The Augmentation Architecture Human + Machine Intelligence

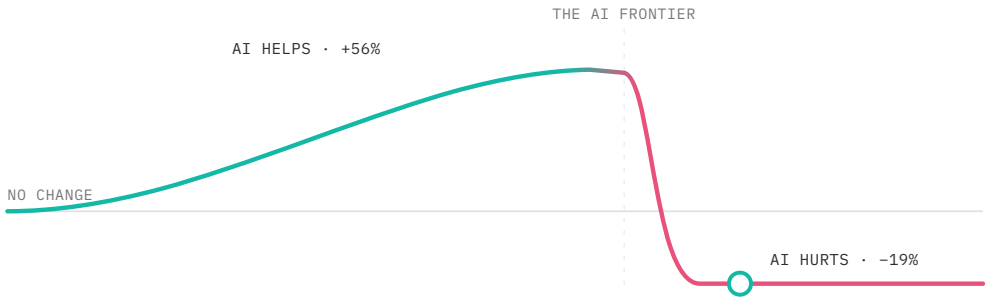
James E. Coleman

STRATEGIC DESIGN · WHITE PAPER SERIES · 2025

ABSTRACT *Generative and predictive AI reshape the intelligence gap, the translation problem, and the trust deficit all at once. Advantage does not accrue to whoever automates the most — it accrues to organizations that design for complementarity, augmenting human judgment rather than substituting for it. Automate-versus-augment is an architecture choice, and its consequences compound.*

FIG. 1 — AI LIFTS YOU — UNTIL IT DROPS YOU BEYOND THE FRONTIER: −19%

Follow the dot. On work AI understands, performance **climbs** — up to +56%. The moment it crosses **the frontier** of what it knows, it falls off a cliff: **19% more wrong**.



A TOOL THAT CUTS BOTH WAYS



40%

On professional writing, generative AI cuts the time by about 40%.[1]



14%

Support agents resolve 14% more issues an hour with it at their side.[2]



55.8%

Developers using Copilot finish tasks 55.8% faster.[3]



−19%

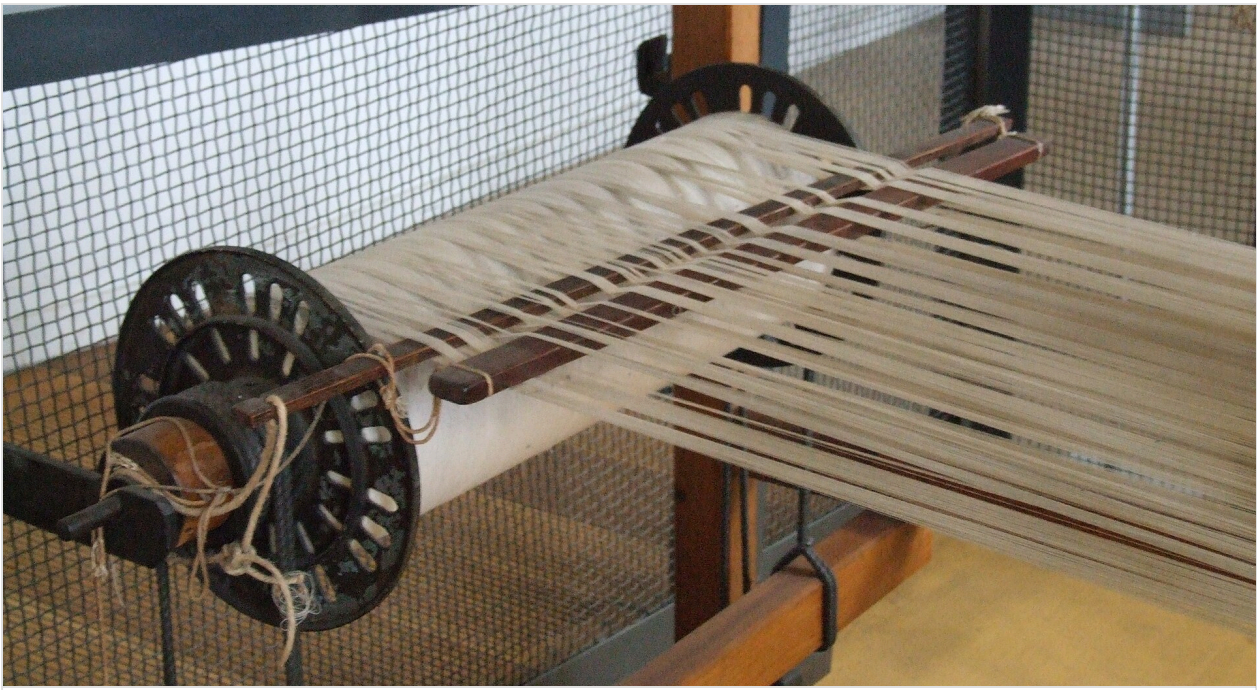
But **past the frontier** of what the model actually knows, the same tool leaves people 19% *less* likely to be right.[4]

§ 01 The Augmentation Choice

Substitution or complementarity. *The same tool, two architectures.*

A generative model dropped into a workflow can be pointed in one of two directions. It can be used to substitute for human work — to replicate a task a person used to perform and remove the person — or to augment it, taking over the parts machines do well so that human attention concentrates on the parts they do not. The technology is identical in both cases. The architecture is not, and the architecture is where the value and the power end up.

Erik Brynjolfsson calls the default drift toward pure substitution the *Turing Trap*: because our benchmark for machine intelligence is human-like performance, we reflexively build systems that imitate and replace people rather than extend them.[5] The trap is not that automation fails — it is that it succeeds narrowly while forfeiting the larger prize. Systems designed to match human capability compete with labor and concentrate the gains among those who own the machines; systems designed to augment create new capability and distribute the gains among those who wield them.



A Jacquard loom — the first machine run by stored instructions, and the ancestor of human-plus-machine work.

Photo: Clem Rutter · CC BY 3.0

> *"The threat of mass unemployment is overblown. The real risk is more subtle: an excessive focus on automation. When machines merely substitute for human labor, they do little to grow the economic pie — while concentrating wealth and power in fewer hands."*
>
> — BRYNJOLFSSON, "THE TURING TRAP," DAEDALUS 2022

This reframes the three problems the earlier papers in this series set out. The intelligence gap (WP 1.0), the valley of death between discovery and deployment (WP 2.0), and the speed-versus-trust tension (WP 3.0) are each, in part, a story about where human judgment sits relative to machine capability. AI does not resolve these problems by removing people from them. It resolves them only where it is architected to raise the ceiling on what people can judge, translate, and trust — and it worsens them wherever it quietly removes the human from a loop that still needed one.

§ 02 What Machines Predict, What Humans Judge

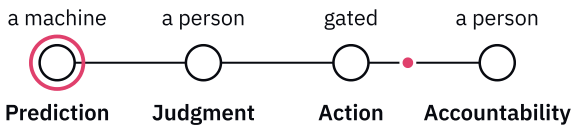
AI makes prediction cheap. *It does not make judgment free.*

The most durable framing of what current AI actually does comes from Agrawal, Gans, and Goldfarb: machine intelligence is, economically, a drop in the cost of *prediction* — using available information to fill in information you don't have.[6] When the price of anything falls, we use more of it, and we use more of its complements. The complement to prediction is judgment: the specification of what we are trying to achieve, the relative value of different outcomes, and who is accountable when a prediction is wrong.[7]

A decision is not a single act. It decomposes into components, and the augmentation choice is really the choice of which components to hand to the machine. Done well, the split is not a compromise — cheaper prediction makes human judgment *more* valuable, not less, because judgment is now the binding constraint on decision quality.

FIG. 5.1 — ANATOMY OF A DECISION

STATE 1 OF 4



WHAT WILL HAPPEN?

Prediction

What will happen, and what is the missing information? Forecasting demand, classifying an image, estimating risk, drafting a first pass, retrieving the relevant precedent. Falling in cost, improving in quality, and increasingly the machine's work.

WHAT DO WE WANT?

Judgment

What do we actually want, and what are the trade-offs between outcomes? Setting the objective and the payoffs a prediction is evaluated against. Irreducibly human in any domain where values, not just facts, are contested.

COMMITTING RESOURCES

Action

Committing resources on the basis of prediction plus judgment. Can be automated where stakes are low and reversibility is high; should stay gated by a human where neither holds.

WHO ANSWERS FOR IT?

Accountability

Who answers for the outcome, and to whom? Cannot be delegated to a model. A system that automates the action while leaving accountability unassigned has not removed the risk — it has only hidden it.

A diagram of the relation, not a chart of data. A decision in four parts, in the order they happen. The word above each part says who does it: a machine, a person, or either, gated by the stakes.

The design error is to treat a task as atomic — fully automatable or fully manual — when almost every task is a bundle of prediction, judgment, action, and accountability in different proportions. The augmentation architecture disaggregates the task first, then assigns each component to whichever party holds the advantage.

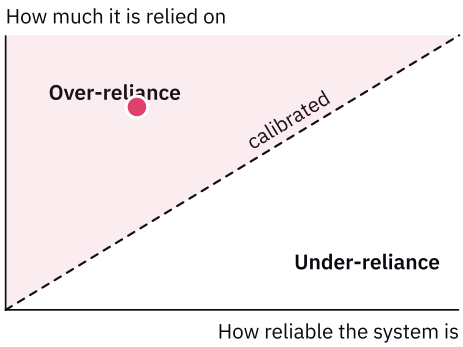
Trusting the machine *is itself a trust problem.*

WP 3.0 argued that trust between parties is an accumulated asset that cannot be compressed below certain thresholds. Augmentation adds a second trust relationship inside the workflow: the human operator's trust in the machine's output. That relationship has the same failure modes as any other — and getting it wrong degrades the very decisions augmentation was meant to improve.

The human-factors literature has studied this for decades under the heading of *appropriate reliance*. Trust in automation should be calibrated — matched to the system's actual reliability in the current context — rather than uniformly high or low.[8] Two symmetric failures follow from miscalibration:

FIG. 5.2 — FAILURE MODES OF RELIANCE

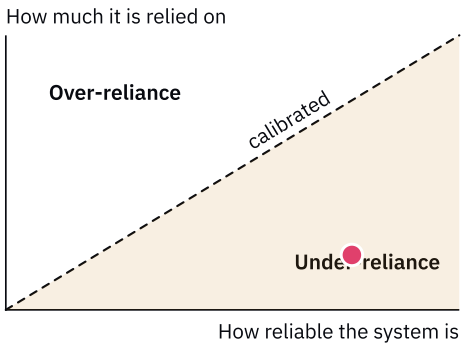
STATE 1 OF 3



AUTOMATION BIAS

Over-reliance

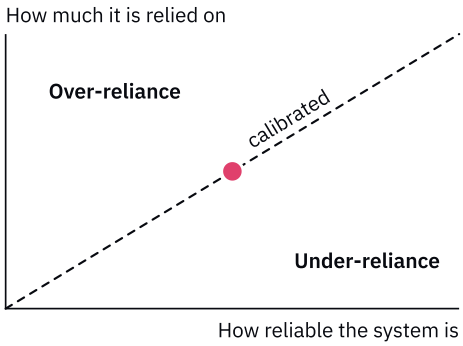
Automation bias: accepting machine output when it is wrong, and failing to notice because the machine was consulted instead of the evidence. Manifests as errors of commission (following a bad recommendation) and omission (missing what the system didn't flag). Worst where output is fluent and confident but unverifiable at a glance.



DISUSE

Under-reliance

Disuse: discounting or disabling a system that is, in fact, more reliable than the human — usually after a salient early failure eroded trust faster than performance warranted. Forfeits the gains entirely.



RELIANCE TRACKS RELIABILITY

Calibration

Reliance tracks reliability. Requires that the system be *legible* — that it expose its confidence, its reasoning, and the boundaries of its competence — so the operator can tell a trustworthy answer from a plausible one.

A diagram of the relation, not a chart of data. The dashed line is where reliance matches reliability. Above it, people lean on the system more than it deserves. Below it, less.

> *"Trust guides reliance when the complexity of the automation makes a complete understanding impractical. Inappropriate reliance — misuse and disuse — follows from miscalibrated trust: trust that exceeds, or falls short of, the automation's true capabilities."*
>
> — LEE & SEE, "TRUST IN AUTOMATION," HUMAN FACTORS 2004

Legibility is therefore not a user-experience nicety; it is the mechanism that makes calibrated reliance possible. A model that cannot communicate *when it should not be believed* forces the operator into one of the two failure modes. Designing for complementarity means designing the interface that lets a human calibrate — the same interface-layer discipline WP 3.0 identified for trust between parties, now turned inward on trust in the tool.

The stakes also set the standard of evidence. The productivity gains cited above come largely from knowledge work — writing, coding, customer support — where a wrong answer is usually cheap to catch and reverse. In regulated, high-consequence domains such as clinical medicine, the same augmentation carries obligations those studies do not test: a clinical decision-support tool is frequently regulated as Software as a Medical Device, is expected to demonstrate safety and effectiveness through prospective validation, and operates where an automation-bias error can harm a patient rather than a document. The complementarity thesis holds, but the burden of proof scales with the cost of being wrong — and a deployment pattern validated on knowledge work does not transfer unexamined to the clinic.

§ 04 Designing the Human-in-the-Loop

Where the human sits *decides what you get.*

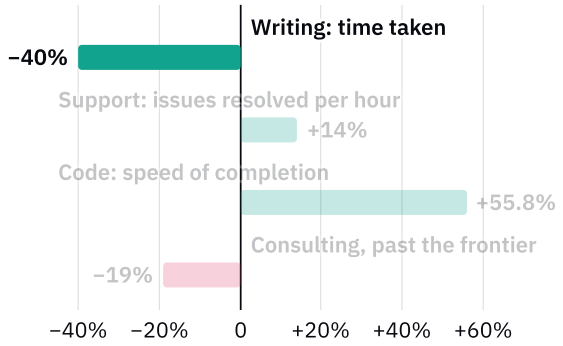
The empirical evidence on generative AI at work is genuinely encouraging — and genuinely uneven, in ways that only make sense through the augmentation lens. In a controlled writing experiment, pro-

fessionals using ChatGPT finished tasks 40% faster and produced work rated 18% higher in quality, with the largest gains going to initially weaker writers and overall dispersion falling.[1] Customer-support agents with an AI assistant resolved 14% more issues per hour — rising to 34% for the least experienced agents, as the model diffused the tacit know-how of the best performers to everyone else.[2] Software developers completed a task 55.8% faster with an AI coding assistant.[3]

A consistent signal runs through these studies: the gains concentrate among novices and lower performers, which is exactly what complementarity predicts — the machine supplies the prediction floor, and human judgment is freed to operate above it. But the same body of evidence carries a warning. In a field experiment with 758 consultants, those using GPT-4 were sharply more productive on tasks *inside* the model's capabilities — yet on a task deliberately placed just *beyond* that boundary, AI-assisted consultants were 19% less likely to reach the correct answer than those working without it.[4] The authors call this boundary the *jagged technological frontier*: capability is uneven and unmarked, and the operator often cannot see which side of it a given task falls on.

FIG. 5.3 — EVIDENCE, AND ITS JAGGEDNESS

STATE 1 OF 4



NOY AND ZHANG

Professional writing

−40% time, +18% quality on professional writing; gains largest for weaker writers, dispersion narrowed. Augmentation lifts the floor.[1]

BRYNJOLFSSON, LI AND RAYMOND

Customer support

+14% issues resolved per hour, +34% for novices; the model propagated the practices of top agents to the rest.[2]

PENG ET AL.

Software development

+55.8% faster task completion for developers using an AI pair-programmer.[3]

DELL'ACQUA ET AL.

Consulting, beyond the frontier

Large gains inside the frontier; −19% likelihood of a correct answer on a task beyond it. Uncritical reliance imports the model's errors wholesale.[4]

Four field studies, one number each, as reported. Green is a gain and red is a loss. The writing result is a fall in time taken, which is a gain. The consulting result is a fall in correct answers, on a task beyond what the model could do.

The design implications follow directly. First, place the human where the frontier is jagged — at the tasks where the model's competence is uncertain and the cost of a confident error is high — and let automation run where the frontier is settled. Second, guard against skill atrophy: if the machine always supplies the prediction, operators lose the expertise needed to catch it when it is wrong, and the whole system slides from augmentation into unmonitored substitution. The human-in-the-loop must stay a *practicing* human, not a rubber stamp. Third, measure complementarity, not adoption: the question is never whether AI was used, but whether the human-plus-machine system outperforms either alone on the outcomes that matter — and on the tasks the frontier makes treacherous.

> *"Intelligence, translation, and trust are accumulated assets. Augmentation compounds them; substitution spends them. The organizations that endure will be those that treat the human-in-the-loop not as a transitional inefficiency to be engineered away, but as the designed seat of judgment the whole architecture is built to protect."*
>
> — JAMES E. COLEMAN, WP 5.0

↳ The series thruline

The gap between knowledge and action is an infrastructure problem, not a resource problem — crossed by systems designed around the people who do the crossing. Here the gap runs from human effort to human-plus-machine capability. The same move recurs across the series: build the system the people who do the crossing currently lack.

§ R References — Primary Sources

Sources *& notes.*

Every quantitative claim in this paper is cited to a primary source and each link below was verified to resolve. The four hero figures (−40% time, +14% issues/hour, +55.8% developer speed, −19% correct beyond the frontier) are drawn directly from the controlled studies cited; no figure on this page is a modeled estimate. Where a study was later published in a peer-reviewed venue, both the working-paper and the journal identifiers are given.

1. [1]Noy S, Zhang W. “Experimental evidence on the productivity effects of generative artificial intelligence.” *Science* 2023; 381(6654): 187–192. Professional writing tasks completed ~40% faster with ~18% higher rated quality; gains largest for lower-ability writers. doi:10.1126/science.adh2586.

2. [2]Brynjolfsson E, Li D, Raymond LR. “Generative AI at Work.” *NBER Working Paper* 31161, 2023 (rev. 2023); forthcoming *Quarterly Journal of Economics*. Access to an AI assistant raised issues resolved per hour ~14% on average and ~34% for novice agents. nber.org/papers/w31161.

3. [3]Peng S, Kalliamvakou E, Cihon P, Demirer M. “The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.” *arXiv preprint* arXiv:2302.06590, 2023. Developers with Copilot completed the programming task 55.8% faster than the control group. arxiv.org/abs/2302.06590.

4. [4]Dell'Acqua F, McFowland III E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR. “Navigating the Jagged Technological Frontier.” *Harvard Business School Working Paper* 24-013, 2023; published in *Organization Science* 2025. Large within-frontier gains; on a task beyond the frontier, AI-assisted consultants were ~19% less likely to produce correct solutions. doi:10.1287/orsc.2025.21838.

5. [5]Brynjolfsson E. “The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence.” *Daedalus* 2022; 151(2): 272–287. Argues that benchmarking AI against human-like performance biases design toward substitution, concentrating wealth and power, rather than augmentation. doi:10.1162/daed_a_01915.

6. [6]Agrawal A, Gans J, Goldfarb A. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, 2018. ISBN 978-1-63369-567-2. Frames machine intelligence as a fall in the cost of prediction that raises the value of its complement, human judgment.

7. [7]Agrawal A, Gans JS, Goldfarb A. “Prediction, Judgment and Complexity: A Theory of Decision Making and Artificial Intelligence.” *NBER Working Paper* 24243, 2018. Formalizes the decomposition of decisions into prediction (machine) and judgment (human). nber.org/papers/w24243.

8. [8]Lee JD, See KA. “Trust in Automation: Designing for Appropriate Reliance.” *Human Factors* 2004; 46(1): 50–80. Defines calibrated trust and appropriate reliance; links miscalibration to automation misuse (over-reliance) and disuse (under-reliance). doi:10.1518/hfes.46.1.50_30392.

SERIES CLOSING STATEMENT

"Intelligence and trust are both accumulated assets. They compound when invested consistently and depreciate rapidly when neglected. The organizations that will define the next era of economic activity are those that treat both as core infrastructure — not as emergent properties of scale, but as designed systems, maintained with intention, evaluated against outcomes."