Παραγωγική Τεχνητή Νοημοσύνη και Εκπαίδευση

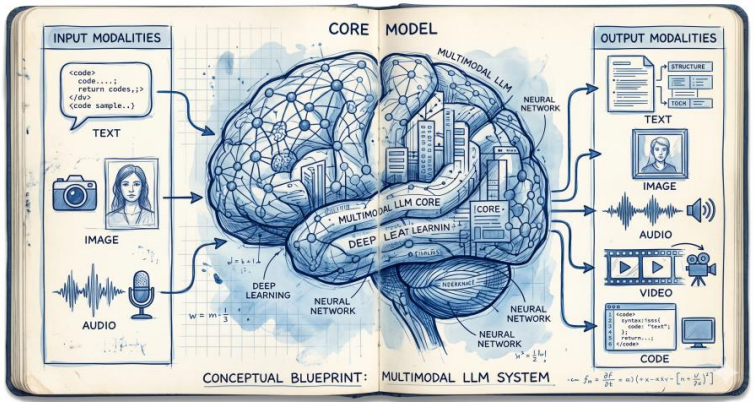
Μια δημιουργική αποδόμηση

Ηλίας Καρασαββίδης
Αναπληρωτής Καθηγητής
Μονάδα ΤΠΕ
Εργαστήριο Θετικών Επιστημών και Τεχνολογίας

Παιδαγωγικό Τμήμα Προσχολικής Εκπαίδευσης
Σχολή Ανθρωπιστικών και Κοινωνικών Επιστημών
Πανεπιστήμιο Θεσσαλίας
ikaras@uth.gr

Περίγραμμα

- Δυνατότητες συστημάτων ΠΤΝ
- Τρόποι "αντίστασης"



01 Δυνατότητες συστημάτων ΠΤΝ

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2026 | Chart: 2026 AI Index report

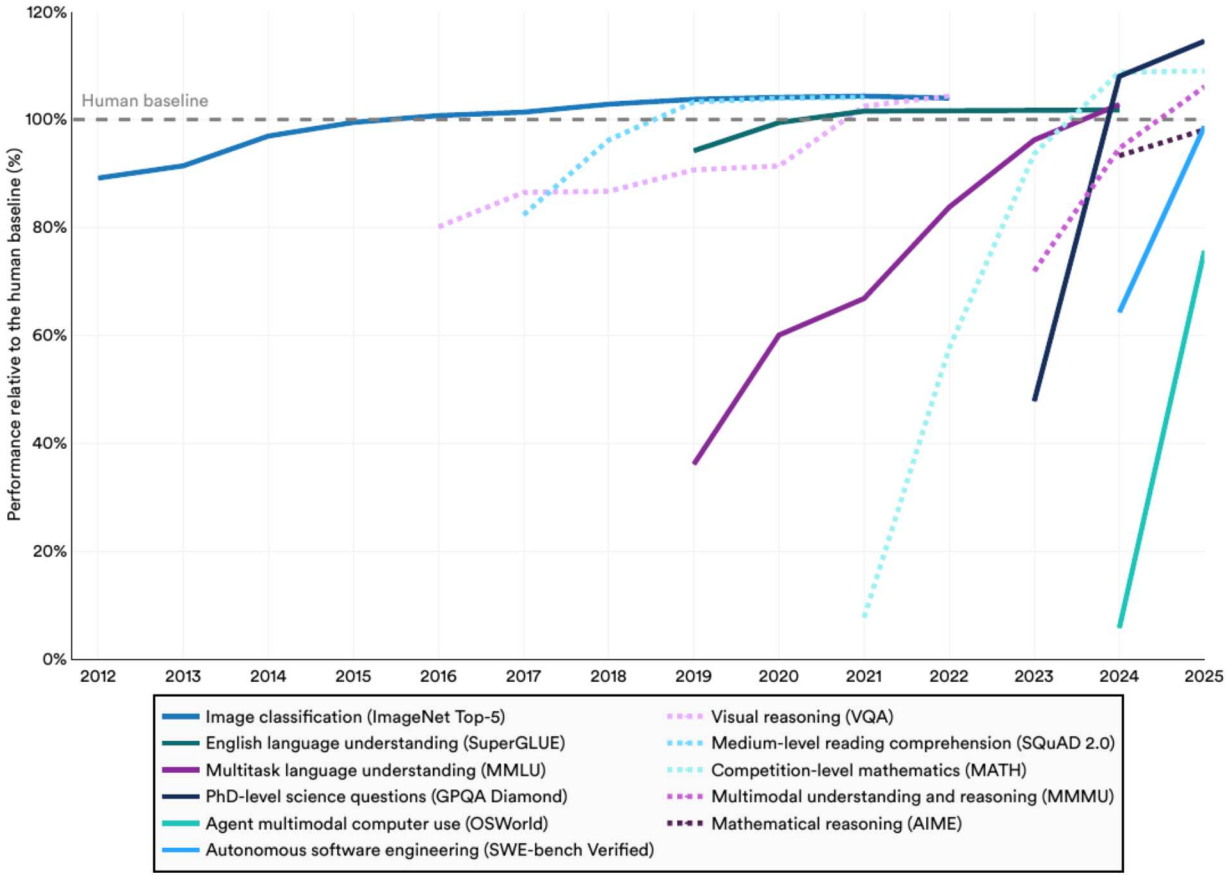
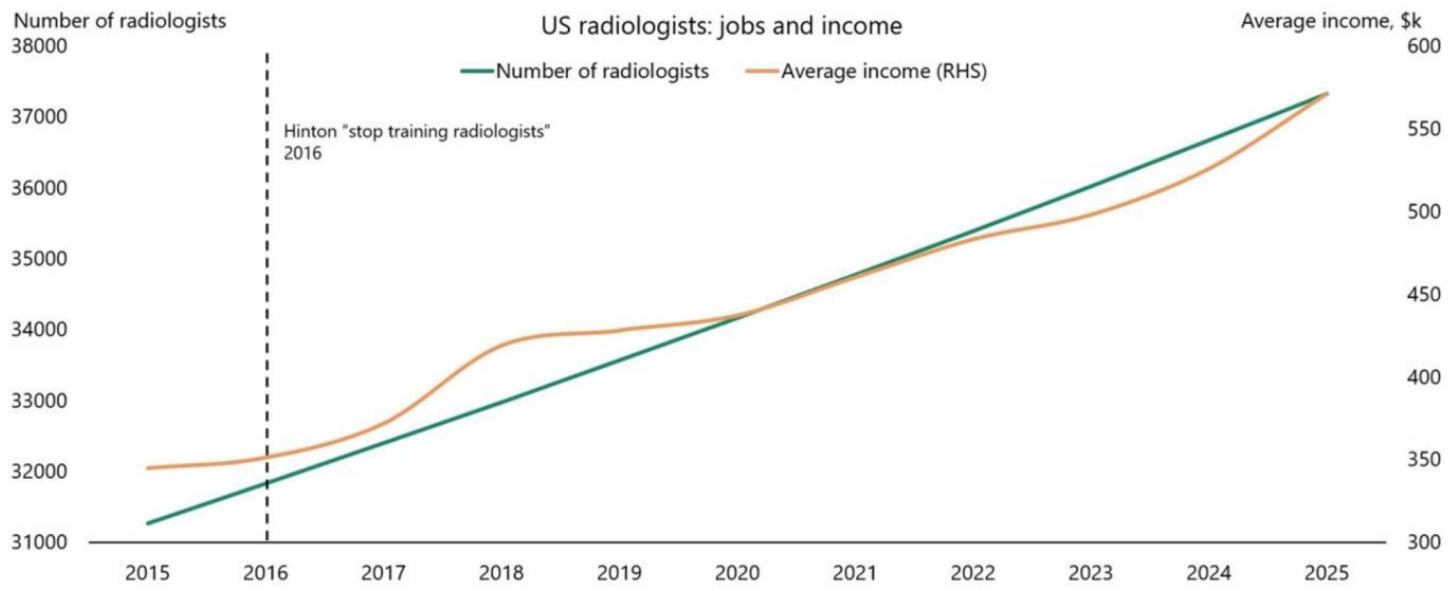


Figure 2.1.1[1]


2015
SELF-DRIVING CARS IN 2 YEARS
CausalPython.io
2016
RADIOLOGISTS OBSOLOLETE IN 5 YEARS
2024
RADIOLOGISTS IN REGULAR CARS DRIVING TO WORK

APOLLO

The radiologist paradox



Sources: CMS Provider Data Catalog (National Downloadable Files), endpoints 30,723 (2014) and 36,024 (2023) per Rosenkrantz et al., AJR 2024: Medscape Physician Compensation Reports 2016–2026, Apollo Chief Economist

Θετική επίδραση!

AI Agents, Productivity, and Higher-Order Thinking: Early Evidence From Software Development

GENERATIVE AI AT WORK*

ERIK BRYNJOLFSSON
DANIELLE LI

Claude Code as an Empirical Economist:
Like Humans but Without the Tails

Serafin Grundl*

Accelerating Scientific Research with Gemini:
Case Studies and Common Techniques

David P. Woodruff[*, †, ‡1,2], Vincent Cohen-Addad[†, ‡1], Lalit Jain[‡1], Jieming Mao[‡1], Song Zuo[†, ‡1], MohammadHossein Bateni[†1], Simina Brânzei[†3,1], Michael P. Brenner[†1,5], Lin Chen[†1], Ying Feng[†6], Lance Fortnow[†7], Gang Fu[†1], Ziyi Guan[†13], Zahra Hadizadeh[†10], Mohammad T. Hajiaghayi[†1,14], Mahdi JafariRaviz[†14], Adel Javanmard[†1,4], Karthik C. S.[†8], Ken-ichi Kawarabayashi[†12], Ravi Kumar[†1], Silvio Lattanzi[†1], Euiwoong Lee[†9], Yi Li[†15], Ioannis Panageas[†10], Dimitris Paparas[†1], Benjamin Przybocki[†2], Bernardo Subercaseaux[†2], Ola Svensson[†13], Shayan Taherijam[†10], Xuan Wu[†15], Eylon Yogev [†16], Morteza Zadimoghaddam[†1], Samson Zhou[†11], Yossi Matias[1], James Manyika[1], and Vahab Mirrokni[*, †, ‡1]

[1]Google Research
[2]Carnegie Mellon University
[3]Purdue University
[4]University of Southern California
[5]Harvard University
[6]MIT
[7]Illinois Institute of Technology
[8]Rutgers University
[9]University of Michigan
[10]University of California, Irvine
[11]Texas A&M University

Θετική επίδραση*

- Μεγαλύτερη ταχύτητα συγγραφής κώδικα, αλλά ταυτόχρονη **αύξηση συνθετότητας** (He et al., 2025)
- Αύξηση παραγωγικότητας, αλλά σε νέα έργα οι εργαζόμενοι είχαν **20% μικρότερη πιθανότητα** να επινοήσουν σωστές λύσεις (Dell'Acqua et al., 2023)

Αρνητική επίδραση

- Ενώ οι προγραμματιστές εκτιμούσαν ότι η χρήση ΠΤΝ θα συντόμευε τον χρόνο ολοκλήρωσης του έργου κατά 20%, στην πράξη αποδείχτηκε ότι η ενσωμάτωση συστημάτων ΠΤΝ επέφερε **χρονική καθυστέρηση** της τάξης του 19% (Becker et al., 2025)

Αρνητική επίδραση

- Προβληματική συνέργεια ανθρώπων - συστημάτων ΠΤΝ (Bean et al., 2026)
  - Αυτόνομα, τα συστήματα ΠΤΝ είχαν ακρίβεια 95% σε διάγνωση
  - Όταν τα ίδια συστήματα ΠΤΝ χρησιμοποιήθηκαν για διάγνωση από ανθρώπους, η ακρίβεια **μειώθηκε** στο 34%

Αρνητική επίδραση

- Το επίπεδο αυτοματοποίησης εργασίας από απόσταση (RLI) με πράκτορες ανέρχεται στο… **2.5%** (Mazeika et al., 2025)
- Αποτελέσματα αξιολόγησης εφαρμογής πρακτόρων από 9 συστήματα αιχμής σε έργα τραπεζικών επενδύσεων (Lau et al., 2026)
  - Το ποσοστό επιτυχίας των συστημάτων ήταν **< 50%**
  - Οι τραπεζίτες αξιολόγησαν τον βαθμό ετοιμότητας των προτάσεων για απευθείας παράδοση σε πελάτες με…**0%**

JAMA Network | Open.

Original Investigation | Health Informatics

Large Language Model Performance and Clinical Reasoning Tasks

Arya S. Rao, BA; Kaiz P. Esmail, BA; Richard S. Lee, BS; Sharon Jiang, BS, MEng; Bianca Arraiza Carlo; Jasleen Gill, MS; Praneet Khanna, BLA; Ezra Kalmowitz, MBE; Basile Montagnese, BE; Kimia Heydari, BA; Qiao Jiao, MS; Ethan Bott, BS; Dan Nguyen, BS; Grace Wang, BS; Michael Hood, MD; Adam B. Landman, MD; Marc D. Succi, MD

Abstract

IMPORTANCE Large language models (LLMs) are increasingly marketed for clinical use, yet their ability to replicate full-spectrum clinical reasoning remains uncertain. Existing evaluations often rely on multiple-choice examinations that do not reflect the complexity of patient care.

OBJECTIVES To evaluate the longitudinal clinical reasoning ability of state-of-the-art LLMs and to introduce a multidimensional, clinically meaningful benchmark for clinical-grade artificial intelligence (AI).

DESIGN, SETTING, AND PARTICIPANTS In this cross-sectional study, performance was evaluated using standardized clinical vignettes from the January 2025 update of MSD Manual vignettes. A total of 21 off-the-shelf LLMs, including recently released GPT-5, Claude 4.5 Opus, Gemini 3.0 Flash and Pro, and Grok 4, were evaluated. Models were assessed by medical student scorers in triplicate across sequential stages of the standard clinical workflow. Analyses were performed from January to December 2025.

MAIN OUTCOMES AND MEASURES The primary outcome was the Proportional Index of Medical Evaluation for LLMs (PrIME-LLM) score, defined as the normalized polygonal area representing balanced accuracy across 5 domains of clinical reasoning as follows: differential diagnosis, diagnostic testing, final diagnosis, management, and miscellaneous clinical reasoning questions. Analyses including analyses of variance, *t* tests, and regression models were used to compare AI model performance and demographic associations.

RESULTS LLMs were tested across 29 clinical vignettes (representing 16 254 responses in total). PrIME-LLM scores ranged from 0.64 (range, 0.63-0.65) (Gemini 1.5 Flash) to 0.78 (range, 0.77-0.79) (Grok 4), with reasoning-optimized models outperforming nonreasoning models and GPT models scoring highest overall. Differential diagnosis was less accurate than diagnostic testing, while final diagnosis, management, and miscellaneous reasoning were more accurate. Failure rates exceeded 0.80 (range, 0.90-1.00) for differential diagnosis in all models but were less than 0.40 (range, 0.09-0.39) for final diagnosis. Multimodal performance was robust; most LLM models showed improved accuracy with image inputs.

CONCLUSIONS AND RELEVANCE In this cross-sectional study of 21 LLMs, frontier LLMs achieved high accuracy on final diagnoses but performed poorly in generating differential diagnoses and navigating uncertainty relative to other reasoning stages. The PrIME-LLM framework provided greater separation than raw accuracy, revealing critical reasoning gaps obscured by traditional benchmarks. Thus, despite version-based improvements and advantages in reasoning-optimized models, off-the-shelf LLMs have not yet achieved the intelligence required for safe deployment and remain limited in demonstrating advanced clinical reasoning.

JAMA Network Open. 2026;9(4):e264003. doi:10.1001/jamanetworkopen.2026.4003

Key Points

Question Can off-the-shelf large language models (LLMs) demonstrate reliable performance across the clinical workflow?

Findings In this cross-sectional study of 21 frontier LLMs tested on 29 standardized clinical vignettes, Grok 4 and other reasoning-optimized models achieved the highest scores, while Gemini 1.5 Flash performed lowest. Differential diagnosis consistently showed the weakest performance, while final diagnosis and management had stronger performances.

Meaning These findings suggest that despite progress, current LLMs remain limited in early diagnostic reasoning and cannot yet be relied on for unsupervised patient-facing clinical decision-making.

+ Invited Commentary
+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Δεν έχουν φτάσει ακόμα στο **απαιτούμενο επίπεδο** 'ευφυίας' για ασφαλή χρήση

Χαρακτηρίζονται από **περιορισμούς** ως προς την υλοποίηση προχωρημένων συλλογισμών με κλινικά δεδομένα

Published in Transactions on Machine Learning Research (01/2026)

Large Language Model Reasoning Failures

Peiyang Song *†
California Institute of Technology, Stanford University
psong@caltech.edu

Pengrui Han *
Carleton College
barryhan@carleton.edu

Noah Goodman
Stanford University
ngoodman@stanford.edu

Reviewed on OpenReview: https://openreview.net/forum?id=vnX1WHMNmz

Abstract

Large Language Models (LLMs) have exhibited remarkable reasoning capabilities, achieving impressive results across a wide range of tasks. Despite these advances, significant reasoning failures persist, occurring even in seemingly simple scenarios. To systema
and address these shortcomings, we present the ***first comprehensive to reasoning failures in LLMs***. We introduce a novel categorizati
distinguishes ***reasoning*** into embodied and non-embodied types, with
subdivided into informal (intuitive) and formal (logical) reasoning. In para
soning ***failures*** along a complementary axis into three types: fundament
to LLM architectures that broadly affect downstream tasks; application-
that manifest in particular domains; and robustness issues characteriz
performance across minor variations. For each reasoning failure, we provid
analyze existing studies, explore root causes, and present mitigation st
ing fragmented research efforts, our survey provides a structured persp
weaknesses in LLM reasoning, offering valuable insights and guiding futu
building stronger, more reliable, and robust reasoning capabilities. We ad
comprehensive collection of research works on LLM reasoning failures,
tory at https://github.com/Peiyang-Song/Awesome-LLM-Reasoning-Fa
an easy entry point to this area.

Mirage: The Illusion of Visual Understanding

Mohammad Asadi[1,*], Jack W. O'Sullivan[2,3,*], Fang Cao[2], Tahoura Nedaee[4], Kamyar Rajabalifardi[1], Fei-Fei Li[5,†], Ehsan Adeli[3,5,6,†], Euan Ashley[2,3,†]

[1]Department of Electrical Engineering, Stanford University, CA, USA
[2]Division of Cardiology, Department of Medicine, Stanford University, CA, USA
[3]Department of Biomedical Data Science, Stanford University, CA, USA
[4]Department of Biology, Stanford University, CA, USA
[5]Department of Computer Science, Stanford University, CA, USA
[6]Department of Psychiatry and Behavioral Sciences, Stanford University, CA, USA

*Equal contributions
†Equal senior contributions

Abstract

Multimodal AI systems have achieved remarkable performance across a broad range of real-world tasks, yet the mechanisms underlying visual–language reasoning remain surprisingly poorly understood. We report three findings that challenge prevailing assumptions about how these systems process and integrate visual information. First, Frontier models readily generate detailed image descriptions and elaborate reasoning traces, including pathology-biased clinical findings, for images never provided; we term this phenomenon *mirage reasoning*. Second, without any image input, models also attain strikingly high scores across general and medical multimodal benchmarks, bringing into question their utility and design. In the most extreme case, our model achieved the top rank on a standard chest X-ray question-answering benchmark without access to any images. Third, when models were explicitly instructed to guess answers without image access, rather than being implicitly prompted to assume images were present, performance declined markedly. Explicit guessing appears to engage a more conservative response regime, in contrast to the *mirage* regime in which models behave as though images have been provided. These findings expose fundamental vulnerabilities in how visual–language models reason and are evaluated, pointing to an urgent need for private benchmarks that eliminate textual cues enabling non-visual inference, particularly in medical contexts where miscalibrated AI carries the greatest consequence. We introduce B-Clean as a principled solution for fair, vision-grounded evaluation of multimodal AI systems.

Αστοχίες

Science News *from research organizations*

This AI knew the answers but didn't understand the questions

Date: April 30, 2026

Source: Science China Press

Summary: For decades, psychologists have debated whether the human mind can be explained by one unified theory or must be broken into separate parts like memory and attention. A recent AI model called Centaur seemed to offer a breakthrough, claiming it could mimic human thinking across 160 different cognitive tasks. But new research is challenging that bold claim, suggesting the model isn't truly "thinking" at all—it's just memorizing patterns.

Share:

RELATED TOPICS

Mind & Brain

- Intelligence
- Language Acquisition
- Child Development
- Literacy

Computers & Math

- Statistics
- Artificial Intelligence
- Computer Modeling

FULL STORY

A cutting-edge AI model that appeared to mimic human thinking may actually just be memorizing answers. New tests reveal it struggles with true understanding, exposing a major gap in today's AI systems. Credit: AI/

PERSPECTIVE

National Science Open
5: 20250053, 2026
https://doi.org/10.1360/nso/20250053

Information Sciences

Can Centaur truly simulate human cognition? The fundamental limitation of instruction understanding

Wei Liu[1] & Nai Ding[1,2,*]

[1]*Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 310013, China;*
[2]*State Key Lab of Brain-Machine Intelligence, MOE Frontier Science Center for Brain Science & Brain-machine Integration, Zhejiang University, Hangzhou 310013, China*

*Corresponding author (email: ding_nai@zju.edu.cn)

Received 16 September 2025; Revised 18 November 2025; Accepted 5 December 2025; Published online 11 December 2025

Traditionally, in psychology, the human mind is divided into modules, such as attention and memory, and each module or submodule, such as top-down attention or working memory, is separately studied and modeled. Whether the human mind could be explained by a unified theory remains unclear. Recently, Binz *et al*. [1] made an important step toward building a unified model, i.e., Centaur, that can predict the human behavior in 160 psychological experiments. Centaur is built by fine-tuning a large language model (LLM) on cognitive tasks and its performance can generalize to held-out participants and unseen tasks, leading the authors to conclude that a single model may comprehensively capture many aspects of human cognition. Although Centaur has reached remarkable performance and provides a valuable tool for cognitive research, it is well-known that LLMs often achieve high performance on fine-tuned tasks and similar tasks by exploiting subtle statistical cues that may even be unnoticeable to humans [2,3]. In other words, the high performance of fine-tuned LLM is sometimes the consequence of overfitting.

To reveal whether the high performance of an LLM is attributable to overfitting, one method is to test whether the LLM performance reduces to the chance level when the input to LLM no longer contains information necessary to perform the task [4,5]. If the LLM still performs above the chance level after crucial information is removed, it is evidence that the LLM bypasses task instructions and directly infers the results based on superficial statistical cues in the answer. The input to Centaur included two parts. One part is the task instruction and the other part is the procedure text. A recent study has shown that the performance of Centaur remains much higher than the baseline cognitive models when the crucial information is removed from the instruction [6]. It remains possible, however, that Centaur successfully infers the task instruction based on the remaining instruction and the procedure text. Therefore, we tested three conditions that either completely removed task information or replaced the task instruction with a misleading instruction (Figure 1a).

We tested Centaur on three conditions.

Διαπιστώσεις

- Η ρητορική απέχει από την πραγματικότητα
- Υπάρχουν αλληοαναιρούμενα οφέλη (trade-offs)

Ωστόσο...

Η πίεση για εισαγωγή σε εκπαιδευτικά συστήματα είναι ασφυκτική


OECD
English
OECD > Publications > OECD Digital Education Outlook 2026
OECD Digital Education Outlook 2026
Exploring Effective Uses of Generative AI in Education
Report
More info
OECD Digital Education Outlook • 19 January 2026
Summary
Support materials

AI Hub for Education

The Evidence Base on AI in K-12: A 2026 Review

The existing research on the impacts of AI on students and teachers

Lily Fesler JP Martinez Claeys Chris Agnew Susanna Loeb

Stanford | SCALE Initiative
Accelerator for Learning

Google

November 2025

AI and the Future of Learning

Ben Gomes
Chief Technologist
Learning & Sustainability

Lila Ibrahim
Chief Operating Officer
Google DeepMind

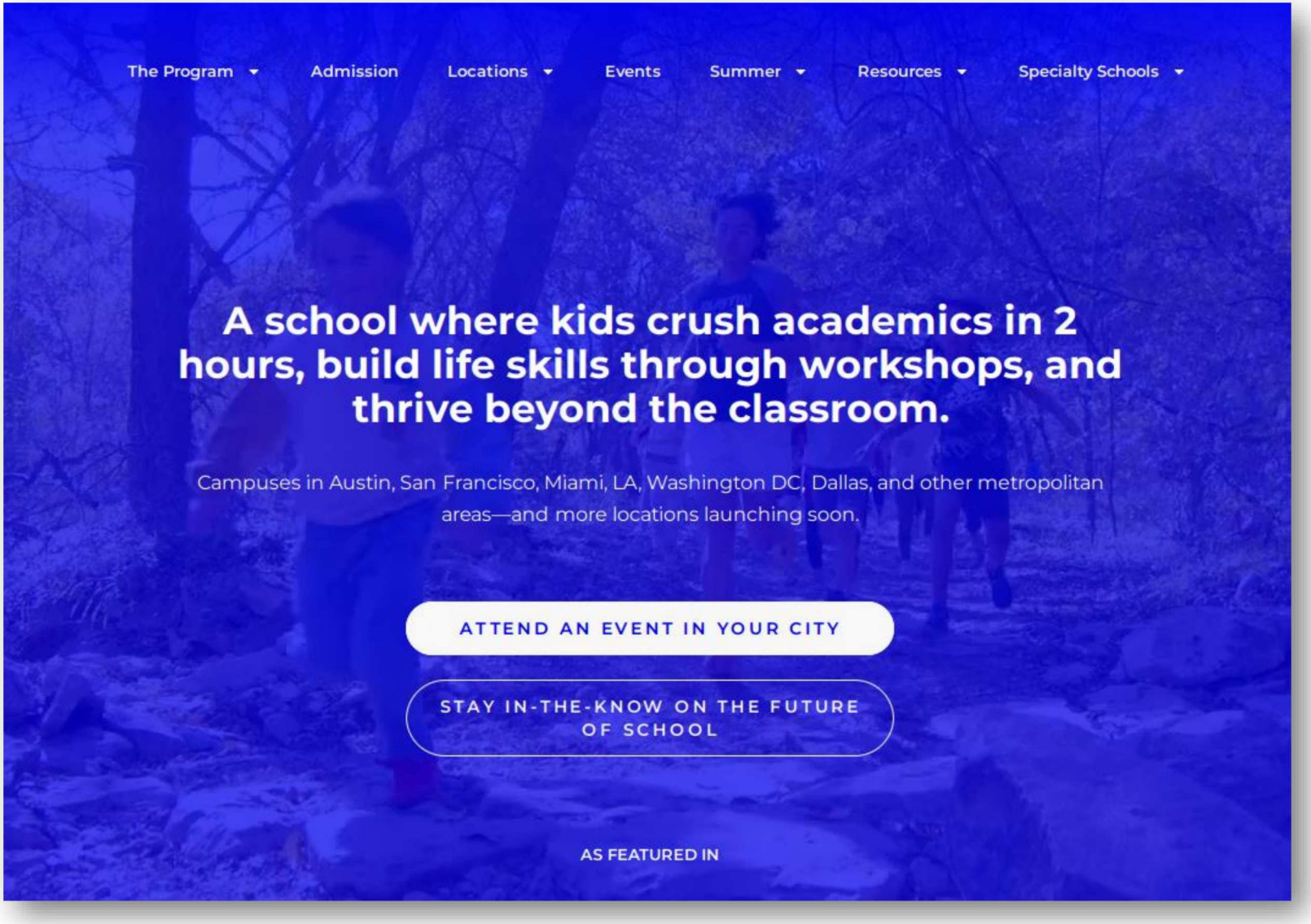
Yossi Matias
VP & GM
Google Research

Christopher Phillips
VP & GM
Education

James Manyika
SVP
Research, Labs, Technology & Society

Anthropic and Teach For All launch global AI training initiative for educators

Jan 21, 2026



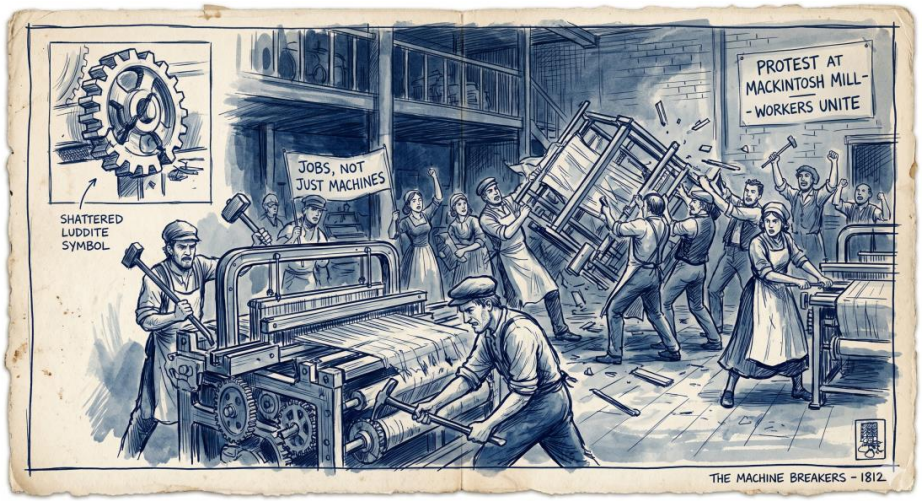
https://alpha.school

Ministers to trial AI tutoring in England's schools

Government claims pilot 'could support up to 450,000 children a year on free school meals to access one to one tutoring'

Freddie Whittaker
More from this author

3 min read | 26 January 2026, 10:30 pm

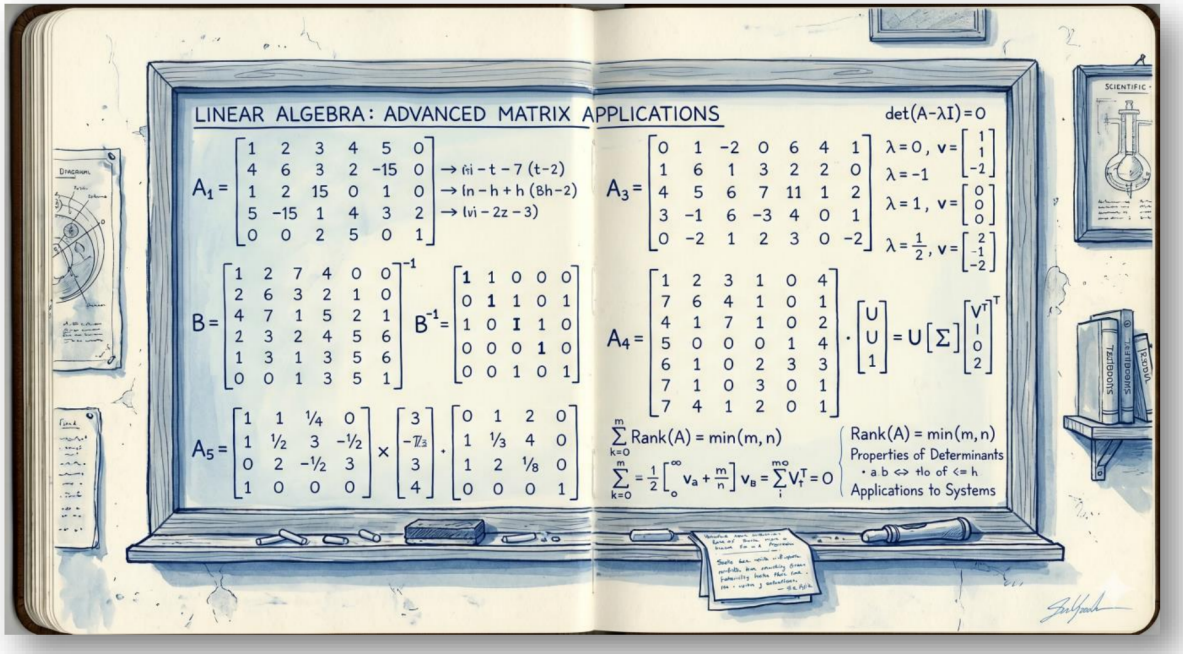


02 Τρόποι αντίστασης

Κριτική Ανάγνωση Εμπειρικής Έρευνας

Μετα-αναλύσεις

- Στατιστικός τρόπος σύνθεσης πρωτογενών μελετών
- Το μέγεθος διαφοράς (effect size) είναι η μετρική που χρησιμοποιείται



Received: 25 January 2023 | Accepted: 17 April 2023
DOI: 10.1111/bjet.13334

British Journal of Educational Technology BERA

REVIEW

Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis

Gökoğlu, S., & Erdoğdu, F. (2025). The effects of GenAI on learning performance: A meta-analysis study. *Educational Technology & Society*, *28*(3), 263-280. https://doi.org/10.30191/ETS.202507_28(3).TP04

The effects of GenAI on learning performance: A meta-analysis study

Seyfullah Gökoğlu[1*] and Fatih Erdoğdu[2]
[1]Bartın University, Türkiye // [2]Zonguldak Bülent Ecevit University, Türkiye // gokogluseyfullah@gmail.com // fatiherdogdu67@gmail.com

Educational Psychology Review (2025) 37:110
https://doi.org/10.1007/s10648-025-10085-5

META-ANALYSIS

Effects of Artificial Intelligence on Educational Functioning: A Review and Meta-Analysis

GeckHong Yeo[1,2] · Jennifer E. Lansford[3]

Received: 28 December 2024 / Accepted: 16 October 2025 / Published online: 19 Nove

Abstract
Burgeoning integration of AI into educational settings could have tions for students' performance. This systematic review and meta- the effects of different types of AI and four levels of learning—co utilization, meta-cognition, and psychological functioning—yi with 464 effect sizes that met criteria for inclusion. AI had large cognition, $r=0.530$, $p<0.001$ [95%CI:0.447 to 0.613] and psycl ing, $r=0.514$, $p<0.001$ [95%CI:0.246 to 0.720], moderate effe utilization, $r=0.417$, $p<0.001$ [95%CI:0.305 to 0.747], and sma

Education and Information Technologies (2025) 30:16211–16239
https://doi.org/10.1007/s10639-025-13420-z

Exploring the impact of generative artificial intelligence on students' learning outcomes: a meta-analysis

Yinkun Zhu[1] · Qiwen Liu[1] · Li Zhao[1]

Received: 23 September 2024 / Accepted: 28 January 2025 / Published online: 26 February 2025

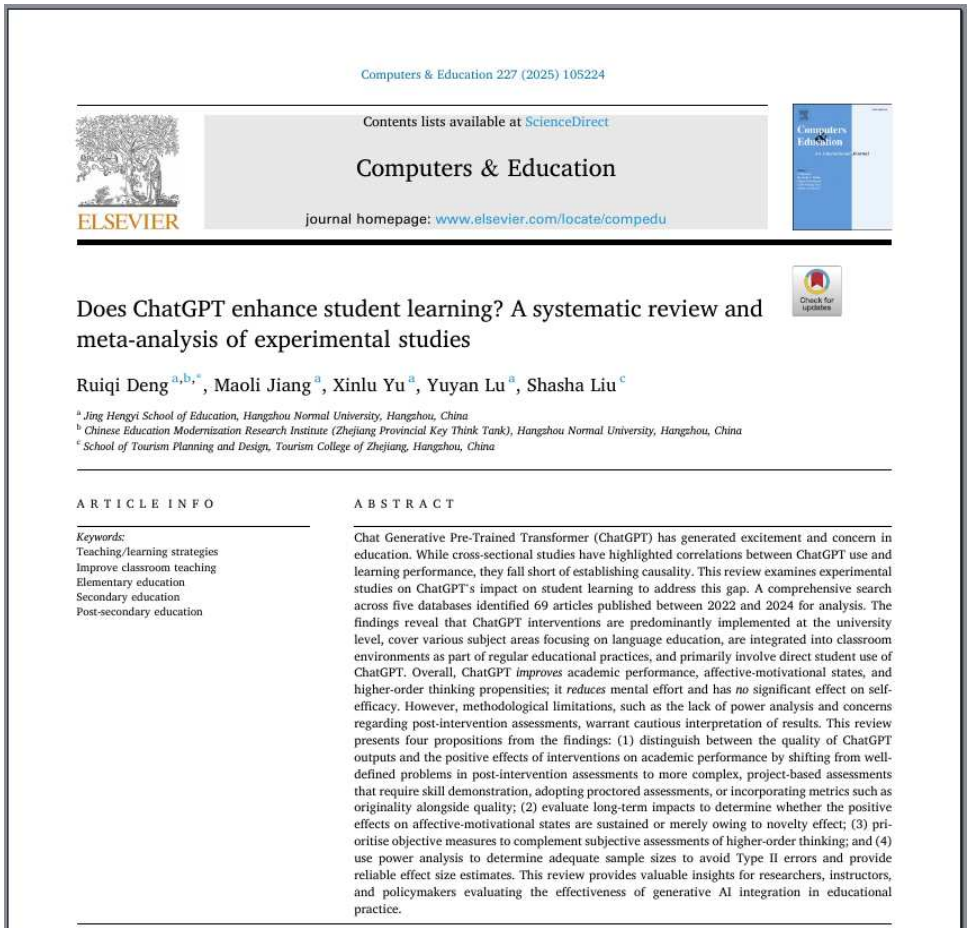
Abstract
Generative artificial intelligence (GAI) has brought new ideas for optimizing students' learning. Despite increasing attention on the effects of GAI on learning outcomes (LO), research results are inconsistent. While GAI's educational benefits are qualitatively described, there is substantial debate about its actual impact on students' LO. The study sought to quantify GAI's impact on students' LO, evalu- oderating factors: n, and knowledge ere selected from mpleted the litera- a-analytic method ur moderating fac- ts' LO ($g=0.392$), $g=0.347$), and in- al types have no howing significant

Education and Information Technologies
https://doi.org/10.1007/s10639-025-13735-x

The impact of generative artificial intelligence on students' higher order thinking: Evidence from a three-level meta-analysis

Xinxiao Nie[1,2] · Yuan Tian[1,2,3] · Mengjie Liu[1,2] · Di Wu[1,2] · Yunxiao Guo[1,2]

Received: 7 December 2024 / Accepted: 21 July 2025



Computers & Education 227 (2025) 105224

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies

Ruiqi Deng [a,b,*], Maoli Jiang [a], Xinlu Yu [a], Yuyan Lu [a], Shasha Liu [c]

[a] Jing Hengyi School of Education, Hangzhou Normal University, Hangzhou, China
[b] Chinese Education Modernization Research Institute (Zhejiang Provincial Key Think Tank), Hangzhou Normal University, Hangzhou, China
[c] School of Tourism Planning and Design, Tourism College of Zhejiang, Hangzhou, China

ARTICLE INFO

Keywords:
Teaching/learning strategies
Improve classroom teaching
Elementary education
Secondary education
Post-secondary education

ABSTRACT

Chat Generative Pre-Trained Transformer (ChatGPT) has generated excitement and concern in education. While cross-sectional studies have highlighted correlations between ChatGPT use and learning performance, they fall short of establishing causality. This review examines experimental studies on ChatGPT's impact on student learning to address this gap. A comprehensive search across five databases identified 69 articles published between 2022 and 2024 for analysis. The findings reveal that ChatGPT interventions are predominantly implemented at the university level, cover various subject areas focusing on language education, are integrated into classroom environments as part of regular educational practices, and primarily involve direct student use of ChatGPT. Overall, ChatGPT *improves* academic performance, affective-motivational states, and higher-order thinking propensities; it *reduces* mental effort and has *no* significant effect on self-efficacy. However, methodological limitations, such as the lack of power analysis and concerns regarding post-intervention assessments, warrant cautious interpretation of results. This review presents four propositions from the findings: (1) distinguish between the quality of ChatGPT outputs and the positive effects of interventions on academic performance by shifting from well-defined problems in post-intervention assessments to more complex, project-based assessments that require skill demonstration, adopting proctored assessments, or incorporating metrics such as originality alongside quality; (2) evaluate long-term impacts to determine whether the positive effects on affective-motivational states are sustained or merely owing to novelty effect; (3) prioritise objective measures to complement subjective assessments of higher-order thinking; and (4) use power analysis to determine adequate sample sizes to avoid Type II errors and provide reliable effect size estimates. This review provides valuable insights for researchers, instructors, and policymakers evaluating the effectiveness of generative AI integration in educational practice.

Μεγάλο μέγεθος διαφοράς* (g=0.71)

“… *it is crucial to highlight that most reviewed studies* ***did not explicitly state*** *whether ChatGPT was permitted during post-intervention assessments. This implies that the quality of ChatGPT’s output is not distinguished from the effect of the intervention itself*” (p. 18)

Humanities & Social Sciences Communications

REVIEW

https://doi.org/10.1057/s41599-025-04787-y OPEN

Check for updates

The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis

Jin Wang[1] & Wenxiang Fan[1,2]✉

As a new type of artificial intelligence, ChatGPT is becoming widely used in learning. However, academic consensus regarding its efficacy remains elusive. This study aimed to assess the effectiveness of ChatGPT in improving students' learning performance, learning perception, and higher-order thinking through a meta-analysis of 51 research studies published between November 2022 and February 2025. The results indicate that ChatGPT has a large positive impact on improving learning performance ($g = 0.867$) and a moderately positive impact on enhancing learning perception ($g = 0.456$) and fostering higher-order thinking ($g = 0.457$). The impact of ChatGPT on learning performance was moderated by type of course ($Q_B = 64.249$, $P < 0.001$), learning model ($Q_B = 76.220$, $P < 0.001$), and duration ($Q_B = 55.998$, $P < 0.001$); its effect on learning perception was moderated by duration ($Q_B = 19.839$, $P < 0.001$); and its influence on the development of higher-order thinking was moderated by type of course ($Q_B = 7.811$, $P < 0.05$) and the role played by ChatGPT ($Q_B = 4.872$, $P < 0.05$). This study suggests that: (1) appropriate learning scaffolds or educational frameworks (e.g., Bloom's taxonomy) should be provided when using ChatGPT to develop students' higher-order thinking; (2) the broad use of ChatGPT at various grade levels and in different types of courses should be encouraged to support diverse learning needs; (3)

Μεγάλο μέγεθος διαφοράς* (g=0.86)

Retraction Note | Open access | Published: 22 April 2026

Retraction Note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis

Jin Wang & Wenxiang Fan

Humanities and Social Sciences Communications **13**, Article number: 528 (2026) | Cite this article

The Original Article was published on 06 May 2025

Retraction Note to: *Humanities and Social Sciences Communications* https://doi.org/10.1057/s41599-025-04787-y, published online 06 May 2025

The Editor has decided to retract this paper owing to concerns regarding discrepancies in the meta-analysis. These issues ultimately undermine the confidence the Editor can place in the validity of the analysis and resulting conclusions. The authors have not responded to correspondence regarding this retraction.

Author information

Authors and Affiliations

Jing Hengyi School of Education, Hangzhou Normal University, Hangzhou, China
Jin Wang & Wenxiang Fan

Chinese Education Modernization Research Institute of Hangzhou Normal University (Zhejiang Provincial Key Think Tank), Hangzhou, China
Wenxiang Fan

*Ανάκληση άρθρου!

Διαπιστώσεις

- Η τρέχουσα εμπειρική έρευνα βρίσκεται ακόμα σε εμβρυϊκό στάδιο
- Θετική επίδραση ΠΤΝ σε διάφορες νοητικές δεξιότητες…
- ….υπάρχουν μεθοδολογικά και άλλα ζητήματα

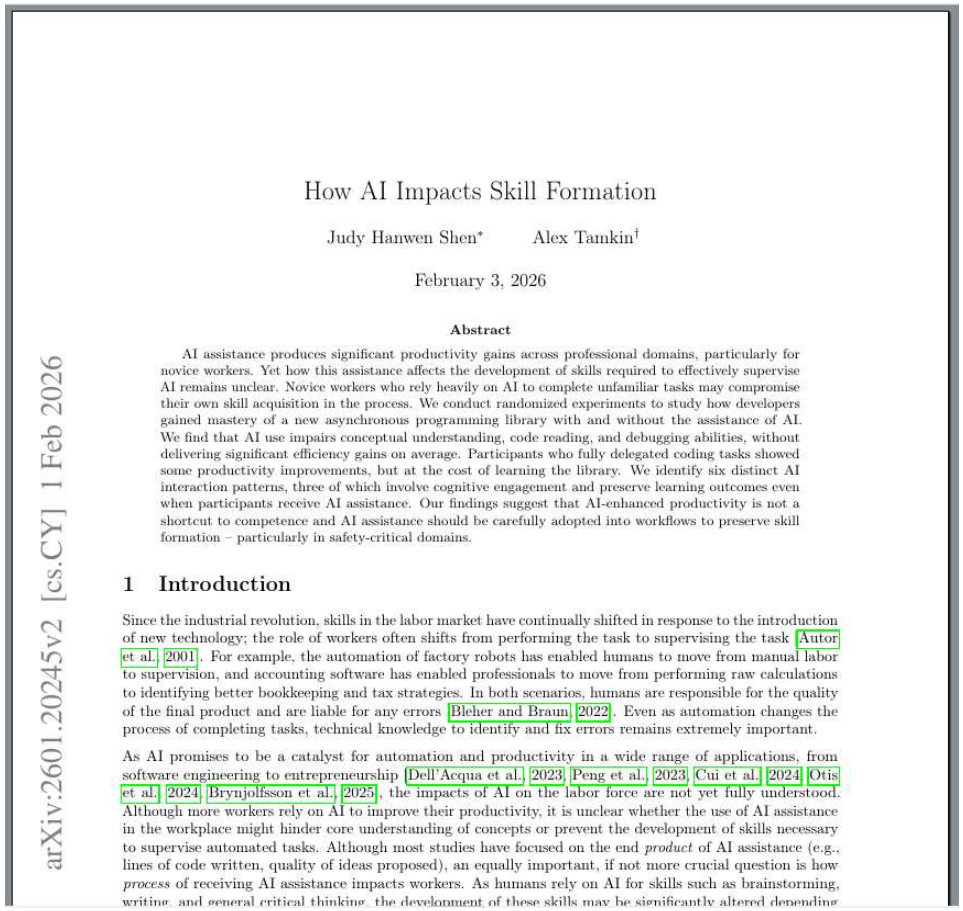
Αξιοποίηση
Θεωρητικού
Πλαισίου
Πολιτισμικής-
Ιστορικής
Θεωρίας

Ποια είναι η γνωστική επίδραση των συστημάτων ΠΤΝ στα εκτελούμενα έργα;

Επίδραση ΠΤΝ

- Ενώ βραχυπρόθεσμα η επίδοση βελτιώνεται, μακροπρόθεσμα **μειώνεται** σημαντικά, όταν δεν είναι προσβάσιμα τα συστήματα ΠΤΝ (Liu et al. 2026)
- Η χρήση συστημάτων ΠΤΝ οδηγεί σε σοβαρή **υπερεκτίμηση** της επίδοσης (Fernandes et al., 2025)

Επίδραση ΠΤΝ



arXiv:2601.20245v2 [cs.CY] 1 Feb 2026

How AI Impacts Skill Formation

Judy Hanwen Shen* Alex Tamkin†

February 3, 2026

Abstract

AI assistance produces significant productivity gains across professional domains, particularly for novice workers. Yet how this assistance affects the development of skills required to effectively supervise AI remains unclear. Novice workers who rely heavily on AI to complete unfamiliar tasks may compromise their own skill acquisition in the process. We conduct randomized experiments to study how developers gained mastery of a new asynchronous programming library with and without the assistance of AI. We find that AI use impairs conceptual understanding, code reading, and debugging abilities, without delivering significant efficiency gains on average. Participants who fully delegated coding tasks showed some productivity improvements, but at the cost of learning the library. We identify six distinct AI interaction patterns, three of which involve cognitive engagement and preserve learning outcomes even when participants receive AI assistance. Our findings suggest that AI-enhanced productivity is not a shortcut to competence and AI assistance should be carefully adopted into workflows to preserve skill formation – particularly in safety-critical domains.

1 Introduction

Since the industrial revolution, skills in the labor market have continually shifted in response to the introduction of new technology; the role of workers often shifts from performing the task to supervising the task [Autor et al., 2001]. For example, the automation of factory robots has enabled humans to move from manual labor to supervision, and accounting software has enabled professionals to move from performing raw calculations to identifying better bookkeeping and tax strategies. In both scenarios, humans are responsible for the quality of the final product and are liable for any errors [Bleher and Braun, 2022]. Even as automation changes the process of completing tasks, technical knowledge to identify and fix errors remains extremely important.

As AI promises to be a catalyst for automation and productivity in a wide range of applications, from software engineering to entrepreneurship [Dell'Acqua et al., 2023, Peng et al., 2023, Cui et al., 2024, Otis et al., 2024, Brynjolfsson et al., 2025], the impacts of AI on the labor force are not yet fully understood. Although more workers rely on AI to improve their productivity, it is unclear whether the use of AI assistance in the workplace might hinder core understanding of concepts or prevent the development of skills necessary to supervise automated tasks. Although most studies have focused on the end *product* of AI assistance (e.g., lines of code written, quality of ideas proposed), an equally important, if not more crucial question is how *process* of receiving AI assistance impacts workers. As humans rely on AI for skills such as brainstorming, writing, and general critical thinking, the development of these skills may be significantly altered depending

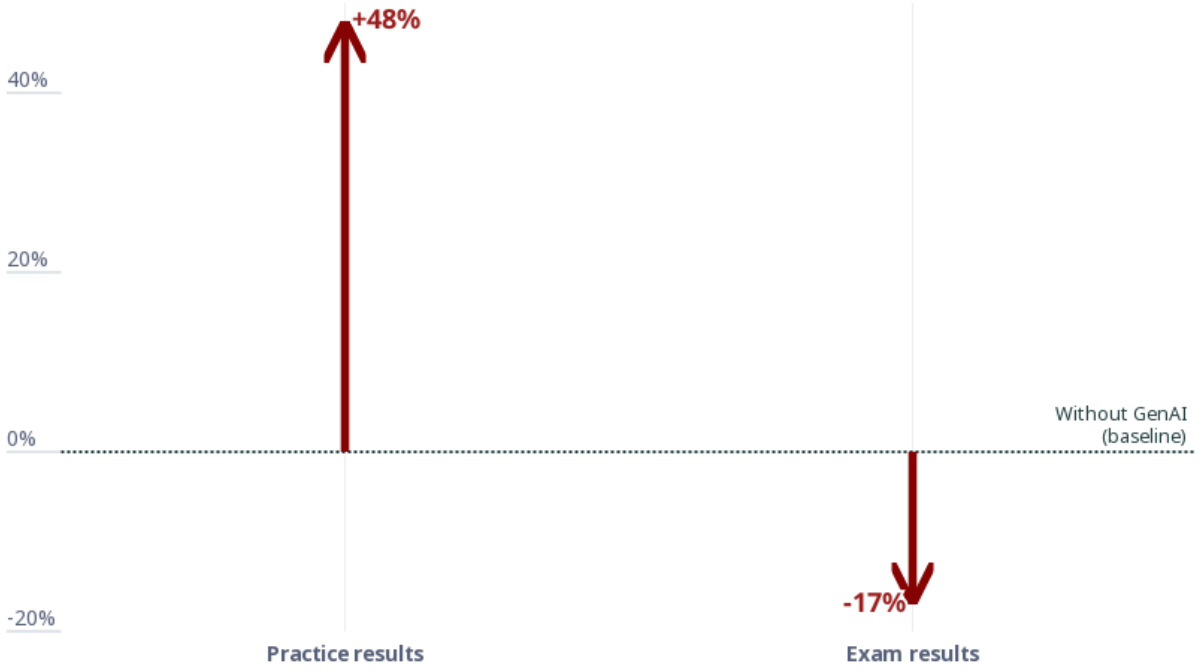
- Η χρήση συστημάτων ΠΤΝ σε πλαίσιο μάθησης (Shen et al., 2026)
  - **δυσχεραίνει** (α) την εννοιολογική κατανόηση, (β) τις ικανότητες ανάγνωσης κώδικα και (γ) εκσφαλμάτωσης
  - δεν οδηγεί σε υψηλότερα επίπεδα αποδοτικότητας

Επίδραση ΠΤΝ

Successfully performing a task with GenAI does not automatically lead to learning

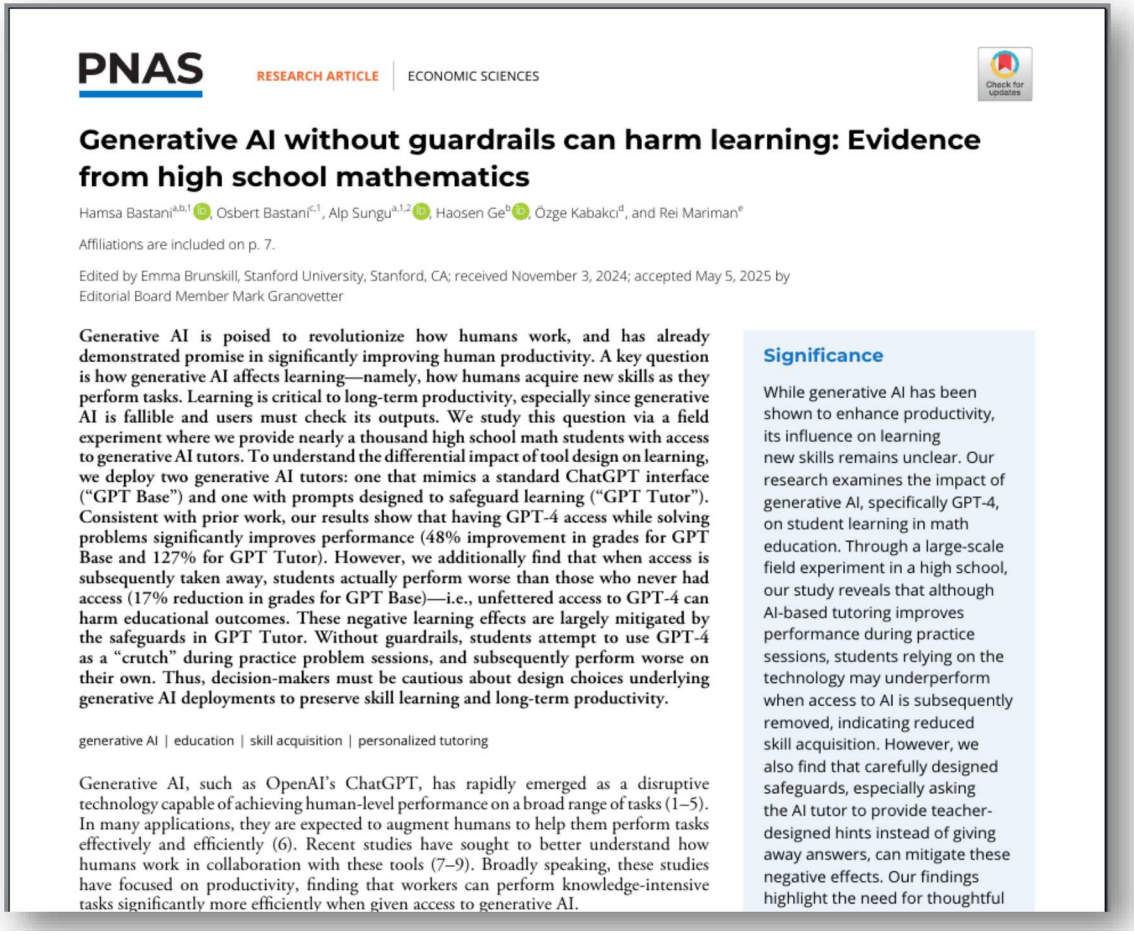
Emerging evidence suggests that while general-purpose GenAI tools can enhance students' performance on tasks, they do not necessarily lead to learning gains. Offloading cognitive tasks to general-purpose chatbots creates risks of metacognitive laziness and disengagement that may deter skill acquisition in the long run. Several studies indicate that although students with access to general-purpose GenAI tools produce higher-quality outputs than their peers, this advantage disappears – and sometimes reverses – in exams when access is removed. In contrast, educational GenAI tools designed or used with an intentional pedagogical purpose tend to show sustained improvements in learning.

Student performance when practicing maths with generic GenAI



Randomised controlled trial of high school students in Türkiye in the 2023-24 school year.
Source: OECD Digital Education Outlook 2026, Figure 1.5.

Επίδραση ΠΤΝ



PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES

Generative AI without guardrails can harm learning: Evidence from high school mathematics

Hamsa Bastani[a,b,1], Osbert Bastani[c,1], Alp Sungu[a,1,2], Haosen Ge[b], Özge Kabakcı[d], and Rei Mariman[e]

Affiliations are included on p. 7.

Edited by Emma Brunskill, Stanford University, Stanford, CA; received November 3, 2024; accepted May 5, 2025 by Editorial Board Member Mark Granovetter

Generative AI is poised to revolutionize how humans work, and has already demonstrated promise in significantly improving human productivity. A key question is how generative AI affects learning—namely, how humans acquire new skills as they perform tasks. Learning is critical to long-term productivity, especially since generative AI is fallible and users must check its outputs. We study this question via a field experiment where we provide nearly a thousand high school math students with access to generative AI tutors. To understand the differential impact of tool design on learning, we deploy two generative AI tutors: one that mimics a standard ChatGPT interface ("GPT Base") and one with prompts designed to safeguard learning ("GPT Tutor"). Consistent with prior work, our results show that having GPT-4 access while solving problems significantly improves performance (48% improvement in grades for GPT Base and 127% for GPT Tutor). However, we additionally find that when access is subsequently taken away, students actually perform worse than those who never had access (17% reduction in grades for GPT Base)—i.e., unfettered access to GPT-4 can harm educational outcomes. These negative learning effects are largely mitigated by the safeguards in GPT Tutor. Without guardrails, students attempt to use GPT-4 as a "crutch" during practice problem sessions, and subsequently perform worse on their own. Thus, decision-makers must be cautious about design choices underlying generative AI deployments to preserve skill learning and long-term productivity.

generative AI | education | skill acquisition | personalized tutoring

Generative AI, such as OpenAI's ChatGPT, has rapidly emerged as a disruptive technology capable of achieving human-level performance on a broad range of tasks (1–5). In many applications, they are expected to augment humans to help them perform tasks effectively and efficiently (6). Recent studies have sought to better understand how humans work in collaboration with these tools (7–9). Broadly speaking, these studies have focused on productivity, finding that workers can perform knowledge-intensive tasks significantly more efficiently when given access to generative AI.

Significance

While generative AI has been shown to enhance productivity, its influence on learning new skills remains unclear. Our research examines the impact of generative AI, specifically GPT-4, on student learning in math education. Through a large-scale field experiment in a high school, our study reveals that although AI-based tutoring improves performance during practice sessions, students relying on the technology may underperform when access to AI is subsequently removed, indicating reduced skill acquisition. However, we also find that carefully designed safeguards, especially asking the AI tutor to provide teacher-designed hints instead of giving away answers, can mitigate these negative effects. Our findings highlight the need for thoughtful

Η πρόσβαση σε ΠΤΝ κατά την επίλυση προβλημάτων **βελτιώνει** σημαντικά την επίδοση (48%-127%)

Η μετέπειτα επίλυση προβλημάτων χωρίς ΠΤΝ επιφέρει **μείωση** της επίδοσης (17%)

Επίδραση ΠΤΝ

- Η **συνέργεια** ανθρώπων και συστημάτων ΠΤΝ δεν είναι δεδομένη (Vaccaro et al., 2024)
- Οι μαθητές τείνουν περισσότερο να **βασίζονται** αντί να μαθαίνουν όταν χρησιμοποιούν συστήματα ΠΤΝ (Darvishi et al., 2024)
- Διαπιστώνεται **απόκλιση** μεταξύ της προσλαμβανόμενης ωφέλειας των συστημάτων ΠΤΝ και της αξιοπιστίας τους (Thesen & Park, 2025)

΄Εμφαση σε Βασικές Γνωστικές και Μεταγνωστικές Δεξιότητες

Μηχανική προτροπών

- Υπάρχει η διαδεδομένη πεποίθηση ότι η εκμάθηση της διαδικασίας μηχανικής προτροπών (prompt engineering)
  - θα αποτελέσει την κύρια δεξιότητα αλφαβητισμού (core literacy skill)
  - θα είναι απαραίτητη για την αποτελεσματική καθοδήγηση των συστημάτων ΠΤΝ

Μηχανική προτροπών

- Αν δεν γνωρίζω τις έννοιες, αδυνατώ
  - να **διατυπώσω** κατάλληλη προτροπή για να κατευθύνω το μοντέλο
  - να **αξιολογήσω** το αποτέλεσμα της δημιουργίας του μοντέλου

pro(mpt)stitutes

Educational Psychology Review (2026) 38:45
https://doi.org/10.1007/s10648-026-10118-7

INTERVENTION STUDY

Think First, ChatGPT Later: Guiding Human–AI Collaboration for Learning Gains in Independent Human Creativity

Sarah Shi Hui Wong[1] · Sophia Xuefei Qiu[2]

Received: 1 August 2025 / Accepted: 8 January 2026

Abstract

Generative artificial intelligence (AI) tools such as ChatGPT can boost creative performance, but do these boosts translate into learning gains? This study examined whether the benefits of ChatGPT for creativity persist even when its assistance is removed, and how people can effectively use ChatGPT to enhance their learning and independent creativity. University students (N = 196) solved a creative product improvement task either independently (human-only group) or using ChatGPT freely (general-AI group) or using ChatGPT in a guided way (regulated-AI group). Specifically, the regulated-AI group used a novel "think first, ChatGPT later" approach—they first generated their own ideas, then collaborated with ChatGPT to improve, develop, and evaluate them. Thereafter, all groups independently solved a creative product invention task. On the first task, the general-AI group produced more creative solutions than the human-only and regulated-AI groups. But without ChatGPT assistance on the second task, the general-AI group's creativity declined to levels comparable to the human-only group. In striking contrast, despite a lack of performance gains on the first task, the regulated-AI group outperformed both the human-only and general-AI groups in independent creativity on the second task. Process analyses revealed that the general-AI group most often simply dictated ChatGPT to directly generate the solutions. Conversely, the regulated-AI group more frequently collaborated with ChatGPT to improve their self-generated ideas, in turn mediating their later advantage over the general-AI group in independent

Δημιουργικότητα

- Η χρήση ΠΤΝ οδηγεί σε πιο δημιουργικές προσεγγίσεις
- Η θετική επίδραση **εξανεμίζεται**, όταν αφαιρεθεί η πρόσβαση σε ΠΤΝ
- Η θετική επίδραση **διατηρείται**, όταν έχει επιχειρηθεί πρώτα η δημιουργικότητα χωρίς ΠΤΝ

Professional Software Developers Don't Vibe, They Control: AI Agent Use for Coding in 2025

RUANQIANQIAN (LISA) HUANG*, UC San Diego, USA
AVERY REYNA*, Independent, USA
SORIN LERNER, Cornell University, USA
HAIJUN XIA, UC San Diego, USA
BRIAN HEMPEL†, UC San Diego, USA

The rise of AI *agents* is transforming how software can be built. The promise of agents is that developers might write code quicker, delegate multiple tasks to different agents, and even write a full piece of software purely out of natural language. In reality, what roles agents play in professional software development remains in question. This paper investigates how *experienced* developers use agents in building software, including their motivations, strategies, task suitability, and sentiments. Through field observations (N=13) and qualitative surveys (N=99), we find that while experienced developers value agents as a productivity boost, they retain their agency in software design and implementation out of insistence on fundamental software quality attributes, employing strategies for controlling agent behavior leveraging their expertise. In addition, experienced developers feel overall positive about incorporating agents into software development given their confidence in complementing the agents' limitations. Our results shed light on the value of software development best practices in effective use of agents, suggest the kinds of tasks for which agents may be suitable, and point towards future opportunities for better agentic interfaces and agentic use guidelines.

1 Introduction

> *I've been a software developer and data analyst for 20 years and there is no way I'll EVER go back to coding by hand. That ship has sailed and good riddance to it.*
> - Developer in Our Survey (S28)

AI is rapidly changing the practice of programming. Already, about half of professional software developers are using AI tools daily [30]. Large language models (LLMs) are particularly good at writing code, and are becoming more skillful every year. Originally, in 2021, LLMs only provided coding assistance as super-charged autocomplete [12]. But more recently, their capabilities have advanced to accessing, modifying, and testing whole codebases in autonomous, step-by-step actions—we are now in the *agentic* coding era. There are many open questions about how capable these agents are and how best to use them. Anecdotally, we sometimes hear from people that they tried it once and it didn't work out so well. But this contrasts with what one reads on social media: some online users claim to use dozens of agents at once to autonomously construct massive software (*e.g.*, [32, 41]), a claim so intriguing but potentially incredulous that it is parodied [16]. *What is really happening?*

Human studies of agentic coding are emerging but still sparse. A notable randomized trial found that experienced open source maintainers were actually slowed down by 19% when allowed to use AI [4], and an agentic system deployed in an issue tracker saw only 8% of its invocations resulting in complete success (a merged pull request) [31]. These results suggest that perhaps agentic AI is not

Οι προχωρημένοι μηχανικοί λογισμικού **ελέγχουν** τους πράκτορες αντί να προγραμματίζουν με βάση το vibe

Πολιτισμική-Ιστορική Θεωρία

- Παρέχει τα εννοιολογικά εργαλεία για αναπαράσταση του ζητήματος
- Η ανθρώπινη νόηση συγκροτείται στη βάση εργαλείων
- Τα συστήματα ΠΤΝ μπορούν να προσεγγιστούν ως **διαμεσολαβητικά εργαλεία**

Διαμεσολάβηση

- Επίδραση εργαλείου στις εμπλεκόμενες νοητικές λειτουργίες
  - Κατάργηση
  - Εισαγωγή νέων

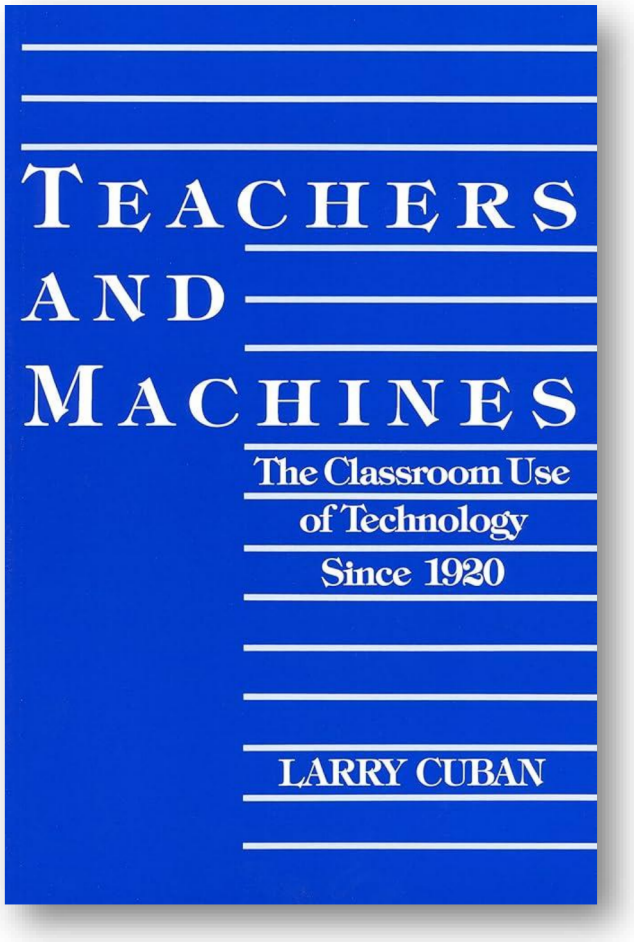
Μετασχηματισμός

- Η χρήση νοητικών εργαλείων **μετασχηματίζει** το εκτελούμενο έργο
- Μέρος (ή και το σύνολο) του απαιτούμενου νοητικού έργου υλοποιείται από το εργαλείο
- Ο χρήστης κάνει **λιγότερα** καθώς το εργαλείο κάνει περισσότερα

Προηγούμενη Εμπειρία με Επαναστατικές Τεχνολογίες

Μαθαίνουμε από την ιστορία

- Καμιά προγενέστερη τεχνολογία δεν έφερε την πολυαναμενόμενη "επανάσταση" στην εκπαίδευση
  - Ραδιόφωνο
  - Φίλμ
  - Τηλεόραση
  - Υπολογιστής
  - Διαδίκτυο
  - Ψηφιακά Παιχνίδια



aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

The Sisyphean Cycle of Technology Panics

Perspectives on Psychological Science
1–15

DOI: 10.1177/1745691620919572
www.psychologicalscience.org/PPS


Amy Orben
MRC Cognition and Brain Sciences Unit and Emmanuel College, University of Cambridge

Abstract

Widespread concerns about new technologies—whether they be novels, radios, or smartphones—are repeatedly found throughout history. Although tales of past panics are often met with amusement today, current concerns routinely engender large research investments and policy debate. What we learn from studying past technological panics, however, is that these investments are often inefficient and ineffective. What causes technological panics to repeatedly reincarnate? And why does research routinely fail to address them? To answer such questions, I examined the network of political, population, and academic factors driving the *Sisyphean cycle of technology panics*. In this cycle, psychologists are encouraged to spend time investigating new technologies, and how they affect children and young people, to calm a worried population. Their endeavor, however, is rendered ineffective because of the lack of a theoretical baseline; researchers cannot build on what has been learned researching past technologies of concern. Thus, academic study seemingly restarts for each new technology of interest, which slows down the policy interventions necessary to ensure technologies are benefiting society. In this article, I highlight how the Sisyphean cycle of technology panics stymies psychology's positive role in steering technological change and the pervasive need for improved research and policy approaches to new technologies.

Keywords

digital-technology use, social media, screen time, well-being, adolescents

In 1941, Mary Preston published "Children's Reactions to Movie Horrors and Radio Crime" in *The Journal of Pediatrics*. The American pediatrician had studied hundreds of 6- to 16-year-old children and concluded that more than half were severely addicted to radio and movie crime dramas, having given themselves "over to a habit-forming practice very difficult to overcome, no matter how the aftereffects are dreaded" (pp. 147–148). Most strikingly, Preston observed that many children consumed these dramas "much as a chronic alcoholic does drink" (p. 167). Preston therefore voiced severe concerns about the children's health and future outcomes: Children who consumed more radio crime or movie dramas were more nervous and fearful and suffered from worse general health and more disturbed eating and sleep.

To truly understand these claims, one needs to consider Preston's work in the context of her time. The decade preceding her work saw both broad social and technological changes; the explosive growth in popu-

by 1940 (Dennis, 1998). In 1936, about nine in 10 New York households owned a household radio, and children in these homes spent between 1 and 3 hr a day listening to these devices (Dennis, 1998). This rapid rise in popularity sparked concerns not limited to Mary Preston's article. A *New York Times* piece considered whether listening to the radio too much would harm children and lead to illnesses because the body needed "repose" and could not "be kept up at the jazz rate forever" (Ferrari, as cited in Dennis, 1998). Concerns voiced by the Director of the Child Study Association of America noted how radio was worse than any media that came before because "no locks will keep this intruder out, nor can parents shift their children away from it" (Gruenberg, 1935). This view was mirrored in a parenting magazine published at the time:

> Here is a device, whose voice is everywhere. . . . We may question the quality of its offering for our children, we may approve or deplore its entertainments

Πρόταση

- Αλφαβητισμός σε ΠΤΝ
- Ρυθμιστικό πλαίσιο σε εθνικό επίπεδο
- Έμφαση σε βασικές γνωστικές και μεταγνωστικές δεξιότητες
- Χρήση 'ανοικτών' συστημάτων ΠΤΝ
- Υιοθέτηση επιτυχημένων πρακτικών από άλλους τομείς (π.χ. σκάκι)

Σκέψη

"*Δεν υπάρχει βασιλικός δρόμος για τη Γεωμετρία*"
(Ευκλείδης)

"*Learning is not a spectator sport*" (Koedinger, 2015)

"*Effort is the algorithm*" (Muller, 2025)

"*Whoever does the work does the learning*" (Holt, 2024)

Karasavvidis, I. & Dafermos, M. (in press). A cultural-historical perspective on the intelligent and dialogic capabilities of generative artificial intelligence systems. In L. Smirnova, P. Birtill, & L. Waddington (Eds.). *Sociocultural Perspectives on AI-Enhanced Education*. Springer.