THE GOAL: Misalignment between the measured metrics and the nebulous concept of critical thinking in joint decision-making

ELENA SERGEEVA†, MIT, USA
ANASTASIA SERGEEVA‡, University of Luxembourg, Luxembourg

The increasing adoption of Generative AI in knowledge-heavy jobs workflows raises questions about its impact on the human knowledge worker's critical thinking (CT) skills and decision-making. Some people think we can mitigate the effects AI has on the human decision-making process by creating AI-based interventions that would encourage more thoughtful decision-making. Creating such interventions, however, requires a precise definition of what critical thinking is and how to measure it. Additionally, encouraging certain modes of thinking in a controlled environment doesn't necessarily translate to people applying them in a decision-making process. This position paper clarifies the problems stemming from both the definitional and the practical CT measurement issues, discusses what characteristics a good proxy measure of CT should have, and talks about human response to 'easy automation'.

Additional Key Words and Phrases: Critical thinking, AI-assisted decision making

1 Introduction

The majority of the practitioners in the broader AI field are familiar with the general model of decision-making, where the goal of the decider is to find a solution that corresponds to the maximum of some sort of utility function. Picking the maximum expected utility (a goal that is frequently implicitly assumed by AI engineers during the design process) is generally known under the umbrella term 'Rational choice theory' in Economics and Psychology [11]. The eponymous Rationality in this model is expressed as the internal coherence of a set of preferences based on which a decider makes the decision: that is, the decision is made based on the corresponding values of the utility function alone, and the values assigned to the choices under this model are: a) Independent from each other b) Self-consistent in the variety of situations where the choice might present itself, and fundamentally reflect the same underlying set of preferences of the decider.

In practice, however, this kind of modelling does not necessarily explain the observed phenomena in human decision-making. It is generally accepted in the field that humans operate under relatively strict computational constraints when making a decision; as such, the decision-making process constitutes a balancing act between finding an acceptable solution and the computational complexity of the decision evaluation. This broader framework is known as 'Bounded rationality' [24]. Where does critical thinking fit into this kind of view of decision-making? To answer this, we should first define what critical thinking is.

2 What is Critical Thinking ?

The idea of promoting critical thinking to make humans better at making decisions as the main goal of the K-12 education, entered the US discourse in the 1980s as a part of an education liberalization trend [7]. Paradoxically, despite

This paper was presented at the 2025 ACM Workshop on Human-AI Interaction for Augmented Reasoning (AIREASONING-2025-01). This is the authors' version for arXiv.
*Both authors contributed equally to this research.
†Both authors contributed equally to this research.

Authors' Contact Information: Elena Sergeeva, MIT, Cambridge, USA, elenaser@mit.edu; Anastasia Sergeeva, University of Luxembourg, Luxembourg, Luxembourg.

the decades of debates, the exact definition of what counts as critical thinking remains ill-defined [14]. According to Robert Ennis, the concept of critical thinking originates in the works of the philosopher John Dewey on 'reflective thinking' [7]. Ennis defines critical thinking as 'reasonable reflective thinking focused on deciding what to believe or do'[7]. This definition does not specify what characterizes the 'reasonability' or 'reflectivity' of thinking: the author tries to give examples of some of the observable characteristic of the process like 'seeking and offering clear reason' and 'seriously considering other points of views', but the examples given are explicitly stated not to be exhaustive, so a person can still be engaged in critical reasoning without explicitly exhibiting any of those characteristics to the observer. Similarly Peter Facione, provides the following definition: to be considered 'critical', thinking must 'be a purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based' .[9] (p.2) The commonly used Bloom taxonomy [3] of 'intellectual abilities and skills' tries to split the thinking process into a set of sub-processes starting from the 'less-reflective' things like idea recall and ending with the higher-level concepts of judgment (or critical evaluation) and creation. Later works [2] that tried to update the taxonomy to represent the process of thinking and judging in a more hierarchical fashion used the following stages to describe the process: *Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating*. Which stages correspond to the actual engagement in critical thinking and even the existence of a separate from the domain knowledge 'critical thinking skill' remains hotly debated [12, 19].

3 Measuring Critical Thinking

As we have shown in the previous section of this paper, the definitions of critical thinking are often broad enough to make any meaningful measurement difficult. However, as the work around the concept was mostly done within the framework of educational psychology, we can take another, more practical approach and specify critical thinking as a set of skills that can be verified through the use of formalized measurement procedures and have proven correlative relationships with real-world behavior and approaches to tasks. According to [25], there are at least 30 tests fully or partly dedicated to the assessment of critical thinking skills. They are based on different theoretical foundations, define different sets of skills as 'critical thinking skills' and test them by applying measurement instruments with varying degrees of freedom: from choosing an answer from a multiple-choice question to writing a detailed critical essay dedicated to the topic. Selected examples are presented in Table 1. This highly operationalized approach to critical thinking testing has been criticized from multiple points of view. With respect to easy to score, multiple-choice options tests, one can doubt their ability to assess the actual reasoning process [13] [22]. With respect to evaluating more sophisticated output assessments such as problem essays and sentence completion the output evaluation (grading using a rubric guide) itself comes into question due to its inherent subjectivity and potential lack of specific critical skills assessment expertise among the graders [21]. Additionally, formal critical thinking education does not result in big changes in test performance [14, 21], which points to either the general inability of the current programs to teach critical thinking as a skill or the test itself not measuring critical thinking improvement in students correctly.

4 'Right' kind of thinking, its proxies and AI

Now that we have a (self-admittedly) vague idea of what critical thinking is and how it hasbeen measured by psychologists and education experts before, we should decide on an appropriate proxy variable that we can measure and optimize our AI intervention for. Critical thinking describes the process a human is engaged in, not the result of the process itself. What is often measured in the AI-based intervention research studies is the result based proxy: often, a user being

Table 1. Critical Thinking Tests: a sample, descriptions adapted from criticalthinking.net

Test	Form of Assessment	What is being tested (per the authors)	Avaliability
CCTST[10]	Multiple-choice	Ability to analyze, infer, explain, evaluate, and interpret information.	Commercial
CLA[15]	Written response to task	Data Literacy, critical reading and evaluation , critiquing an argument.	Commercial
Cornell CT Z[8]	Multiple-choice	Deduction, fallacies identification, judging the reliability of supporting statements.	Commercial
The Ennis-Weir CT Essay Test [28]	Essay response	Understanding the problem, reasons and assumptions; presenting the position, providing qualitative reasoning; avoiding the use of emotion to persuade, irrelevance and dealing with credibility problems.	Free

persuaded by the system output that a certain statement is correct or not correct [5] or self-reported engagement in 'critical thinking' aligned processes [17]. Alternatively one can imagine picking one of the critical thinking measuring tests' performance as a target metric to use to asses changes in an AI user's critical thinking.

Are those kind of proxy variables good proxies ?

(1) With respect to 'Resulting belief' goal as a proxy: There are several theories on how humans can get persuaded to make a certain decision. Many of them acknowledge that persuasion doesn't have to rely on the person thinking about the fact critically. For example, the Elaboration Likelihood Model (ELM)[20], states that persuasion can happen through two different routes: the central route, which involves high levels of critical thinking and requires more effort from the user, or the peripheral route, which requires less effort and less critical thinking. Similarly, the persuasive tactics listed by Cialdini [4] exploit human biases and heuristic decision making. For example, the scarcity persuasion tactic exploits the human fear of missing the chance to grab precious and rare assets before it is too late, or the consistency persuasion tactic exploits a person's beliefs about themselves to leverage desired behavior. Imagine an LLM-based bot that has a set of consensus-factual claims it wants to persuade user to believe in. It can lie, invent facts and use all of the persuasion techniques to achieve the end result. Would more people believing in those claims after the interaction with this bot reflect them engaging in more critical thinking?

(2) With respect to 'Reported engagement in activity': Self-reported experiences are notoriously hard to summarize, due to the differences in individual human definitions. Would more people reporting engaging in more critical thinking after a given intervention mean they actually change the way they think about things? For the cases where we have both the test results and self-reported perception measures, the two often tell very different stories. For example older people report the same or lesser levels of cognitive and memory failures as their younger counterparts when interviewed using The Cognitive Failures Questionnaire [6]. At the same time the actual performance on the laboratory memory tests negatively correlates with age [6]. With respect to the Critical thinking attitude vs. skills tests one can say that there is almost no correlation between being 'critical thinking open' and 'critical thinking able' [14].

(3) With respect to 'Performance on standardized test for critical thinking': As summarized in the previous section, those tests have a lot of issues. Additionally, the connection between the test performance and the actual thinking outcomes 'in the wild' remains a hot topic in educational psychology [16].

Based of the listed issues, we invite the community to think about a better proxy metric that :

(1) Does not solely rely on the distant end result to assess the changes in the process
(2) Does not solely rely on human self-reported changes in the process
(3) Allows us to measure both intermediate and final changes in the decision making process

5 Concluding remarks: Critical thinking and the constrained use of AI

Imagine we have constructed an AI-based intervention that somehow improves human critical thinking as defined by that elusive better critical thinking metric. Will a human who became 'better' at critical thinking apply their ability to 'critically think' 'in the wild' when using AI models? Let us get back to the bounded rationality model of decision-making: humans make decisions to satisfy a certain set of 'good enough' constraints with the minimum computational effort required from the human. Critical thinking is inherently computationally expensive, if a joint system produces an acceptable (with respect to any given goal) decision without it, what would motivate a human to engage in it? People will always gravitate towards the least effort satisfactory solution, so if the system provides a satisfactory answer the user will be very unlikely to extend additional human effort to modify or engage with the answer beyond accepting it.

One can look at the adoption of the use of AI clinical scribes to see the way AI-automation unfolds in a knowledge-heavy reasoning-heavy domain of medicine. Note-taking (the process of producing structured record of a clinical encounter) is a part of the clinicians' training process [1]. The process of writing is widely understood as a set of specific hierarchical highly structured thinking processes which writers engage in during the act of composing.[18] During the process of writing doctors have to 'process' all of the encounter data, extract the relevant data and often 'reason through' to the diagnosis and treatment itself. That said, a lot of medical professionals find the process of writing those notes burdensome/tiresome and time consuming [23]. As such, AI-scribes are being very rapidly incorporated into the clinician workflow [26], despite relatively frequent important omissions and hallucinations in the produced summaries and possible prompt-induced summarizations biases. Defenders of the AI scribes point towards the reduction of what they call cognitive drudgery and already degraded quality of clinical notes narratives as the reason this part of the medical practice should be automated. A boundary between cognitive drudgery, 'mindless' tasks, and outsourcing reasoning is not as obvious as we would like it to be, however. While not yet deployed in medical practice, models that boast near clinical level of accuracy on diagnostic dialogue interactions [27] and management plan generation are being developed quite rapidly. How can we deploy AI in a way that won't make the human side of the system reason less? By deciding not to automate reasoning and gravitating towards automating reasoning for light tasks instead. To do that, we would need to develop an appropriate way to rank the tasks on that scale. One way to do it would be to compare the difference between the way an expert and a novice would approach that task: if the process of completing the task remains largely unchanged with respect to both the steps one would take and the eventual result one would get, it would be relatively safe to automate. You can also rely on the simple heuristics lists like

(1) Task is mostly retrieval or surface-level review;
(2) Task requires minimal working memory load;
(3) Task does not involve diagnosis, clinical reasoning or planning.

One can imagine defining and analyzing the specific nature of the task before making a decision about its possible automaton and then proceeding with designing the models for very specific tasks where the automation is limited to the repetitive inherently reasoning light tasks where the correctness of the output can be assessed by a knowledge professional without a lot of cognitive investment. For example:

- Good: "Find all the mentions of a certain type of drug class in the 100 pages of patient medical record and provide the end-user with links to the points in the texts where they are mentioned" [Nature of the task: Direct token extraction, end user with knowledge can easily assess if the system is correct or not by very limited context reading]
- Bad: "Ask the user to accept or decline an automatically produced diagnosis that abstracts 100 pages of the patient record and provides the user with an accessible post-hoc explanation as to why." [Nature of the task: Clinical reasoning, end user with relevant knowledge would have to look through the whole patient record and reason through it to assess the result, clinical abstraction risks, doctors not considering the alternatives and differing to the generated result]

Alternatively, we can just accept parts of the reasoning ‘outsourcing’ as something inevitable and focus on the joint human-AI system performance with respect to any given task exclusively.

Acknowledgments

The research was supported by the Luxembourg National Research Fund (REMEDIS, REgulatory and other solutions to MitigatE online DISinformation (INTER/FNRS/21/16554939)).

References

[1] Sari Altschuler, Ian Huntington, Maria Antoniak, and Lauren F Klein. 2024. Clinician as editor: notes in the era of AI scribes. *The Lancet* 404, 10468 (2024), 2154–2155.

[2] Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition.* Addison Wesley Longman, Inc.

[3] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* Longman New York.

[4] Robert Cialdini. 2001. Principles of persuasion. *Arizona State University, eBrand Media Publication* 3, 7 (2001), 2580–2592.

[5] Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (2024), eadq1814.

[6] Joost CF de Winter, D Dodou, and PA Hancock. 2015. On the paradoxical decrease of self-reported cognitive failures with age. *Ergonomics* 58, 9 (2015), 1471–1486.

[7] Robert H Ennis. 2015. Critical thinking: A streamlined conception. In *The Palgrave handbook of critical thinking in higher education.* Springer, 31–47.

[8] Robert H Ennis, Jason Millman, and Thomas N Tomko. 2005. Cornell critical thinking tests: Administration manual. *California: Critical Thinking Company* (2005).

[9] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[10] Peter A Facione. 1990. The California Critical Thinking Skills Test–College Level. Technical Report# 1. Experimental Validation and Content Validity. (1990).

[11] Steven L Green. 2002. Rational choice theory: An overview. In *Baylor University Faculty development seminar on rational choice theory.* Baylor University Waco, TX, USA, 1–72.

[12] Richard L Gregory. 1972. Seeing as thinking: An active theory of perception. *Times Literary Supplement* 23 (1972), 707–708.

[13] Leo Groarke. 2022. What's Wrong with the California Critical Thinking Skills Test? *Critical Thinking Education and Assessment, 2nd ed.* (2022).

[14] Donald Hatcher and Kevin Possin. 2021. Thinking Critically about Critical-Thinking Assessment. *Critical Thinking and Reasoning: Theory, Development, and Practice* (2021).

[15] Richard H Hersh. 2007. Collegiate Learning Assessment (CLA): Defining Critical Thinking, Analytical Reasoning, Problem Solving and Writing Skills. *nd). http://www. teaglefoundation. org/learning/pdf/hersh_ctdefinitions. pdf* (2007).

[16] Emily R Lai. 2011. Critical thinking: A literature review. *Pearson's Research Reports* 6, 1 (2011), 40–41.

[17] Hao-Ping Hank Lee, 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. (2025).

[18] Richard Menary. 2007. Writing as thinking. *Language sciences* 29, 5 (2007), 621–632.

[19] Jennifer Wilson Mulnix. 2012. Thinking critically about critical thinking. *Educational Philosophy and theory* 44, 5 (2012), 464–479.

[20] Richard E Petty, John T Cacioppo, Richard E Petty, and John T Cacioppo. 1986. *The elaboration likelihood model of persuasion.* Springer.

[21] Kevin Possin. 2013. A serious flaw in the Collegiate Learning Assessment [CLA] test. *Informal Logic* 33, 3 (2013), 390–405.

[22] Kevin Possin. 2014. Critique of the Watson-Glaser critical thinking appraisal test: The more you know, the lower your score. *Informal Logic* 34, 4 (2014), 393–416.

[23] Lisa S Rotenstein, Nate Apathy, A Jay Holmgren, and David W Bates. 2023. Physician note composition patterns and time on the EHR across specialty types: a national, cross-sectional study. *Journal of General Internal Medicine* 38, 5 (2023), 1119–1126.

[24] Herbert A Simon. 1990. Bounded rationality. *Utility and probability* (1990), 15–18.

[25] Critical Thinking. 2025. How Can Critical Thinking Skills Be Tested? https://criticalthinking.net/how-can-critical-thinking-skills-be-tested/ Accessed: 2025-03-02.

[26] Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery* 5, 3 (2024), CAT–23.

[27] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654* (2024).

[28] Patrice Holden Werner. 1991. The Ennis-Weir critical thinking essay test: An instrument for testing and teaching. *Journal of Adolescent & Adult Literacy* 34, 6 (1991), 494.