Proxy Metrics, Evaluative Capacity, and the Hidden Costs of AI Engagement

Toward a Theory of How Organizations Lose Judgment While Gaining Productivity

Why do experienced practitioners sincerely believe AI is improving their work when independent measurement reveals persistent gaps between perceived and actual outcomes, and why does this misjudgment resist correction through expertise, feedback, or organizational learning?

Vikram Bapat, Ecosystems, Platforms & Strategy Research Group, Institute for Manufacturing, University of Cambridge. Supervisor: Dr Florian Urmetzer.

EXISTING ACCOUNTS

Each framework explains something. None reaches the mechanism.

Institutional logics (Thornton et al. 2012)
Explains how competing logics restructure what counts as success. Does not ask whether the technology constitutes the confirming evidence that selects between logics.

Sociomateriality (Orlikowski 2007)
Recognizes mutual constitution of social and material. Assumes practitioners can evaluate their own reconstitution using criteria unaffected by that reconstitution.

Dynamic capabilities (Teece 2007)
Assumes the sensing apparatus survives reconfiguration.

Goodhart's Law (Chrystal & Mizen 2003)
Explains metric corruption through strategic gaming. The 39-point METR gap persisted through sincere belief, not optimization.

INTEGRATION GAP

No existing theory integrates all four:

1. Technology constitutes new observables more legible than what the organization is accountable to
2. Engagement transforms the evaluator such that pre-engagement criteria become structurally unreliable
3. Institutional feedback loops reinforce proxy optimization through competing logic asymmetry
4. Field-level discourse constitutes the evaluative frame before any organization tests it operationally

The evaluative continuity assumption persists because no account has specified why, mechanistically, it should fail.

THEORETICAL ARCHITECTURE

Four resources, sequentially linked. The contribution is the integration.

Performativity (MacKenzie 2006, Callon 2007)
Constitutes which criteria are legitimate before engagement begins. Field-level discourse performs proxy metrics into evaluative vocabulary.

Form/function (Faulkner & Runde 2019)
Fertile form enables positioning drift. The same AI tool produces different proxy-criterion gaps in different organizations.

Transformative experience (Paul 2014)
The practitioner who set pre-engagement criteria is not the practitioner evaluating post-engagement outcomes.

Institutional logics (Thornton et al. 2012)
Market logic makes proxy metrics legible. Professional logic makes accountable criteria legible. AI engagement systematically elevates one while occluding the other.

AI engagement may constitute attractive proxy metrics that displace the criteria organizations are accountable to, while eroding the evaluative capacity needed to detect the substitution.

If this holds, the mechanism operates through sincere belief, not strategic gaming. This is the constitutive distinction from Goodhart's Law.

THE PATTERN: PATTERNED, PERSISTENT, RESISTANT TO EXPERTISE

39pp
Perception-reality gap
Developers perceived 20% speedup. Objective measurement: 19% slower. Sincere belief, not gaming.
METR 2025 (RCT, N=16)

75%
Scaffolded, not internalized
AI closed 75% of education gap during execution. Gap reappeared in full when AI removed.
Cruces et al. 2026 (NBER)

94%
Idea overlap
AI-generated ideas: 94% overlap. Human ideas: 100% unique. Individual quality up, collective diversity collapsed.
Meincke, Collins & Evans 2025

Null
Effect at population scale
Widespread adoption, confident practitioners. Administrative records: null effects on earnings and hours.
Humlum & Vestergaard 2025

+34%
Novices gain, experts drift
Novice agents +34%. Expert agents: negligible gain, quality decline, increased AI adherence over time.
Brynjolfsson et al. 2025 (N=5,172)

90/14
Confidence decoupled from outcomes
90% of daily AI users confident. 14% achieve consistently positive outcomes. 37% of time saved lost to rework.
Workday 2026 (N=3,200)

RESEARCH STAGE

THEORY (CURRENT)
Theoretical framework in development
Integrating four theoretical resources with a curated evidence constellation across three levels of analysis. Full literature synthesis complete across six traditions. Paper drafted and under revision.

EMPIRICAL (NEXT)
Semi-structured interviews
Two populations: frontline technology practitioners across professional domains to surface criteria shift, and boundary activity performers to surface translation work. Frontline first.

FALSIFICATION
Diagnostic, not predictive
If proxy-criterion divergence self-corrects without intervention, braking fails as theorized. If the perception gap closes with experience, evaluator transformation is disconfirmed.

Failure is self-announcing. Degradation is self-concealing. If the framework holds, the mechanism that produces the pattern is the same mechanism that conceals it.