

# 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<sup>1</sup> Judy Hoffman<sup>1</sup> Jamie Morgenstern<sup>1</sup>

# Autonomous Vehicle Solutions are at Two Extremes

Very comfortable



**Serious safety lapses led to Uber's fatal self-driving crash, new documents suggest**

Comfort

**Problem: Need better common sense and reasoning**

**My Herky-Jerky Ride in General Motors' Ultra-Cautious Self Driving Car**

GM and Cruise are testing vehicles in a chaotic city, and the tech still has a ways to go.

Not comfortable



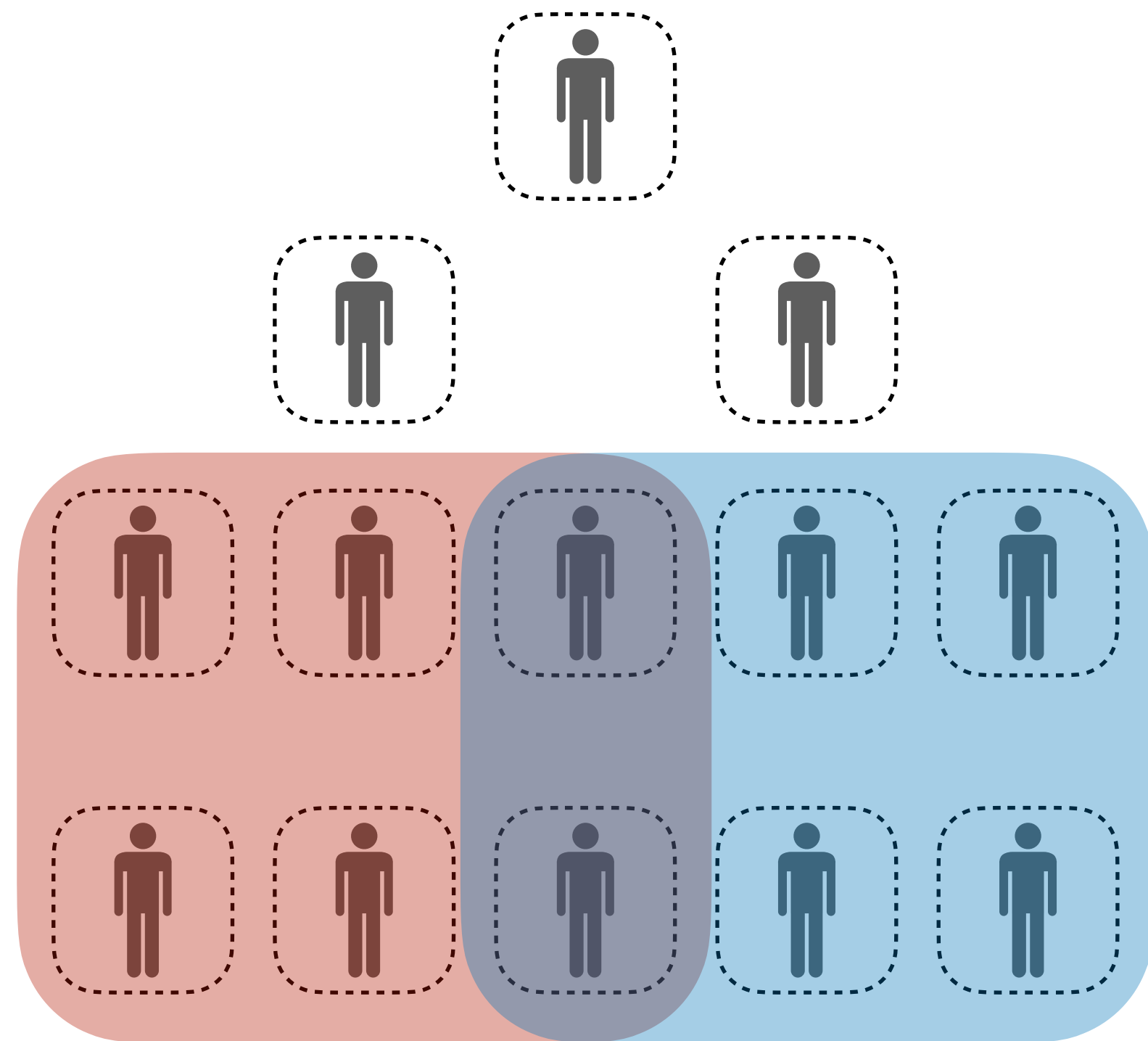
Not cautious

Very cautious

Cautious

# 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

