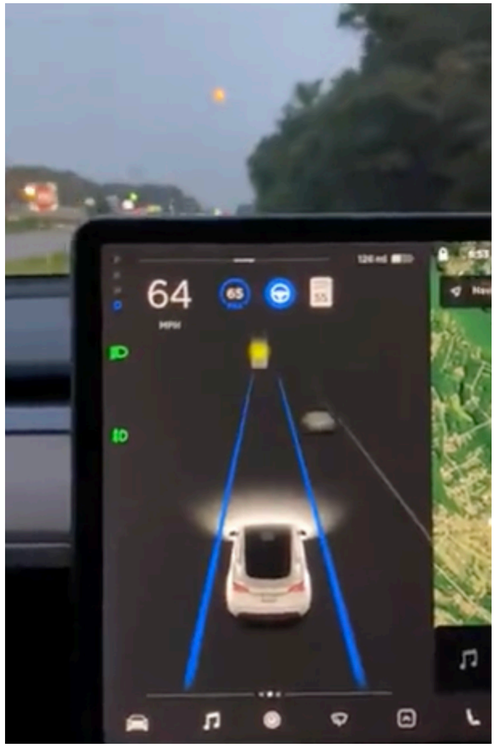
Accountability Layers

Stress-testing Using Explainable AI for Safety-critical Systems

Leilani H. Gilpin
Assistant Professor
Dept. of Computer Science & Engineering
UC Santa Cruz
lgilpin.com

Complex Systems Fail in Complex Ways

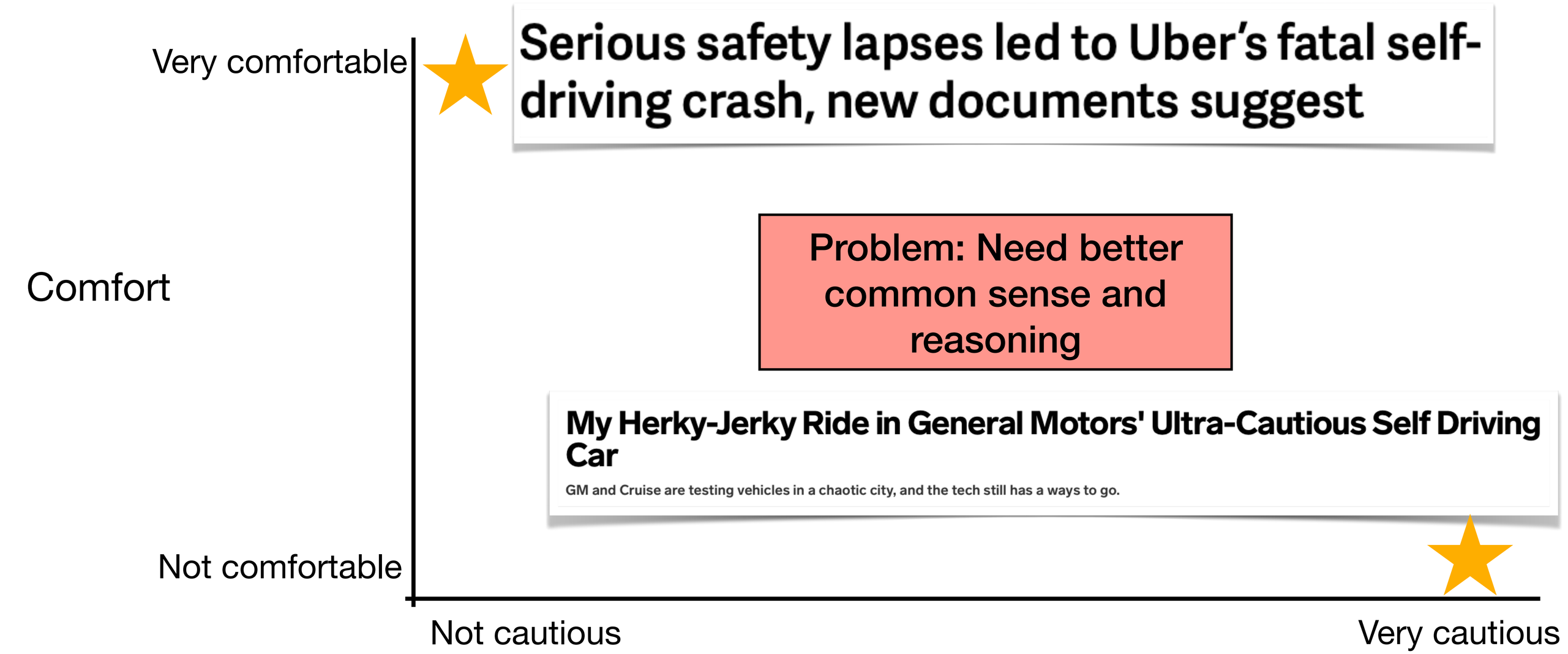
K. Eykholt et al. “Robust Physical-World Attacks on Deep Learning Visual Classification.”



Predictive Inequity in Object Detection

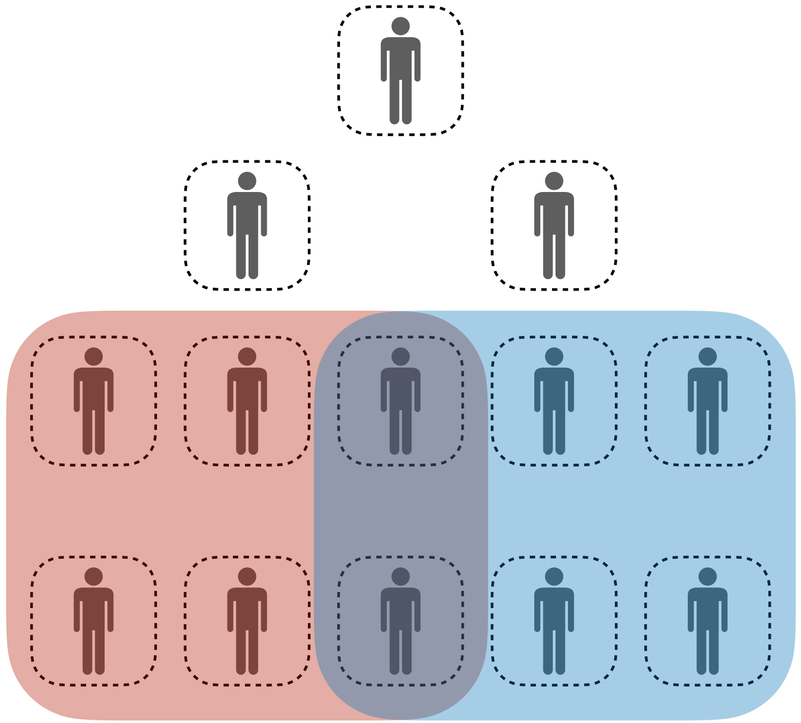
Benjamin Wilson [1] Judy Hoffman [1] Jamie Morgenstern [1]

Autonomous Vehicle Solutions are at Two Extremes



Architecture Inspired by Human Organizations

Communication and Sanity Checks



Local Sanity Checks

Synthesizer to reconcile inconsistencies between parts.

1. Hierarchy of overlapping committees.
2. Continuous interaction and communication.
3. When failure occurs, a story can be made, combining the members' observations.

“Explanations” in Layers