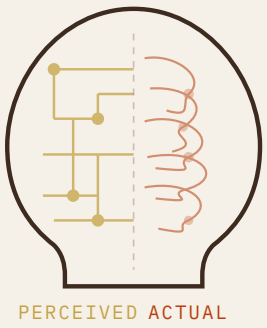
AI + ETHICS WORKSHOP · SPRING 2026

Why Are People Miscalibrated on AI?

How identity-protective cognition distorts our views

Teddy Wright & **Valen Cole** · University of Utah



01 DEFINING AI

What Are We Actually Forming Views About?

"AI" is not one thing. It is a set of capabilities entering distinct domains — labor, education, creative work, relationships, information systems, healthcare — each producing different effects. **Treating AI as monolithic guarantees miscalibration.** A person right about AI in one domain can be completely wrong in another. Further, AI is fundamentally input-output: outcomes are largely a function of the user and context it enters, not properties of the technology itself. Both hype and fear tend to misattribute to AI what belongs to the human system.

02 THE IMPORTANCE

Why It Matters

AI is restructuring how we work, learn, create, and communicate. Miscalibration costs in both directions: **overestimating** capabilities leads to over-reliance and hype-driven policy; **underestimating** leads to missed opportunities and failure to prepare. When your model of AI doesn't match reality, your decisions don't match the world you're living in.

03 A MODEL FOR VIEW FORMATION

How Views Form

AI challenges the predictive models people rely on to navigate the world — models of how value is created, how learning works, what expertise means. These models are entangled with **identity, livelihood, and wellbeing**, making them costly to update. When updating is costly, people optimize for stability over accuracy, distorting both the information they seek and how they process it.

View On AI = *Optimization Target* (Information Quality × Processing Quality)

Multiplicative — failure in any component compromises the whole. Identity preservation as the target systematically distorts both inputs.

THE CAUSAL CHAIN

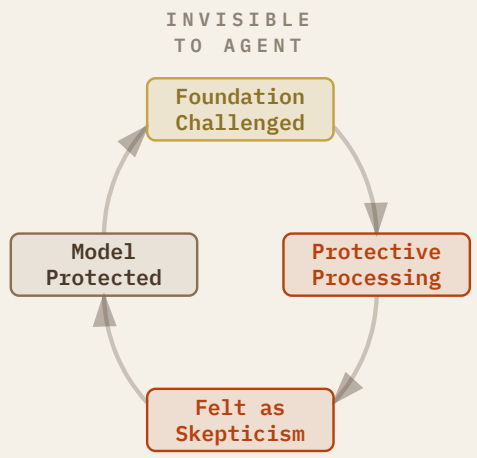
AI disrupts predictive models → Updating is costly *(identity · livelihood · wellbeing)* → Optimize for stability over accuracy → Distorts seeking & processing → Miscalibrated views

04 THE HARD PROBLEM

The Self-Concealing Property

The optimization problem has a feature that makes it uniquely hard to fix: **it is invisible to the person running it.** Identity-protective cognition doesn't feel like bias. It feels like rigor. The professor dismissing AI capability data isn't thinking *"I'm protecting my identity"* — they're thinking *"I'm being appropriately skeptical."* These two states produce identical felt experiences.

> *You cannot introspect your way out because the introspection itself runs on the compromised optimization target. Smarter people aren't immune — they produce more sophisticated rationalizations.*



05 INTERVENTIONS

What To Do About This

If introspection is compromised, corrections must be structural.

01 — Evaluate Methods, Not Just Outputs
Assess work against criteria without knowing whether AI was involved.

02 — Interrogate Comfort
When your analysis confirms your existing model, that is precisely when to apply scrutiny.

03 — Interpersonal Calibration
Someone with a different identity investment can see your blind spots. Build disagreement into process, not just culture.

04 — Decompose Before Evaluating
Break "AI" into specific capabilities and domains before forming any view.