The Human-Development Conversion Constraint

What Should Count as Progress in AI-Rich Societies?

Micheal Charles Preble
Independent Researcher
Manuscript v1.7

Abstract

Rapid advances in artificial intelligence are commonly described as technological progress, yet the growth of artificial capability and the growth of human capability are not equivalent phenomena. Artificial systems may become more capable, more widely available, and more deeply embedded in everyday life while the substantive capabilities, agency, resilience, opportunities, and lived conditions of the people living alongside them change only weakly, unevenly, or adversely. This paper argues that artificial capability and human development should therefore be treated as distinct trajectories and that the social significance of technological progress under advanced AI should be evaluated partly through the relationship between them. The central philosophical problem is not whether AI can improve outcomes, but whether improvement in the capability of an increasingly intelligent environment should be interpreted as evidence of improvement in the human condition.

The paper develops a human-development conversion constraint to address this problem. The constraint does not treat capability-oriented evaluation of technology as novel; rather, it builds explicitly on the capability approach and on prior work showing that technologies can alter the meaning and structure of human capabilities themselves. Its narrower claim is that public or institutional claims of AI-driven human progress are not warranted by improvements in system capability, adoption, or output alone. Such claims require separate evidence that technological capability has become a substantively valuable, sufficiently distributed, and sufficiently durable expansion of human possibility.

The argument also rejects the assumption that meaningful human enhancement requires biological or technological modification of the individual. Work on extended cognition and human-AI coupling already establishes that practical human capability can be expanded through external technological resources rather than exclusively through capacities internal to the biological organism. Once this premise is accepted, however, a further question emerges: whether technologically extended functioning should count as increased human progress merely because performance improves. The paper argues that technological availability, practical access, effective use, substantive benefit, distribution, and durability must remain analytically distinguishable if claims about human advancement are to be justified.

The resulting account does not propose another general theory of AI welfare, distributive justice, or human enhancement. Its contribution is narrower: it states a constraint on the validity of progress claims in AI-rich environments. Where improved AI-mediated performance is accompanied by losses in human capacities or powers that remain necessary for meaningful agency in the relevant domain, the performance gain is not sufficient evidence of human progress. The paper therefore treats conversion failure-not technological limitation-as the central object of evaluation.

Keywords: artificial intelligence; human development; capability approach; progress accounting; human-AI interaction; extended cognition; learning; automation; digital inequality

1. Introduction

Artificial intelligence is increasingly evaluated through measures of what machines can do. Researchers compare benchmark performance, organizations measure productivity and adoption, governments assess technological readiness, and investors track economic value. These measures provide important information about technological change, but they do not establish that the people living within an increasingly capable technological environment are experiencing corresponding forms of progress. A society may become substantially more successful at producing artificial capability without becoming equally successful at producing education, opportunity, resilience, institutional intelligibility, meaningful participation, or substantive freedom for the people living within it.

The distinction becomes increasingly important as artificial intelligence moves from occasional tool use toward persistent participation in social and institutional life. Artificial systems now contribute to education, research, professional work, commerce, public administration, communication, health navigation, and everyday decision-making. Under such conditions, artificial capability matters not only as a property of discrete artifacts but increasingly as a feature of the environment within which human beings act, learn, compete, encounter institutions, and pursue valued ends. The more deeply artificial systems become embedded in ordinary activity, the less adequate it becomes to evaluate technological progress without separately examining the human conditions produced within that environment.

Greater environmental capability does not automatically imply greater human development. A student may produce substantially better work with artificial assistance without developing a corresponding ability to understand or evaluate the reasoning involved. An organization may increase productivity while its workers become less capable of explaining or challenging processes for which they continue to bear responsibility. A public institution may become faster through automation while simultaneously becoming more difficult for citizens to understand, contest, or navigate when automated processes fail. In each instance, technological performance and human advancement can move together, but nothing in the existence of improved output requires them to do so.

This possibility is consistent with several established research traditions. Capability-oriented accounts of human development distinguish resources from the substantive freedoms people are actually able to exercise, and contemporary work applying this tradition to artificial intelligence emphasizes that AI may expand capabilities while also affecting agency, autonomy, equity, and well-being (Lengfelder et al., 2025). Research on human-AI systems similarly distinguishes successful technologically mediated performance from retained human competence and shows that the benefits of human-AI collaboration vary substantially by task and configuration. Sociotechnical scholarship, meanwhile, emphasizes that the consequences of AI are mediated by institutional design, power, inequality, and social context rather than determined by technological capability alone.

The present paper does not claim that these individual distinctions are new. Its argument begins once they are taken seriously together. If artificial capability and human development are analytically distinct, then the growth of the former cannot itself constitute evidence of equivalent growth in the latter. The philosophical problem is therefore one of progress accounting: what additional conditions would justify describing technological advancement as human advancement when increasingly capable artificial systems become part of the environment through which human action is performed?

This paper argues that artificial capability and human development should be treated as separate trajectories. These trajectories may be strongly coupled, weakly coupled, or move in conflicting directions. Machine capability may rise rapidly while human developmental conditions rise slowly, remain comparatively flat, become more unequally distributed, or deteriorate in selected respects. Conversely, social and institutional arrangements may generate significant human gains from relatively modest technological improvements by converting available capability more effectively into education, agency, resilience, and opportunity.

The argument developed here is normative rather than causal. It proposes that the social significance of increasing artificial capability should be evaluated partly according to the extent to which that capability is converted into broadly distributed and durable improvements in substantive human possibility. This does not require every technological advance to improve every dimension of human life, nor does it assume that AI is the sole cause of social change. It rejects only the inference that improvement in artificial capability can itself stand as sufficient evidence of improvement in the human condition.

The accounting boundary follows the object of the claim. Evidence that a model has become more capable can support a claim about model capability, while evidence that a firm has become more productive can support a claim about organizational performance. A claim about human progress, however, requires evidence about changes in the human condition. Artificial intelligence makes this distinction increasingly important because it is becoming embedded in the environments through which people learn, work, communicate, and encounter institutions.

The remainder of the paper develops this argument in several stages. Section 2 distinguishes machine capability from human progress and formulates the human-development conversion problem. Section 3 examines technological enhancement without requiring equivalence between human and machine capability, while Section 4 treats AI as an increasingly environmental source of practical capability. Sections 5 through 7 clarify the accounting boundary, identify relevant human-side outcomes, and examine the distributive conditions of progress. Section 8 illustrates how the constraint applies to claims about intelligibility and analytical literacy. Section 9 states the Human-Development Conversion Constraint formally, Section 10 addresses major objections, and Section 11 concludes by returning to the question of what an AI-rich society should count as progress.

2. Human Progress and Machine Progress

Technological capability and human capability have never been identical. Transportation systems extend the distance over which people can act without increasing unaided human running speed, while written language expands memory and communication without biologically increasing human storage capacity. Telecommunications permit action and communication across distances that human sensory organs could never bridge unaided, and modern institutions make highly specialized knowledge available to individuals who could not independently reproduce it. Civilization has repeatedly expanded practical human capability by altering the external environment rather than by requiring corresponding biological transformation.

Artificial intelligence belongs partly within this historical pattern, although it introduces an important complication. Many earlier technologies extended relatively identifiable functions such as transportation, storage, calculation, or communication. Contemporary AI increasingly participates in interpretation, synthesis, prediction, planning, evaluation, problem formulation, and other activities associated with judgment and cognition. As artificial systems become involved in these domains, the relationship between external technological capability and human capability becomes more difficult to interpret because improved performance no longer reveals clearly what the person understands, retains, directs, or contributes.

The fact that capability may be external does not make that capability less relevant to human agency. Work on extended cognition has long challenged organism-bound accounts of cognitive functioning (Clark & Chalmers, 1998), and recent scholarship applies those insights directly to generative AI and human-AI coupling. Clark (2025), for example, argues that human beings characteristically construct hybrid thinking systems that incorporate non-biological resources, while Hernández-Orallo (2025) argues that assessment in an AI-rich environment must account for people operating within increasingly interdependent human-machine systems. The present paper accepts that external technological resources can legitimately expand human capability and does not treat biological self-sufficiency as the appropriate standard for human advancement.

The philosophical problem begins after technological extension is granted. Improved extended functioning does not by itself establish that the human condition has improved. Equivalent observed performance can be produced by materially different configurations of retained human capacity, external capability, authority, dependence, and institutional support. A high-quality output therefore reveals something about the functioning of the overall configuration but does not, without further evidence, identify what has changed in the human participant.

This underdetermination is especially apparent in education. Two students may produce equally sophisticated essays while one can explain the argument, identify weaknesses, and apply the underlying reasoning to a novel problem, whereas the other cannot. Both may benefit from the use of AI, and both may have produced a legitimate artifact, but the human-development interpretation differs substantially. Output quality alone cannot determine whether the technology contributed to durable learning, effective extension, temporary substitution, or some combination of these states.

Professional work produces similar ambiguity. Classic human-factors research already warns that automation can leave human operators responsible for abnormal situations while reducing the practice and situational engagement needed to respond effectively (Bainbridge, 1983; Parasuraman et al., 2000). Two analysts may generate comparably accurate reports while differing dramatically in their ability to recognize model failure, challenge inappropriate assumptions, evaluate evidence, or explain limitations to decision-makers. The resulting artifact may satisfy an organizational performance criterion in both cases, but the human capability embedded within the process is not equivalent. Where responsibility remains human, such differences may become consequential even when short-term output quality remains high.

Institutional scale magnifies the same problem. Research on automated public decision systems has documented how technological efficiency can coexist with new burdens, opacity, and unequal exposure to automated classification (Eubanks, 2018). A public agency may automate a decision process and achieve substantial improvements in processing speed, cost, and consistency while leaving citizens less able to understand why decisions were made, identify errors, obtain meaningful recourse, or function when automated channels fail. The technological system may improve according to conventional operational measures even as the practical relationship between individuals and the institution becomes less intelligible or governable. Human progress therefore cannot be inferred from system performance without specifying what human conditions are being assessed.

None of this implies that every displaced human skill should be preserved. Such a standard would make ordinary technological development difficult to defend and would confuse human advancement with the maintenance of historical forms of labor. Some capabilities can reasonably migrate into reliable external systems, particularly where retaining them internally imposes substantial cost without meaningful developmental benefit. The important question is not whether capability remains biologically internal, but what kind of human possibility results from the configuration through which the capability is exercised.

Artificial capability, practical access, effective use, AI-assisted functioning, retained human capacity, and substantive human development should therefore remain analytically distinguishable. Movement from one condition to the next is possible but not guaranteed. Technological capability may exist without broad access, access may exist without effective use, effective use may increase performance without producing durable capability, and improved performance may coexist with fragile dependence, unequal distribution, or reduced practical authority. Treating these stages as interchangeable obscures precisely the transitions that determine whether technological progress has become human progress.

The resulting problem can be described as a human-development conversion problem. This framing is continuous with the capability approach's longstanding distinction between resources and achieved functionings, and it also inherits a complication identified by Coeckelbergh (2011): information technologies may alter the meaning and structure of the capabilities being evaluated rather than functioning as neutral means toward fixed human ends. The

present paper therefore does not treat conversion as a simple pipeline from an unchanged resource to an unchanged capability. Nor does the conversion constraint freeze a pre-AI list of human powers and demand that every one of them remain intact. The claimant must instead specify which human powers remain necessary for meaningful agency within the technologically altered practice itself, and then show that the asserted performance gain has not been purchased by materially eroding those powers. Technological capability supplies new resources and possibilities to an environment, but the relevant human outcome depends upon what happens when those resources encounter people, institutions, opportunities, constraints, existing inequalities, and technologically altered practices. The analytical value of this formulation does not lie in claiming novelty for the individual stages involved. It lies in refusing to infer later human-development states from earlier technological ones and in making the quality of conversion itself an object of evaluation.

3. Enhancement Without Equivalence

One influential response to accelerating artificial intelligence understands the relationship between human beings and increasingly capable machines in competitive terms. If artificial systems become faster, more knowledgeable, more reliable, or more effective across an expanding range of cognitive tasks, humans may appear progressively deficient when evaluated on the same dimensions. From that perspective, meaningful human enhancement can begin to look like a problem of reducing the performance gap between biological humans and artificial systems through genetic modification, neurotechnology, brain-computer interfaces, pharmacological enhancement, or other forms of direct human alteration.

Such a framing confuses technological equivalence with human advancement. Human beings have never needed to reproduce the physical or informational properties of their most powerful technologies in order to benefit from them. The user of an aircraft does not become capable of biological flight, the user of a library does not internalize every book, and the participant in a communications network does not acquire the physical capacity to transmit information across continents. Human possibility frequently expands because the surrounding environment changes in ways that make new forms of action practically available.

Artificial intelligence can extend this historical pattern into domains closer to cognition. Individuals may gain access to translation, tutoring, planning, analysis, technical explanation, research assistance, and problem-solving capacities that previously required considerable time, money, education, or specialist support. In such cases, effective human capability may increase even when the relevant computational capacity remains external. Contemporary work on extended cognition and AI-mediated enhancement already supports this general possibility, so the claim should be treated as a foundation rather than as the distinctive contribution of the present paper.

Once external technological extension is accepted, however, a further philosophical consequence follows. Human enhancement need not be evaluated by asking whether a biological person has become technologically equivalent to the artificial systems around them. A person may become substantially more capable of acting in the world because the environment supplies reliable cognitive support, just as transportation, writing, and institutional expertise extend action without reproducing their functions biologically. The appropriate standard of advancement therefore shifts from competitive equivalence toward substantive human possibility.

This shift also avoids an unstable defense of human value. Human worth need not depend upon retaining comparative advantage over artificial systems at every economically valuable task. If artificial systems outperform people across more domains, a conception of human value based primarily on task superiority will become increasingly difficult to sustain. A more durable account evaluates whether technology expands people's opportunities for understanding, choice, creation, participation, responsibility, relationship, and pursuit of valued ends rather than whether people remain superior performers.

Environmental extension nevertheless cannot automatically be equated with enhancement. External systems may increase immediate output while increasing fragility, dependency, inequality, or institutional domination. They may broaden nominal possibilities while narrowing meaningful alternatives, or they may produce apparent capability that disappears when access, pricing, model behavior, institutional policy, or provider availability changes. A technologically richer environment can therefore make people more effective in some respects while leaving them more vulnerable in others.

The relevant philosophical distinction is consequently not simply between internal and external capability. An externally supplied capability may substantially increase agency, while an internally possessed competence may be practically useless if institutions prevent its exercise. What matters is whether the arrangement expands substantive human possibility under conditions that allow people to make meaningful use of the capabilities available to them. The location of capability matters less than the human condition produced through access to it.

This reframing offers an alternative to the assumption that increasingly capable AI necessarily requires increasingly machine-like humans. The objective need not be technological competition, nor need it be technological withdrawal. A third possibility is to design technological environments in which advances in artificial capability increase the practical freedom and effective capability of human beings without requiring biological equivalence between human and machine. Under this view, the measure of advancement is not the quantity of technology incorporated into the body but the expansion in substantive possibility produced by the environment within which the person acts.

4. Artificial Intelligence as Environmental Capability

As artificial intelligence becomes embedded throughout social institutions, understanding it solely as a collection of discrete products becomes increasingly inadequate. A specialized system used occasionally can reasonably be evaluated primarily through its immediate function, but a technological class that mediates education, work, communication, access to expertise, public administration, finance, commerce, and institutional navigation begins to influence the conditions under which ordinary agency occurs. AI in this sense becomes environmental not because it disappears into the background completely, but because the quality of the surrounding artificial infrastructure increasingly shapes what people are practically able to do.

The term environmental capability is used here descriptively rather than as a claim of conceptual priority or legal classification. It does not imply that artificial intelligence is necessarily a public utility, that governments should operate frontier models, or that every citizen must receive identical technological services. It identifies a functional shift in the importance of technological access. When the quality of AI assistance materially affects a person's ability to learn, navigate institutions, acquire expertise, organize work, or solve complex problems, that assistance increasingly contributes to the structure of practical opportunity.

This makes simple adoption an inadequate measure of social integration. Two societies may report similar rates of AI use while producing substantially different human consequences. One may use artificial systems to broaden educational support, reduce administrative burdens, improve disability access, strengthen public-service navigation, and increase the capabilities of small businesses and households. Another may concentrate the most consequential advantages inside large firms, elite educational institutions, professional organizations, or expensive subscription tiers while widespread consumer access produces comparatively little developmental effect.

Differences also occur among individuals who nominally possess the same technology. Effective benefit may vary according to literacy, language, connectivity, cost, model quality, institutional support, disability access, professional authority, and the ability to identify when a system should or should not be trusted. Formal access therefore provides limited evidence about substantive capability. A society can reduce the first-order digital divide while preserving a second-order divide in the quality, reliability, and practical usefulness of the artificial capability available to different populations (van Dijk, 2005).

This creates the possibility of a new form of capability stratification. If high-quality artificial assistance materially improves learning, planning, analysis, translation, technical instruction, institutional navigation, and access to expertise, unequal access to such assistance may affect more than consumer convenience. It may alter who is able to learn efficiently, exploit opportunities, recover from setbacks, navigate complex systems, and exercise rights. The relevant inequality is not simply between people who use AI and people who do not, but between people who inhabit differently capable informational environments.

Section 2's conversion distinctions also apply at societal scale. Technological readiness or adoption can support claims about availability and use, but they do not by themselves establish substantive, distributed, or durable human benefit.

Once artificial intelligence becomes sufficiently embedded in the environment of ordinary action, the question of what machines can do becomes inseparable from the question of what people become able to do because those machines exist. This framing does not collapse the machine and the person into a single object of evaluation. Rather, it requires their trajectories to remain distinct. The more intelligent the environment becomes, the more important it becomes to determine whether increased environmental intelligence is actually enlarging substantive human possibility.

5. Why the Accounting Boundary Must Be Human

The requirement that human-progress claims rest on human-side evidence rather than system-level evidence is compatible with, but distinct from, digital-commons and benefit-sharing arguments. Huang and Siddarth (2023), for example, examine the public informational resources on which generative AI depends and ask how value derived from those resources should be governed or returned. Those questions matter, but they are not necessary premises for the conversion constraint defended here. The paper can therefore leave ownership, taxation, compensation, and collective title open while still insisting that claims about human progress be supported by evidence about human development.

Artificial intelligence nevertheless makes the accounting problem unusually salient because it increasingly mediates cognition, judgment, education, administration, communication, and access to expertise. When a technology becomes part of the environment through which people think and act, its social significance cannot be read directly from the capability of the artifact. The evaluative burden shifts toward the condition of the human beings situated within the technologically altered practice.

The relevant accounting boundary is consequently defined by the claim being made. If the claim concerns model capability, model evidence is sufficient. If the claim concerns organizational productivity, organizational evidence may be sufficient. If the claim concerns human progress, however, evidence about human possibilities, powers, and durable conditions becomes indispensable. The conversion constraint is intended to police that inference rather than to settle the distributive ownership of the technology itself.

6. Human-Side Outcomes of AI-Mediated Change

The social return from artificial capability cannot be reduced to a single economic measure. Productivity, wages, employment, investment, and aggregate output remain important components of technological evaluation, but they do not exhaust the conditions under which people learn, participate, choose, understand institutions, recover from difficulty, or pursue valued ends. A society can become wealthier while important dimensions of substantive human capability remain stagnant or become more unequally distributed. Economic improvement therefore provides evidence about one dimension of social progress rather than a complete account of human development.

The category of human-side outcomes identifies this wider evaluative target. It refers to observable improvements in substantive human conditions associated with the successful diffusion and organization of artificial capability. Such improvements might include better education, wider access to expertise, lower administrative burden, enhanced

disability support, stronger household resilience, improved navigation of health and government systems, greater economic mobility, or increased capacity to solve complex personal and collective problems.

Existing research already demonstrates that the relationship between AI and social development is heterogeneous rather than uniform. Nahar (2024), for example, models the effects of AI-based innovation on sustainable-development outcomes across multiple countries and reports varying impacts across development dimensions and national contexts. Grewal et al. (2024) similarly identify preservation and growth of human capability, social belonging, and equitable sharing of AI benefits as major societal challenges. Such work supports the proposition that AI's human consequences are multidimensional, but it does not itself establish the normative principle developed here.

Human-side outcomes should therefore not be understood as another catalog of possible AI benefits. Their analytical purpose is to identify the human side of the conversion relationship and to insist that this side be measured independently from underlying technological capability. A society may possess extremely capable systems and report substantial adoption while still producing weak developmental gains. Another society may use less advanced systems more effectively through education, institutional design, accessibility, competition, or integration into ordinary life and thereby generate stronger human-side outcomes.

The concept is intentionally plural because no single index can settle what constitutes human development across all societies and circumstances. Different communities may reasonably assign different weight to education, economic security, health, civic participation, resilience, autonomy, or other dimensions of valued human functioning. The argument does not require universal agreement about a comprehensive theory of flourishing. It requires only that human-development consequences be examined explicitly rather than inferred from machine capability, adoption, or aggregate economic value.

Durability is equally important. Immediate convenience may disappear when a provider changes terms, a model is withdrawn, an institution reorganizes a workflow, or users lose capacities required to supervise increasingly automated processes. Durability therefore concerns not only persistence over time but also whether meaningful capability remains usable or recoverable under ordinary changes in providers, prices, interfaces, institutional workflows, and technological availability. Short-term performance improvements can also conceal path-dependent changes in dependence or retained competence. Human advancement therefore should not be inferred from temporary gains without asking whether the corresponding expansion in substantive possibility persists under ordinary technological and institutional change.

These qualifications do not transform the philosophical principle into a causal model. Artificial intelligence operates amid education systems, markets, demographic change, institutional quality, social norms, regulatory structures, and many other variables. Strong causal attribution will often require longitudinal designs, natural experiments, policy discontinuities, or other empirical methods capable of separating AI-related effects from broader social changes. The philosophical requirement is simply that promised human benefits eventually be sought in human outcomes rather than permanently inferred from technological potential.

The resulting question is therefore not whether AI is generally beneficial or harmful. It is whether growth in artificial capability is accompanied by durable changes in the substantive possibilities and lived conditions of the people among whom that capability is deployed. Asking the question does not prejudge the answer. It prevents machine progress and human progress from being collapsed into a single time series before their relationship has been examined.

7. Distribution as a Constituent of Progress

Human-development gains cannot be assessed adequately through population averages alone. Technological transformation may raise mean productivity, convenience, or access while concentrating the most consequential improvements among groups that already possess wealth, education, institutional authority, or technical expertise. Questions of distributive justice and technological inequality are already central to AI scholarship, so the present paper does not claim that distribution has been overlooked. Its narrower claim is that distribution should be treated as a condition on the inference from technological benefit to broad human progress.

Access is particularly difficult to represent as a binary variable. Two individuals may both use artificial intelligence while receiving substantially different levels of practical benefit because of differences in model quality, subscription cost, language support, educational background, connectivity, accessibility, professional context, or institutional support. Counting users therefore reveals relatively little about how much effective capability has been generated. Formal availability may expand while capability inequality persists or even deepens.

Institutions amplify these differences. Schools may integrate AI into strong pedagogy or introduce it without sufficient educational structure, workplaces may use artificial systems to expand worker judgment or primarily to intensify monitoring, and public agencies may use AI to make services more intelligible or to construct decision systems that become more difficult to challenge. The same underlying technological capability can therefore produce substantially different developmental consequences depending on organizational design and power.

Critical scholarship on AI and development reinforces the importance of these institutional differences. Iazzolino and Stremlau (2024), for example, argue that AI-for-social-good discourse can obscure the political economy through which private actors shape development priorities and epistemic infrastructure. Their analysis pressures any simple assumption that deployment of advanced technology automatically constitutes social improvement. The present argument takes this concern seriously by requiring that distribution and institutional effects remain visible within judgments of technological progress.

An adequate account of human progress must preserve the conversion distinctions introduced earlier without aggregating them prematurely into a single numerical index. A system may be widely accessible while producing highly concentrated developmental gains, and an aggregate gain may remain too fragile to support the broader progress claim.

Distribution does not require identical systems or identical outcomes for everyone. Complex societies legitimately permit specialization, variation, and unequal use of specialized resources where these differences do not undermine basic opportunity or participation. The narrower requirement is that claims about social progress remain sensitive to who actually receives the capabilities allegedly generated by technological advancement. If artificial intelligence is described as transformative for society as a whole, the distribution of that transformation cannot be treated as an afterthought.

8. Illustrative Human-Side Claims: Intelligibility and Analytical Literacy

The conversion constraint can be applied to human-side outcomes that are neither productivity measures nor claims about retained task skill. One example is institutional intelligibility. Existing work on high-risk inferences and predictive privacy emphasizes that institutions can acquire substantial predictive power over individuals while those individuals possess limited capacity to inspect, understand, or contest consequential classifications (Wachter & Mittelstadt, 2019; Mühlhoff, 2023). A claim that AI improves institutional participation should therefore require evidence that affected people can understand relevant decisions, identify material errors, obtain correction, or otherwise navigate the process more effectively, rather than inferring progress from faster or more accurate institutional prediction alone.

A second example is analytical literacy. Reviews of AI literacy in education already treat understanding, critical evaluation, and responsible use as important competencies (Casal-Otero et al., 2023; Yim & Su, 2025). Under the conversion constraint, a claim that an AI-rich environment expands civic or educational agency would require evidence that people become better able to interpret probabilistic claims, recognize uncertainty, question model outputs, or obtain meaningful assistance when they cannot do so themselves. The paper does not propose a new literacy curriculum; these examples simply illustrate how the constraint changes the evidence required by a human-progress claim.

The same logic applies to other non-economic outcomes such as accessibility, resilience, or civic participation. In each case, the evaluator must identify the human functioning or freedom being claimed, specify what would count as improvement within the altered practice, and distinguish direct evidence of that improvement from evidence that the surrounding technical system became more capable.

9. The Human-Development Conversion Constraint

The preceding analysis supports a narrower claim than a general theory of technological justice. Artificial capability and human development are distinct trajectories, and improvement in the former does not by itself warrant a claim about improvement in the latter. The relevant issue is inferential validity: what evidence is required before a statement about better AI-mediated performance can support a statement about human progress?

The Human-Development Conversion Constraint is the following: A claim that increased artificial capability constitutes human progress is warranted only to the extent that the capability is converted into substantively valuable human possibilities whose distribution and durability are adequate to the scope of the progress claim. The constraint deliberately leaves plural questions of value open, but it does not leave the inference empty. A claimant must identify the human functioning or freedom said to have improved, show that the improvement is practically available to the relevant population, and provide evidence that the gain is not merely a transient artifact of assistance whose surrounding human conditions have materially worsened.

A stronger failure condition follows in domains where humans continue to bear consequential responsibility. If AI-mediated performance improves while the human capacities or powers necessary to understand the stakes, identify material error, contest the process, or exercise the responsibility assigned to them deteriorate substantially, the performance gain is not sufficient evidence of human progress in that domain. This is not a prohibition on automation or skill migration. It is a prohibition on using system success as a proxy for human advancement when the human conditions necessary for meaningful agency have moved in the opposite direction.

The constraint therefore distinguishes several possible outcomes. Conversion succeeds when artificial capability expands substantively valuable human possibilities under conditions adequate to the claim being made. Conversion remains unproven when performance rises but relevant human outcomes are unmeasured. Conversion fails when an asserted human gain depends upon a dimension that has materially deteriorated or when the gain is so narrowly distributed or fragile that the scope of the progress claim exceeds the evidence. These categories do not produce a universal index, but they do rule out specific invalid inferences.

This formulation also permits genuine technological dependence. Human beings routinely expand capability through reliable external systems, and unaided reproduction of every function is not required for progress. The relevant standard is therefore not independence but whether the technologically extended arrangement preserves the human capacities and practical powers that remain necessary for agency, responsibility, and continued participation in the particular domain. Which capacities are necessary will vary by context, and that variation should be made explicit rather than hidden inside a universal notion of skill retention.

Education provides an empirical illustration of the constraint. Bastani et al. (2025) conducted a preregistered field experiment with nearly one thousand high-school mathematics students. Access to a standard GPT-4 interface improved performance on assisted practice problems by 48%, yet students in that condition scored 17% below the no-AI control on the subsequent unassisted examination; a teacher-grounded tutor with learning safeguards removed the negative exam effect, although it did not produce a positive unassisted advantage over the control group. A separate 30-month working paper following 26,811 Chinese secondary students reported the same directional split at larger scale: after AI adoption, homework scores rose by 18% and homework time fell by 30%, while closed-book examination scores fell by 20% within six months (Strömberg et al., 2026). Strömberg and colleagues further report that the losses were concentrated among the roughly 80% of users whose time-and-score pattern was consistent with outsourcing more of the homework process, while students who kept homework time closer to the non-AI baseline experienced smaller losses.

These studies do not establish a universal theory of AI and learning, and the relevant educational standard must itself be made explicit. Here the chosen human-side outcome is learning understood partly through performance that transfers beyond the assisted practice context, rather than fluency with the tool alone. Under that criterion, the observed increase in assisted homework or practice performance cannot by itself support a claim of improved human learning when unassisted assessment moves in the opposite direction. The guarded-tutor result is equally important because it shows that the constraint does not condemn AI assistance as such: design choices can preserve the performance benefit while avoiding at least some of the measured learning penalty.

The purpose of the constraint is therefore not merely to recommend that evaluators "look at both sides." It determines which kind of evidence is required by the kind of progress claim being made and identifies conditions under which a system-level gain cannot validly carry a human-development conclusion. A companion methodological paper, Beyond Performance: A Diagnostic Method for Evaluating Human-Advancement Claims in AI-Mediated Systems (Preble, 2026), develops that decision procedure in operational form. The present paper establishes the philosophical constraint that makes such a methodology necessary.

10. Objections

10.1 The accounting boundary is arbitrary

A first objection holds that the paper simply chooses to privilege human outcomes after technological improvement has already been demonstrated at the level of the model, firm, or institution. If productivity rises or service quality improves, why should a further human-development test be required before the change can count as progress?

The answer depends on the scope of the claim. A productivity improvement is genuine progress in productivity, and a model improvement is genuine progress in model capability. The conversion constraint becomes relevant only when those narrower achievements are generalized into claims about human or social progress. The accounting boundary is therefore not selected arbitrarily; it follows the object of the proposition being defended.

This distinction also prevents the constraint from becoming a comprehensive theory of technological value. Different levels of analysis can register real gains simultaneously, and no human-development test is required to describe a model as more accurate or a firm as more efficient. What the constraint blocks is the unsupported movement from one level of analysis to another.

10.2 Markets already perform the relevant conversion

A second objection holds that successful technologies generate human benefit precisely because people choose to purchase, adopt, and use them. If artificial systems increase productivity, reduce cost, attract widespread adoption, and generate high consumer demand, market behavior already provides evidence that the technology is improving human conditions. Additional human-development accounting may therefore seem redundant or paternalistic.

Market evidence is relevant but incomplete. Adoption reveals willingness to use a technology under existing alternatives and constraints, but it does not establish whether use produces durable capability, whether benefits are broadly distributed, whether important externalities accompany those benefits, or whether people become dependent upon systems that later affect their practical freedom. Many socially important outcomes also occur through schools, households, workplaces, public agencies, and civic institutions that cannot be evaluated adequately through individual consumer choice alone.

The principle therefore does not reject markets as mechanisms of technological conversion. Competitive markets may prove highly effective at transforming artificial capability into human benefit in many domains. The claim is only that market success should not be treated as a complete measure of human-development success. A technology may be profitable, widely adopted, and useful while still producing uneven or fragile developmental consequences that deserve separate evaluation.

10.3 Human development is too plural to support a common standard

A third objection concerns pluralism. Human flourishing is contested, and different societies legitimately disagree about which forms of life, capability, and well-being deserve emphasis. A principle connecting AI progress to human development may therefore appear to require an overly comprehensive account of the good or to invite institutions to impose a particular conception of human flourishing.

The objection properly limits the ambition of the argument. No universal ranking of all human outcomes is required, and the principle should not be used to erase legitimate variation in social values. Capability-oriented approaches already provide one way of preserving pluralism by focusing on substantive opportunities rather than mandating identical outcomes (Sen, 1999; Nussbaum, 2000). The present argument requires only that societies distinguish machine capability from changes in the human condition and make explicit the human dimensions against which technological success is being assessed.

Pluralism complicates measurement but does not eliminate the distinction between technological advancement and human advancement. Different societies may prioritize different developmental dimensions while still agreeing that machine capability and human progress should not be treated as synonyms. The principle therefore permits disagreement about the content of human development while preserving the requirement that some account of human development be examined independently.

10.4 Causal attribution may be impossible

Human societies change for many reasons simultaneously, and the effects of AI may be difficult to isolate. If educational performance improves after widespread AI adoption, the change may also reflect funding, demographic shifts, curriculum reform, or labor-market incentives. If mobility declines, housing, taxation, institutional quality, or macroeconomic conditions may explain more than technological adoption. A human-development conversion principle may therefore appear to demand causal knowledge that empirical research cannot reliably provide.

This objection constrains methodology rather than the philosophical distinction. Some contexts may support natural experiments, longitudinal comparison, policy discontinuities, or other methods capable of estimating causal contribution, whereas other contexts may justify only descriptive association or bounded attribution. The appropriate

evidentiary standard will depend upon the domain and data available. The principle does not require researchers to claim more causal certainty than their designs support.

Difficulty identifying causality does not justify collapsing the variables. On the contrary, maintaining artificial capability and human development as separate trajectories is a prerequisite for investigating their relationship at all. Where attribution remains uncertain, uncertainty should remain visible rather than being replaced by an assumption that technological progress constitutes human progress automatically.

10.5 The constraint may be empirically premature

A serious objection is temporal rather than causal. General-purpose technologies often require years or decades before their developmental effects become visible in institutional statistics, and frontier AI diffusion remains recent. A demand for demonstrated human-side improvement could therefore condemn technologies before education systems, labor markets, legal institutions, and ordinary practices have had time to reorganize around them.

The objection is strongest against immediate aggregate judgment, but it does not defeat the constraint. Where developmental effects are not yet observable, the appropriate conclusion is uncertainty rather than a declaration of human progress based on model capability or adoption. The constraint therefore imposes evidentiary patience in both directions: early technological success does not prove human advancement, while the absence of mature longitudinal evidence does not prove failure.

This temporal limitation also clarifies the role of leading indicators. Where long-run human-development measures are unavailable, researchers may examine shorter-horizon evidence concerning learning transfer, authority, contestability, resilience, accessibility, or other domain-specific outcomes, while labeling those findings appropriately. The principle becomes empirically idle only if it is interpreted as requiring a complete social-development time series before any provisional judgment can be made.

10.6 Human-development metrics can themselves be gamed

A second hard objection concerns Goodhart effects and measurement politics. If governments, firms, or institutions begin evaluating AI through "human-side outcome" metrics, actors may optimize the chosen indicators while undermining the underlying values the indicators were intended to represent. A framework designed to prevent false inference from machine metrics could therefore generate a new class of false inference from human metrics.

This risk is real and argues against a single composite score. The conversion constraint is intentionally decomposed because no one measure should be allowed to stand as a sufficient statistic for human progress across domains. Evaluators should report the human outcome being claimed, the evidence used, the populations affected, and material countervailing effects rather than compressing them prematurely into a universal index.

The objection also strengthens the requirement that measures remain contestable and revisable. Indicators should function as evidence within an argument rather than as automatic verdicts. A ministry that raises a "human capability score" by redesigning the metric without improving the underlying functioning would violate the constraint rather than satisfy it, because the claimed conversion must remain connected to substantively valuable human possibilities rather than to the metric alone.

10.7 Positional goods and unequal compounding

Broadly distributed capability gains do not necessarily translate into better access to positional goods such as selective education, scarce employment, housing, prestige, or institutional power. If AI raises everyone's writing performance while the number of elite opportunities remains fixed, relative competition may simply reset at a higher technological baseline. The resulting expansion in one capability may be real without producing equivalent improvement in the scarce outcomes people seek.

This objection shows why distribution cannot be understood merely as equal access to a tool. Human-development evaluation must distinguish absolute gains in functioning from relative changes in opportunity and institutional power. A broad increase in capability may still count as a real gain, but claims about mobility, equality, or opportunity require separate evidence about those outcomes rather than being inferred from the capability gain itself.

The problem is sharper when higher-status groups can compound superior models, data, institutional support, and professional authority into durable advantages while median users receive only basic assistance. Such a society may experience widespread adoption and even rising average capability while institutional power becomes more concentrated. The conversion constraint does not by itself prescribe a tax, public model, or procurement regime, but it requires progress claims to expose rather than average away that divergence.

10.8 Cross-border and externalized conversion

Human-development gains within one population can coexist with harms imposed on another. AI systems may increase domestic productivity, administrative capacity, or security while contributing to surveillance, labor extraction, influence operations, or other externalized burdens elsewhere. A purely national accounting frame could therefore classify a local gain as progress while ignoring the conditions through which it was produced.

The appropriate scope of the conversion claim must therefore be explicit. A claim of progress for a particular institution or population may be defensible even when broader global effects are contested, but it should not be generalized beyond the population and consequences actually evaluated. Claims about social or civilizational progress carry a correspondingly wider evidentiary burden.

This does not require a complete theory of global justice within the present paper. It requires scope discipline. The larger the progress claim, the broader the population and external effects that must be considered before the claim is warranted.

11. Conclusion

The rapid expansion of artificial intelligence creates a measurement problem that is also a philosophical problem. Measures of machine capability tell us increasingly sophisticated things about what artificial systems can accomplish, while adoption, investment, productivity, and economic value reveal important consequences of their diffusion. None of these measures, however, establishes what has happened to the substantive possibilities of the people living within an increasingly intelligent environment. A society can become technologically more capable without becoming proportionally more capable in the human sense.

Artificial capability and human development should therefore remain separate trajectories. The conversion distinctions developed in this paper prevent improvements in capability, adoption, or output from silently carrying conclusions about human benefit, distribution, or durability. Making that relationship explicit is necessary if claims about human advancement are to carry more than rhetorical force.

This separation also changes the meaning of human enhancement. Increasingly capable machines do not imply that biological humans must become technologically equivalent to them in order to remain valuable or capable. Human beings have always expanded effective capability through tools, institutions, infrastructures, and shared knowledge, and artificial intelligence can become another powerful form of environmental extension. The relevant question is not whether humans can reproduce artificial capability internally but whether artificial capability expands what humans can meaningfully understand, choose, pursue, and become.

The accounting boundary follows the object of the claim. Evidence that a model, firm, or institution improved may justify conclusions about model capability, productivity, or administrative performance, but claims of human progress require evidence about the human condition. Artificial intelligence makes this distinction increasingly important

because it is becoming embedded in the environments through which people learn, work, communicate, and encounter institutions.

The Human-Development Conversion Constraint captures that requirement by refusing to treat growth in artificial capability as equivalent growth in human progress. Its social significance should instead be evaluated partly according to whether increasing artificial capability is converted into broadly distributed and durable expansions of substantive human possibility. The principle does not supply a single metric, prescribe one institutional mechanism, or resolve every dispute about flourishing. It establishes the proper direction of inference.

An AI-rich society will therefore need conceptual and empirical tools for measuring not only what increasingly capable machines can do, but what their growing capability changes in the lives and possibilities of the people who live among them. The central philosophical question is not whether artificial intelligence will become extraordinarily capable, because increasing capability already appears likely. The more consequential question is whether increasing artificial capability will be converted into extraordinary human possibility, and whether societies will retain enough conceptual clarity to recognize the difference when it is not.

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

Casal-Otero, L., Catalá, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. International Journal of STEM Education, 10, Article 29. https://doi.org/10.1186/s40594-023-00418-7

Clark, A. (2025). Extending minds with generative AI. Nature Communications, 16, Article 4627. https://doi.org/10.1038/s41467-025-59906-9

Clark, A., & Chalmers, D. J. (1998). The extended mind. Analysis, 58(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Coeckelbergh, M. (2011). Human development or human enhancement? A methodological reflection on capabilities and the evaluation of information technologies. Ethics and Information Technology, 13, 81-92. https://doi.org/10.1007/s10676-010-9231-9

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

Grewal, D., Guha, A., & Becker, M. (2024). AI is changing the world: For better or for worse? Journal of Macromarketing, 44, 870-882. https://doi.org/10.1177/02761467241254450

Hernández-Orallo, J. (2025). Enhancement and assessment in the AI age: An extended mind perspective. Journal of Pacific Rim Psychology, 19. https://doi.org/10.1177/18344909241309376

Huang, S., & Siddarth, D. (2023). Generative AI and the digital commons. arXiv. https://doi.org/10.48550/arXiv.2303.11074

Iazzolino, G., & Stremlau, N. (2024). AI for social good and the corporate capture of global development. Information Technology for Development, 30(4), 626-643. https://doi.org/10.1080/02681102.2023.2299351

Lengfelder, C., Tapia, H., & Biggeri, M. (2025). Navigating AI with a human development compass - Shaping tomorrow's capabilities. Journal of Human Development and Capabilities, 26(3), 439-448. https://doi.org/10.1080/19452829.2025.2520011

Mühlhoff, R. (2023). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. Big Data & Society, 10. https://doi.org/10.1177/20539517231166886

Nahar, S. (2024). Modeling the effects of artificial intelligence (AI)-based innovation on sustainable development goals (SDGs): Applying a system dynamics perspective in a cross-country setting. Technological Forecasting and Social Change. https://doi.org/10.1016/j.techfore.2023.123203

Nussbaum, M. C. (2000). Women and human development: The capabilities approach. Cambridge University Press.

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

Preble, M. C. (2026). Beyond performance: A diagnostic method for evaluating human-advancement claims in AI-mediated systems [Manuscript].

Sen, A. (1999). Development as freedom. Alfred A. Knopf.

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

van Dijk, J. (2005). The deepening divide: Inequality in the information society. SAGE Publications.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review, 494-620. https://doi.org/10.7916/cblr.v2019i2.3424

Yim, I. H. Y., & Su, J. (2025). Artificial intelligence literacy education in primary schools: A review. International Journal of Technology and Design Education, 35, 2175-2204. https://doi.org/10.1007/s10798-025-09979-w