Beyond Performance

A Diagnostic Method for Evaluating Human-Advancement Claims in AI-Mediated Systems

Micheal Charles Preble
Independent Researcher
Manuscript v2.3 Final

Abstract

Claims that artificial intelligence has improved human performance increasingly appear in education, professional work, public administration, and other domains. Yet improvement in an AI-mediated system does not by itself establish that the human participant has advanced. A person may produce stronger outputs while losing capacities that remain necessary for meaningful agency in the altered practice, while formal oversight may persist after practical authority has migrated elsewhere. At the same time, a reduction in unaided performance does not automatically establish human decline, because reliable external systems can expand genuine human capability. The methodological problem is therefore inferential: what evidence is required before a claim about improved AI-mediated performance can support a stronger claim about human advancement?

This paper develops a diagnostic method for evaluating such claims. The method begins from three configurations-autonomous capability, governed extended capability, and ungoverned dependency-and uses a three-pattern diagnostic to classify observed change as coupled improvement, performance-only acceleration, or regressive decoupling. Each progress claim must declare, before evaluation, the human functioning at issue, the relevant population and role, the scope of the claim, the required contrast, the durability window, and the failure signs that would defeat the claim. The method then evaluates five human-side conditions: governance-critical capacity, authority in practice, governability and continuity, distribution, and time.

The method is applied to published educational evidence. Bastani et al. (2025) report a 48% improvement on AI-assisted mathematics practice in a standard GPT condition alongside a 17% decline on the later unassisted examination; a guarded tutor removed the penalty but did not produce a positive unassisted learning gain over control. Strömberg, Lei, and Wu (2026) report homework scores rising by 18% and homework time falling by 30% after AI adoption, while closed-book examination performance declined by 20% within six months, with losses concentrated among the majority of users whose behavior was consistent with greater homework outsourcing. Under a declared transfer-beyond-assistance criterion, these cases support an assisted-performance claim but do not support a general claim of improved human learning.

The paper does not propose a universal index or a validated instrument. Its contribution is a pre-validation decision procedure for determining whether a progress claim succeeds, remains unproven, or fails given the evidence available.

Keywords: human-AI interaction; human advancement; progress claims; AI-assisted performance; retained capacity; meaningful control; governability; longitudinal evaluation; human-AI teaming

1. From Performance Claims to Human-Advancement Claims

AI evaluation routinely begins with output. Researchers measure accuracy, speed, quality, completion, productivity, or user performance with and without artificial assistance. These measures are necessary when the question concerns whether a tool or human-AI configuration performs well. They become insufficient when the conclusion shifts from a proposition about system performance to a proposition about the human participant.

The inferential problem can be stated narrowly. An AI-mediated configuration may improve while the human participant becomes more capable, less capable, or differently capable in ways that the artifact alone cannot reveal. Human-AI research already provides strong reasons to reject simple performance assumptions. Vaccaro, Almaatouq, and Malone (2024), in a meta-analysis of 106 experiments and 370 effects, found that human-AI combinations performed worse on average than the better of human or AI alone, with important task-level variation. Work on automation likewise shows that performance gains can coexist with out-of-the-loop problems, reduced practice, and weakened capacity to intervene when abnormal conditions arise (Bainbridge, 1983; Parasuraman et al., 2000).

The opposite mistake is equally important. A decline in unaided performance does not automatically establish human regression, because technological extension can create genuine practical capability that no individual could reproduce independently. The relevant distinction is therefore not assistance versus independence, but the condition of the human within the altered practice. A method for evaluating human advancement must be able to distinguish productive technological extension from performance substitution and from fragile or ungoverned dependency.

The method developed here evaluates claims, not persons in the abstract. Its object is inferential validity: whether the evidence supports moving from the proposition that an AI-mediated configuration improved to the stronger proposition that a human or population advanced. The procedure is designed to return one of three dispositions-succeed, unproven, or fail-relative to a declared claim and evidence set.

2. Configurations and the Diagnostic Patterns

The method begins with three configurations that describe how capability may be arranged. Autonomous capability refers to a function the person can perform without live external assistance. Governed extended capability refers to a function materially expanded by AI while the person retains the capacities and practical powers needed to direct, interpret, contest, and sustain the arrangement. Ungoverned dependency refers to high or improved functioning that depends on an external system the person cannot meaningfully evaluate, challenge, substitute, or recover from when those powers remain necessary in the altered practice.

These configurations do not establish a moral ranking by themselves. A person can legitimately become more dependent on reliable technology, and a reduction in unaided skill can be an acceptable consequence of specialization or technological progress. What matters is whether the practice still preserves the human capacities and powers that the domain treats as necessary for meaningful agency, responsibility, or participation.

Observed change is classified using three primary patterns. Coupled improvement occurs when AI-mediated performance rises and the relevant human-side conditions improve or remain robust. Performance-only acceleration occurs when output improves but the human-side evidence is flat, absent, or insufficient. Regressive decoupling occurs when AI-mediated performance rises while a declared human-side condition materially deteriorates.

A separate residual category, human conversion gain, may be useful in future work for cases in which human outcomes improve substantially without an equivalent increase in underlying frontier-model capability. It is not part of the three-pattern performance/human-side diagnostic because it tracks a different relationship: human gain relative to model-capability growth rather than human trajectory relative to AI-mediated performance trajectory. The present paper does not exercise that residual category.

The diagnostic is intentionally prior to measurement aggregation. It does not assign a universal score to human advancement. Instead, it structures the decision by asking whether the human trajectory and the AI-mediated performance trajectory are actually coupled in the dimension the claim invokes.

3. The Evaluation Protocol

Before measurement begins, the claimant must specify six elements. First, the human functioning or freedom at issue must be named, such as learning, professional judgment, ability to contest an administrative decision, or effective access to a service. Second, the population and role must be identified, because gains for managers, students, citizens, clinicians, or workers cannot be assumed to generalize across roles. Third, the scope of the claim must be declared, distinguishing a task-performance claim from a claim about learning, control, organizational capability, or broad social progress. Fourth, the contrast needed to support the claim must be specified, such as tool-on versus tool-off, delayed transfer, withdrawal, provider switching, or role-disaggregated comparison. Fifth, a durability window must be chosen that is appropriate to the claim. Sixth, failure signs must be declared in advance so that the method does not redefine success after the results are known.

The human-side evidence is organized into five conditions. Governance-critical capacity concerns the human capacities necessary to direct, interpret, verify, contest, or responsibly rely upon the AI-mediated process within the altered practice. Authority in practice concerns whether the human can actually change objectives, reject recommendations, escalate uncertainty, or alter consequential outcomes under real working conditions. Governability and continuity concern whether the external system can be meaningfully understood at the level required by the domain, corrected, challenged, substituted, exited, and recovered from. Distribution and role concern who receives the gains and who receives new error burdens, dependency, or loss of discretion. Time concerns whether the arrangement remains beneficial as skill, reliance, switching costs, and institutional structure change.

These conditions are not co-equal requirements for every claim. A narrow assisted-performance claim may require none of them beyond output evidence, while a human-learning claim requires a human-side capacity contrast and a meaningful time horizon. A meaningful-control claim requires observed evidence of authority rather than documentary rights alone. A broad social-progress claim requires distributional and longitudinal evidence that would be unnecessary for a narrow task claim.

The procedure can be summarized as a simple sequence: declare the claim → specify the required human-side evidence → classify the observed relationship → return succeed, unproven, or fail. The method therefore uses the following decision table.

Claim type	Required contrast	Human-side minimum	Fail if	Unproven if
Assisted performance	Treated vs. control, or tool-on vs. tool-off	None beyond the stated output measure	The declared output measure does not improve	No valid performance contrast
Human learning or skill	Delayed unaided, transfer, or equivalent domain-valid measure	Governance-critical capacity relevant to the declared learning objective	Assisted performance rises while the declared human capability materially declines	No delayed or transfer measure
Meaningful control	Observed override, escalation, correction, or decision-right evidence	Authority in practice plus relevant governability	Responsibility remains human while practical influence over the decision becomes materially ineffective	Only formal or documentary rights are measured
Claim of governed extension	Assisted function plus continuity or recovery contrast appropriate to the domain	Capacity sufficient for the altered practice, authority, and governability	Functioning depends on a system the person cannot meaningfully evaluate, correct, substitute, or recover from where those powers are necessary	Continuity and governability are not tested
Broad social progress	Population- and role-disaggregated outcomes over a declared window	Distribution and time, plus any domain-specific human-side condition	Mean improvement coexists with material deterioration in a relevant subgroup or role that contradicts the scope of the claim	Only aggregate or short-term evidence is available

A claim succeeds when the required contrast supports the stated human-side outcome and no predeclared failure condition is triggered. A claim is unproven when the available design cannot answer the human-side question being

asked, even if system performance improves. A claim fails when evidence directly contradicts the declared human-side criterion or when a veto condition necessary to the claim is triggered. This procedure does not require a composite score because the disposition follows from the relationship between the claim, the required evidence, and the observed failure signs.

4. Worked Application: AI and Secondary-School Learning

The source studies already report the empirical split between assisted task performance and later unassisted performance; the method's contribution is not to rediscover that split, but to prevent an assisted-performance result from being generalized into a human-learning claim that the study's own human-side evidence does not support. Education is a useful test because the difference between assisted performance and learning can be measured directly. The relevant claim must nevertheless be declared before the evidence is interpreted. In the present application, human learning is defined as improvement that transfers beyond the live assistance context, as evidenced by delayed or unassisted performance on the relevant mathematics tasks. This is not the only defensible conception of learning in an AI-rich environment, but it is a legitimate one, and the method requires that such a criterion be stated rather than smuggled in after the results are known. For these cases, material deterioration is defined directionally rather than by a universal effect-size threshold: assisted performance moving upward while the predeclared transfer measure moves downward is sufficient to trigger failure of the human-learning claim.

Bastani et al. (2025) provide the first scored case. In a preregistered field experiment involving nearly one thousand high-school mathematics students, access to a standard GPT-4 interface improved performance on assisted practice problems by 48%. On the later unassisted examination, however, students in that condition scored 17% below the no-AI control. Under the method developed here, the claim "standard GPT assistance improved practice performance" succeeds, because the relevant assisted-performance contrast is positive. The stronger claim "standard GPT assistance improved human learning" fails under the declared transfer-beyond-assistance criterion because the required human-side contrast moved in the opposite direction.

The guarded tutor condition receives a different disposition. Bastani et al. report that a teacher-grounded tutor with learning safeguards removed the negative examination penalty but did not produce a positive unassisted learning advantage over the control group. The claim "the guarded tutor avoided the measured learning penalty" succeeds narrowly on the reported examination outcome. The claim "the guarded tutor improved human learning" remains unproven, because eliminating a negative transfer effect is not equivalent to demonstrating a positive learning gain.

Strömberg, Lei, and Wu (2026) provide a second instantiation at larger scale in a 30-month working paper following 26,811 Chinese secondary students. After AI adoption, homework scores rose by 18% and homework time fell by 30%, while closed-book examination scores declined by 20% within six months. The authors further report that the losses were concentrated among the roughly 80% of users whose time-and-score pattern was consistent with greater homework outsourcing, while students who maintained homework time closer to the non-AI baseline experienced smaller losses.

Under the same declared learning criterion, the claim "AI adoption improved homework performance" succeeds as an assisted-performance claim. The broader claim "AI adoption increased student capability" fails for the outsourcing-majority pattern because the human-side transfer measure deteriorated while the assisted artifact improved. For the time-maintaining minority, the claim is unproven under the human-learning criterion: the reported losses were smaller than for the outsourcing-majority group, but the available data include no subgroup-level transfer estimate that isolates this group's unassisted performance from the aggregate decline. Per the decision table's own standard for a human-learning claim, the absence of that delayed or transfer measure at the subgroup level means the claim can be neither affirmed nor defeated on the evidence reported - it is unproven, not merely suggestive of a milder pattern.

These cases show why the method is not equivalent to a general warning about deskilling. The system can improve task output without supporting a learning claim, and a safeguarded design can change the disposition. The method therefore evaluates the claim attached to the evidence rather than treating AI assistance itself as either beneficial or harmful.

5. A Second Claim Type: Meaningful Control in Public Decision Systems

A different domain illustrates why the method cannot be reduced to skill retention. Consider an AI-supported public decision process in which administrative throughput improves while final legal responsibility remains assigned to a human reviewer. The relevant human-side question is not whether the reviewer can reproduce the model's analysis unaided, but whether the reviewer retains practical authority to identify material problems, reject or alter recommendations, escalate uncertainty, and provide effective recourse where required.

Meaningful-human-control research already supplies much of the relevant measurement vocabulary. Santoni de Sio and van den Hoven (2018) distinguish tracking and tracing conditions for meaningful control. Siebert et al. (2023) argue that responsibility should be commensurate with a person's actual ability and authority to control the system. Davidović (2023) emphasizes that the appropriate form of control depends upon the purpose control is intended to serve, while Tsamados, Floridi, and Taddeo (2024) argue that advanced AI may require collaborative rather than merely supervisory forms of control.

The present method uses those constructs to determine the validity of a progress claim. A claim that "AI improved administrative throughput" can succeed on processing-time or accuracy evidence alone. A claim that "AI improved human-controlled administration" requires observed evidence that reviewers can exercise relevant authority under real workflow conditions. If the human remains legally responsible while override, escalation, or correction become practically ineffective, the stronger human-control claim fails even if throughput rises.

This example also shows the role of structural governability. A formal right to override does not establish practical authority if time pressure, institutional incentives, interface defaults, or the absence of alternative workflows make exercise of that right unrealistic. In a public-service application, Eubanks (2018) provides a broader warning that automated administrative systems can increase classification power while creating new burdens and asymmetries for affected populations. The method therefore treats documentary rights as insufficient evidence when the claim concerns meaningful control in practice.

6. Relation to Existing Frameworks

The method is deliberately narrower than the literatures from which its components are drawn. Vaccaro et al. (2024) evaluate when human-AI combinations outperform their components and show that complementarity cannot be assumed. Gonzalez et al. (2026) develop a broader complementarity framework spanning cognition, task structure, trust, team composition, institutions, and equity. These approaches primarily concern the performance and functioning of human-AI teams, whereas the present method asks whether a stronger human-advancement claim is warranted by the evidence.

Meaningful-control and autonomy frameworks occupy a second neighboring area. Buijsman, Carter, and Bermúdez (2025) develop an autonomy-preserving approach to AI decision support centered on skilled competence, role specification, failure indicators, and reflective practice. Santoni de Sio and van den Hoven (2018), Siebert et al. (2023), Davidović (2023), and Tsamados et al. (2024) provide overlapping accounts of control, responsibility, and collaborative agency. The present method does not replace those frameworks; it uses their constructs as evidence when the progress claim being tested concerns human authority or governed extension.

The closest pressure comes from longitudinal delegation research. Huang and Vishnoi (2026), in a formal preprint, model path dependence under adaptive AI delegation and show how short-run performance gains can coexist with long-run skill loss. Di Santi (2026), in a conceptual preprint, proposes metrics for distinguishing cognitive amplification from cognitive delegation, including dependency and cognitive-drift measures; notably, the tested regimes in that work do not demonstrate genuine amplification under the proposed metric framework. These studies directly motivate the need for longitudinal and human-side contrasts.

The residual contribution of the present method is therefore procedural rather than ontological. It does not claim that capacity, control, distribution, or time are newly discovered dimensions. It specifies how those dimensions are used to evaluate the validity of a progress claim, including what must be declared before measurement, what contrast is required, when a claim remains unproven, and when evidence is strong enough to defeat the claim.

7. Limitations and Validation

The method is not a validated instrument. The thresholds that distinguish material from immaterial deterioration will vary by domain, and the meaning of a governance-critical capacity must be specified within the technologically altered practice rather than imported from unaided human performance. A school may reasonably treat transfer beyond live assistance as central to learning, while a professional domain may treat reliable tool-mediated performance as legitimate so long as verification and responsibility remain intact.

The method also depends on predeclaration. If researchers are free to redefine the relevant human functioning after observing the data, the succeed/unproven/fail dispositions become vulnerable to post hoc rationalization. Empirical validation should therefore examine whether independent evaluators applying the same declared claim and failure criteria reach similar dispositions across cases.

Future work should test the method across domains that present different inferential structures. Education is useful because transfer can be measured directly. Public administration tests practical authority and recourse. Professional work can test responsibility alignment, continuity, and skill change under repeated use. Cross-domain validation should focus not on forcing one metric across all contexts, but on whether the protocol reliably prevents unsupported movement from system-level evidence to human-advancement conclusions.

The method also inherits the limitations of the evidence it evaluates. Bastani et al. provide a strong transfer contrast but do not settle long-run effects on educational pathways, switching costs, or broader forms of agency. Strömberg et al. identify a large-scale association between usage patterns and later examination performance, but the time-and-score pattern is not itself a complete mechanism of cognitive change. The method therefore narrows claims rather than pretending that one study can establish a complete human-development trajectory.

8. Conclusion

AI-mediated performance and human advancement are different dependent variables. The practical consequence of that distinction is not another list of ethical dimensions but a requirement on evidence. A claim about human advancement must declare the human functioning at issue, the population and scope of the claim, the required contrast, the relevant time window, and the signs that would defeat the claim before system-level improvement is interpreted as progress.

The diagnostic then classifies the relationship between the AI-mediated performance trajectory and the human-side trajectory. Coupled improvement supports a stronger advancement claim, performance-only acceleration leaves the claim unproven, and regressive decoupling defeats it where the declared human-side criterion deteriorates.

The educational evidence shows why this distinction matters. Bastani et al. and Strömberg et al. both report cases in which assisted work improves while later unassisted performance declines. The correct conclusion is not that AI assistance is inherently harmful, because guarded designs and different usage patterns can change the outcome. The correct conclusion is that an assisted-performance gain cannot carry a human-learning claim when the declared learning measure moves in the opposite direction.

The methodological rule is therefore simple enough to apply but strict enough to reject unsupported claims: improvement in an AI-mediated system warrants a claim of human advancement only when the evidence required by that human claim is separately observed, and the claim must remain unproven or fail when the relevant human-side evidence is absent or contradictory.

References

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779. https://doi.org/10.1016/0005-1098(83)90046-8

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Buijsman, S., Carter, S. E., & Bermúdez, J.-P. (2025). Autonomy by design: Preserving human autonomy in AI decision-support. Philosophy & Technology, 38. https://doi.org/10.1007/s13347-025-00932-2

Davidović, J. (2023). On the purpose of meaningful human control of AI. Frontiers in Big Data, 5, Article 1017677. https://doi.org/10.3389/fdata.2022.1017677

Di Santi, E. (2026). Cognitive amplification vs cognitive delegation in human-AI systems: A metric framework [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.18677

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Gonzalez, C., Donahue, K., Goldstein, D. G., Heidari, H., Jalali, M. S., Schelble, B. G., Singh, A., & Woolley, A. (2026). Toward a science of human-AI teaming for decision making: A complementarity framework. PNAS Nexus, 5. https://doi.org/10.1093/pnasnexus/pgag030

Huang, L., & Vishnoi, N. K. (2026). Path dependence under adaptive AI delegation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.02950

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30(3), 286-297. https://doi.org/10.1109/3468.844354

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

Siebert, L. C., Lupetti, M., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D. A., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. AI and Ethics, 3, 241-255. https://doi.org/10.1007/s43681-022-00167-3

Strömberg, D., Lei, V., & Wu, Y. (2026). The generative AI learning penalty: Evidence from Chinese secondary education (CEPR Discussion Paper No. DP21577).

Tsamados, A., Floridi, L., & Taddeo, M. (2024). Human control of AI systems: From supervision to teaming. AI and Ethics, 5, 1535-1548. https://doi.org/10.1007/s43681-024-00489-4

Vaccaro, M., Almaatouq, A., & Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8, 2293-2303. https://doi.org/10.1038/s41562-024-02024-1