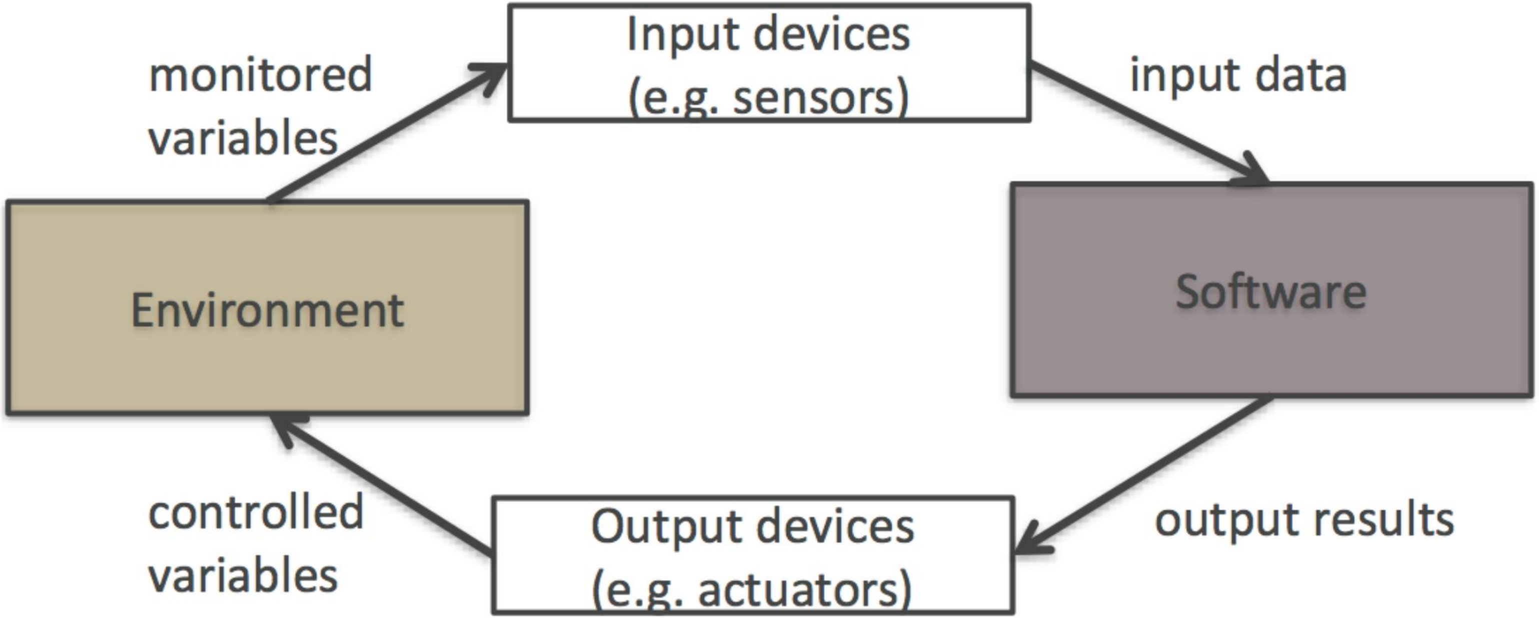
Recitation 5

Requirements

World & the Machine

- Software/AI alone cannot establish system requirements/goals
- Environmental assumptions are critical
- Key takeaway: Think about the environment and assumptions; they are important to satisfy system goals; they can help you identify risks and mitigations
- Concepts
  - Requirements are expressed in terms of world phenomena
  - Shared phenomena: interface between world and machine (actions, events, data flow, etc.)
  - Assumptions are expressed in terms of world and shared phenomena
  - Specifications are expressed in terms of machine and shared phenomena
- Example: What constitute parts of the environment for a COVID-19 detection mobile app?

World & the Machine



What Could Go Wrong?

- Missing/incorrect environmental assumptions [machine works as per spec]
- Wrong/violated specification [machine doesn’t work as per spec]
- Inconsistency in assumptions and specification / requirements
- Feedback loop: behavior of the machine affects the world, which in turn affects input to the machine, and so on.
- Data drift: behavior of the world changes over time, causing assumptions to become invalid
- Adversaries: Bad actors deliberately manipulate inputs / violate assumptions

Identify Assumptions & Potential Problems

- Amazon product recommendations
- Predictive policing
- Screening applicants for Masters program

Amazon Product Recommendations

- Requirements (in the world)
  - Recommend products that the user would like (and is more likely to purchase)
- Specifications (for the machine)
  - Recommend highly rated products up front or higher in the list
  - Return list of products with the same category as items in purchase history higher in the list
- Assumptions (about the world and shared phenomena)
  - ??
- Problems
  - ??

Amazon Product Recommendations

- Assumptions
  - Information about products from vendors are accurate
  - Product ratings are authentic and represent the real quality of that product
  - Products are tagged with the appropriate category by vendors
- Problems
  - What if the ratings are tampered with?
  - What if products are labeled incorrectly?
  - We recommend based on product type => User purchases those products => …
  - New product / product types based on the latest trend
  - Should we recommend just based on product type?

Predictive Policing

- Requirements (in the world)
  - Decide where to allocate police patrol by looking at crime rate of neighborhoods to proactively prevent crimes
- Specifications (for the machine)
  - Based on historical crime rate in each neighborhood, return the top few neighborhoods where police need to patrol
- Assumptions (about the world and shared phenomena)
  - ??
- Problems
  - ??

Predictive Policing

- Assumptions
  - Arrests are valid and appropriate (no bias, with proper reason)
  - All arrests are logged in the system properly
- Problems
  - Feedback loop (Police increase the frequency of patrol in neighborhood X => More arrests made in neighborhood X => New crime data fed back to model => …
  - What if the data about people arrested was tampered with

Applicant Screening

- Requirements (in the world)
    - Pick top 50 applicants from the applicant pool to send admit letters to for a Masters program
- Specifications (for the machine)
    - Given a list of documents for each candidate, recommend top 50 candidates who are most suited to the program by assigning a score to each candidate (return the ones with higher scores)
- Assumptions (about the world and shared phenomena)
    - ??
- Problems
    - ??

Applicant Screening

- Assumptions
  - Documents are authentic, and reflect reality
  - Application data are correct
  - Preferences of the department are known
  - All people who get admits will accept the offer
  - Department staff do not send offers outside the candidates recommended
- Feedback loop
  - We pick based on past candidate profiles => We end up having people with few backgrounds/profiles in the program (less diverse)

Thank You!