Capabilities for Better ML Engineering

Chenyang Yang, Rachel Brower-Sinning, Grace A. Lewis, Christian Kästner, Tongshuang Wu

Carnegie Mellon University

SafeAI

# Models' Safety Issues in Production Systems

Photos from sunny weather
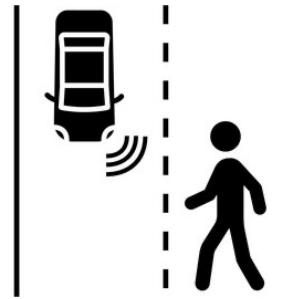
Pedestrian detection models

Models' Safety Issues in Production Systems

Test data
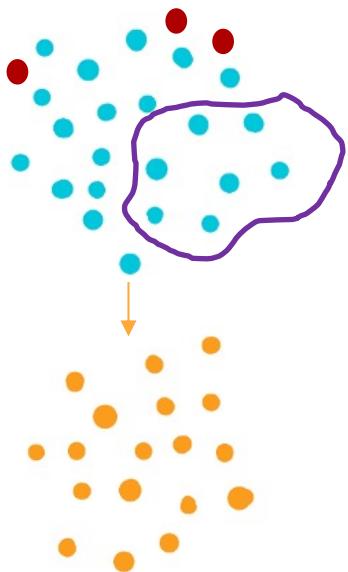
Distribution shift

Important outliers

Concept variations

We need to go beyond test data!

Beyond Accuracy: A Scattered Landscape

Model evaluation & data augmentation

Data slicing

Perturbations

Counterfactuals

...

Model qualities

Accuracy

Robustness

Fairness

Generalizability

...

Only models very specific kinds of phenomena/attack model

Capability: A Unifying Framework

Capabilities: decomposing requirements into **fine-grained specifications** of behaviors expected of an ML model

Detect pedestrians ➡️

Robust to extreme weather

Recognize wheelchair users

Fair to different age groups

Capability: A Unifying Framework

Detect pedestrians...

in extreme weather → **Robustness**

using wheelchairs

of different body sizes → **Generalizability**

in rural area

wearing costumes → **Fairness**

on a scooter

of different skin colors

Capability: A Unifying Framework

Detect pedestrians...

in extreme weather

using wheelchairs

of different body sizes

in rural area

wearing costumes

on a scooter

of different skin colors

Perturbations

Slicing

Counterfactuals

Capability: A Unifying Framework

Detect pedestrians…

in extreme weather

using wheelchairs

of different body sizes

in rural area

wearing costumes

on a scooter

of different skin colors

Use capabilities in…

model testing & debugging

data collection & documentation

model design & development

model documentation

model deployment

…

Do our data/model reflect the expected capabilities?

Capability: A Research Agenda

Identification → Assessment → Instantiation / Communication

Capability: Open Questions

Detect pedestrians…

in extreme weather

using wheelchairs

of different body sizes

in rural area

wearing costumes

on a scooter

of different skin colors

Identification → Assessment → Instantiation

Assessment → Communication
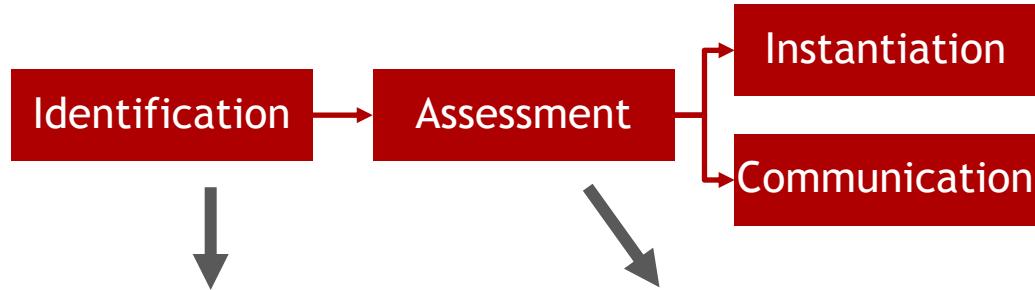
How to find capabilities? What capabilities should we care?

Requirements elicitation? Domain knowledge reuse? Human-AI interaction?

Capability: Open Questions

Detect pedestrians...

    in extreme weather

    using wheelchairs

    of different body sizes

    in rural area

    wearing costumes

    on a scooter

    of different skin colors

Identification → **Assessment** → **Instantiation** / **Communication**
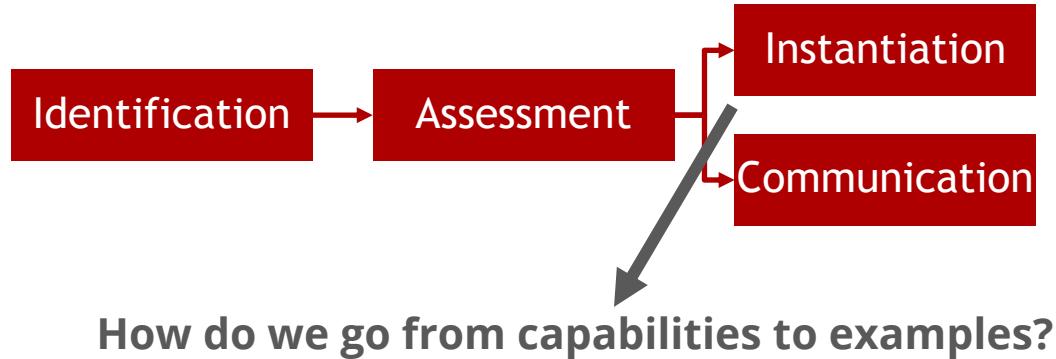
How do we go from capabilities to examples?

Strategies selection? Trade-offs?

Capability: Open Questions

Detect pedestrians...

in extreme weather

using wheelchairs

of different body sizes

in rural area

wearing costumes

on a scooter

of different skin colors

Identification → Assessment → Instantiation

Communication

How could capabilities be communicated across different stakeholders?

Language? Interface? Conflict resolution?

ML engineers, software engineers, users, regulation agencies...

Takeaways

Capability is a **unifying framework** for scattered work on **ML specifications**.

Capability is a **useful abstraction** to think about in **ML engineering**, especially in safety-critical systems.

Many open questions in using capabilities:

Identification → Assessment → Instantiation

Assessment → Communication

Detect pedestrians...

in extreme weather

using wheelchairs

of different body sizes

in rural area

wearing costumes

on a scooter

of different skin colors