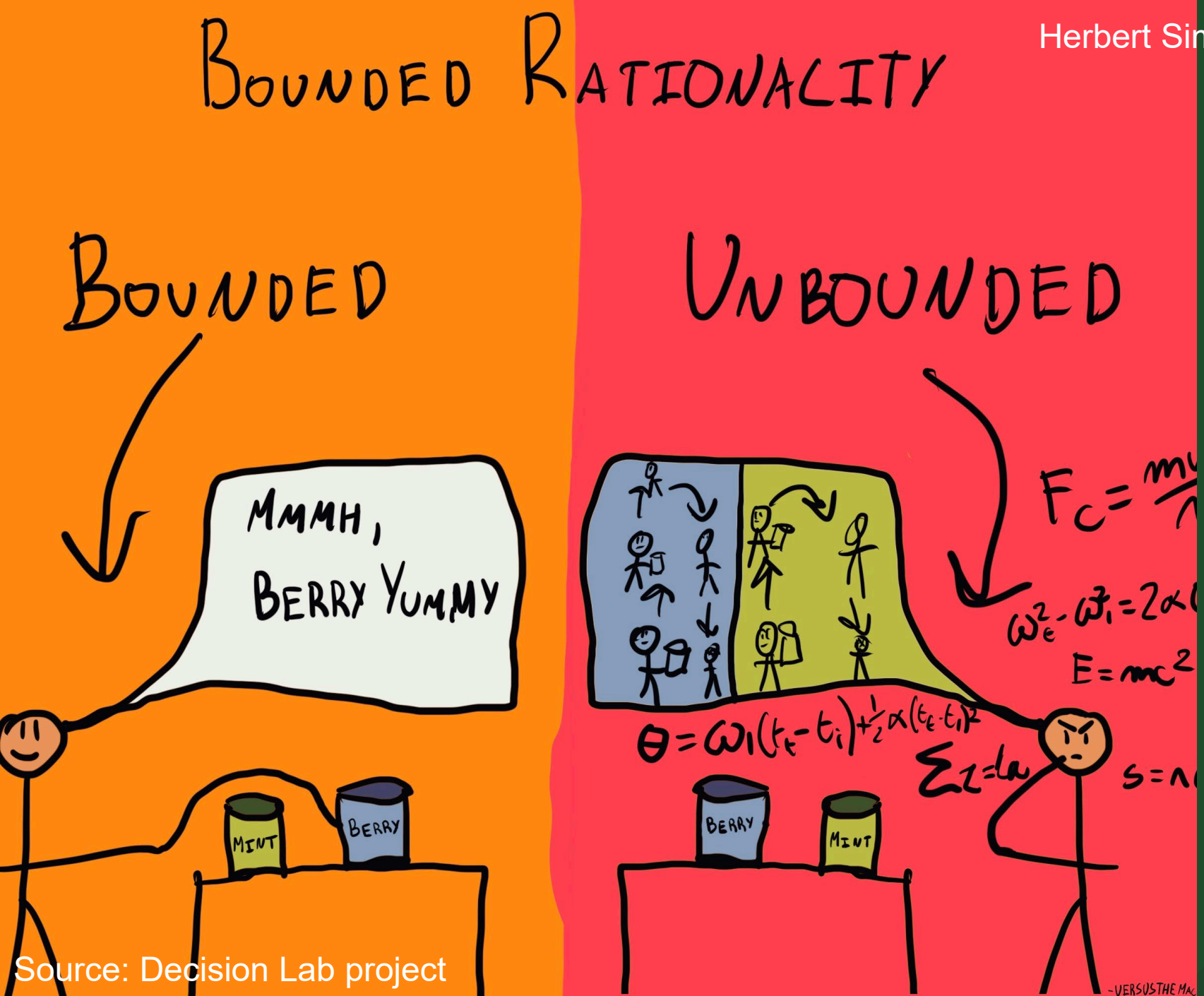
	Human Behaviour	Rational Behaviour
Thinking (Mental Process)	**1. Thinking Humanly** Machines that think intelligently like humans	**3. Thinking Rationally** Machines that think rationally
Acting (Action)	**2. Acting Humanly** Machines that perform activities that human consider intelligent	**4. Acting Rationally** Machines that act rationally

Source: Adapted from Russell and Norvig (2010), Figure 1.1, p.2.

Şimşek, Ö. (2020). Bounded Rationality for Artificial Intelligence. In R. Viale (Ed.), *Routledge Handbook of Bounded Rationality* (pp. 338-348). Routledge.

*Russell, Stuart and Peter Norvig. 2010. Artificial Intelligence: A Modern Approach, Third Edition. New Jersey: Prentice Hall.

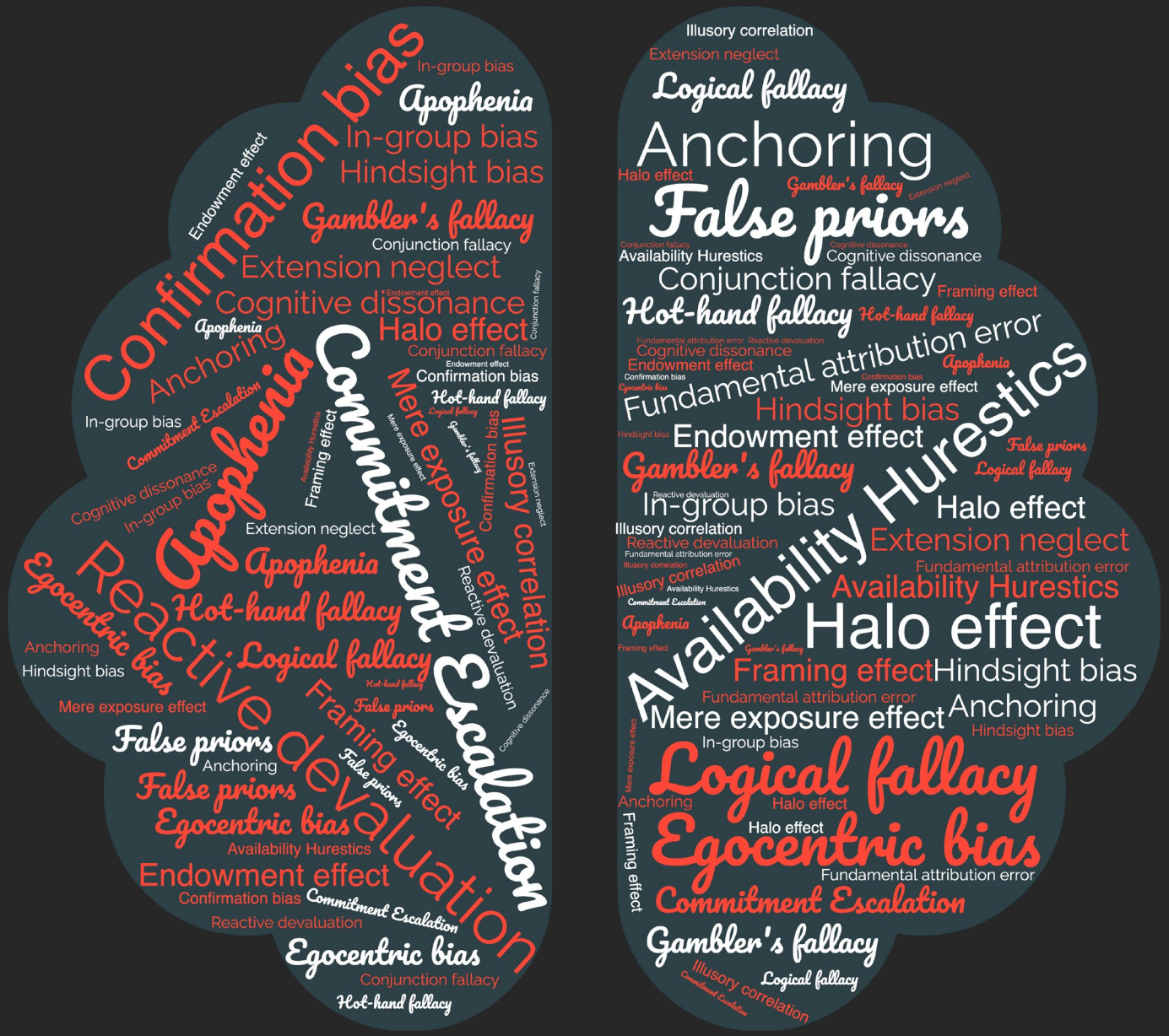
Herbert Simon's principle of bounded rationality



We are motivated to satisfice vs. optimize

Our cognitive limits influence the approach and application of AI systems

Source: Decision Lab project



Amos Tversky and
Daniel Kahneman views on cognitive bias

Continuity of biases in AI framework development and applications

Commitment and Escalation of Commitment

Confirmation

Anchoring

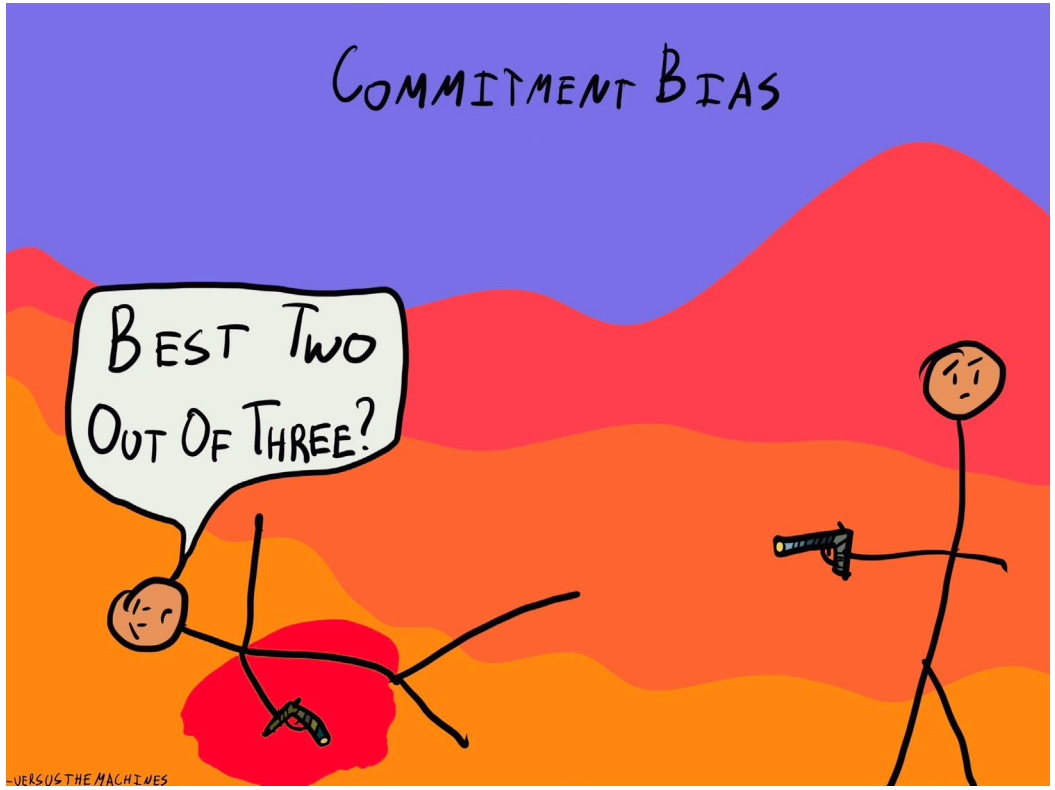
Framing Effect

Hindsight and overconfidence

Availability Heuristics

Commitment

- Can lead to training and monitoring pathways that align with what has been done or observed in the past.
- Compromised ability to remain neutral
- continued commitment leads to escalation of commitment biases (locked in effects)

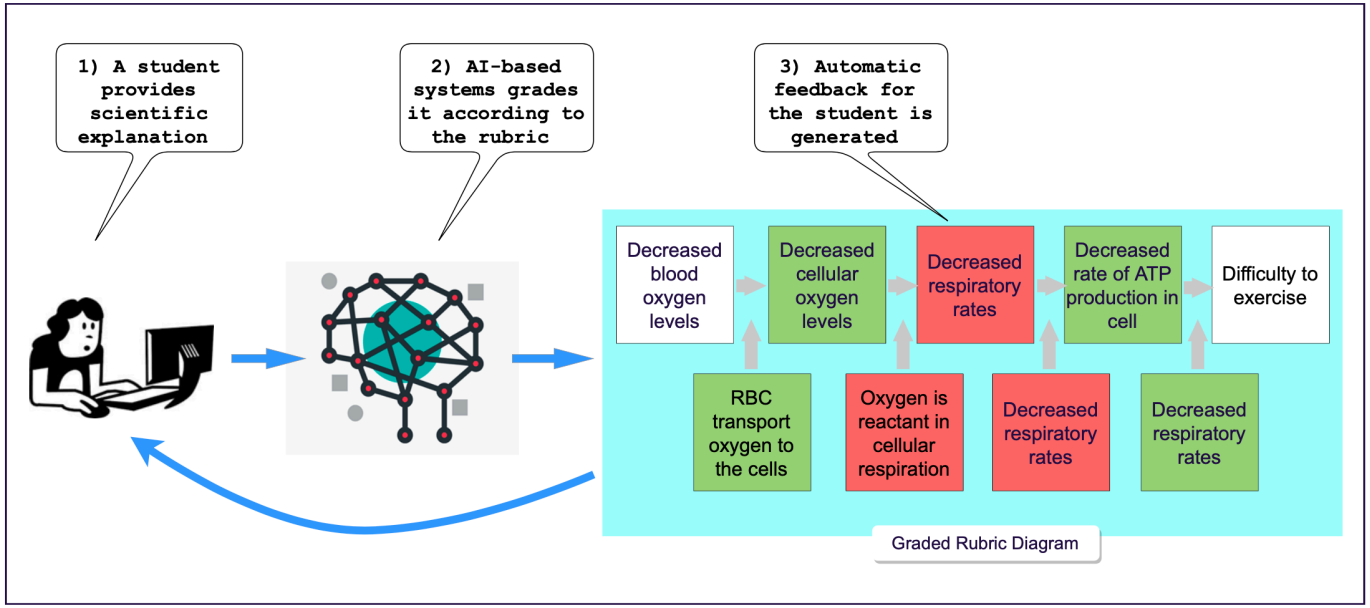


Source: Decision Lab project

our tendency to remain committed to our past behaviors, particularly those exhibited publicly, even if they do not have desirable outcomes.

Confirmation Bias

- can lead to labelling errors, labels assigned on prior beliefs rather than objective conclusions
- **confounding biases (correlation testing) - omitted variables**
- can lead to human-AI operators overriding predictions



Nazaretsky, T., Cukurova, M., Ariely, M., & Alexandron, G. (2021, September). Confirmation bias and trust: Human factors that influence teachers' attitudes towards AI-based educational technology. In CEUR Workshop Proceedings (Vol. 3042).

our underlying tendency to notice, focus on, and give greater credence to evidence that fits with our existing beliefs. Leads us to poor decisions as it distorts our reality from which we draw evidence.

Anchoring

- can distorts perception and use of AI system
- can cause users to formulate skewed perceptions of predictions, anchoring to the first answer they are given

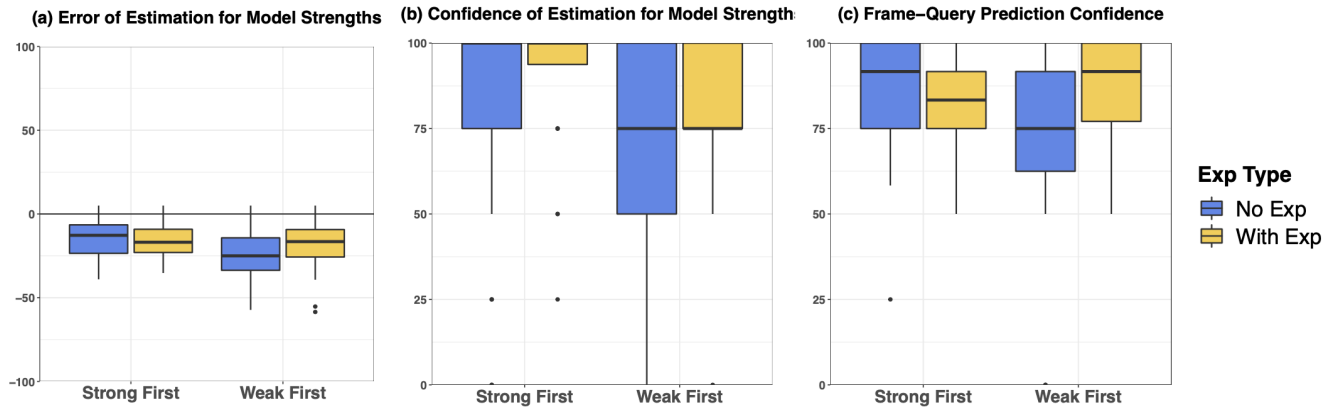


Figure 5: Mental model metrics. (a) Participants' error of estimation for component accuracy (below 0 is underestimation). (b) Percentage of components for which participants rated as being confident in their estimation. (c) Percentage of frame-query pairs for which participants felt confident in their predictions. The last two plots are based on strength-detection (as described in Section 4.2)

Nourani, Mahsan, et al. "Anchoring bias affects mental model formation and user reliance in explainable AI systems." *26th International Conference on Intelligent User Interfaces*. 2021.

We rely too heavily on the first piece of information we are given about a topic

Framing effect

- can cause training to focus on the way data is coded vs the data itself – COMPAS*
- can cause problem formation issues

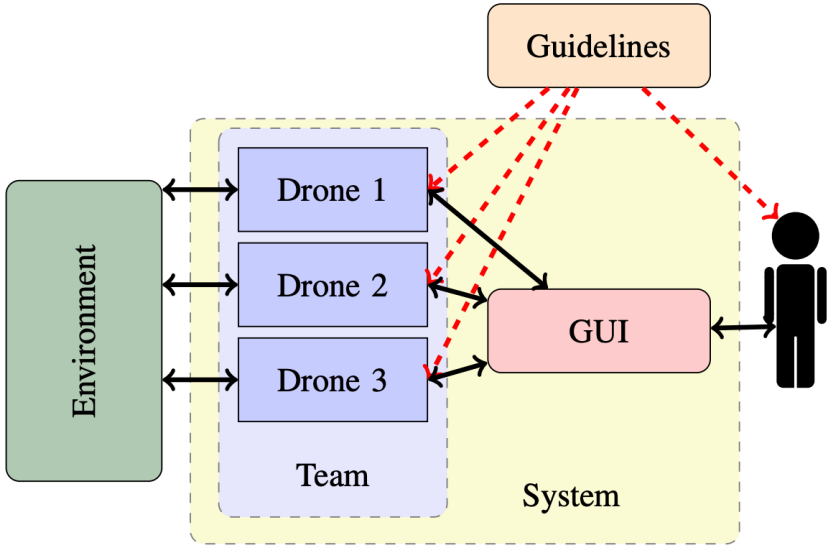


Fig. 1. Framework architecture

P. E. U. Souza, C. P. Carvalho Chanel, F. Dehais and S. Givigi, "Towards human-robot interaction: A framing effect experiment," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, 2016, pp. 001929-001934, doi: 10.1109/SMC.2016.7844521.

Framing bias refers to the observation that the manner in which data is presented can affect decision making

* Dressel, J., Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1); https://advances.sciencemag.org/content/4/1/eaao5580.

Hindsight and overconfidence

reliable source of information

- can lead to an overconfidence in our ability to predict consequences of decisions
- can lead to source information errors (data sets), and interpretation errors, along with trust (AI rejection of true values)

Table 1. Important factors leading to knowledge bias.

Factors Leading to Knowledge Bias	Description
Experimental bias	Inherent bias in experiment leading to inaccurate outcomes, and pre-existing beliefs leading to wrong perceptions such as hindsight
Problems with information reliability	Synthesizing systems based on false or partially accurate data
Limited expert knowledge	Domain experts may have limited knowledge of their own domain that will limit the knowledge programmed into the system
Shallow information	Implicit knowledge contained in systems such as electronic health records may be shallow and may not include the necessary details

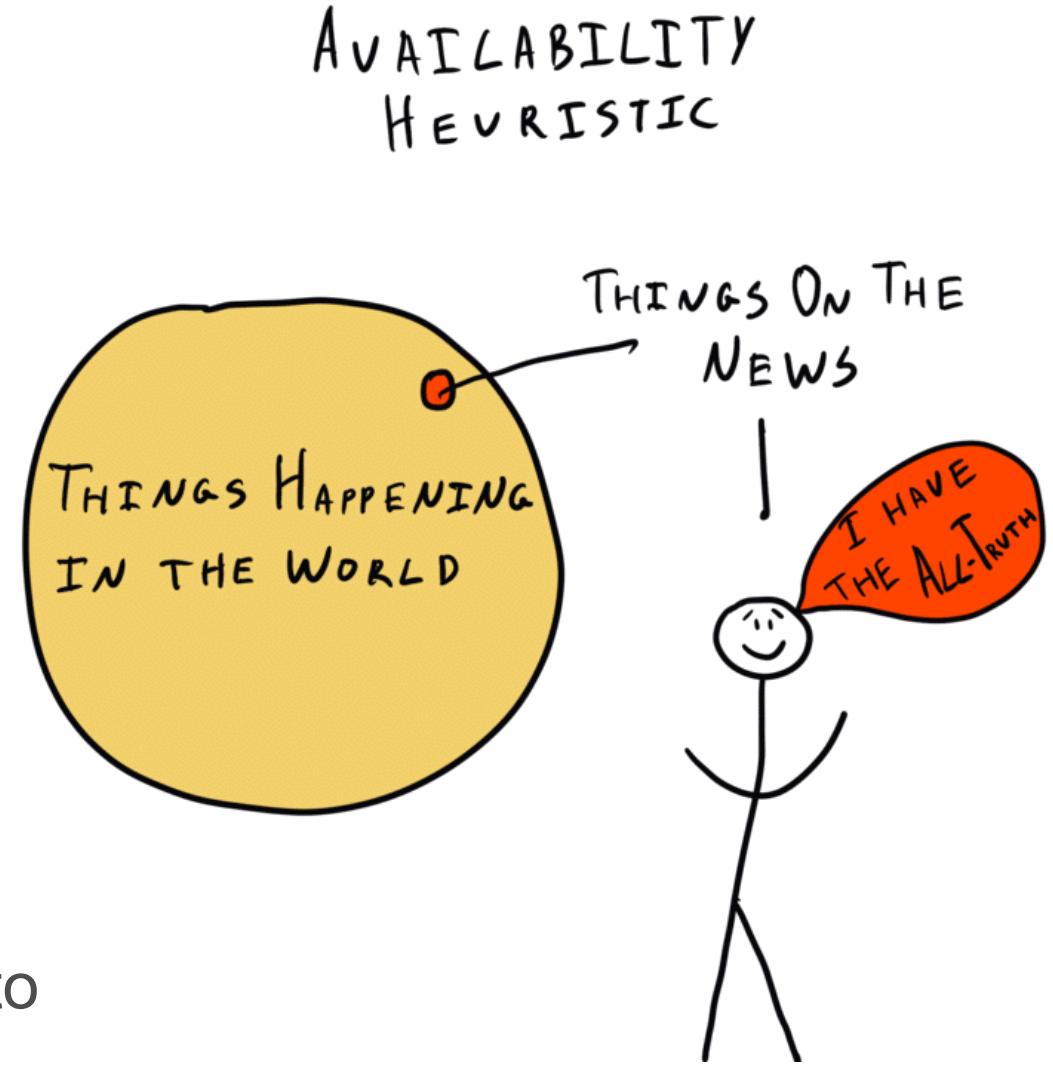
Gurupur, Varadraj, and Thomas TH Wan. "Inherent bias in artificial intelligence-based decision support systems for healthcare." Medicina 56.3 (2020): 141.

Hindsight - our tendency to look back at an unpredictable event and think it was easily predictable. It is also called the 'knew-it-all-along' effect.

Overconfidence – tendency for excessive certainty in one's answers/knowledge

Availability heuristic

- can lead to selection bias, association bias in training models
- AI models trained in mis-proportional data sets can exhibit availability heuristics (weighting problems)



Source: unknown

Availability heuristic describes our tendency to use information that comes to mind quickly, a mental shortcut to improve efficiency.

Enabling/Improving the Human-Machine Team

Human Augmentation

Physical Enhancement | Cognitive Enhancement | Spatial Enhancement

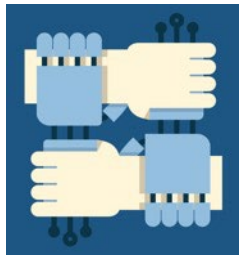
Teaming

Team Dynamics

Task Sharing

Advanced Concepts of Teaming

Assurance



Trust and Bias Calibration

Explainability and Robustness

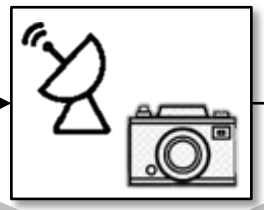
Security and Resilience

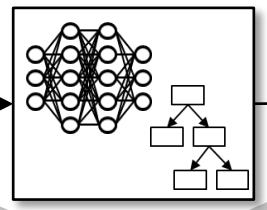
HMT Technologies should aim to enable or improve one or more elements of the Human-Machine Team in order to improve mission outcomes

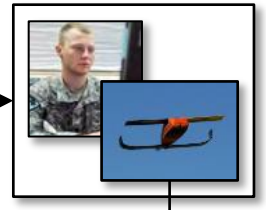
Challenges in Trustworthy AI and Autonomy

Critical need exists to address the trustworthiness of AI for National Security

Environment → **Sensors** → **AI Models** → **Decision Makers**







Closed-Loop Feedback

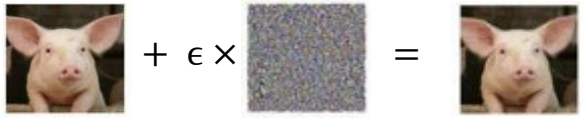
Resilience to Adversary

- Physical-world adversarial examples
- Traditional forms of deception

AI Model: "85 MPH"

Adversarial Robustness

Original Image + $\epsilon \times$ Optimized, Digital Noise = Adversarial Example



AI Model: "Pig"

AI Model: "Airliner"

Transparency

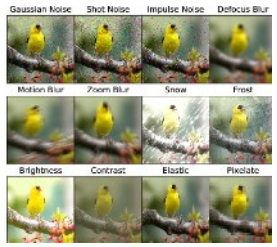
- Lacking human logic / context
- Unexplainable outputs
- Unclear decision process

Robustness and Adaptability

- Biases
- Distribution Shift
- Anomalies
- Novel classes

Robustness to Sensor Degradations

- Noise
- Distortion
- Occlusions
- Data dropout



Safety

- Unbounded behaviors
- Operating outside envelope
- Single point of failure

Real world patterns of cognitive biases

The continuity of supervisory AI-Human teaming biases

Data Biases

Sampling bias – lack of true representative datasets

Patterns of bias distributed throughout data

Humans

Data

Use

Design

Application Biases

human-AI teaming issues

Harmful repurposing of AI models

Untrustworthy AI

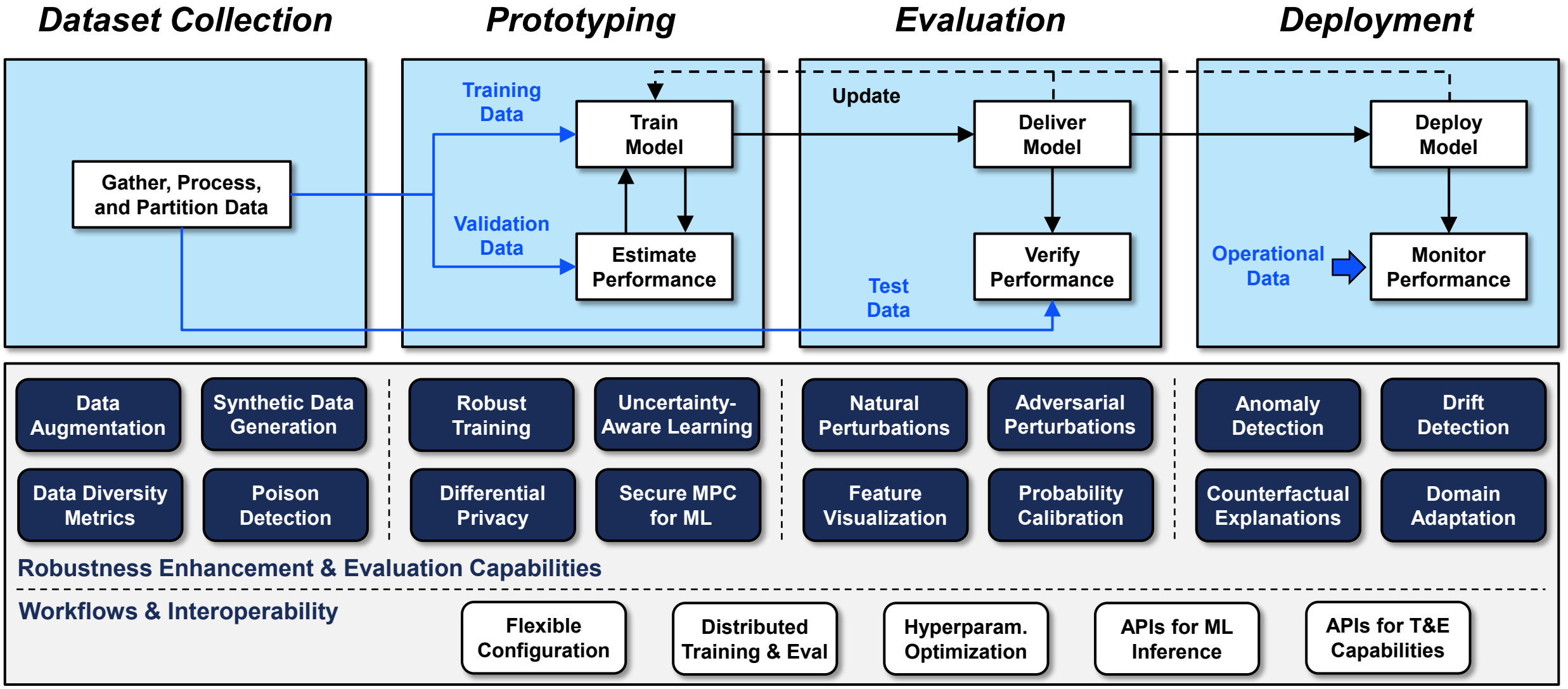
Biased AI design and deployment

Biased deployments

Imbalances in problem formation

Biased requirements, modelling, and testing

Trustworthy AI techniques should span the full life cycle



* MPC = multi-party computation

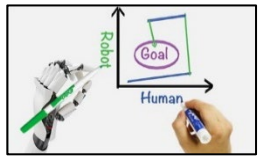
Sample Effects of Bias Propagation

- **Predictive Policing - Extreme confirmation bias causing self-reinforcing training data**
  - Prediction algorithms need many positive and negative examples of each offense and over a range of communities
  - Vast majority of data comes from 'high-crime' communities (not higher socio-economic communities)
  - AI runs risk of confirming human and algorithmic bias of high crime areas while overlooking low crime area offenses

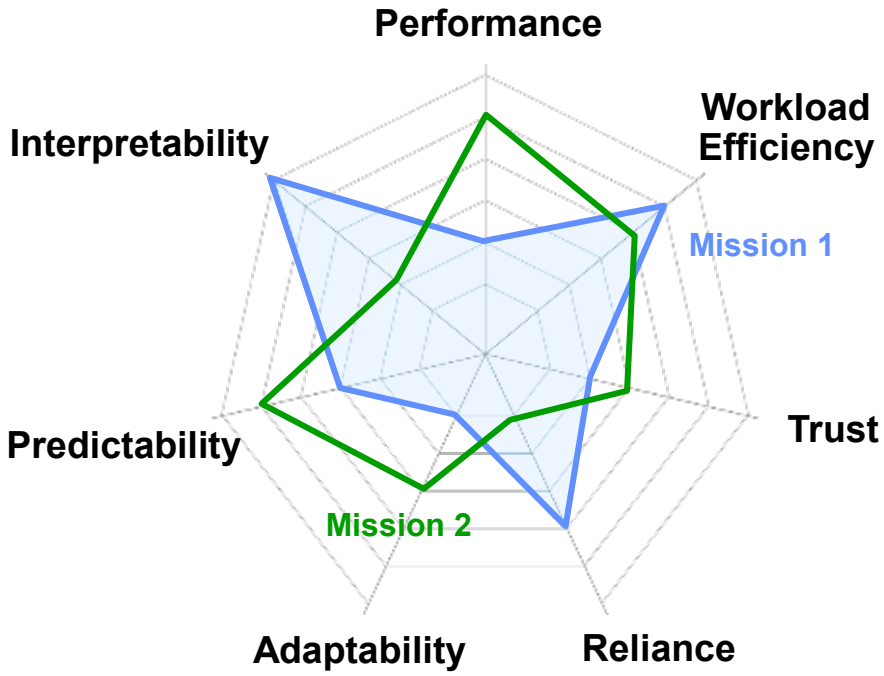
- **Facial Recognition for Security Checkpoints – Bias favoring accuracy for particular subgroups**
  - Overall reported 'accuracy rates' may not be subdivided by ethnicities
  - High confidence built based on one group leads to bias against populations with a high error rate with potential arrest of innocent people

- **Effects of AI hype - A bias that AI is 'smarter' than most people**
  - Many may suspend critical thinking in favor of algorithmic results
  - Danger in favoring black-box AI usage in critical decision making, but the algorithms are not in charge

Metrics for Effective Human-AI Systems



Effectiveness Web



Metric	Description
Performance	How well did the human-AI system complete intended tasks?
Workload Efficiency	Did the system decrease the human cognitive workload compared to baseline?
Trust	How much does the human trust the system to aid in completing intended tasks?
Reliance	How often does the human choose to use the system to complete intended tasks?
Adaptability	Can the AI adapt to new/unexpected situations?
Predictability	Does the AI act in a way consistent with the human mental model?
Interpretability	Does the human understand how and why the AI makes decisions?

Chang, K. C. et al, "A Multidimensional Analysis of Effectiveness for Human-AI Teaming Systems," LL Tech Report (2020)

Metrics allow for benchmarking of Human-AI systems and the means to evaluate the effectiveness of AI implementations

Thank you