Reflections on Safety and Artificial Intelligence

Eric Horvitz

Exploratory Technical Workshop on Safety and Control for AI
Carnegie Mellon University
Pittsburgh, PA
June 27, 2016
AI & Safety

Constellation of methods referred to as Artificial Intelligence will touch our lives more closely and intimately

AI moving into high-stakes applications

- Healthcare
- Transportation
- Finance
- Public policy
- Defense

→ Much to do on principles, methods, and best practices
Safe AI Systems

Relevance of Multiple Subdisciplines
Relevance of Multiple Subdisciplines

Safe AI Systems

Planning

Control Theory

Machine Learning

Sensor Fusion
Relevance of Multiple Subdisciplines

Safe AI Systems

- Metareasoning
- Planning
- Control Theory
- Machine Learning
- Sensor Fusion
Relevance of Multiple Subdisciplines

Safe AI Systems

- Mechanism Design
- Metareasoning
- Multiagent Systems
- Planning
- Control Theory
- Machine Learning
- Sensor Fusion
Safe AI Systems

Relevance of Multiple Subdisciplines

- Mechanism Design
- Metareasoning
- Multiagent Systems
- Robust Optimization
- Sensor Fusion
- Machine Learning
- Verification
- Planning
- Control Theory
- Security
Relevance of Multiple Subdisciplines

Safe AI Systems

- Mechanism Design
- Metareasoning
- Multiagent Systems
- Robust Optimization
- Sensor Fusion
- HCI
- Security
- Verification
- Planning
- Control Theory
- Machine Learning
Relevance of Multiple Subdisciplines

Safe AI Systems

Multiagent Systems

Mechanism Design

Metareasoning

Planning

Control

Machine Learning

Sensor Fusion

Robust Optimization

HCI

Security

Verification

Control Theory

Machine Learning

Metareasoning

Multiagent Systems

Mechanism Design

Safe AI Systems

Safety-Critical Systems
safety
ˈsāftē/

noun
1. the condition of being protected from or unlikely to cause danger, risk, or injury
safety-critical ˈsāftēˌkridək(ə)l/

adjective
1. systems whose failure could result in loss of life, significant property damage, or damage to the environment.

2. designed or needing to be fail-safe for safety purposes.
fail-safe  

**noun**
device or practice that, in the event of a failure, responds or results in a way that will cause no harm, or at least minimizes harm.

**adjective**
incorporating some feature for automatically counteracting the effect of an anticipated possible source of failure.
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

Fail-safe practice

Full-power throttle on arrested landing
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

Fail-safe plan
Free return trajectory
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

✓ Mechanism
✓ Practice
✓ Plan

Fail-safe plan
Free return trajectory
Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

✓ Mechanism
✓ Practice
✓ Plan
Fail-safe

George Westinghouse, 1869

Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

- Mechanism  - Monitoring
- Practice
- Plan
AI in the Open World

Growing interest in issues & directions with AI in real-world settings

Grappling with uncertainty and more general incompleteness

AAAI President’s address (2008), “Artificial Intelligence in the Open World.”

AAAI President’s address (2016), “Steps Toward Robust Artificial Intelligence.”

E. Horvitz. Artificial Intelligence in the Open World, AAAI President’s Address, Chicago, IL, July 2008.

T. Dietterich, Steps Toward Robust Artificial Intelligences, AAAI President's Address, Phoenix, AX. February, 2016.
Special Considerations with AI

Open-world complexity $\rightarrow$ incomplete understanding

Uncertainties & poor-characterization of performance
Poor operating regimes, unfamiliar situations
Special Considerations with AI

Open-world complexity $\rightarrow$ incomplete understanding

- Uncertainties & poor-characterization of performance
- Poor operating regimes, unfamiliar situations

Rich ontology of failures

- Numerous failure modalities
- New attack surfaces (e.g., *machine learning attack*)
- Self-modification & gaming (e.g., *modify reward fcn*)
- Unmodeled influences
Special Considerations with AI

Open-world complexity $\rightarrow$ incomplete understanding
Uncertainties & poor-characterization of performance
Poor operating regimes, unfamiliar situations

Rich ontology of failures
Numerous failure modalities
New attack surfaces (e.g., machine learning attack)
Self-modification & gaming (e.g., modify reward fcn)
Unmodeled influences

Challenges of transfer across time & space
Challenge of coordinating human-machine collaborations
Operational opacity
AI & Open-World Complexity

Frame problem

*How to tractably derive consequences of an action?*

Qualification problem

Understanding preconditions required for actions to have intended effects

Ramification problem

Understanding all important effects of action
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
Decision making under uncertainty & incompleteness

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

AI & Open-World Complexity
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness
Direction: Learn about abilities & failures

Deep learning about deep learning performance

Fang, et al., 2015

Caption: a man holding a tennis racquet on a tennis court

Quality score $[0,1]$

$$s = \frac{e^{W \cdot f}}{1 + e^{W \cdot f}}$$
Direction: Learn about abilities & failures

\[ p(\text{fail} | E, t) \]

Performance → Predictive model of confidence → Successes & failures

Toyama & H. 2000
Direction: Learn about abilities & failures

Performance
Successes & failures

Predictive model of confidence

$p(\text{fail} \mid E, t)$

Target Ground Truth
Inferred State

Toyama & H. 2000
Direction: Learn about abilities & failures

Performance

Successes & failures

Predictive model of confidence

$p(\text{fail} \mid E, t)$

Inferred State

Inference Reliability

Inference Reliability Indicator 1

... 

Inference Reliability Indicator $n$

Target Ground Truth

Toyama & H. 2000
Direction: Robustness via analytical portfolios

Toyama & H. 2000
Direction: Robustness via analytical portfolios

Toyama & H. 2000
Direction: Robustness via analytical portfolios

Unmodeled situations in open world

Perceptual modalities

- back. subtract
- color based
- motion decay

Joint inference

- facing away
- jolted camera
- periph. distraction
- lights out

Toyama & H. 2000
Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences
Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences
Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences

![Graph showing comparison of rates of ischemic stroke with Warfarin and Dabigatran at different rates of ICH with Warfarin. The graph compares Warfarin with Dabigatran at 110 mg and 150 mg doses.](image)
Direction: Robust optimization to minimize downside

Robust optimization under uncertain parameters
Risk-sensitive objective
e.g., conditional-value-at-risk budget

Methods trade upside value for reducing probability of costly outcomes

Tamar, 2015; Chow, et al., 2014; per Dietterich, AAAI lect. 2016
Direction: Learn about unknown unknowns

Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

Human engagement
Direction: Learn about unknown unknowns

Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

Human engagement

“Beat the Machine” (Attenberg, Ipeirotis, Provost 2015)
Direction: Learn about unknown unknowns

Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

Human engagement

“Beat the Machine” (Attenberg, Ipeirotis, Provost 2015)
Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

Krumm, H., 2006
Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

Krumm, H., 2006
Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

Predict previously unseen destination

Krumm, H., 2006
Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

Predict previously unseen destination

Destinations, $E_1, \ldots, E_n, t$

Krumm, H., 2006
Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

Predict previously unseen destination

Krumm, H., 2006
Direction: Joint modeling of key dimensions of error

Example: Learn about errors of perception & control

Probabilistic models of control $\varphi_{\text{roll}}$

Probabilistic models of sensing $\varphi_{\text{obstacle}}$

Sadigh & Kapoor, 2016
Direction: Joint modeling of key dimensions of error

Proposed trajectory

φ_roll
φ_obstacle

Sample 1
Sample 2
Sample n

Sadigh & Kapoor, 2016
Direction: Joint modeling of key dimensions of error

Trajectory safe if:

\[
\frac{\sum \sqrt{\text{roll}}}{\sum \sqrt{\text{obstacle}}} > 1 - \epsilon
\]

Sadigh & Kapoor, 2016
Direction: Joint modeling of key dimensions of error

Sadigh & Kapoor, 2016
Direction: Joint modeling of key dimensions of error

Value of refining models & system
- Value of additional data
- Value of enhancing sensors
- Value of better controller

Sadigh & Kapoor, 2016
Direction: Joint modeling of key dimensions of error

Sadigh & Kapoor, 2016

\[ p > 1 - \epsilon \]
Direction: Joint modeling of key dimensions of error

Sadigh & Kapoor, 2016

Fail-safe

\[ p > 1 - \epsilon \]

(video)
Direction: Joint modeling of key dimensions of error

```cpp
bool AvoidCarCrash(double[] x, double[] y, double[] t, double mu_x, ...
..., double mu_y, double mu sx, double mu sy, double sigma_sq,
..., double Thresh)
{
    // Sample location and velocities for the other vehicle
    x_other = Gaussian(mu_x, sigma_sq);
    y_other = Gaussian(mu_y, sigma_sq);
    sx_other = Gaussian(mu sx, sigma_sq);
    sy_other = Gaussian(mu sy, sigma_sq);

    bool isSafe = true;
    for (int i = 0; i < x.GetLength(0); i++)
    {
        // Compute distances to the ego vehicle at each time step
        Xdistance = x[i] - (x_other + time[i]*sx_other);
        Ydistance = y[i] - (y_other + time[i]*sy_other);

        // Safety invariants that require min threshold distance
        SafeInX = (Xdistance > Thresh) || (Xdistance < -Thresh);
        SafeInY = (Ydistance > Thresh) || (Ydistance < -Thresh);
        isSafeNow = (SafeInX || SafeInY)

        isSafe = isSafe && isSafeNow;
    }

    return isSafe;
}
```

Sadigh & Kapoor, 2016
Safe AI Systems

- Verification
- Security
- Cryptography
Direction: Verification, security, cryptography

Static analysis
Run-time verification
Whitebox fuzzing
Cybersecurity to protect attack surfaces
Appropriate use of physical security, isolation
Encryption for data integrity, protection of interprocess comms.
Direction: Runtime verification

Difficult to do formal analysis of large-scale system

→ Analysis & execution considers info. from running system

Satisfy or violate desired properties?

Identify problem, future problem

Engage human

Take fail-safe action
Direction: Metalevel analysis, monitoring, assurance
Direction: Metalevel analysis, monitoring, assurance

Environment

AI system

State'

Action
Direction: Metalevel analysis, monitoring, assurance

Diagram:
- **Environment’**
- **AI system**
- Arrows: Action, Reward, State’

Legend:
- **State**
- **Direction:** Metalevel analysis, monitoring, assurance
Direction: Metalevel analysis, monitoring, assurance
Direction: Metalevel analysis, monitoring, assurance

Environment

AI system

Reward

Reinforcement

State'

Action

Learning
Direction: Metalevel analysis, monitoring, assurance
Direction: Metalevel analysis, monitoring, assurance
Direction: Metalevel analysis, monitoring, assurance

Adversary

Environment'

State'

Perception

Reward

Reinforcement

AI system

Learning

Self-modification

Action

e.g., see: Amodei, Olah, et al., 2016
Direction: Metalevel analysis, monitoring, assurance

Reflective analysis
- Operational faithfulness
- Ensure isolation, detect mods
- Identify external meddling

Adversary

AI system

Environment'

State'
Perception
Reinforcement

Reward

Learning
Direction: Metalevel analysis, monitoring, assurance

Run-time verification
Static analysis

Adversary

Environment'

State'
Perception
Reinforcement

AI system

Reward

Reflective analysis
- Operational faithfulness
- Ensure isolation, detect mods
- Identify external meddling
Direction: Human-machine collaboration

Models of human cognition
Transparency of state, explanation
Mastering coordination of initiatives
Direction: Human-machine collaboration

China Airlines 006 (Feb 1985)

747 dives 10,000 in 20 seconds. 5g, supersonic.

Air France 447 (June 2009)

Unrecoverable stall.
Direction: Human-machine collaboration

China Airlines 006 (Feb 1985)
  747 dives 10,000 in 20 seconds. 5g, supersonic.

Air France 447 (June 2009)
  Unrecoverable stall.
Direction: Human-machine collaboration
Rich spectrum of autonomy
How to best work together for safety?

Kamar, Hacker, H., 2012
Direction: Human-machine collaboration
Rich spectrum of autonomy
How to best work together for safety?

Kamar, Hacker, H., 2012
Direction: Human-machine collaboration
Rich spectrum of autonomy
How to best work together for safety?

Kamar, Hacker, H., 2012
Direction: Human-machine collaboration
Direction: Human-machine collaboration

Infer challenges with machine competency
Direction: Human-machine collaboration

Infer challenges with machine competency       Infer human attention
Direction: Human-machine collaboration

Infer challenges with machine competency

Infer human attention
Direction: Human-machine collaboration

Continual prediction of trajectories

Infer human attention

\[ p(\text{attention state} | E) \]
Direction: Human-machine collaboration

Continual prediction of trajectories

Infer human attention

\[ p(\text{attention state}|E) \]

Time
Direction: Human-machine collaboration

Continual prediction of trajectories

Infer human attention
Direction: Human-machine collaboration

Safety-assuring mixed-initiative planner
- Driver’s attention over time
- Latency of human input
- Latency tolerance of situation
- Cost & influence of alerting driver
- Custom language, ongoing dialog

Gain driver attention $t$
Slow to defer need $t'$
Implement failsafe $t''$
Direction: Develop Best Practices for Safe AI

- Phases of study, testing, reporting for rolling out new capabilities in safety-critical domains (akin to FDA clinical trials, post-marketing surveillance)
- Disclosure & control of parameters on failure rates, tradeoffs, preferences
- Transparency & explainability of perception, inference, action
- System self-monitoring & reporting machinery
- Isolation of components in intelligence architectures
- Detecting & addressing feedback of system’s influence on self
Direction: Develop Best Practices for Safe AI

- Standard protocols for handoffs, attention, awareness, warning, in human-machine collaborations
- Policies for visible disclosure of autonomy to others (e.g., indication to others that a car is currently on automated policy)
- Fail-safe actions & procedures given predicted or sensed failures
- Enhancing robustness via co-design of environment & systems
- Testing for drift of assumptions, distributions in domains
- Special openness & adherence to best practices for data, learning, decision making for applications in governance & public policy
Direction: Address concerns about “superintelligences”

Addressing concerns of public

Significant differences of opinion, including experts

Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence

Google-owned Boston Dynamics released a video showing a 6’ tall 320-lb humanoid robot named Atlas running freely in the woods

By Michael Sainato • 08/19/15 12:30pm
Addressing concerns of the public

Significant differences of opinion, including among experts

Speculations Concerning the First Ultraintelligent Machine

“...[A]n ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind.” I.J. Good (1965)
Direction: Address concerns about “superintelligences”

Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. To do so would of course meet with great opposition, unless we have advanced greatly in religious toleration from the days of Galileo. There would be great opposition from the intellectuals who were afraid of being put out of a job. It is probable though that the intellectuals would be mistaken about this. There would be plenty to do, in trying to say, i.e. in trying to keep one’s intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control in the way that is mentioned in Samuel Butler’s ‘Brewton’.
Direction: Address concerns about “superintelligences”

“Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the conse-
quence. For it seems possible that once the machine thinking method had started, it would not take long to outstrip our feeble powers.

...they would be able to converse with each other to sharpen their wits.

At some stage therefore, we should have to expect the machines to take control in the way that is mention in Samuel Butler’s Erewhon.”

Alan Turing, 1951
Direction: Address concerns about “superintelligences”

Addressing concerns of public

Significant differences of opinion, including experts

• Do we understand possibilities?

• What kind of research should done proactively?

• Can we “backcast” from imagined poor outcomes

• Designs of clear ways to thwart possibilities, ease concerns