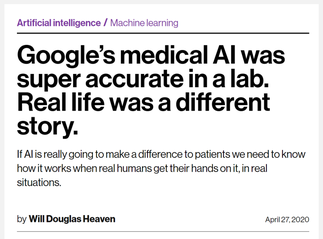
Motivation

- An AI system's decisions depend on its model (hand-crafted or machine-acquired from data)
- Impossible to anticipate all situations and create a perfect model
- Models are almost always incomplete
  - Unavailability of information during system design
  - Designer overlooked the information as unrelated to the system's core functionality
  - Model is too complex and had to be simplified to quickly generated solution
- When operating with incomplete knowledge, a deployed AI system may behave in an unexpected, undesirable manner → affects reliability
- Undesirable behavior ranges from mild and tolerable events to safety-critical failures
- Addressing this practical problem is critical due to the broad societal impacts of AI


Artificial intelligence / Machine learning

Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

by **Will Douglas Heaven** April 27, 2020


STAYING SAFE

Alphabet's Eric Schmidt: The design of AI should "avoid undesirable outcomes"

December 22, 2015

By **Mike Murphy** Technology editor


Grading on a Curve? Why AI Systems Test Brilliantly but Stumble in Real Life

EDMUND L. ANDREWS May 13, 2020

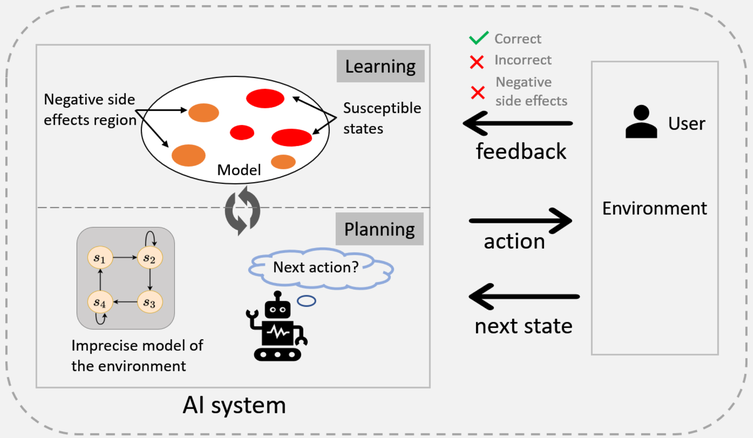

Another Roomba ran over dog poop and then proceeded to 'clean' the house

Research Goals

Develop techniques to improve the reliability of deployed AI systems

1. How to enable AI systems to be cognization of their limitations and side effects?
2. How can AI systems learn to identify sources of errors and adapt their behavior to overcome these limitations?
3. How to effectively leverage human input to accelerate adaptation, without excessively relying on humans?
4. How to improve solution quality without full model revisions, which are expensive and hard to verify?
5. What are the conditions under which bounded-optimality can be achieved?

Solution approaches based on decision theory, automated planning, reinforcement learning, and machine learning.



Behavior adaptation