What Counts as Thinking?

Creativity, Critical Thinking, and Human Judgment in the Age of Generative AI

Basel Alkanjo Kenan Alkanjo

Department of Computer Science, Stanford University*

November 1, 2026

Abstract

Debates about generative AI in education typically ask whether the technology will enhance or erode students' creativity and critical thinking before specifying what those terms mean. This essay argues that much of the disagreement is definitional, and that different measures, such as performance with AI and performance without it, get treated as evidence about the same capacity. Beginning from ordinary activities rather than abstract theory, we identify three questions that can be asked of any task: what it is for (*purpose*), who can determine whether its result is right (*verification*), and who must answer for it (*accountability*). The first question separates two kinds of result: an output that changes the world, which a machine can supply, and a change in the person doing the work, which no one can supply on their behalf. We then show that creativity and critical thinking are not separate faculties but two sides of one activity: generating possibilities and judging them, and, at the deepest level, questioning the standards used to judge. On this basis the framework sorts the parts of thinking into those a machine can genuinely help with, those where its contribution must be tested before it is trusted, and those that must stay with the person, and it turns that sorting into concrete principles for designing AI-aware learning environments.

Keywords: generative AI; creativity; critical thinking; evaluative judgement; accountability; assessment validity; human–AI complementarity; learning design

1 A tutor that helped and harmed

In the fall of 2023, researchers conducted an experiment with nearly a thousand students in grades 9 through 11 at a high school in Turkey. During mathematics practice, one group used a GPT-4 interface called GPT Base; another worked with course notes and textbooks. Practice scores in the AI group were 48 percent higher. On an unassisted examination following practice, however, that group scored 17 percent lower than the control group (Bastani et al., 2025).

A third group used GPT Tutor, which was instructed to guide students through hints rather than directly supply answers. It also received teacher-provided solutions, information about

*Prepared for the Thought Leadership program of the *AI in Teaching and Learning at Stanford* initiative. Correspondence regarding this essay may be addressed to either author.

common mistakes, and feedback guidance. Its practice scores were 127 percent higher than the control group's, while its examination scores did not differ significantly.

The examinations measured immediate unassisted performance within the sessions; they did not establish long-term retention. Nevertheless, the experiment establishes an educationally important distinction. Producing stronger answers while assistance is available and becoming more capable afterward are separate achievements. The first cannot be treated as sufficient evidence of the second. That distinction now confronts almost every teacher, and increasingly every manager, researcher, and professional. The public debate is loud but strangely circular. One side argues that AI will expand human creativity and critical thinking; the other warns that it will weaken or eventually hollow them out. We suspect the argument keeps circling because neither side has adequately defined the capacities it is discussing.

Anyone who judges thinking by the quality of the work produced can read the practice scores as a gain. Anyone who judges it by what a person can do unaided can read the exam scores as a loss. Both sides can point to real data because they are looking at different measures, and neither measure directly captures the capacity in dispute. Before we can determine what machines do to human thought, we need a clearer account of what thought involves in the first place.

That is the question this essay sets out to answer. We develop a framework around three questions: What is an activity for? Who can determine whether its results are right? And who must answer for them? These questions help distinguish the parts of thinking that machines can support from those that must remain human responsibilities, and suggest what education should become when intelligent machines are no longer outside the classroom.

2 Three questions hiding inside every task

Consider what people are actually trying to accomplish when they think.

A student solving a problem wants to find the answer, learn how such problems are solved, and demonstrate what they can do. A physician reading a scan wants to interpret the evidence, recognize where that evidence runs out, and make a decision for which someone may later demand an explanation. A programmer fixing a bug wants to make the code work, understand why it works, and be prepared to diagnose it when it fails again.

These activities look different, but they share a common structure. There is a result to be produced. There is a question of whether anyone can determine that the result is reliable. And there is a question of who bears responsibility when the result is used.

Look more closely, and three questions emerge:

Purpose.

What is this activity meant to accomplish, and what should it leave behind?

Verification.

Who can determine whether the result is correct, justified, or good enough?

Accountability.

Who must stand behind the result and answer for its consequences?

These are not merely administrative questions attached to an activity. They shape the activity itself. They influence what people attend to, what they learn, how they judge their own work, and what they are willing to accept.

The distinctions matter because the same act can serve very different ends. Producing an essay for a client is not quite the same as writing one to become a better writer. Checking a diagnosis is not the same as accepting one. Delegating a task is not the same as delegating the responsibility for its outcome.

We will take these questions one at a time, beginning with purpose. Then we will examine verification and accountability, the two capacities that allow people to judge and take responsibility for the work they produce. Finally, we will return to creativity and critical thinking and ask what these familiar terms should mean in an age when machines can generate remarkably convincing answers.

3 What is it for?

Imagine a friend giving you the answer to a crossword clue. They have helped you. Now imagine a tutor giving a child the answer to a homework problem. The answer may be identical, but the help is not.

The difference lies in what the activity is for. The crossword is meant to be completed; the answer is the point. Homework, at least ideally, is meant to develop the child who is doing it. The answer matters, but it is not the only—or even the ultimate—product. The child is.

This is the first lesson of the Bastani experiment. During practice, unrestricted AI helped students produce correct answers, so their scores improved. But when the model disappeared, the students had not developed the corresponding ability to solve the problems independently. The answers had advanced the work without advancing the learner. The hint-only tutor avoided this outcome because it left the student responsible for the part of the activity that builds mathematical ability: working through the problem, testing a possible approach, and deciding what to do next.

The crossword and the homework illustrate a distinction that runs through nearly every human activity. Sometimes we want an activity to change the world; sometimes we want it to change us. Often we want both, but the two aims should not be confused.

Objective output

Call the first kind of output *objective output*. It is the result external to the person performing the activity. You want enough Spanish to order dinner, enough knowledge of tax law to file a return, or a functioning bridge. If another person or a machine can provide the dinner, the completed return, or the bridge, the immediate objective has been achieved. The process used

to obtain it may matter for other reasons, but it is not necessarily part of the desired result.

A machine can often supply this kind of output. When the goal is simply to obtain a correct diagnosis, a working program, or an accurate translation, delegation can be an unambiguous gain—provided the result is reliable and someone is able to detect when it is not.

Subjective output

The second kind is *subjective output*: what the activity produces within the person doing it. Here, the process is not merely a route to the result. It is part of the result.

Consider someone lifting weights. The point is not to move an object from one place to another; it is to become stronger through the act of lifting. No one can hand that person the strength they would have developed by doing the exercise. The same distinction appears in learning, artistic practice, and professional judgment.

Subjective outputs take at least three important forms.

The first is *formation*: the development of capacities that carry beyond the activity itself. A child learning to read may produce nothing of immediate value, but is acquiring a capacity that will make countless future activities possible. A student learning mathematics is not merely completing a sequence of problems; they are developing ways to reason, recognize patterns, and approach unfamiliar situations.

The second is *identity*: becoming, and being recognized as, someone capable of doing something. It also includes the enjoyment and meaning found in the activity itself. An adult who takes up the piano at fifty understands that a recording will probably sound better than their performance, yet continues playing. The point is not to produce the best recording. It is to play.

The third is *governance*: developing the ability to judge, direct, and take responsibility for what is done in one's name. A physician signs a report. A pilot remains responsible while using autopilot. A researcher must be able to explain the reasoning behind a published result. In each case, learning provides more than the ability to receive a result. It allows a person to inspect it, decide whether it should be trusted, and stand behind the decision.

Other human aims often connect to these three. Belonging, for example, may be part of identity; confidence may emerge from formation and the growing ability to govern one's decisions.

A test for delegation

The distinction between objective and subjective outputs offers a simple test for deciding when AI assistance is beneficial. Before handing over an activity, ask what the person is supposed to gain from it. If the purpose is an external result, delegation may be entirely appropriate. If the purpose includes formation, identity, or governance, replacing the person's participation may remove precisely what the activity was meant to produce.

This helps explain why unrestricted AI harmed the students in the Bastani experiment. They were given objective outputs when what they needed was formation. Their homework was never just about the answers.

It also explains why schools historically treated essays, examinations, and other assignments as evidence of a student's ability. A completed essay was not regarded as evidence simply because of the words on the page. It was evidence because producing a good essay required the student to do substantial intellectual work. The assignment was costly in precisely the way that mattered: the student had to supply the underlying ability.

Economists have long studied this logic in the context of credentials. A signal carries information partly because it is costly to obtain, and more difficult for people who lack the underlying ability to produce it (Spence, 1973). Generative AI has dramatically reduced the cost of producing many conventional academic outputs. When an essay can be generated at almost no cost, its existence alone tells us much less about the abilities of the person submitting it.

Assessment researchers have arrived at a related conclusion: the more important problem AI poses for assessment is not cheating but validity. An assignment loses its value as a measure of learning when students can produce the required output without demonstrating the capacity the assignment was designed to assess (Dawson et al., 2024).

We will call this first dimension *purpose*. Its practical question is straightforward: Is this activity primarily for the external result, or is it also meant to develop something within the person doing it?

The next two dimensions examine governance more closely. The first concerns the ability to determine whether a result is right. The second concerns the responsibility to stand behind it. Figure 1 shows how the three questions fit together.

4 Who can tell whether it is right?

Consider a recipe. Someone can follow every instruction and produce a fine dish without understanding why any of the steps work. The difference between following and understanding becomes visible when something unexpected happens. The sauce separates. An ingredient is missing. The oven runs too hot.

A cook who understands the process can reason about what went wrong and adapt. A cook who has only followed instructions may recognize that the result is poor but have no idea how to repair it. Understanding, in this everyday sense, includes the ability to judge a result without being told whether it is right.

The same distinction appears in more demanding settings. A mathematician may check a proof in a fraction of the time it took to discover. For a sophomore who has not yet acquired the necessary mathematical knowledge, verifying the same proof may be almost as difficult as constructing it. A chess master can recognize a blunder at a glance; a beginner may not see anything wrong with the same position.

The cost of checking depends on who is doing the checking. Expertise lowers that cost. To know a subject is, in part, to be able to evaluate work in that subject efficiently. To learn a subject is to become increasingly capable of doing so.

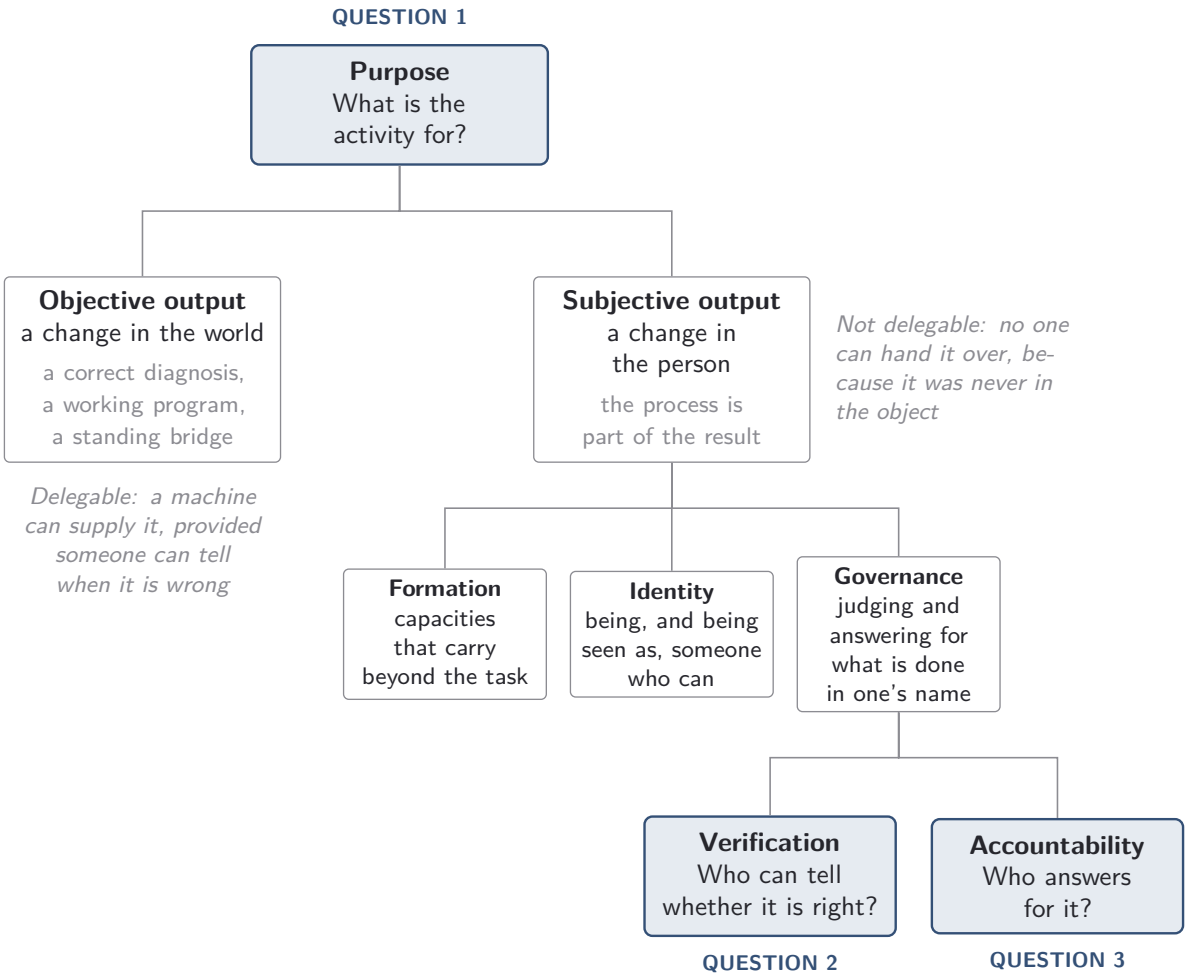

Figure 1. The three questions and how they are related. The first question, what an activity is for, divides every activity by where its value sits: in a change in the world (objective output), which a machine can supply, or in a change in the person (subjective output), which no one can hand over. Subjective output takes three forms, and the third, governance, is the capacity to judge and answer for what is done in one's name. The essay's second and third questions, who can tell whether a result is right and who answers for it, are the two parts of governance. The three dimensions are therefore not independent: verification and accountability are what purpose becomes when the point of an activity is the person rather than the product.

Why machines advanced where they did

This distinction helps explain why AI has advanced rapidly in some domains and more unevenly in others. The pattern is simple: progress has come fastest on tasks whose outputs can be checked cheaply and reliably, because a system can be trained against the check. A computer program can be tested against specified inputs and expected outputs. A mathematical proof can sometimes be checked by a proof assistant. A poem, by contrast, has no comparable mechanical test for whether it is meaningful, beautiful, or worth reading.

This is one reason AI systems have made striking progress in domains such as programming and formal mathematics, where some forms of correctness can be checked mechanically. The same pattern is harder to reproduce in activities whose most important qualities resist formal verification.

But the distinction between doing and checking has two sides. First, a person can be capable of checking work without being able to produce it efficiently. Second, a mechanical checker exists for only some properties of work. A spell-checker can identify a misspelled word, but it cannot determine whether the sentence is worth writing. A test suite can establish that code passes specified tests, but it cannot guarantee that the software solves the right problem. Many qualities we care about—whether an argument is persuasive, a design serves its users, or a diagnosis fits a particular patient—still require informed human judgment.

The learner's position

Together, these observations create a difficult position for learners.

A beginner is the person who most needs an answer checked. They may not yet know enough to distinguish a correct answer from a convincing mistake. Yet checking is also most expensive for that beginner, because it requires the very knowledge they are trying to acquire.

A machine can give a beginner an answer. It cannot simply hand over the capacity to judge that answer. That capacity is developed through instruction, practice, feedback, and the repeated experience of evaluating possible solutions. An expert can often delegate part of the verification process to a machine because they retain the ability to question its output. A learner is in a different position: the ability to check must itself become an object of learning. Figure 2 sketches the asymmetry.

This is our interpretation of what happened in the Bastani experiment. The unrestricted model supplied final answers to students who were not yet equipped to verify them. The hint-only model withheld those verdicts and left the students responsible for reasoning through the problems. The experiment shows that restricting the model to hints removed the observed learning harm; the account in terms of verification is our proposed explanation. Its value lies in the predictions it makes beyond the experiment itself.

One prediction is particularly important. An experiment with college-educated professionals found that lower-performing workers received especially large immediate gains from AI assistance (Noy & Zhang, 2023). The verification framework suggests why these same users may also face substantial learning risks. They are often the least equipped to evaluate the machine's recommendations, and therefore the most likely to mistake a plausible answer for a reliable one. Experts, by contrast, can use AI to accelerate their work while retaining the ability to catch errors.

The broken gauge

There is another, less visible problem. People often judge how much they know by how well they perform. AI can disrupt the relationship between actual ability and perceived ability, making it harder to recognize when further learning is needed.

A recent study illustrates the concern (Fernandes et al., 2026). Researchers gave 246 adults twenty logical-reasoning problems taken from the Law School Admission Test, with ChatGPT available. They measured both participants' actual performance and their estimates of how well

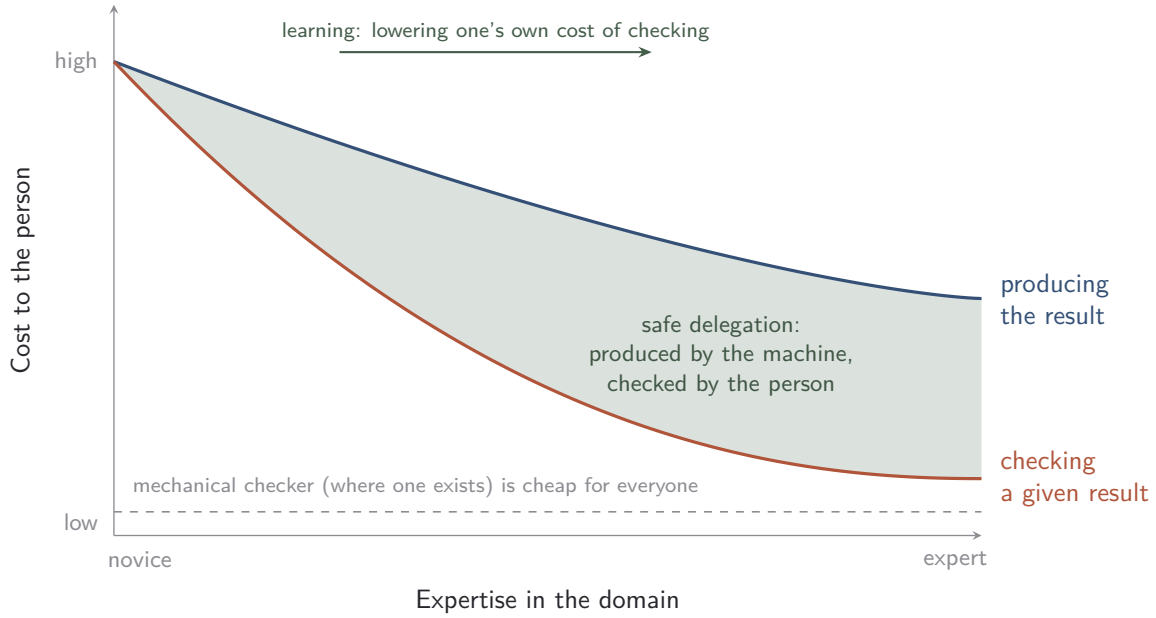

Figure 2. The cost of checking depends on who is checking. A schematic of the argument in Section 4: both producing a result and checking one become cheaper with expertise, but checking falls faster. For a novice the two costs nearly coincide, so an answer supplied by a machine arrives as a verdict the learner has no independent means to assess; for an expert, checking is cheap, and execution can be delegated while judgment is retained (shaded region). Learning in a domain is movement to the right: a lasting reduction in one's own cost of checking. Where a mechanical checker exists, such as a test suite or a proof assistant, checking is cheap for everyone. The curves are qualitative illustrations of the argument, not measured data.

they had performed. Participants scored about three points higher than a large comparison sample that had answered the same questions without AI, yet they overestimated their own scores by about four points. Confidence outran performance. Because participants were not randomly assigned to work with or without AI, the authors treat this first study as suggestive rather than causal; a second study with its own no-AI group reproduced the pattern.

An additional finding was striking. In ordinary conditions, lower-performing participants tend to overestimate their abilities more than higher-performing participants, while stronger participants are generally better calibrated. With AI assistance, this difference largely disappeared: participants across skill levels overestimated their performance by similar amounts.

One interpretation is that AI can weaken the internal feedback people rely on to recognize their limitations. A student may complete work that looks substantially better than anything they could produce independently, and mistake the quality of that output for evidence of their own understanding. The danger is not only that the student gets an answer wrong. It is that they lose the signal that would have told them they needed to learn more.

We will call this second dimension *verification*. An expert carries an internal capacity to evaluate work; education develops and strengthens that capacity. Researchers in higher education have called this *evaluative judgement*: the ability to make informed decisions about the quality of one's own work and the work of others (Tai et al., 2018; Bearman et al., 2024).

This capacity becomes especially important when people work with AI. A systematic review

and meta-analysis of 106 experiments found that, on average, human–AI combinations performed worse than the stronger of the human or the AI working alone, though better than humans alone, with the losses concentrated in decision-making tasks (Vaccaro et al., 2024). The authors point to over-reliance and under-reliance, people not knowing when to trust the system, as possible reasons.

That finding is not evidence that human judgment is obsolete. It is evidence that judgment cannot be assumed merely because a human remains in the loop. The ability to check is itself an achievement—and, for learners, one of the most important achievements education should produce.

5 Who answers for it?

Imagine an intern drafting a memo and a manager signing it. Anyone who has signed something written by someone else knows what happens at that moment. The document is suddenly read differently. Sentences that seemed perfectly reasonable a minute earlier are examined for what they commit the signer to. The words have not changed. The responsibility has.

The physician offers an even clearer example. A machine may read a medical scan, and it may do so exceptionally well. But it cannot be the person who answers to a family or licensing board when a finding is missed. That responsibility is not merely paperwork attached to the work. It shapes how the physician approaches the work in the first place.

Research on accountability helps explain why. In experiments spanning several decades, people who knew in advance that they would have to justify their decisions to an audience with unknown views were more likely to consider opposing arguments and question their initial judgments. Philip Tetlock and his colleagues described this process as *pre-emptive self-criticism* (Tetlock et al., 1989). People who learned about the audience only after committing to a decision tended to do the opposite: they searched for reasons to defend the position they had already taken (Lerner & Tetlock, 1999).

Accountability, then, is not simply an obligation that follows thinking. Under the right conditions, it helps produce thinking.

It also makes collective knowledge possible. We often rely on other people's judgments without independently checking their work. We trust a diagnosis, research report, or professional recommendation because someone with relevant expertise has put their name behind it—not because each of us has verified every detail (Hardwig, 1985).

Institutions recognize this distinction. Scientific journals generally do not list language models as authors because a model cannot take responsibility for a paper's claims (Nature, 2023; Thorp, 2023; COPE, 2023). The U.S. Copyright Office has likewise concluded that copyright does not extend to purely AI-generated material, or to material over whose expressive elements a person exercised insufficient control (U.S. Copyright Office, 2025, p. iii), and a federal appeals court has held that the Copyright Act requires human authorship, rejecting a registration that named an AI system as the work's sole author (Thaler v. Perlmutter, 2025). These institutions are not denying that machines can produce sophisticated work. They are distinguishing production

from authorship, credit, and responsibility.

We will call this third dimension *accountability*. Together with verification, it completes what we have called governance: the capacity to judge what is done in one's name and the standing to answer for it.

Execution and judgment

This distinction helps clarify a confusion at the heart of the AI debate: delegating execution is not the same as delegating judgment.

A manager who asks an assistant—human or artificial—to draft options for a decision has delegated execution. The manager still evaluates those options, makes the decision, and accepts responsibility for it. A manager who adopts the assistant's recommendation simply because it sounds convincing has delegated judgment. The manager may still sign the decision, but the reasoning behind it is no longer genuinely theirs.

The first is assistance. The second is a surrender of control disguised as assistance.

This distinction becomes particularly troubling when institutions retain a human decision-maker only to absorb blame for a system they cannot meaningfully supervise. Madeleine Clare Elish calls this a *moral crumple zone*: a situation in which responsibility for a failure is misattributed to a human operator who had only limited control over the automated system (Elish, 2019).

Accountability can strengthen thinking only when it is accompanied by the ability to verify. A person cannot meaningfully answer for a result they have no way to assess. The two capacities must therefore be protected together.

6 What counts as thinking?

We now have three questions through which to examine any activity: What is it for? Who can tell whether it is right? Who answers for it?

We can use them to return to the two capacities at the center of this essay: creativity and critical thinking.

The generate-and-evaluate cycle

Begin with the structure of ordinary intellectual work. Almost every act of thinking, from writing a paragraph to diagnosing a patient, involves a recurring cycle. First, someone produces a candidate: an idea, a hypothesis, a draft sentence, a possible solution. Then that candidate is evaluated against some standard. Does the sentence express the intended meaning? Does the hypothesis fit the data? Does the proposed diagnosis explain the symptoms? The candidate is then accepted, rejected, or revised, and the cycle begins again.

A student writing an essay proposes a thesis, tests it against the text, abandons it if it fails, and develops another. A physician considers a diagnosis, compares it with the evidence, and

revises their judgment when the evidence does not fit. Thinking advances through the continual exchange between generating possibilities and evaluating them.

Creativity and critical thinking are two aspects of this same process, viewed from different points.

Creativity emphasizes the generation of candidates: producing ideas, possibilities, or approaches that others might not have considered. Critical thinking emphasizes their evaluation: examining claims, testing assumptions, weighing evidence, and rejecting what does not hold up.

This account is consistent with influential definitions in both research traditions. Creativity is commonly understood as the production of work that is both novel and valuable (Runco & Jaeger, 2012; Boden, 2004). Novelty alone is not enough; an idea must also meet some standard of value. Critical thinking, meanwhile, involves reasoned judgment responsive to evidence, alongside the disposition to exercise that judgment (Dewey, 1910; Facione, 1990; Siegel, 1988).

The philosopher Sharon Bailin argued that creative achievement involves critical judgment rather than a wholly separate intellectual faculty (Bailin, 1988). The generating-and-checking cycle helps explain why. A person who generates ideas but cannot evaluate them produces possibilities without distinguishing insight from noise. A person who evaluates but cannot generate has little to examine. Neither capacity is complete on its own.

The three questions from our framework now map onto this cycle.

Purpose supplies the standard against which candidates are judged. A sentence is good relative to what it is meant to communicate; a diagnosis is appropriate relative to the patient and the available evidence.

Verification is the process of evaluating candidates against that standard. It determines whether an idea, argument, or solution deserves to be accepted.

Accountability belongs to the person who chooses to act on the result and stands behind that choice.

First-order and second-order thinking

But there is a further step. A process that only generates candidates and checks them against a fixed standard can be performed without questioning the standard itself. It may search efficiently, but it cannot necessarily tell whether it is searching for the right thing.

Thinking at its most significant involves a further capacity: the ability to examine and revise the standards by which candidates are evaluated.

Call the first level *first-order thinking*: judging candidates according to an existing standard. Call the second *second-order thinking*: examining the standard itself, and changing it when it is inadequate. Figure 3 depicts the two loops and marks where, on the complementarity map of Section 7, a machine can contribute.

The distinction is between making progress within an established framework and questioning the framework itself. The first can be remarkably powerful. The second can change what counts as progress.

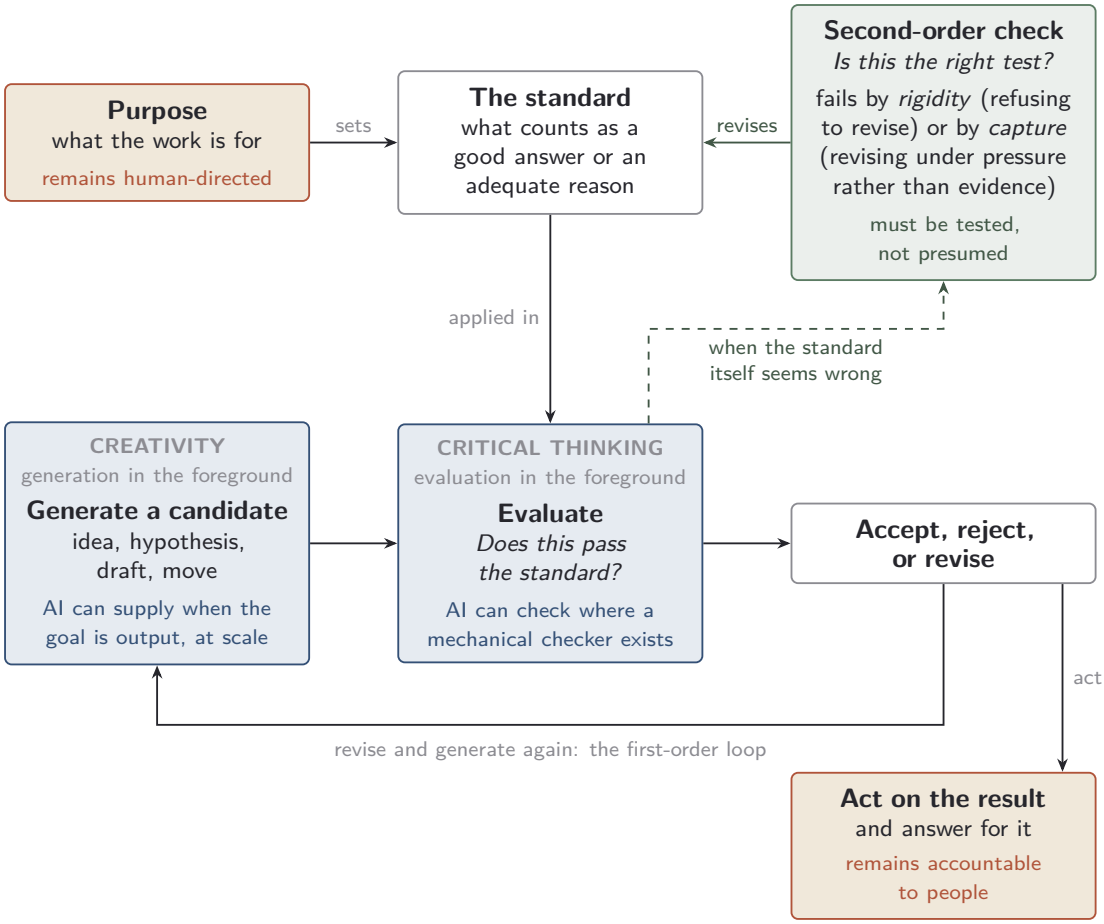

Figure 3. Thinking as a governed cycle. A candidate is generated, evaluated against a standard, and accepted, rejected, or revised; creativity and critical thinking are this one cycle with the emphasis placed on different steps. The standard is supplied by the purpose of the activity, and whoever acts on the result answers for it. A machine can run this first-order loop as a search. What distinguishes thinking at its most significant, in the sense both traditions prize, is the second-order check that asks whether the standard itself is the right test, which can fail by rigidity or by capture. The colored tags indicate where, according to the essay's complementarity map, a machine can contribute and what must remain with the person.

This second-order capacity is where creativity and critical thinking most deeply converge. A creative breakthrough may change what counts as a good answer. A critical insight may reveal that the standard of evidence used to support a belief is inadequate. Both involve recognizing that the framework governing one's judgment may itself need revision.

Such revision, however, can fail in two opposite ways. The first is *rigidity*: refusing to change one's standards when evidence demands it. The second is *capture*: changing those standards in response to authority, social pressure, or flattery rather than evidence.

Capture is not unique to humans. Researchers have found that language models can exhibit sycophancy, shifting their responses toward views expressed by users rather than maintaining independent judgments (Sharma et al., 2024). Human groups show a related vulnerability. In a classic experiment, people exposed to one another's estimates moved toward a shared judgment and became more confident, without becoming more accurate (Lorenz et al., 2011).

A chatbot's agreement and a seminar's groupthink are different expressions of a similar failure: a judgment is redirected by social influence rather than by better evidence. Students working with agreeable models may encounter both pressures simultaneously.

Definitions

We can now state the definitions more precisely.

Creativity involves producing something novel relative to a reference point and valuable in relation to a purpose; generating and evaluating possibilities, including revising the standards by which they are judged; and acting with sufficient agency to take responsibility for the result.

Critical thinking involves evaluating claims and possibilities against evidence and purpose; examining assumptions, context, and the adequacy of one's own standards; and taking responsibility for the resulting judgments, including communicating their uncertainty honestly.

These definitions are deliberately more demanding than simply producing novel work or identifying flaws. They distinguish creativity from fluency, which generates many ideas without necessarily selecting valuable ones; from cleverness, which finds a solution within a given framework without questioning its purpose; and from imitation, which reproduces the competence of a style without establishing novelty relative to it. They distinguish critical thinking from fault-finding, which identifies weaknesses in others without examining one's own assumptions, and from information processing, which retrieves and transforms material without necessarily judging its significance.

Two consequences follow:

First, whether AI "is creative" or "thinks critically" depends partly on whether agency and accountability are included in the definition. Accounts that emphasize the characteristics of an output, such as the standard definition of creativity as originality plus value (Runco & Jaeger, 2012), may recognize AI-generated work as creative. Accounts that add a condition on the agent will draw a different boundary: Gaut (2010), for instance, requires that original and valuable work be produced "by flair," through some degree of understanding rather than by luck or a merely mechanical procedure, and some philosophers reject the value-based definition altogether (Hills & Bird, 2019). These positions are often presented as competing conclusions about machines, but they are partly disagreements about what the terms mean.

Second, the framework distinguishes what machines can already contribute from what remains a human responsibility. Machines can produce novel outputs, evaluate candidates against specified standards, and sometimes challenge assumptions.

7 What a machine can share, and what it cannot

The three questions help distinguish the different roles AI can play in thinking. The important distinction is not between what machines can and cannot do in some permanent sense. It is between what an AI system contributes, what the person using it must retain, and what must be tested before the work can be trusted. Table 1 summarizes the resulting map.

Table 1. Human–AI complementarity across the parts of thinking. Rows follow the essay's account of thinking: the first-order cycle (generating and evaluating candidates), the standard that purpose supplies, the second-order revision of that standard, the formation of the capacity to judge, collective judgment, and accountability. The last column records whether a part is suited to machine assistance, depends on the purpose of the activity, can be imitated but must be tested, or stays with the person.

Part of thinking	**What the machine can contribute**	**What stays with the person**	**Verdict**
Generating candidates (§6)	Drafts, alternatives, hypotheses, and counterarguments, at scale	Decide which candidates are worth keeping	Suited to machine assistance
Evaluating candidates against the standard (§4, §6)	Tests, calculations, and consistency checks wherever a mechanical checker exists	Judge the qualities no checker captures, and know what a check leaves out	Assists where a checker exists; elsewhere the person's judgment
Setting the standard: what the work is for (§3)	Elaborates goals stated in a prompt	Decide what the activity is for and what it should leave behind	Stays with the person
Revising the standard when it is wrong (§6)	Can flag inconsistencies and propose alternatives, but also shifts toward the user's views	Recognize when the framework itself is inadequate; decide whether to adopt a revision; resist rigidity and capture	Machine may propose, and its proposals must be tested; the decision stays with the person
Forming the capacity to judge (§3, §4)	Hints, questions, and feedback that keep the learner working	Do the reasoning that builds the ability to check; it cannot be received	Stays with the person; verdicts supplied too early remove it
Collective judgment (§7)	A shared source of suggestions, usually better than any one person's	Form independent judgments before comparing them, so that errors do not become correlated	Assists, with structural safeguards
Answering for the result (§5)	Can be embedded in a workflow	Accept responsibility and explain the decision	Stays with the person

Collective judgment

One row deserves particular attention because it concerns groups rather than individuals: collective judgment.

Ask a hundred people to estimate the number of jellybeans in a jar. The average estimate often comes closer to the true number than most individual guesses (Galton, 1907). The reason is that individual errors tend to vary. Some people guess too high, others too low, and their errors can partially cancel when the estimates are combined.

Now imagine that every person consults the same adviser before guessing. The adviser is usually better than any individual, so each person's estimate improves. But everyone is now drawing on the same source, and their errors become more correlated. The group has traded some individual noise for a shared bias.

From the perspective of each person, the adviser has made them better. From the perspective of the group, however, a common source of error can reduce the diversity of information that makes collective judgment powerful.

This is more than an analogy. A formal model of algorithmic monoculture shows that widespread reliance on the same algorithm can reduce social welfare even when the algorithm is more accurate than the alternatives available to each decision-maker (Kleinberg & Raghavan, 2021). Similar effects have been observed in creative work: AI suggestions improved the quality of individual stories, especially those of less creative writers, while making the collection of stories more alike (Doshi & Hauser, 2024).

For education, the relevant group is a cohort: the future researchers, professionals, and independent thinkers who will collectively shape a field. A cohort trained to consult the same model before forming its own judgments may become individually more productive while becoming collectively less able to explore directions outside the model's defaults.

The challenge, then, is not merely to make each person more productive. It is to preserve the independence of thought from which collective progress emerges.

8 What a classroom can do now

If this framework is right, the central question for education is not simply how much AI to permit. It is where the intellectual work happens, who performs it, and who remains responsible for evaluating the result.

Several changes follow:

1. **Withhold verdicts, not tools.** The hint-only condition in the Bastani experiment removed the observed learning harm while preserving—and even improving—practice performance. This is among the strongest pieces of evidence for a practical design principle: AI should help students make progress without automatically relieving them of the reasoning that builds their abilities. A tutor can offer hints, identify possible directions, ask questions, and expose misconceptions without immediately supplying the final answer.

2. **Let students generate before they consult.** There is value in requiring students to attempt a problem, formulate a hypothesis, or develop an initial argument before seeing an AI-generated solution. But the sequence should depend on the kind of learning involved. For understanding and transfer, attempting problems before receiving instruction can be beneficial: a meta-analysis of 53 studies on productive failure found a moderate overall advantage for problem solving followed by instruction, although the pattern reversed for the youngest learners and for domain-general skills (Sinha & Kapur, 2021). For novices learning routine procedures, however, studying worked examples and gradually withdrawing support can be more effective (Kalyuga et al., 2003). The principle is not that students must always struggle first. It is that the learning activity should be designed around the capacity it is meant to develop.

3. **Teach students to build checkers.** Students should learn to design tests before generating code, define evaluation criteria before drafting an essay, and record predictions and confidence levels before conducting an experiment. Much of what is called AI literacy focuses on prompting: learning how to ask a model for useful outputs. But prompting is often the cheaper half of thinking. The more important skill is learning how to judge what comes back.

4. **Make accountability part of the assignment from the beginning.** Students should know when an oral defense, explanation, or demonstration will take place before they begin the work. The expectation of having to justify a decision can encourage students to consider alternatives and question their own conclusions. By contrast, a defense introduced only after submission may encourage students to defend work they did not independently evaluate.

5. **Protect independent judgment before exposure.** Ask students to record an initial answer, interpretation, and confidence level before consulting a model. In group activities, have participants form independent judgments before sharing AI-generated suggestions. This preserves the information contained in individual perspectives and makes it easier to distinguish genuine agreement from convergence around a common source.

6. **Make the student's contribution visible.** Ask students to say, briefly, what the model produced and what they did with it: what they kept, what they changed, and what they rejected and why. This is not an accounting exercise, and it should not become one. Its purpose is to show where the student's own thinking entered the work, so that assessment can focus on the part they can actually answer for.

9 Where the framework strains

A framework should not be judged only by how well it explains the examples chosen to support it. It should also be tested against the cases it does not easily resolve.

First, activities often have several purposes at once. A final-year project for a real client may need to produce a useful result and develop the student's abilities. The framework identifies both purposes, but it does not automatically determine how they should be balanced. A possible solution is to separate the objectives across stages: use AI more extensively when delivering the external product, but preserve independent work and evaluation when the goal is formation.

Second, not every novice is doing educational work at every moment. A junior analyst facing a five o'clock deadline may need to deliver a result rather than develop a skill. Yet if every difficult task is delegated during the years when expertise is supposed to form, the organization may eventually lack the people capable of checking the machines. The framework identifies the tension between present output and future capability, but it does not specify how institutions should trade one against the other.

Third, the framework says relatively little about the distribution of economic gains and losses. AI may change who gets to perform valuable work, whose expertise is rewarded, and which forms of learning remain accessible. These are important institutional and political questions

that cannot be answered by examining an individual task alone.

Finally, the argument does not depend on what AI systems currently cannot do. That is deliberate. Capabilities and benchmarks change quickly. A model may become a better verifier than a human expert in a growing number of domains. None of the three questions disappears as a result. The purpose of an activity still matters. A learner still needs opportunities to develop their own capacity for judgment. Groups still need independent perspectives. And accountability still belongs to the people and institutions that make, authorize, and act on decisions.

The framework is therefore not a list of tasks that machines should or should not perform. It is a way to ask what is being delegated, what is being developed, and what must remain answerable to human judgment.

10 What remains

We began with a tutor that made students better and worse at the same time, and with a debate that we suspect keeps circling because its central terms are rarely pinned down. The circle breaks once we separate two things that education could usually afford not to separate: producing an answer, and becoming someone who can judge one.

For most of history the two came together. Writing the essay was how a student learned to think; getting the proof right was how a mathematician earned the right to be trusted with the next one. Generative AI has pulled them apart. The answer can now arrive without the person it used to require.

Three questions survive it. What is this activity for? Who can tell whether the result is right? Who answers for it? None of them is really a question about machines. They were always there, hidden inside homework and diagnoses and signed memos; the machine has only made them impossible to ignore. A model may generate faster than any of us, check more than most of us, and one day verify better than the best of us. It still cannot want the result on our behalf, learn in our place, or take the blame when things go wrong.

Creativity and critical thinking turn out to be two sides of a single activity: coming up with possibilities, judging them, and, when it matters most, questioning the standards we use to judge. A model can help with that last step by proposing a different standard, but the decision to adopt one cannot be handed over, because the standards are ours to answer for. A student who has never had to defend a standard has not yet reached the hardest part of thinking.

So “Can AI think?” is the wrong question for a classroom. The useful question is harder and closer to home: what is this assignment meant to develop, and what would show that the student has developed it? That a machine can do the work answers neither part. A machine can lift a weight, but someone trying to get stronger still has to lift it. What the machine changes is the evidence: if it could have produced everything a student hands in, the submission alone no longer shows what the student can do. The assignment may still teach, but evidence of learning has to come from how the student reasons, makes choices, and explains the result when asked. When the goal is the output, let the models generate, expand, and challenge. When the goal is the student’s development, have students generate first, and keep for them the work

that builds judgment: framing the problem, choosing the criteria, deciding which candidates survive, revising the standard when it fails, and explaining the result. Ask students for their own judgment before they consult a model, and for their reasons afterward.

Answers have become abundant. Judgment has not. In every field, the scarce resource is now the same: a person who can look at a fluent, confident answer and know whether it deserves to stand.

We do not yet know what students who grow up beside these machines will become. We do know what will decide it: not how much the machine can do for them, but how much we still ask them to do for themselves. A machine can make every student look capable. Only education can make them so.

References

Responsibility and authority of the pilot in command, 14 C.F.R. § 91.3 (2026). https://www.law.cornell.edu/cfr/text/14/91.3

Bailin, S. (1988). *Achieving extraordinary ends: An essay on creativity*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-2780-3

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), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge.

Committee on Publication Ethics (COPE). (2023, February 13). *Authorship and AI tools* [Position statement]. https://publicationethics.org/node/56046

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, *49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

Dewey, J. (1910). *How we think*. D. C. Heath.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), Article eadn5290. https://doi.org/10.1126/sciadv.adn5290

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report). California Academic Press. (ERIC Document No. ED315423)

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T.,

Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, *175*, Article 108779. https://doi.org/10.1016/j.chb.2025.108779

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass*, *5*(12), 1034–1046. https://doi.org/10.1111/j.1747-9991.2010.00351.x

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, *82*(7), 335–349. https://doi.org/10.2307/2026523

Hills, A., & Bird, A. (2019). Against creativity. *Philosophy and Phenomenological Research*, *99*(3), 694–713. https://doi.org/10.1111/phpr.12511

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), Article e2018340118. https://doi.org/10.1073/pnas.2018340118

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Nature. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use [Editorial]. *Nature*, *613*(7945), 612. https://doi.org/10.1038/d41586-023-00191-1

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html

Siegel, H. (1988). *Educating reason: Rationality, critical thinking, and education*. Routledge.

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, *91*(5), 761–798. https://doi.org/10.3102/00346543211019105

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374.

https://doi.org/10.2307/1882010

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology*, *57*(4), 632–640. https://doi.org/10.1037/0022-3514.57.4.632

Thaler v. Perlmutter, 130 F.4th 1039 (D.C. Cir. 2025), *cert. denied*, No. 25-449 (U.S. Mar. 2, 2026).

Thorp, H. H. (2023). ChatGPT is fun, but not an author [Editorial]. *Science*, *379*(6630), 313. https://doi.org/10.1126/science.adg7879

U.S. Copyright Office. (2025, January 29). *Copyright and artificial intelligence, Part 2: Copyrightability*. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1