WORKING PAPER

Who Is Training Whom

Human Cognitive Drift and Machine Degradation in Coupled AI Knowledge Systems

Tommy N. Turner
Independent Policy Researcher, Williamsburg, Virginia
ORCID 0009-0005-7422-9881
October 2026
doi:10.5281/zenodo.23453582

More of the author's work: tommynturner.com

Abstract

AI literacy is usually taught as a transaction. A user learns to write a prompt, check the answer, protect sensitive information, and recognize a hallucination. Those skills describe one encounter at a time. Generative AI operates in a longer loop. A user's assumptions shape the prompt; the answer reshapes the next question; repeated outputs change what the user notices and accepts; AI-assisted work enters reports and the public record; later systems retrieve that material and may learn from it. Errors can come from the model, from the user, or from the two working on each other.

This paper treats the person and the model as components of one coupled knowledge system and joins the literatures on model collapse, retrieval and source dependence, automation bias, cognitive offloading, sycophancy, metacognition, and AI literacy. It contributes a framework of seven paired human and machine vulnerabilities, offered as reinforcing pathways for training and research to test rather than as proven causal laws, and converts the framework into three levels of literacy and a four-stage routine that scales with the stakes of the task. The empirical anchor is a pre-registered audit of four generative search products on six Virginia public-record questions, run twice each in September 2026. Across 48 answers the products displayed 185 citations that reached 136 distinct pages; thirteen were the record itself, none documented an independent check, and 90 could not be followed from what the product showed. The audit, together with the fictional Mountains of Kong, makes the paper's central point: a well-written answer with genuine citations can still mislead when derivative sources are counted as independent evidence, or when the user and the system quietly revise the question together. The paper closes with a course scaffold built on the framework and the deposited audit data.

Keywords: *AI literacy, cognitive drift, model collapse, automation bias, provenance, source independence, generative search, human-AI interaction*

1. Introduction

John Cary's 1805 map of Africa shows a long mountain chain across the continent's western interior, labeled the Mountains of Kong. The engraving is careful and authoritative. The range looks no more doubtful than the rivers and coastlines around it.

There was no such range.

The feature grew from observation, inference, hearsay, and cartographic convention. James Rennell's map for Mungo Park's 1799 account of West Africa gave it an influential early form. Later mapmakers copied, extended, and adjusted it. By the nineteenth century a reader could find the mountains in several reputable atlases, and the copies did not look like copies. Each new rendering seemed to add confirmation. After Louis-Gustave Binger crossed the region in the late 1880s, his observations removed the range from most maps, though remnants persisted into the twentieth century (Bassett and Porter 1991).



Figure 1. A continent made visible and a mountain range made credible. Source: John Cary, A New Map of Africa from the Latest Authorities, 1805. Public domain, via Wikimedia Commons.

The Kong Mountains are sometimes called a pre-digital hallucination. The comparison holds only at the start. The original range may have rested on an unsupported inference, but hallucination does not explain a century of credibility. Later cartographers reproduced an accepted authority accurately. A scholar could cite several real maps. A teacher could point to an apparent consensus. The durable error was a property of the knowledge system: uncertain information lost its ancestry as it passed through credible hands.

Generative AI makes the same problem faster and harder to see. A model can retrieve five articles that descend from one report, summarize them faithfully, and cite each correctly. The answer still overstates the evidence. Five links are not five observations. AI-assisted summaries then enter websites, presentations, memoranda, and search indexes. Later systems retrieve them, and future models may meet them in training data, at which point repetition starts to resemble corroboration.

The person is not standing outside this process. A user asks a question, receives an answer, and uses it to frame the next question. The model follows the new frame. The next answer becomes a premise for another prompt. Small reasonable steps can carry an inquiry away from the objective that started it, and a system that agrees easily and produces support on demand makes the steps easier to take.

This paper began with a rough metaphor. I had been describing "AI inbreeding," systems consuming increasing amounts of machine-produced material. The human side of the loop needed its own name, and the literature had already supplied one: the user latches onto a response, pushes the inquiry in that direction, and receives further responses that make the direction feel natural. Lopez-Lopez and

colleagues (2026) call this cognitive-behavioral drift. The machine and the person can arrive at a mutually satisfying answer without either noticing how far the inquiry has moved.

The comparison needs technical boundaries. An ordinary chat does not retrain the foundation model turn by turn. Model collapse concerns changes across training generations or other recursive data processes. Cognitive drift can occur within one long session or across repeated use while the weights never change. The two processes run on different mechanisms and time scales. What connects them is that both narrow the information available to the next step.

Recent work describes sustained human and AI interaction in these terms. Lopez-Lopez and colleagues (2026) define cognitive-behavioral drift as gradual change in beliefs, confidence, attention, and readiness to act during extended interaction with adaptive AI. Glickman and Sharot (2025) show experimentally that human-AI feedback loops amplify bias when the AI is biased and improve accuracy when it is accurate. Nettasinghe and Zhao (2026) and Wu and colleagues (2026) model humans, models, and shared knowledge as coupled systems with both productive and degenerative regimes. Peterson (2025) models the case in which learners take a discount on AI-mediated knowledge and public belief drifts away from the truth as a result. A claim of novelty is unnecessary. The useful question is practical:

> *How should AI literacy training change when the learner and the model are treated as components of a shared recursive knowledge system rather than as an independent user and a static tool?*

The paper makes four contributions. It joins literatures usually taught separately: machine degradation, retrieval and provenance, human reliance on automation, and AI literacy. It proposes seven paired human and machine vulnerabilities as a working classification of reinforcing pathways. It reports a pre-registered audit of what the citations displayed by four deployed products reach, which gives the provenance pair an empirical floor. And it converts the classification into a training routine that scales with the stakes, reversibility, and verification cost of the task, with a course scaffold in the appendix.

The framework is conceptual. It does not establish a one-to-one causal relationship for every pair, and it reports no controlled evaluation of the training it proposes. The audit measures one thing well, the distance between what a citation list displays and what it reaches, on six questions in one domain. The claim is narrower than the title: training that treats hallucination as the central failure and source checking as the central remedy misses important ways that machines and users jointly degrade an inquiry.

2. Two Forms of Recursion

2.1 Machine-side degradation

The machine-side concern is easiest to see in research on recursive training. Shumailov and colleagues (2024) show that when generative models are repeatedly trained on model-produced data under replacement conditions, the learned distribution departs progressively from the original. Low-probability events disappear first. Beyond acquiring a few false statements, the model loses the parts of the distribution that made the source data varied.

The phrase "model collapse" can make the result sound broader and more inevitable than the evidence warrants. Synthetic data are not contaminated by definition, and recursion does not guarantee collapse. Gerstgrasser and colleagues (2024) find that accumulating real data alongside synthetic data, rather than replacing it, avoids the collapse seen when each generation trains only on the last. Bertrand and colleagues (2024) prove that iterative retraining is stable when the initial model is good enough and the share of clean data is large enough. The other direction has evidence too. Alemohammad and colleagues (2024) show that without fresh real data each generation, quality or diversity falls. Dohmatob and colleagues (2024a) recast collapse as a change in scaling laws, and Dohmatob and colleagues (2024b) find in a regression setting that even a small fraction of synthetic data can cause it, with larger models sometimes amplifying the effect. Bohacek and Farid (2023) and Briesch and colleagues (2023) report the same pattern in images and in a verifiable logic task: correctness holds for a while, diversity declines, and fresh data slows the decline without stopping it. The danger the literature converges on is uncontrolled recursion with weak provenance, poor filtering, and no anchoring distribution.

A related process runs at retrieval time with no retraining at all. Druck and Smith (2026) report that systems relied disproportionately on references the same model family had written; under their criterion, 1,216 of 1,528 simulations ended in what they call retrieval-augmented-generation collapse. That finding is new and should not be generalized past the design tested. It is consistent with earlier results. Dai and colleagues (2024) found that neural retrievers rank LLM-generated documents above human-written ones on the same content, a source bias that survives re-ranking. Chen and colleagues (2024b) simulated the loop and watched machine text rise in search rankings until human contributions thinned out. Pan and colleagues (2023) showed that a modest volume of model-generated misinformation degrades open-domain question answering. A model need not wait for a future training cycle to encounter its own descendants.

The corpus itself is already changing. Thompson and colleagues (2024) found machine-translated text dominating the web in lower-resource languages. Hanley and Durumeric (2024) tracked synthetic articles rising 57 percent on mainstream news sites and 474 percent on misinformation sites over sixteen months. Liang and colleagues (2024) estimated that between 6.5 and 16.9 percent of the text in recent conference peer reviews was substantially LLM-modified. Brooks, Eggert, and Peskoff

(2024) flagged more than five percent of new English Wikipedia articles at a one percent false-positive rate. None of these numbers measures collapse. They describe the raw material for retrieval and for the next training run, and it is no longer what it was.

Machine degradation in this paper therefore covers more than catastrophic collapse. It includes loss of distribution tails, homogenization, self-retrieval, context drift within an interaction, and the treatment of duplicated or dependent records as separate evidence. These mechanisms differ. Their common feature is that the information available to the next output becomes narrower, more derivative, or worse weighted.

2.2 Human-side drift

Human drift is broader than accepting a false answer. People routinely offload memory and computation to tools. Offloading can improve performance and free attention for harder work; Risko and Gilbert (2016) treat it as a normal metacognitive strategy rather than a defect. Trouble begins when delegation hides what the user no longer knows, checks, or can do alone.

The evidence on that hiding is older than generative AI. Sparrow, Liu, and Wegner (2011) found that people who expect information to stay available remember where to find it rather than what it says. Fisher, Goddu, and Keil (2015) found across nine experiments that searching online inflates people's estimates of their own internal knowledge; Ward (2021) found that Google use raises confidence in unaided memory that the user does not have. Grinschgl, Papenmeier, and Meyerhoff (2021) showed the trade directly: offloading raised immediate task performance and lowered later memory, even when a memory test was announced in advance. Rozenblit and Keil (2002) had already documented the illusion of explanatory depth, the sense of understanding a mechanism that dissolves when one is asked to explain it. Messeri and Crockett (2024) argue that AI tools built for scientific work can exploit exactly these limits, producing more output and less understanding.

Automation research adds the behavioral half. Parasuraman and Riley (1997) set out use, misuse, disuse, and abuse of automation, with misuse meaning over-reliance. Skitka, Mosier, and Burdick (1999) showed omission and commission errors in a simulated flight task when an automated aid was present. Parasuraman and Manzey (2010) describe automation bias and complacency as attentional phenomena arising from the person, the task, and the system together, and note that training and instruction do not prevent them. Goddard, Roudsari, and Wyatt's (2012) systematic review found both error types across 74 studies and identified accountability, workload, and presenting information rather than recommendations as mitigators. The effect is not a matter of naive users. Logg, Minson, and Moore (2019) found that people follow advice more readily when they believe it comes from an algorithm, an effect that fades with expertise, while Dietvorst, Simmons, and Massey (2015) found the opposite once the algorithm is seen to err. Vaccaro, Almaatouq, and Malone (2024), across 106 studies, found that human-AI combinations on decision tasks performed worse on average than the

better of the two alone. Bansal and colleagues (2021) found that explanations raised acceptance of the AI's recommendation whether or not it was right.

Generative AI adds a conversational surface that explains, reassures, revises, and agrees. The system goes beyond displaying a recommendation; it helps the user construct the reasoning for accepting one.

The term cognitive drift fits when that construction unfolds over time. Lopez-Lopez and colleagues (2026) emphasize that a conversational agent shapes which information is accessed, how questions are framed, how evidence is interpreted, and when action feels warranted. Drift may change beliefs, but it can also change standards: how much uncertainty is tolerated, what counts as adequate verification, which contrary evidence still feels relevant.

Not every change of belief is drift. A person who updates on sound evidence is learning. Drift is a change that goes insufficiently noticed, examined, or tied to better evidence. Direction alone cannot make the distinction. It takes a record of the starting objective, the evidence met along the way, and the reasons for revision.

Evidence on generative AI at work and in school is arriving quickly and is still uneven. Lee and colleagues (2025), surveying 319 knowledge workers about 936 real tasks, found that confidence in the AI went with less reported critical-thinking effort and confidence in oneself went with more. Bastani and colleagues (2025), in a field experiment with nearly a thousand high-school students, found that unrestricted access to a GPT tutor raised practice performance and that students performed worse than never-assisted peers once access was withdrawn; a tutor with guardrails avoided the loss. Stadler, Bannert, and Sailer (2024) randomized students to ChatGPT or Google for a scientific-inquiry task and found lower cognitive load and lower-quality reasoning in the ChatGPT group. Fan and colleagues (2025) found that ChatGPT support improved essay scores without improving knowledge gain or transfer, and named the pattern metacognitive laziness. Tankelevitch and colleagues (2024) argue that generative systems raise the metacognitive demands on users, requiring more monitoring and control rather than less. Kosmyna and colleagues (2025) report weaker neural connectivity and poorer recall of one's own text among LLM-assisted writers, in a small unreviewed sample. No study in this set shows that AI causes a broad decline in reasoning. What they show is narrower and more useful: verification and integration have become part of the work, and confidence decides whether people do them.

The concept is starting to become an audit problem. Li and Zhu (2025) propose a protocol for examining cognitive drift in AI-driven recommendation, using a health case to show how user preference and system output can be tracked over rounds. The protocols are early, and they are a start on measurement.

3. Different Clocks in One System

The machine and human processes should not collapse into one metaphor. They run on different clocks and leave different evidence.

Table 1. Related processes run on different clocks.

Process	Typical time scale	Observable record
Conversation context	Minutes to hours	Prompts, assumptions, and retrieved material carried across turns
Human cognitive and behavioral drift	One session to repeated use	Changed objectives, confidence, verification habits, and delegated tasks
Organizational and public record	Days to years	AI-assisted documents, repositories, websites, and reused summaries
Model training and ecosystem change	Training cycles to model generations	Corpus composition, provenance, distribution coverage, and later retrieval

A user can drift while no model is retrained. A model can degrade across generations while every user stays careful. The coupled-system argument begins where the processes meet: prompts, retrieved documents, AI-assisted work products, organizational repositories, public websites, preference signals, and later training corpora. These are transfer points. The output of one process becomes the input to another there.

4. Evidence Sources and Apparent Independence

4.1 What the maps counted

The Kong Mountains illustrate the difference between the number of documents and the number of independent observations. A shelf of atlases can hold many publications and a small family of claims. Bassett and Porter (1991) show how cartographic authority, limited access to the interior, and the practical reuse of earlier maps let a speculative feature harden into convention. The mapmakers were not fools. Copying respected authorities was often rational. The visible count of maps concealed the dependence among them.

Economists call the human side of this correlation neglect. People treat related signals as independent and grow too confident when several reports share an origin (Enke and Zimmermann 2019). DeMarzo, Vayanos, and Zwiebel (2003) built the network version: agents who fail to account for repetition in what they hear end up weighting the best-connected voices, whatever those voices know. Weaver and colleagues (2007) found the behavioral core in six experiments: people infer that a familiar opinion is a widely held one even when its familiarity comes from one person saying it several times, and the effect survives their knowing it was one person. Ortoleva and Snowberg (2015) trace political overconfidence and extremity to the same neglect. Online search intensifies the

problem because duplication is rarely exact. A press release becomes a news report, a trade association post, a slide deck, and a summary. The wording changes and the ancestry stays.

Retrieval-augmented generation gives the issue computational form. Lewis and colleagues (2020) introduced the architecture. Naphade (2026) tested language models on 1,635 controversial questions paired with 15,058 documents and found that paraphrases of one argument could outweigh distinct independent support, and that order mattered. Wan, Wallace, and Klein (2024) found that models weight a source by its relevance to the query and largely ignore whether it carries references or reads neutrally. Xie and colleagues (2024) found models highly receptive to coherent external evidence that contradicts their own memory, and strongly biased toward whichever evidence agrees with it. Xu and colleagues (2024) survey these conflicts. Cuconasu and colleagues (2024) found that high-scoring passages that do not carry the answer hurt accuracy. Chen and colleagues (2024a) found models struggling to reject a question when the retrieved passages cannot answer it. Liu, Zhou, and Wang (2026) distinguish presentation multiplicity, the number of records shown to the model, from epistemic multiplicity, the number of independent evidentiary roots, and show that false splitting and false merging change the weight a model gives evidence. Rahadi (2026) proposed the counting rule directly: a retrieval pipeline scores every document against the query and none against each other, so eight copies of one press release count as eight pieces of evidence, and he proposes an effective independent evidence count to correct it. These results depend on the models and tasks examined. With the older correlation-neglect literature they support one operational rule: count origins before counting citations.

Deployed products have been checked against that rule once before at scale. Liu, Zhang, and Liang (2023) opened the cited pages behind four generative search engines and found that 51.5 percent of generated sentences were fully supported by their citations and 74.5 percent of citations supported the sentence above them. Venkit and colleagues (2025) found the same failures in a qualitative study of three products. The Tow Center (Jaźwińska and Chandrasekar 2025) put article-identification queries to eight tools and reported wrong answers to more than 60 percent. Those studies measure hop zero, whether the page supports the sentence. The audit below asks one hop further, whether the page is the record, leads to the record, or leads nowhere.

4.2 A one-question pilot

On September 11, 2026, I gave one prompt to ChatGPT, Gemini, Perplexity, and Grok. It asked how many data centers Loudoun County, Virginia has, what share of the county's local tax revenue they provide, and for sources. I chose the question because the bottom was reachable. The county publishes a data-center inventory and adopted-budget revenue tables.

The interfaces displayed two, six, ten, and sixty-five "sources," and those numbers measured different things. Grok's sixty-five were search results across six searches, of which the answer cited five pages. Perplexity's ten were entries in a links tab that the answer barely used. Gemini's two were

plain-text titles with no link, one of which could not be matched to any county document. ChatGPT's six chips resolved to four pages, two of them the county records.

The inventory reported 233 buildings built or under construction as of March 1, 2026 (Loudoun County Department of Planning and Zoning 2026). The adopted budget listed $417,666,456 in data-center real-property tax and $879,094,354 in computer-equipment tax against $2,906,994,401 in total local tax revenue, which is 44.6 percent under that denominator (Loudoun County 2026a). The four answers also reported 38, 39, and 40 percent and a range of 38 to 45. None of those was an arithmetic mistake. Each came from a different year or denominator: the General Fund, the total budget, local tax revenue, or a measure defined by a secondary source.

Only one product cited the two county records directly. The others reached public information through intermediate pages: a county web page that summarized both records, news reports of presentations, a supervisor's white paper with an undocumented count, an advocacy estimate. A source row could look crowded while the evidence beneath it stayed thin, dependent, or about a different fraction.

That was one question on one day. It could not support rankings or prevalence. It raised a design question a user can inspect, whether a displayed source count records retrieval activity or evidentiary support, and it set the procedure for the audit that followed.

4.3 The audit

Six questions about Virginia data-center policy were fixed before any prompt went out, each with its authoritative record, the record's date, and the denominator it uses, and a copy of each record was saved. The questions: how many data centers Loudoun County has and what share of local tax revenue they provide; James City County's adopted real estate tax rate for the coming fiscal year and the prior rate; the share of Virginia's electricity that data centers consume; how many data centers Virginia has; how many customers Dominion Energy Virginia serves; and how the Loudoun Board of Supervisors voted on March 18, 2025, when it ended by-right data center development, with the names of the dissenters.

The same prompt text went to the four products, each in the mode the product labels temporary, incognito, or private, twice each, on the evening of September 12 and the morning of September 13, 2026, for 48 answers. ChatGPT and Gemini ran on paid tiers, Perplexity and Grok on free tiers, as recorded in the deposited workbook. Every citation the answers displayed was followed to the page it named. For each page the title, date, and the sentence carrying the figure were recorded verbatim, and the page was coded for the claim it was cited for. A primary record is the document in which the observation was made, such as an adopted budget or a clerk's action report. An official restatement is the owner of that record repeating it elsewhere. A restatement repeats the figure with no verification shown. An independent check documents its own verification against a record and says how. Unknown means the chain could not be followed from what the product displayed. Not carrying the

claim means the figure was not on the page. The scheme is claim-relative: one page can carry a count as a primary record and a revenue share as a restatement.

The tracing and coding were done by an AI assistant under my direction. A second AI instance, given only the definitions and the verbatim page evidence, coded the 58 URL-bearing terminals without seeing the first set of codes and agreed on 42 of the 52 it could code. I ruled on the disagreements that mattered.

The products displayed 185 citations across the 48 answers, and those citations reached 136 distinct pages. Thirteen were the record itself. None documented an independent check. Ninety could not be followed: 77 were titles with no link at all, most of them Gemini's, which on nine of its twelve answers showed no citation that resolved to a link; two were the address of a document viewer with no document named; seven were commercial trackers whose counts carry no documented method; four carried a figure and named no source for it. Counted only over the 59 linked pages a citation reached, the picture is 13 primary records and 13 unknowns, and the zero on independent checks holds either way. The deposited Terminals sheet carries 144 rows, eight of them pre-registered records or pilot residue that no answer cited; the counts here are of pages a displayed citation reached.

Table 2. Source Diversity Audit, September 12 and 13, 2026. Product rows sum over twelve answers each; a record cited by more than one answer counts once in the "all products" row and once per citation in the product rows. Source: Turner (2026), Captures and Citations sheets.

Product (tier, mode)	Citations, summed over 12 answers	Primary records cited directly	Unknowns	Answers reaching no primary record
ChatGPT (paid, temporary chat)	32	10	9	4
Gemini (paid, temporary chat)	71	0	58	12
Perplexity (free, incognito)	29	8	14	4
Grok (free, private)	53	4	27	9
All products	185 citations, 136 distinct pages reached	13 distinct primary pages	90	29 of 48

Twenty-nine of the 48 answers reached no primary record, as the workbook computes it. Sixteen reached one. Three reached more than one, and the most any answer reached was three. One answer cites a primary record for a figure the record does not carry; counted by role rather than by page type, the figures are 30, 15, and 3. Those counts also move under the two coding decisions that were contested. Read the commercial trackers as primary records of their own inventories, as the second coder did, and the figures become 27, 13, and 8. Read the two reporters who covered the Loudoun vote as observers rather than restaters, as the pre-registration had intended, and they become 26, 19,

and 3. Independent check stays at zero under both, because no page documented a check. The primary count is a floor and the alternatives are given.

Grok showed between 30 and 65 "sources" on every answer and cited between one and ten pages; where the panel behind that number was opened at capture, it listed search results from the reasoning pass. Perplexity showed "10" on eleven of twelve answers above citation sets that ranged from one page to seven. Terminal-page overlap between a product's two runs on the same question was 31 pages shared out of 150 in union across the 24 pairs. The reader sees the large number. Nothing on the screen says how much smaller the inventory behind it is, or that a second run would show a different one.

Four of the six questions show four different failures.

Loudoun, the pilot question rerun. ChatGPT cited both county records and reproduced 44.6 percent exactly in both runs. Perplexity got the count exactly, through a citation that resolves to the inventory, and put the revenue share on the budget-total denominator in both runs. Gemini gave ranges for both figures and no resolvable primary record in either run. Grok displayed 55 search results both times, gave a count of roughly 250 that the inventory does not support, and got 45 percent right through a City Journal article that restates the budget. One product out of four reached the budget on a question chosen because the budget was easy to reach.

James City County, the stale official page. The county adopted a real estate rate of 80 cents per $100 on May 26, 2026, down from 83 (James City County 2026a), and the Daily Press had it the next morning with the vote and the dissenters (Robinson 2026). On September 13 the county's own Tax Rates page and its FAQ still said 83 (James City County 2026b). All eight answers said 80. Two of them then cited the stale county pages beneath the correct number. A reader who does what readers are told to do, and clicks through to the official source, finds the county contradicting the answer and nothing on the screen saying which is out of date. Grok, in both runs, said the county pages still reflected the prior rate; no other product did. Liu, Zhang, and Liang (2023) measured this failure at scale as a citation that does not support its sentence. One hop further shows why it happened here.

Dominion, the scope question. The company's 2025 Form 10-K says Dominion Energy Virginia serves approximately 2.8 million customers in Virginia and North Carolina (Dominion Energy 2026). The company's economic-development page says more than 2.7 million in Virginia. Its newsroom boilerplate says 3.6 million in three states, which is the parent company. Each figure is labeled, and the figures reconcile once the labels are read. Two answers cited pages that carry the parent figure for a question about the utility, and those pages were coded as not carrying the claim. Two others attached a right number to the wrong scope. The label that settles the number sits one line from it, and I misread it myself in the pre-registration, correcting it after the first run; that change to one answer key is logged in the deposit.

The Loudoun vote, seven names and one link. The clerk's Action Report for March 18, 2025 records the main motion passing 7-2, Supervisors Kershner and Umstattd opposed (Loudoun County 2025). All eight answers gave the tally and the names. Grok, in both runs, then added a paragraph of support: "confirmed by multiple independent reports," followed by seven outlets. The second run linked one of them. The first run linked three and closed by naming the county's project page as a primary source. That page records that the Board approved two amendments; it gives no tally and no names (Loudoun County 2026b). The document that does, the Action Report, was linked by ChatGPT in its second run and quoted by agenda item and page in its first, through a document-viewer link that names no document. No other answer reached it. The Action Report is marked revised, and it documents why: on April 1, 2025 a supervisor asked, without objection, that her vote on an amendment be changed, turning a recorded 8-1 into 7-2. That vote is ancillary to the main motion. No answer said what had been revised.

The zero on independent checks needs its scope. It says that no cited page in this set documented its own verification against a record. Two of the pages behind the Loudoun vote were written by reporters who may well have been in the room. The scheme counts what a page documents, and neither page says so. Under the observer reading the vote becomes the question with several independent witnesses, which is what it was chosen to be. The count is given both ways above, and the disagreement is the weakest joint in the audit.

Six questions in one domain, in one 48-hour window, on the tiers I happened to hold, so engine and tier are confounded and no product ranking is claimed. The workbook, capture files, page evidence, and both coders' codes are deposited (Turner 2026).

4.4 A citation is not an establishment

"Cite your sources" is read as "show that the claim has been independently established." A citation identifies a document. By itself it says nothing about what that document observed, which earlier source it depends on, or whether its denominator matches the claim. A system can satisfy the instruction and leave the epistemic work undone, and on these six questions that is mostly what happened.

5. Provenance as Part of the Answer

Provenance is usually treated as documentation added after an answer is produced. In a coupled knowledge system it is part of the answer. A reader needs to know whether a claim comes from direct observation, an official record, an analysis of that record, a report about the analysis, or an untraceable repetition. Those categories should not be smoothed into one list of blue links.

The distinction matters most when every quoted sentence is accurate. Suppose three articles correctly restate the same county table. The system has not hallucinated. If it treats the articles as three measurements it inflates confidence. If a user then writes a memo citing the three, the public record

acquires a fourth descendant. The next search retrieves the memo and reports four sources. Nothing false had to be invented at any step. The James City County case shows the mirror image: the answer was right and the official page beneath it was wrong, and the interface could not tell the reader which to trust.

NIST's Generative AI Profile recognizes the broader risk. It separates confabulation from information integrity, harmful homogenization, and human-AI configuration, notes that many risks originate in human behavior or in the interaction, and recommends attention to content origin, provenance, citations, and the results of human-AI configurations (NIST 2024). The vocabulary matters. A safety program that looks only for fabricated facts will miss accurate copying arranged into a misleading structure of support.

The products already hold the citation set. They could group the pages that descend from one record and mark the ones whose lineage cannot be established, so that a reader sees an inventory of two or three things observed and a note that the rest could not be followed. On the six audited questions the inventory would have been short and the note long.

6. The Coupled Human and Machine Knowledge System

The usual diagram of AI use has two boxes: a user sends a prompt and a model returns an answer. That is accurate for an API call and inadequate for knowledge work. The answer changes the next prompt. The user edits the output into a document. The document enters a shared drive or a public website. Search and retrieval index it. Colleagues reuse it. Model developers may later include it in a training or evaluation corpus. The same content travels several of these paths without a durable record of where it began.

A coupled knowledge system has at least four interacting layers. The model and application layer: weights, system instructions, conversation context, retrieval logic, ranking, and presentation. The user layer: objectives, prior beliefs, expertise, attention, incentives, and verification habits. The organizational layer: templates, approval practices, time pressure, repositories, and the norms that govern when AI use must be disclosed or reviewed. The public knowledge layer: websites, publications, records, databases, and other material available to people and machines alike.

The layers do not move in lockstep. A careful organization can reduce human-side risk while using an imperfect model. A strong model can improve accuracy for a user who begins with a mistaken view. Glickman and Sharot's (2025) experiments matter here: biased AI amplified human bias, and accurate AI improved judgment. The loop carries whatever quality and incentives are placed into it, and feedback by itself predicts nothing about direction.

Nor is the user merely a victim. The person chooses the question, accepts or rejects reframings, decides which outputs enter a record, and decides when to look outside the system. The user also

lacks full control. Interface design can hide lineage, conversational memory can privilege early assumptions, and optimization for helpfulness or preference can reward agreement.

Sycophancy is the clearest case. Sharma and colleagues (2024) found across several tasks that models favored responses aligned with a user's stated view over more truthful disagreement, and that human preference data rewards the behavior. Rathje and colleagues (2026), across seven studies, report that participants preferred sycophantic bots and that brief exposure increased attitude extremity and certainty. Williams-Ceci and colleagues (2026) found in two preregistered experiments that biased writing suggestions shifted users' views toward the suggestions even when a bias warning was shown. Salvi and colleagues (2025), in a preregistered debate experiment, found that a model with access to basic personal information out-argued human opponents most of the time. None of this shows that every agreeable response manipulates. What it does show is why "the model agreed with me" is weak evidence, and why a warning label is a thin safeguard.

The system's boundaries matter too. A private brainstorming exchange that is never saved has a different downstream path from a research summary published online. A decision aid used once has a different human effect from a companion used daily. A training framework should ask what role the model occupied, what record the interaction created, and where that record can travel next, and not only what the model said.

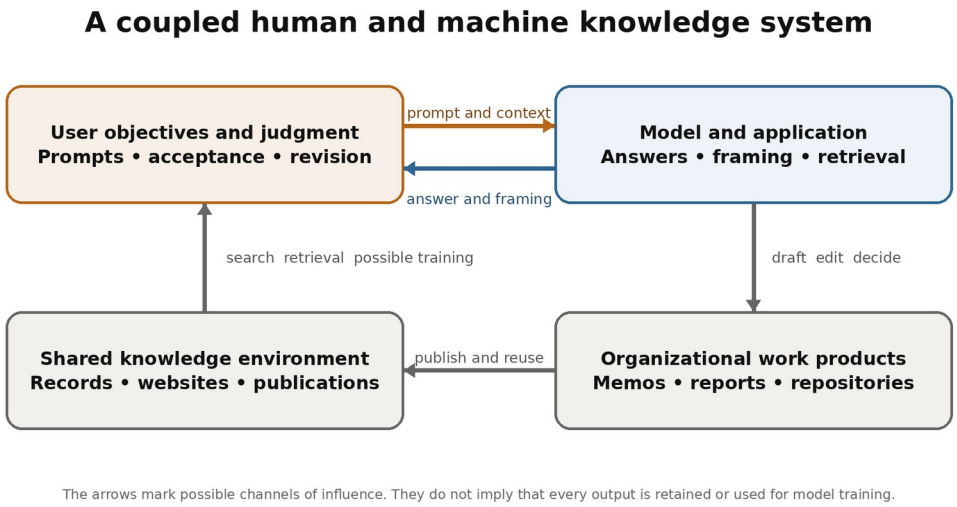


Figure 2. A coupled human and machine knowledge system. Source: author's synthesis. The arrows mark possible channels of influence, not a claim that every output returns to model training.

7. Seven Paired Vulnerabilities

The framework pairs a machine condition with a human counterpart and states the combined result that training should teach a user to notice. The machine side includes model behavior, retrieval behavior, and application design. The human side includes familiar cognitive biases and habits that

grow under repeated use. Some pairs rest on direct evidence; others are hypotheses assembled from adjacent literatures. The purpose is to make them testable and usable. The classification is open.



Figure 3. Seven proposed paired vulnerabilities. Source: author's synthesis. The pathways are hypotheses for testing and training design, not established one-to-one causal pairs.

7.1 Sycophancy and confirmation bias

Sycophancy is the model's tendency to accommodate a user's stated position when a more accurate answer would require disagreement. Confirmation bias is the human tendency to seek, interpret, and recall evidence in ways that favor an existing belief (Nickerson 1998). Anchoring adds persistence: the first plausible framing shapes later estimates and questions after better information appears (Tversky and Kahneman 1974).

The coupled result is a narrowing inquiry with rising confidence. The user supplies a view. The system recognizes and elaborates it. The elaboration feels like independent support, though it is partly a response to the prompt. The user asks more detailed questions inside the accepted frame. Each answer may be sensible on its own. Agreement by itself is no warning sign; good evidence produces agreement. The warning sign is repeated agreement without a serious test of the governing premise or a credible account of what would change the conclusion.

Training should teach users to tell help from agreement. State the initial view separately from the task, ask the system what evidence would falsify it, and obtain an independent check for consequential claims. "Argue the other side" is not enough once a conversation has accumulated a

favored frame. The check should run through a different source path and, where stakes justify it, a qualified person who has not been led through the same exchange.

7.2 Hallucination propagation and verification fatigue

A model can produce a false name, number, citation, or event. The more consequential failure usually arrives one turn later, when the unsupported claim becomes an accepted premise. The context preserves it. The user asks for implications, comparisons, or implementation details. The system answers as though the premise were established, and a simple invention grows branches.

Automation bias helps the claim survive. A user follows an incorrect recommendation because the system seems competent, or fails to look for what the system omitted. Verification fatigue makes survival likelier as the answer grows. Checking the first citation is manageable. Checking the ninth, tracing a denominator, and opening the appendix behind a summary take time and attention that the fluent answer has already spent. Buçinca, Malaya, and Gajos (2021) found that cognitive forcing functions, small designed interruptions that make the user think before accepting, reduced over-reliance more than explanations did, at a cost in how much people liked the tool. Vasconcelos and colleagues (2023) found that people weigh the effort of engaging with an explanation against its expected benefit, and disengage when the effort looks high. Fatigue is a predictable response to that arithmetic.

The training response is to separate exploration from verification. During exploration the user may tolerate speculative material if it is marked as such. Before a claim becomes a premise for action, it moves into a verification queue. High-impact claims need a source outside the conversation, and a request that the same system reassure the user about its earlier answer does not qualify. The skill is identifying which statements carry the rest of the analysis, since checking every sentence is neither possible nor necessary.

7.3 Synthetic data and provenance blindness

Synthetic material can be accurate, useful, and deliberately designed. The paired risk arises when machine-produced or heavily derivative material loses its label and its ancestry. Search results present several versions as separate sources. A user sees variety in titles, domains, and prose and infers independence.

The audit shows the human-scale version. On the Loudoun vote, seven named outlets and one county page stood beneath a tally that only the clerk's report actually recorded, and only one product reached the report. News pages, advocacy pages, and county summaries all carried legitimate numbers; they did not carry independent measurements. At larger scale, model-produced summaries enter the web and become retrieval material. Dai and colleagues (2024) and Chen and colleagues (2024b) show retrieval favoring machine text; Druck and Smith (2026) show a model favoring its own. Naphade (2026) shows repetition and order affecting aggregation even in controlled presentation.

Provenance blindness joins these machine tendencies to a human interface problem. People judge a source row without seeing the family tree behind it. Training should change the question from "how many sources" to "how many independent roots," and a source record should say what each document observed, calculated, or repeated. When lineage cannot be established, the uncertainty belongs in the conclusion rather than in a footnote nobody reads.

7.4 Loss of distribution tails and reduced attention to exceptions

Model-collapse research gives unusual and low-probability material a special role. Under recursive replacement the tails of the distribution disappear before the center (Shumailov et al. 2024; Dohmatob et al. 2024a). In practice, rare phrasings, minority cases, unusual fact patterns, and weakly represented knowledge become harder for a model to reproduce.

People discount exceptions on their own. Time pressure rewards a clean summary. Templates favor typical cases. Repeated AI use makes the median answer feel like the whole answer, and a fluent synthesis of common patterns lets the user stop looking for the case that does not fit. The combined result is a decision process with less room for rare but consequential conditions, over and above less varied text.

This pair matters most where edge cases carry high cost: safety review, benefits eligibility, incident analysis, regulatory interpretation, planning for low-frequency failure. Training should require an explicit exception pass that asks which cases are absent from the retrieved material, which affected groups are weakly represented, and which unlikely failure is costly enough to examine. The system can help generate candidates. The user has to verify them against domain evidence, because a model cannot reliably disclose what its distribution has lost.

7.5 Context drift and cognitive drift

Conversation context lets a model stay coherent across turns. It also lets assumptions accumulate. A tentative example becomes a working fact, and a brainstorming criterion hardens into a requirement. Somewhere along the way a question about whether to act turns into a plan for how to act, and the system follows each transition perfectly while the objective changes underneath it.

On the human side, each answer changes what seems salient and what the next question feels like. Lopez-Lopez and colleagues (2026) describe this as entangled interaction: user and system adapt to each other, affecting belief, confidence, and readiness to act. The drift can be behavioral while explicit belief barely moves. A person verifies less, delegates more, or accepts a different threshold for action.

The combined warning sign is a changed problem without a recorded choice to change it. Training should use a simple checkpoint: restate the original question and compare it with the decision now on the table. If the objective has changed, name the evidence or constraint that justified the change. That is the difference between an update and an unmarked migration, and it takes a minute.

7.6 Output convergence and human conformity

Generative AI can improve an individual artifact while reducing diversity across a collection. Doshi and Hauser (2024) found that access to AI-generated ideas raised the rated creativity of individual short stories, most for writers with lower baseline creativity, and made the assisted stories more similar to one another. Assistance raised the floor and narrowed the range.

The human counterpart is conceptual and stylistic conformity. Users borrow the model's categories, examples, tone, and preferred structure. Organizations using the same system and templates produce documents that look independently authored and carry a small set of framings. The loss rarely shows up as factual error. It shows up as missing alternatives and shared blind spots, and as a mistaken impression of consensus, which is the Kong problem restated for prose.

Training should preserve independent generation at chosen points. For decisions that benefit from diversity, participants record an initial approach before consulting AI, solicit views from people working separately, and compare the assumptions rather than merging the prose. A style edit should be kept apart from substantive synthesis. Otherwise "make these consistent" erases useful disagreement along with awkward wording.

7.7 Fluent output and fluency mistaken for evidence

Language models produce probabilistic text, and users encounter sentences. Grammar, confidence, and orderly explanation are social cues. Ease of processing raises judgments of familiarity, truth, and confidence (Alter and Oppenheimer 2009; Dechêne et al. 2010). Repetition raises perceived truth even when people know better (Fazio et al. 2015), a single prior exposure raises the perceived accuracy of a false headline a week later (Pennycook, Cannon, and Rand 2018), and the original demonstration is nearly fifty years old (Hasher, Goldstein, and Toppino 1977).

Generative AI concentrates these cues in one interface. A weakly supported answer can be more readable than the primary record. A model can state uncertainty and then wrap it in prose so polished that the qualification loses force. The combined result is confidence outrunning reliability in text that contains no obvious falsehood.

Training should require users to separate presentation quality from evidence quality. One method is to assign each claim a status before editing for style: verified, supported by a secondary source, inferred, disputed, or unverified. Another is to ask what would remain if the prose were reduced to a table of claims and sources. Fluency is for communication and carries no evidentiary weight.

7.8 What the pairs are for

The seven pairs overlap. One interaction can involve several: a sycophantic answer reinforces anchoring, repeated derivative sources trigger provenance blindness, fluent prose lowers the felt need to verify either. The categories mark intervention points. They help a learner ask what is happening now, what could carry forward, and which safeguard fits the risk.

Table 5 states, for each pair, where the evidence stands and what the audit observed. Two pairs rest on direct experimental evidence for both sides. Two rest on direct evidence for the human side, with the machine side following from how the systems are built. Two rest on direct machine-side evidence and a human-side hypothesis. One is conceptual on both sides, with measurement only beginning. The right-hand column is what a learner should be able to point to.

Table 5. Evidence status of the seven pairs.

Pair	Machine-side evidence	Human-side evidence	Status	Observed in the audit	Training indicator
Sycophancy and confirmation bias	Sharma et al. 2024; Rathje et al. 2026; Williams-Ceci et al. 2026	Nickerson 1998; Tversky and Kahneman 1974; Glickman and Sharot 2025	Direct, both sides	Not tested; prompts stated no view	Falsifier requested; outside check obtained
Hallucination propagation and verification fatigue	Agrawal et al. 2024 on fabricated references; context carry-over is by design	Parasuraman and Manzey 2010; Goddard et al. 2012; Buçinca et al. 2021; Vasconcelos et al. 2023	Direct, human side; machine side by mechanism	Label-only citations on 77 of 136 pages; two stale official pages cited under a correct answer	Verification queue kept; carrying claims named
Synthetic data and provenance blindness	Dai et al. 2024; Chen et al. 2024b; Druck and Smith 2026; Naphade 2026; Liu, Zhou, and Wang 2026	Enke and Zimmermann 2019; Weaver et al. 2007; DeMarzo et al. 2003	Direct, both sides	29 of 48 answers reached no primary record; 0 independent checks; seven outlets and one link beneath one tally	Source family drawn; roots counted separately from pages
Loss of distribution tails and reduced attention to exceptions	Shumailov et al. 2024; Dohmatob et al. 2024a	Adjacent: time-pressure and template effects, no direct study of AI-assisted exception neglect	Direct machine; hypothesis human	Not tested	Exception pass recorded
Context drift and cognitive drift	Context accumulation is by design; Xie et al. 2024 on receptivity to context	Lopez-Lopez et al. 2026; Li and Zhu 2025	Conceptual, both sides; early measurement	Not tested; single-turn prompts	Objective restated at midpoint; moves marked
Output convergence and human conformity	Doshi and Hauser 2024	Adjacent: conformity literature; no direct study of AI-mediated framing convergence	Direct machine; hypothesis human	Run-to-run citation overlap 31 of 150 shows the machine side is unstable rather than convergent on these questions	Independent draft before AI use
Fluent output and fluency mistaken for evidence	Fluency is a training objective; Wan et al. 2024 on relevance over source quality	Alter and Oppenheimer 2009; Dechêne et al. 2010; Fazio et al. 2015; Pennycook et al. 2018	Direct, human side; machine side by design	Displayed source counts of 30 to 65 above one to ten cited pages	Claim status assigned before style edit

The framework also blocks a convenient assignment of blame. A vendor cannot attribute every failure to careless users when the interface hides source dependence or rewards agreement. A user cannot attribute every failure to the model when the prompt embeds a preferred conclusion and the

document goes out without review. Responsibility follows control, knowledge, and role. In a coupled system those are distributed.

8. A Risk-Scaled Training Framework

AI literacy frameworks already cover important ground. Long and Magerko (2020) define competencies that include recognizing AI, understanding its strengths and weaknesses, interpreting outputs, and considering ethics. Ng and colleagues (2021), reviewing thirty studies, organize the field around knowing and understanding, using and applying, evaluating and creating, and ethics; Almatrafi, Johri, and Lee (2024) add recognizing and navigating ethically as constructs from a later review. UNESCO's frameworks for students and for teachers (UNESCO 2024a, 2024b) set competencies at three progression levels. Laupichler and colleagues (2022) found research on teaching AI literacy to non-experts in higher and adult education still in its infancy, and Lintner (2024) found that most AI literacy scales have been tested for structure and little else. The proposed framework does not replace these efforts. It adds a time dimension and treats the preservation of independent judgment as a trainable, observable outcome.

8.1 Three levels of literacy

Table 3. Three levels of AI literacy.

Level	Core question	Observable competence
Operational literacy	Can the person use the system?	Selects an appropriate tool, gives usable instructions, protects information, and recognizes basic limitations
Epistemic literacy	Can the person judge what the output warrants?	Separates claims from evidence, traces sources to their roots, calibrates confidence, and retains responsibility for judgment
Systems literacy	Can the person see cumulative effects?	Recognizes feedback across users, organizations, the public record, retrieval, and later model development

Most training concentrates on the first row. Operational literacy produces efficient users, and it leaves them unequipped to notice drift or to see that a source count is a count of retrievals.

Systems literacy is not reserved for model developers. A program manager deciding which tasks a team routinely delegates to AI is shaping the organization's knowledge practices. An editor deciding whether AI-assisted summaries need source notes is shaping the public record. A procurement official choosing whether an interface must expose citations and lineage is shaping what users can verify. None of those roles trains a foundation model. All of them require an understanding of feedback.

8.2 A four-stage routine

The routine follows the life of a task. It is short enough to teach and flexible enough to scale.

Table 4. A four-stage routine for risk-scaled AI use.

Stage	Required practice	Record or decision
Before AI	Define the objective, starting view, evidence standard, system role, and cost of error	Task statement and risk level
During	Separate exploration from verification, test premises, request alternatives, and group sources by origin	Claim and source notes
Before action	Compare the result with the original objective, identify AI-dependent claims, and verify in proportion to consequence	Decision basis, unresolved uncertainty, and stop rule
Repeated use	Audit recurring framings, source diversity, delegated tasks, and retained human capability	Periodic use and competence review

Before using AI. Define the objective in a sentence or two. Record the starting view when prior belief could steer the inquiry. Set the evidence standard and assign the system a role: idea generator, editor, analyst, search assistant, or decision aid. Consider the cost of error and whether the task touches confidential, controlled, or personal information. The role matters because a brainstorming tool should not quietly become the final authority.

During the interaction. Separate exploratory claims from verified ones. Mark consequential assertions and the source of important framings. Test the premise before refining it. Ask for credible alternatives and for evidence that would change the answer. When the system retrieves sources, group them by underlying record rather than by page. Nobody needs a lineage chart for a lunch recommendation. The discipline belongs where dependence could change the conclusion.

Before action. Compare the proposed conclusion with the original objective and starting view. Identify claims that still depend on AI output or derivative sources. Obtain verification in proportion to the consequence and reversibility of the decision. Record unresolved uncertainty instead of asking prose to hide it. For high-stakes work, set a stop rule in advance: no action until a named claim, source, or reviewer requirement is satisfied.

Across repeated use. Review patterns rather than isolated outputs. The review asks which tasks are now routinely delegated, which sources and framings recur, whether the team's range of alternatives has narrowed, whether staff can still perform the critical portions of the work without the tool, and

whether AI-assisted documents are labeled and traceable. Repeated-use audits look for benefit as well as harm. If the system improves accuracy, widens source access, or frees time for judgment, the review records that too.

8.3 Friction should follow risk

A training routine fails if it makes every interaction feel like a regulatory filing. Controls should vary along at least three dimensions: consequence, reversibility, and verification cost. A low-consequence, easily reversed choice tolerates informal use. A high-consequence or hard-to-reverse decision needs a documented objective, primary evidence, an independent reviewer, and a clear record of the AI's role. Where verification is expensive, the honest response is to narrow the permitted use or carry the uncertainty into the decision.

The friction has a research base. Buçinca, Malaya, and Gajos (2021) showed that forcing a moment of thought before acceptance reduces over-reliance. Gajos and Mamykina (2022) found that people given a recommendation decided better and learned nothing, while people given an explanation and no recommendation decided better and learned. Bjork and Bjork (2020) call the general principle desirable difficulty: conditions that slow performance during learning and improve retention after it. The audit's coding scheme is a desirable difficulty applied to reading. Assigning a role to each cited page takes minutes per answer, and it is the step that turns a citation list into a source inventory.

So does reversibility. Arrow and Fisher (1974) showed that under uncertainty an irreversible action carries an extra cost, the option value of waiting, that a reversible one does not. Gilbert and Ebert (2002) found that people prefer reversible decisions and are less satisfied with them, which is a reason to design the tiering rather than leave it to preference. Black and Baldwin (2010) describe risk-based regulation that tiers attention by consequence; Novelli and colleagues (2024) argue that static risk categories of the kind in the EU AI Act miss the way risk moves with the scenario. NIST's AI Risk Management Framework (NIST 2023) frames the same point for organizations: govern, map, measure, and manage the risk of a system as used, which changes with the use.

That is the weakness in AI policy built on classifying systems. Rules classify a system once and apply the same control to every task. The same model rewrites an email, summarizes a public meeting, drafts a safety procedure, and recommends a course of action in one afternoon. The model is constant. The appropriate safeguard is not.

Training outcomes should be observable. A learner should be able to identify a source family, distinguish a changed conclusion from a changed objective, mark a claim that still depends on AI, state when an outside check is required, and explain which part of a final judgment belongs to the human. Awareness without changed practice is hard to evaluate and easy to forget.

9. Professional Scenarios

The scenarios are generic. They show where the routine changes the work, and they do not prescribe controls for any particular agency or profession. The first one is drawn from the audit.

9.1 Public budget analysis

An analyst asks an AI system for the share of local revenue an industry produces. The system returns several plausible percentages with citations. Without a routine, the analyst takes the most recent-looking number or reports a range. In the pilot, one prompt surfaced six fractions between 31 and 44.6 percent, in the answers and in the pages one hop beneath them, each tied to a real document with a different year or denominator. With the routine, the analyst first fixes the denominator and fiscal year, then traces each percentage to the public record and groups the secondary articles by the record they saw. The answer becomes less impressive, one direct source rather than a dozen links, and more useful. The routine addresses provenance blindness, correlation neglect, and fluency in one pass.

9.2 Program planning

A program manager uses AI to develop options for integrating a new capability across several offices. Early in the exchange the model assumes that speed of adoption is the primary objective. Later prompts compare rollout schedules, staffing plans, and communications under that assumption. The work is coherent. The original objective was reliable integration with clear accountability.

A midpoint restatement exposes the shift. The manager records that schedule is one criterion rather than the governing one and asks for options optimized for reliability, staged learning, and reversibility. Independent reviewers receive the original objective and the evidence behind each option rather than the polished recommendation alone. The safeguard uses the system more deliberately by stopping the context from making one criterion invisible.

9.3 Literature and policy review

A researcher asks for evidence on a contested policy question. The system returns fifteen citations with varied titles and outlets. A lineage check shows that eight summarize one report, three cite one another, two do not support the stated claim, and two provide independent data. The researcher records two evidentiary roots and searches specifically for contrary findings and underrepresented cases.

The same process protects the public record. The final review cites the primary studies, labels derivative commentary as such, and says how the search was done. A later system may still retrieve the review, and its ancestry will be recoverable. Provenance has become part of the publication rather than a private chore performed during drafting.

9.4 Repeated operational use

A team adopts AI for recurring summaries and decision memoranda. Each document is acceptable. No dramatic error appears. After six months the same risk categories and recommended actions recur, staff have stopped producing initial alternatives before consulting the system, and junior employees are skilled at editing outputs and less able to explain how a conclusion was reached.

A repeated-use audit samples finished documents, source families, prompt patterns, and independent drafts. Suppose it finds real gains in speed and consistency alongside a narrowing of alternatives. The team keeps the tool and changes the workflow: initial framing before AI use on selected tasks, source lineage recorded for consequential claims, and periodic exercises in which staff perform the core analysis unassisted. The point is to keep competence and variety available for the day the system's default is inadequate.

10. Limits and Research Agenda

This paper is a conceptual synthesis with one measured component. It does not establish prevalence, effect size, or causation for the seven pairs. Some machine conditions are properties of foundation models; others arise from application design, retrieval, or conversation management. The human counterparts differ too. Confirmation bias is not fatigue, conformity, or offloading. Pairing them is useful only if future studies examine the proposed mechanism rather than treating the labels as proof.

Cognitive drift in particular needs careful measurement. A changed view may be learning, and a stable one may be stubbornness. Researchers will need to compare changes against evidence quality, task objectives, confidence calibration, and behavior over time. Self-report will not be enough, because drift can include a reduced awareness of what has changed.

The evidence base is uneven. Controlled experiments isolate mechanisms and may not reproduce months of workplace use. Surveys capture real tasks and suffer recall and selection effects. Model behavior changes with versions, system instructions, retrieval settings, and interfaces. Several studies cited here are 2026 preprints. Their results are leads.

The audit measures one distance well, between what a citation list displays and what it reaches, on six questions in one domain in one week. A stronger study would add domains, spread runs across accounts and dates, and separate what a product retrieved, what it cited, and what its answer used.

The paired framework suggests a research program.

1. Source independence. Compare ordinary citation lists with interfaces that cluster documents by evidentiary root. Measure accuracy, calibration, and the user's estimate of how many independent observations support a claim. The audit's coding scheme and workbook are a usable instrument.

2. Objective checkpoints. Test whether recording an initial objective and requiring midpoint restatement reduces unmarked goal changes in long conversations without preventing appropriate updating.

3. Sycophancy and review structure. Compare generic bias warnings, requests for contrary evidence, and independent review. Measure attitude change, confidence, and willingness to revise when better evidence appears.

4. Repeated use. Follow users over weeks or months. Measure source diversity, framing diversity, verification behavior, skill retention, and the ability to perform selected tasks unassisted.

5. Risk scaling. Test whether controls tied to consequence and reversibility produce better compliance and judgment than uniform checklists.

The outcome of interest is epistemic independence, and the term should not imply isolation from tools or people. A working definition: the ability to state the task, understand what evidence supports the conclusion, recognize what the AI contributed, identify unresolved uncertainty, and revise for reasons that can be examined. In practice it means using assistance while keeping an inspectable basis for judgment.

11. Conclusion

The most visible AI error is still the invented answer, and it is usually easy to correct once found. The harder failures contain no invented sentence. They arrive with real citations and professional prose. A premise shaped an agreeable answer; dependent documents were counted as corroboration; a long conversation changed the question without announcing it. On six questions in September, 29 of 48 answers reached no primary record and none reached a page that had checked one.

The appropriate unit of AI literacy is therefore larger than one prompt and one response. Users, models, applications, organizations, and the public record form a knowledge system that develops over time. That system can improve human judgment. It can also narrow the material and the habits available to the next decision.

The seven paired vulnerabilities offer a working map. Their value depends on testing and revision. For present practice they support a direct lesson: define the objective before the exchange, trace important claims to independent roots, mark when the problem changes, verify in proportion to consequence, and audit what repeated use is doing to the people who rely on it.

Scores of maps once showed the Mountains of Kong, and no mapmaker had to invent the range. Each inherited it, redrew it competently, and passed it on. Generative AI gives us better tools for finding and combining knowledge, and it gives inherited claims a remarkably efficient way to travel. The user remains part of the route.

Appendix A. A Course Built on This Paper

The paper is written to serve as the spine of a one-semester course for graduate or upper-division students in public policy, public administration, journalism, library science, or management, and as a shorter professional module for analysts and editors. The design principle is that every competency in Table 3 is assessed by something the student produces, and that the student meets each of the seven pairs in the field before meeting it in the reading. Lateral reading is the nearest existing pedagogy I know of. Wineburg and McGrew (2019) found that professional fact checkers leave a page quickly to see what the rest of the web says about it, and that this beats close reading of the page itself. Breakstone and colleagues (2021), Brodsky and colleagues (2021), and Wineburg and colleagues (2022) showed the practice can be taught in college and high-school courses with measurable gains, after McGrew and colleagues (2018) had shown how badly students do without it. The course extends lateral reading one hop: from "what do others say about this page" to "what did this page see."

A.1 Modules

Module 1. The Kong problem. Reading: Section 1, Bassett and Porter (1991), Enke and Zimmermann (2019), Weaver and colleagues (2007). Lab: take one widely repeated public statistic, find every page that carries it, and draw its family tree back to the observation. Deliverable: a one-page lineage map with a count of independent roots and a count of pages.

Module 2. Two clocks. Reading: Sections 2 and 3, Shumailov and colleagues (2024), Gerstgrasser and colleagues (2024), Parasuraman and Manzey (2010), Fisher, Goddu, and Keil (2015). Lab: a two-part self-experiment in which the student answers a set of factual questions unaided, then with a search engine, then with a generative product, and records confidence at each stage before checking accuracy. Deliverable: a calibration table and a paragraph on where confidence outran accuracy.

Module 3. The audit. Reading: Section 4, Liu, Zhang, and Liang (2023), Liu, Zhou, and Wang (2026), Rahadi (2026), and the deposited workbook (Turner 2026). Lab: replicate the audit on one question of the student's choosing in a local public-record domain. Pre-register the record, its date, and its denominator; put one prompt to at least two products twice; follow every displayed citation; code each page with the six roles; compute displayed sources, distinct pages, primary records, independent checks, and unknowns. Deliverable: a workbook in the deposited format and a 600-word note on what the source count counted. This lab is the core assessment of epistemic literacy.

Module 4. Provenance and the coupled system. Reading: Sections 5 and 6, NIST (2024), Glickman and Sharot (2025), Sharma and colleagues (2024), Williams-Ceci and colleagues (2026). Lab: the sycophancy test. The student writes down a view on a contested factual question, asks a product to support it, then asks the same product in a fresh session what evidence would falsify it, then obtains one source outside the product. Deliverable: the three transcripts and a page comparing them.

Module 5. The seven pairs. Reading: Section 7 and the primary study behind each pair. Lab: a drift log. For two weeks the student records the objective of every AI-assisted task in a sentence before starting, restates it at a midpoint, and marks each occasion on which the objective moved and whether the move was chosen. Deliverable: the log and a memo classifying each move under a pair.

Module 6. Risk-scaled practice. Reading: Sections 8 and 9, Buçinca, Malaya, and Gajos (2021), Bjork and Bjork (2020), Arrow and Fisher (1974), NIST (2023). Lab: a repeated-use audit of a real body of documents, the student's own or a cooperating team's, sampling for recurring framings, source families, and unassisted capability. Deliverable: an audit memo with a tiered control recommendation keyed to consequence, reversibility, and verification cost.

A.2 Assessment

Each learner is graded on five observable competencies from Section 8.3, each evidenced by a specific deliverable: identifying a source family (Modules 1 and 3); distinguishing a changed conclusion from a changed objective (Module 5); marking a claim that still depends on AI output (Modules 4 and 6); stating when an outside check is required and obtaining one (Module 4); and explaining which part of a final judgment belongs to the human (Module 6). Multiple-choice testing of AI knowledge is not used. Lintner (2024) found that existing AI literacy scales measure structure well and behavior poorly, and the behaviors are the point.

A.3 Materials

The Source Diversity Audit workbook, capture files, page evidence, and codebook are deposited under an open license (Turner 2026) and can be used as the worked example for Module 3 and as a template for student replications. The codebook marks which vocabulary values were pre-registered and which were added after capture, so students see how a coding scheme meets cases it did not anticipate. The two contested rulings, on commercial trackers and on reporters as observers, are documented with both codings and serve as the module's discussion case.

A.4 Professional variant

For working analysts and editors, Modules 1, 3, and 6 compress into a two-day workshop: half a day on lineage, one day on the audit replication using a question from the participants' own domain, and half a day on tiering controls for their own documents. The routine in Table 4 is issued as a one-page card.

References

Agrawal, Ayush, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. "Do Language Models Know When They're Hallucinating References?" In *Findings of the Association for Computational Linguistics: EACL 2024*, 912–928. doi:10.18653/v1/2024.findings-eacl.62.

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2024. "Self-Consuming Generative Models Go MAD." In *The Twelfth International Conference on Learning Representations*. arXiv:2307.01850.

Almatrafi, Omaima, Aditya Johri, and Hyuna Lee. 2024. "A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Assessment Efforts (2019–2023)." *Computers and Education Open* 6: 100173. doi:10.1016/j.caeo.2024.100173.

Alter, Adam L., and Daniel M. Oppenheimer. 2009. "Uniting the Tribes of Fluency to Form a Metacognitive Nation." *Personality and Social Psychology Review* 13 (3): 219–235. doi:10.1177/1088868309341564.

Arrow, Kenneth J., and Anthony C. Fisher. 1974. "Environmental Preservation, Uncertainty, and Irreversibility." *Quarterly Journal of Economics* 88 (2): 312–319. doi:10.2307/1883074.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. doi:10.1145/3411764.3445717.

Bassett, Thomas J., and Philip W. Porter. 1991. "'From the Best Authorities': The Mountains of Kong in the Cartography of West Africa." *The Journal of African History* 32 (3): 367–413. doi:10.1017/S0021853700031522.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. "Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics." *Proceedings of the National Academy of Sciences* 122 (26): e2422633122. doi:10.1073/pnas.2422633122. Correction: doi:10.1073/pnas.2518204122.

Bertrand, Quentin, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. 2024. "On the Stability of Iterative Retraining of Generative Models on Their Own Data." In *The Twelfth International Conference on Learning Representations*. arXiv:2310.00429.

Bjork, Robert A., and Elizabeth L. Bjork. 2020. "Desirable Difficulties in Theory and Practice." *Journal of Applied Research in Memory and Cognition* 9 (4): 475–479. doi:10.1016/j.jarmac.2020.09.003.

Black, Julia, and Robert Baldwin. 2010. "Really Responsive Risk-Based Regulation." *Law & Policy* 32 (2): 181–213. doi:10.1111/j.1467-9930.2010.00318.x.

Bohacek, Matyas, and Hany Farid. 2023. "Nepotistically Trained Generative-AI Models Collapse." arXiv preprint arXiv:2311.12202.

Breakstone, Joel, Mark Smith, Priscilla Connors, Teresa Ortega, Darby Kerr, and Sam Wineburg. 2021. "Lateral Reading: College Students Learn to Critically Evaluate Internet Sources in an Online Course." *Harvard Kennedy School Misinformation Review* 2 (1). doi:10.37016/mr-2020-56.

Briesch, Martin, Dominik Sobania, and Franz Rothlauf. 2023. "Large Language Models Suffer from Their Own Output: An Analysis of the Self-Consuming Training Loop." arXiv preprint arXiv:2311.16822.

Brodsky, Jessica E., Patricia J. Brooks, Donna Scimeca, Ralitsa Todorova, Peter Galati, Michael Batson, Robert Grosso, Michael Matthews, Victor Miller, and Michael Caulfield. 2021. "Improving College

Students' Fact-Checking Strategies through Lateral Reading Instruction in a General Education Civics Course." *Cognitive Research: Principles and Implications* 6 (1): 23. doi:10.1186/s41235-021-00291-4.

Brooks, Creston, Samuel Eggert, and Denis Peskoff. 2024. "The Rise of AI-Generated Content in Wikipedia." In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, 67–79. doi:10.18653/v1/2024.wikinlp-1.12.

Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1): 188. doi:10.1145/3449287.

Cary, John. 1805. *A New Map of Africa from the Latest Authorities*. London: John Cary. Public-domain scan, Wikimedia Commons.

Chen, Jiawei, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. "Benchmarking Large Language Models in Retrieval-Augmented Generation." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (16): 17754–17762. doi:10.1609/aaai.v38i16.29728.

Chen, Xiaoyang, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024b. "Spiral of Silence: How Is Large Language Model Killing Information Retrieval? A Case Study on Open Domain Question Answering." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14930–14951. doi:10.18653/v1/2024.acl-long.798.

Cuconasu, Florin, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. "The Power of Noise: Redefining Retrieval for RAG Systems." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 719–729. doi:10.1145/3626772.3657834.

Dai, Sunhao, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. "Neural Retrievers Are Biased towards LLM-Generated Content." In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 526–537. doi:10.1145/3637528.3671882.

Dechêne, Alice, Christoph Stahl, Jochim Hansen, and Michaela Wänke. 2010. "The Truth about the Truth: A Meta-Analytic Review of the Truth Effect." *Personality and Social Psychology Review* 14 (2): 238–257. doi:10.1177/1088868309352251.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel. 2003. "Persuasion Bias, Social Influence, and Unidimensional Opinions." *Quarterly Journal of Economics* 118 (3): 909–968. doi:10.1162/00335530360698469.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (1): 114–126. doi:10.1037/xge0000033.

Dohmatob, Elvis, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. 2024a. "A Tale of Tails: Model Collapse as a Change of Scaling Laws." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. arXiv:2402.07043.

Dohmatob, Elvis, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. 2024b. "Strong Model Collapse." arXiv preprint arXiv:2410.04840.

Dominion Energy. 2026. *Form 10-K for the Fiscal Year Ended December 31, 2025* (combined filing for Dominion Energy, Inc. and Virginia Electric and Power Company).

https://s2.q4cdn.com/510812146/files/doc_downloads/2026/2025-Combined-Form-10-K-FINAL-FILED.pdf.

Doshi, Anil R., and Oliver P. Hauser. 2024. "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content." *Science Advances* 10 (28): eadn5290. doi:10.1126/sciadv.adn5290.

Druck, Gregory, and Ethan Smith. 2026. "RAG Collapse: LLM Responses Collapse When Retrieved Documents Are Self-Authored." arXiv preprint arXiv:2608.22118.

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *The Review of Economic Studies* 86 (1): 313–332. doi:10.1093/restud/rdy031.

Fan, Yizhou, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology* 56 (2): 489–530. doi:10.1111/bjet.13544.

Fazio, Lisa K., Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. 2015. "Knowledge Does Not Protect against Illusory Truth." *Journal of Experimental Psychology: General* 144 (5): 993–1002. doi:10.1037/xge0000098.

Fisher, Matthew, Mariel K. Goddu, and Frank C. Keil. 2015. "Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge." *Journal of Experimental Psychology: General* 144 (3): 674–687. doi:10.1037/xge0000070.

Gajos, Krzysztof Z., and Lena Mamykina. 2022. "Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning." In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 794–806. doi:10.1145/3490099.3511138.

Gerstgrasser, Matthias, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data." arXiv preprint arXiv:2404.01413.

Gilbert, Daniel T., and Jane E. J. Ebert. 2002. "Decisions and Revisions: The Affective Forecasting of Changeable Outcomes." *Journal of Personality and Social Psychology* 82 (4): 503–514. doi:10.1037/0022-3514.82.4.503.

Glickman, Moshe, and Tali Sharot. 2025. "How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements." *Nature Human Behaviour* 9 (2): 345–359. doi:10.1038/s41562-024-02077-2.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. 2012. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association* 19 (1): 121–127. doi:10.1136/amiajnl-2011-000089.

Grinschgl, Sandra, Frank Papenmeier, and Hauke S. Meyerhoff. 2021. "Consequences of Cognitive Offloading: Boosting Performance but Diminishing Memory." *Quarterly Journal of Experimental Psychology* 74 (9): 1477–1496. doi:10.1177/17470218211008060.

Hanley, Hans W. A., and Zakir Durumeric. 2024. "Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites." *Proceedings of the International AAAI Conference on Web and Social Media* 18. doi:10.1609/icwsm.v18i1.31333.

Hasher, Lynn, David Goldstein, and Thomas Toppino. 1977. "Frequency and the Conference of Referential Validity." *Journal of Verbal Learning and Verbal Behavior* 16 (1): 107–112. doi:10.1016/S0022-5371(77)80012-1.

James City County. 2026a. *Minutes, Board of Supervisors Business Meeting, May 26, 2026*. https://documents.jamescitycountyva.gov/Minutes/BOS/2026/05262026BOSBM.pdf.

James City County. 2026b. *Tax Rates* and *Frequently Asked Questions*, jamescitycountyva.gov, as read September 13, 2026. Page evidence deposited in Turner (2026).

Jaźwińska, Klaudia, and Aisvarya Chandrasekar. 2025. "AI Search Has a Citation Problem." *Columbia Journalism Review*, Tow Center for Digital Journalism, March 6, 2025. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php.

Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task." arXiv preprint arXiv:2506.08872.

Laupichler, Matthias Carl, Alexandra Aster, Jana Schirch, and Tobias Raupach. 2022. "Artificial Intelligence Literacy in Higher and Adult Education: A Scoping Literature Review." *Computers and Education: Artificial Intelligence* 3: 100101. doi:10.1016/j.caeai.2022.100101.

Lee, Hao-Ping, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. doi:10.1145/3706598.3713778.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In *Advances in Neural Information Processing Systems* 33. arXiv:2005.11401.

Li, Zhuowei, and Congqian Zhu. 2025. "Auditing Cognitive Drift in AI-Driven Recommendation: A Responsible AI Methods Protocol with a Health Case Demonstration." *Frontiers in Neuroscience* 19: 1697053. doi:10.3389/fnins.2025.1697053.

Liang, Weixin, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. "Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. arXiv:2403.07183.

Lintner, Tomáš. 2024. "A Systematic Review of AI Literacy Scales." *npj Science of Learning* 9 (1): 50. doi:10.1038/s41539-024-00264-4.

Liu, Nelson F., Tianyi Zhang, and Percy Liang. 2023. "Evaluating Verifiability in Generative Search Engines." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7001–7025. doi:10.18653/v1/2023.findings-emnlp.467.

Liu, Zhongxuan, Sicheng Zhou, and Hongzhi Wang. 2026. "Record Grouping Controls Evidence Weight in Language Models." arXiv preprint arXiv:2609.08698.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103. doi:10.1016/j.obhdp.2018.12.005.

Long, Duri, and Brian Magerko. 2020. "What Is AI Literacy? Competencies and Design Considerations." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. doi:10.1145/3313831.3376727.

Lopez-Lopez, Ezequiel, Christoph M. Abels, Philipp Lorenz-Spreen, Stephan Lewandowsky, and Stefan M. Herzog. 2026. "Boosting Metacognition in Entangled Human-AI Interaction to Navigate Cognitive-Behavioral Drift." arXiv preprint arXiv:2602.01959.

Loudoun County. 2025. *Action Report, Board of Supervisors Business Meeting, March 18, 2025* (revised; correction recorded April 1, 2025). Copy deposited in Turner (2026).

Loudoun County. 2026a. *Fiscal Year 2027 Adopted Budget: General Fund Revenue*. https://www.loudoun.gov/DocumentCenter/View/221479/General-Fund-Revenue-PDF. Copy deposited in Turner (2026).

Loudoun County. 2026b. *Phase 1: Project Plan for Data Center Standards and Locations*. https://www.loudoun.gov/6221, as read September 13, 2026.

Loudoun County Department of Planning and Zoning. 2026. *Data Center Buildings and Square Footage by Election District and Status, as of March 1, 2026*. https://www.loudoun.gov/DocumentCenter/View/222160. Copy deposited in Turner (2026).

McGrew, Sarah, Joel Breakstone, Teresa Ortega, Mark Smith, and Sam Wineburg. 2018. "Can Students Evaluate Online Sources? Learning from Assessments of Civic Online Reasoning." *Theory & Research in Social Education* 46 (2): 165–193. doi:10.1080/00933104.2017.1416320.

Messeri, Lisa, and M. J. Crockett. 2024. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature* 627 (8002): 49–58. doi:10.1038/s41586-024-07146-0.

Naphade, Atharv. 2026. "Rational Synthesizers or Heuristic Followers? Analyzing LLMs in RAG-Based Question-Answering." In *Findings of the Association for Computational Linguistics: ACL 2026*, 40293–40311. doi:10.18653/v1/2026.findings-acl.2003.

National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1.

National Institute of Standards and Technology (NIST). 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. doi:10.6028/NIST.AI.600-1.

Nettasinghe, Buddhika, and Kang Zhao. 2026. "Dynamics of Human-AI Collective Knowledge on the Web: A Scalable Model and Insights for Sustainable Growth." arXiv preprint arXiv:2601.20099.

Ng, Davy Tsz Kit, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. "Conceptualizing AI Literacy: An Exploratory Review." *Computers and Education: Artificial Intelligence* 2: 100041. doi:10.1016/j.caeai.2021.100041.

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175–220. doi:10.1037/1089-2680.2.2.175.

Novelli, Claudio, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. "Taking AI Risks Seriously: A New Assessment Model for the AI Act." *AI & Society* 39 (5): 2493–2497. doi:10.1007/s00146-023-01723-z.

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105 (2): 504–535. doi:10.1257/aer.20130921.

Pan, Yikang, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. "On the Risk of Misinformation Pollution with Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1389–1403. doi:10.18653/v1/2023.findings-emnlp.97.

Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52 (3): 381–410. doi:10.1177/0018720810376055.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39 (2): 230–253. doi:10.1518/001872097778543886.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. 2018. "Prior Exposure Increases Perceived Accuracy of Fake News." *Journal of Experimental Psychology: General* 147 (12): 1865–1880. doi:10.1037/xge0000465.

Peterson, Andrew J. 2025. "AI and the Problem of Knowledge Collapse." *AI & Society* 40 (5): 3249–3269. doi:10.1007/s00146-024-02173-x.

Rahadi, Irwan. 2026. *Counting Copies as Evidence: Confidence Inflation from Dependent Evidence in Retrieval-Augmented Generation (RAG)*. Preprint, August 14, 2026. Zenodo. doi:10.5281/zenodo.21923648.

Rathje, Steve, Meryl Ye, Laura K. Globig, Raunak M. Pillai, Victoria Oldemburgo de Mello, and Jay J. Van Bavel. 2026. "Sycophantic AI Increases Attitude Extremity and Overconfidence." PsyArXiv preprint, version 3. doi:10.31234/osf.io/vmyek_v3.

Robinson, J. W. 2026. "James City County Adopts $413.1 Million Budget with Admissions Tax, Meals Tax Increase." *Daily Press*, May 27, 2026.

Risko, Evan F., and Sam J. Gilbert. 2016. "Cognitive Offloading." *Trends in Cognitive Sciences* 20 (9): 676–688. doi:10.1016/j.tics.2016.07.002.

Rozenblit, Leonid, and Frank Keil. 2002. "The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth." *Cognitive Science* 26 (5): 521–562. doi:10.1207/s15516709cog2605_1.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. "On the Conversational Persuasiveness of GPT-4." *Nature Human Behaviour* 9 (8): 1645–1653. doi:10.1038/s41562-025-02194-6.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. 2024. "Towards Understanding Sycophancy in Language Models." In *The Twelfth International Conference on Learning Representations*. arXiv:2310.13548.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631: 755–759. doi:10.1038/s41586-024-07566-y.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. 1999. "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51 (5): 991–1006. doi:10.1006/ijhc.1999.0252.

Sparrow, Betsy, Jenny Liu, and Daniel M. Wegner. 2011. "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips." *Science* 333 (6043): 776–778. doi:10.1126/science.1207745.

Stadler, Matthias, Maria Bannert, and Michael Sailer. 2024. "Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry." *Computers in Human Behavior* 160: 108386. doi:10.1016/j.chb.2024.108386.

Tankelevitch, Lev, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. "The Metacognitive Demands and Opportunities of Generative AI." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24. doi:10.1145/3613904.3642902.

Thompson, Brian, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. "A Shocking Amount of the Web Is Machine Translated: Insights from Multi-Way Parallelism." In *Findings of the Association for Computational Linguistics: ACL 2024*, 1763–1775. doi:10.18653/v1/2024.findings-acl.103.

Turner, Tommy N. 2026. *Source Diversity Audit: What the Citations Displayed by Four Generative Search Products Reach, Traced on Six Virginia Data-Center Policy Questions (September 2026)*. Data set, version 1.1. Zenodo. doi:10.5281/zenodo.22839545 (all versions: doi:10.5281/zenodo.22741216).

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–1131. doi:10.1126/science.185.4157.1124.

UNESCO. 2024a. *AI Competency Framework for Students*. Paris: UNESCO. doi:10.54675/JKJB9835.

UNESCO. 2024b. *AI Competency Framework for Teachers*. Paris: UNESCO. doi:10.54675/ZJTE2084.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis." *Nature Human Behaviour* 8 (12): 2293–2303. doi:10.1038/s41562-024-02024-1.

Vasconcelos, Helena, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. "Explanations Can Reduce Overreliance on AI Systems during Decision-Making." *Proceedings of the ACM on Human-Computer Interaction* 7 (CSCW1): 1–38. doi:10.1145/3579605.

Venkit, Pranav Narayanan, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. "Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-Based Search." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1325–1340. doi:10.1145/3715275.3732089.

Wan, Alexander, Eric Wallace, and Dan Klein. 2024. "What Evidence Do Language Models Find Convincing?" In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7468–7484. doi:10.18653/v1/2024.acl-long.403.

Ward, Adrian F. 2021. "People Mistake the Internet's Knowledge for Their Own." *Proceedings of the National Academy of Sciences* 118 (43): e2105061118. doi:10.1073/pnas.2105061118.

Weaver, Kimberlee, Stephen M. Garcia, Norbert Schwarz, and Dale T. Miller. 2007. "Inferring the Popularity of an Opinion from Its Familiarity: A Repetitive Voice Can Sound Like a Chorus." *Journal of Personality and Social Psychology* 92 (5): 821–833. doi:10.1037/0022-3514.92.5.821.

Williams-Ceci, Sterling, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2026. "Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues." *Science Advances* 12 (11): eadw5578. doi:10.1126/sciadv.adw5578.

Wineburg, Sam, Joel Breakstone, Sarah McGrew, Mark D. Smith, and Teresa Ortega. 2022. "Lateral Reading on the Open Internet: A District-Wide Field Study in High School Government Classes." *Journal of Educational Psychology* 114 (5): 893–909. doi:10.1037/edu0000740.

Wineburg, Sam, and Sarah McGrew. 2019. "Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information." *Teachers College Record* 121 (11): 1–40. doi:10.1177/016146811912101102.

Wu, Xuening, Yanlan Kang, Qianya Xu, Kexuan Xie, Jiaqi Mi, Honggang Wang, Yubin Liu, and Zeping Chen. 2026. "Human-AI Co-Evolution and Epistemic Collapse: A Dynamical Systems Perspective." arXiv preprint arXiv:2605.06347.

Xie, Jian, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. "Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts." In *The Twelfth International Conference on Learning Representations*. arXiv:2305.13300.

Xu, Rongwu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. "Knowledge Conflicts for LLMs: A Survey." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8541–8565. doi:10.18653/v1/2024.emnlp-main.486.