When Copilot Becomes *Autopilot*

Generative AI's Critical Risk to Knowledge Work and a Critical Solution

2024

Advait Sarkar

Microsoft Research, University of Cambridge, University College London

https://advait.org advait@microsoft.com

Collaborators: Xiaotong (Tone) Xu, Ian Drosos, Carina Negreanu, Christian Poelitz, Andy Gordon, Nick Wilson, Neil Toronto, Sean Rintel, Lev Tankelevitch, Richard Banks

I: The Critical Risk

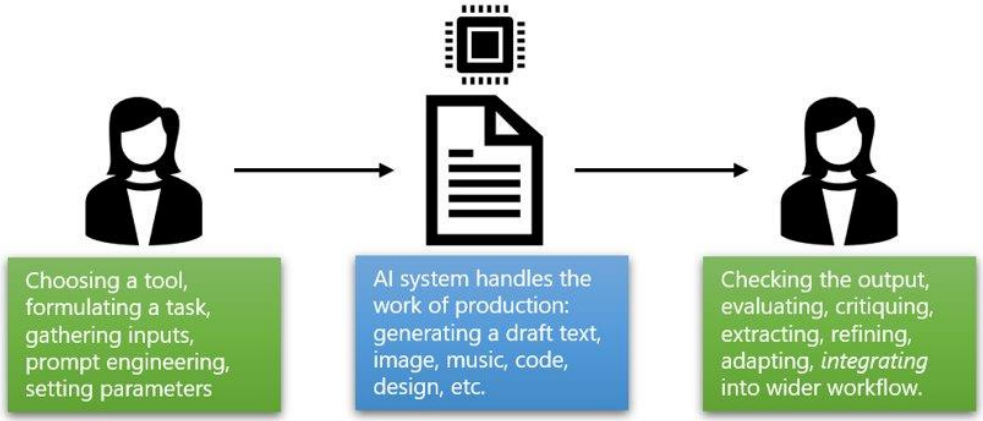
Sarkar, A. (2023). Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work.*

AI shifts knowledge work from material production to *critical integration*

As the labour of *material production* decreases…

… critical integration becomes a new form of creative labour:

Deciding *where and how* to use AI in a workflow	*Critically assessing* AI output and adjusting it to fit the workflow

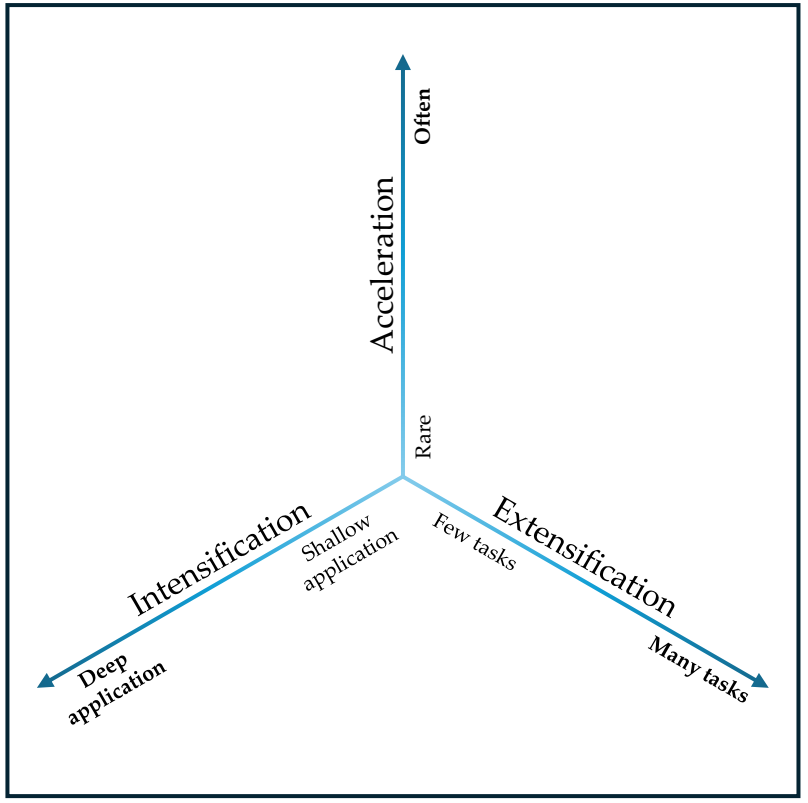


The critical integration "sandwich": when **AI handles production**, **human critical thinking is applied at either end of the process** to complete knowledge workflows.

Sarkar, A. (2023). Will Code Remain a Relevant User Interface for End-User Programming with Generative AI Models?. In *Proceedings of the 2023 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*.

A *Generative Shift*

- A radical *widening in scope and capability* of automation due to generative AI.
- Generative shift comprises:
  - *Intensification*: AI applied more deeply to current tasks
  - *Extensification*: AI applied more broadly to new tasks
  - *Acceleration*: AI applied more frequently across all tasks



The generative shift: three dimensions of AI usage growth in knowledge work.

Sarkar, A. (2023). Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work.*

Sarkar, A. (2024). Intention Is All You Need. *Under review, draft available on request.*

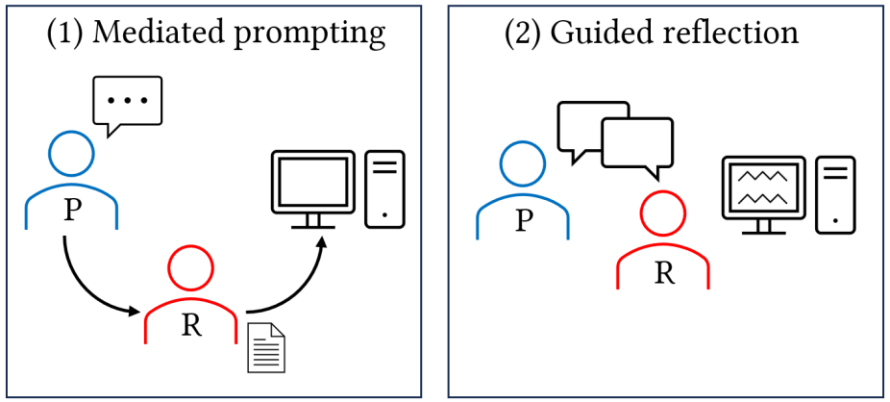
Mechanised convergence: automation makes all work look the same

- A study (n=293) of participants writing short stories with varying degrees of AI assistance found that exposure to GenAI "ideas" leads to a *reduced diversity* of content (Doshi & Hauser, 2023). Participants exposed to even a single GenAI suggestion produce stories similar to the average of the other stories in the same experimental condition.
- A study (n=758) of strategy consultants at BCG examined the effects of ChatGPT use on a set of consultancy tasks (Dell'Acqua et al., 2023). The majority of participants with access to ChatGPT retain a very high amount of its reponse – typically around 90% – in their submitted work. Participants without access to ChatGPT produce ideas with more conceptual variation than those with access, showing that usage of ChatGPT *reduces the range of ideas* generated. The variation across responses produced by ChatGPT is smaller than what human participants produce on their own.
- Large language models have a "homogenization effect" on creative ideation (Anderson et al., 2024). In a creative ideation task, participants produce *less semantically distinct* ideas when using ChatGPT. Moreover, participants feel less responsible for ideas produced with ChatGPT assistance.
- A study (n=115) finds that conversational search built on GenAI *increases selective exposure* compared to conventional search (Sharma et al., 2024). Users engage in more biased information querying with conversational search, and the bias is exacerbated when the model is itself "opinionated" to reinforce the user's views. The authors call this a "generative echo chamber".
- Similarly, a study (n=1506) of co-writing with GenAI found that using an "opinionated" language model affects the opinions expressed in participants' writing and moreover, actually *shifts their opinions*, as measured in a subsequent attitude survey (Jakesch et al., 2023)

Drosos I. and Sarkar, A., et al. (2024, *forthcoming*). "It's like a rubber duck that talks back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. In *Proc. CHIWORK*.

Sarkar, A., et al. (2023). Participatory prompting: a user-centric research method for eliciting AI assistance opportunities in knowledge workflows. In *Proc. Psychology of Programming Interest Group*.

Participatory prompting study: data analysis with Bing Chat



The participatory prompting method: participants have a researcher-mediated interaction with GenAI, incorporating guided reflection on each conversational "turn".

Participants (n=15) completed data analysis tasks "in conversation" with Bing Chat (now Microsoft Copilot for Web)

Participants struggled with critical integration:

- Barriers to intent formation and query formulation
- Inadequate support for evaluating large volume of output and focusing attention
- Attention further divided between verification of intent interpretation, correctness, and quality
- Struggled with complex, qualitative decisions, often beyond their expertise

⇒ strong tendency for accepting and "*satisficing*"

II: The Critical Solution

Sarkar, A. (2024, *forthcoming*). AI Should Challenge, Not Obey. In *Communications of the ACM*.

AI as critic

	AI as assistant	AI as critic
🎯 Aim	Get the job done	**Figure out the job**
🎁 Benefit	Speed and efficiency	**Higher quality work**
🧩 Challenge	Error and hallucination	**Critical thinking**

AI as critic

❌ Not for auto-fixable errors

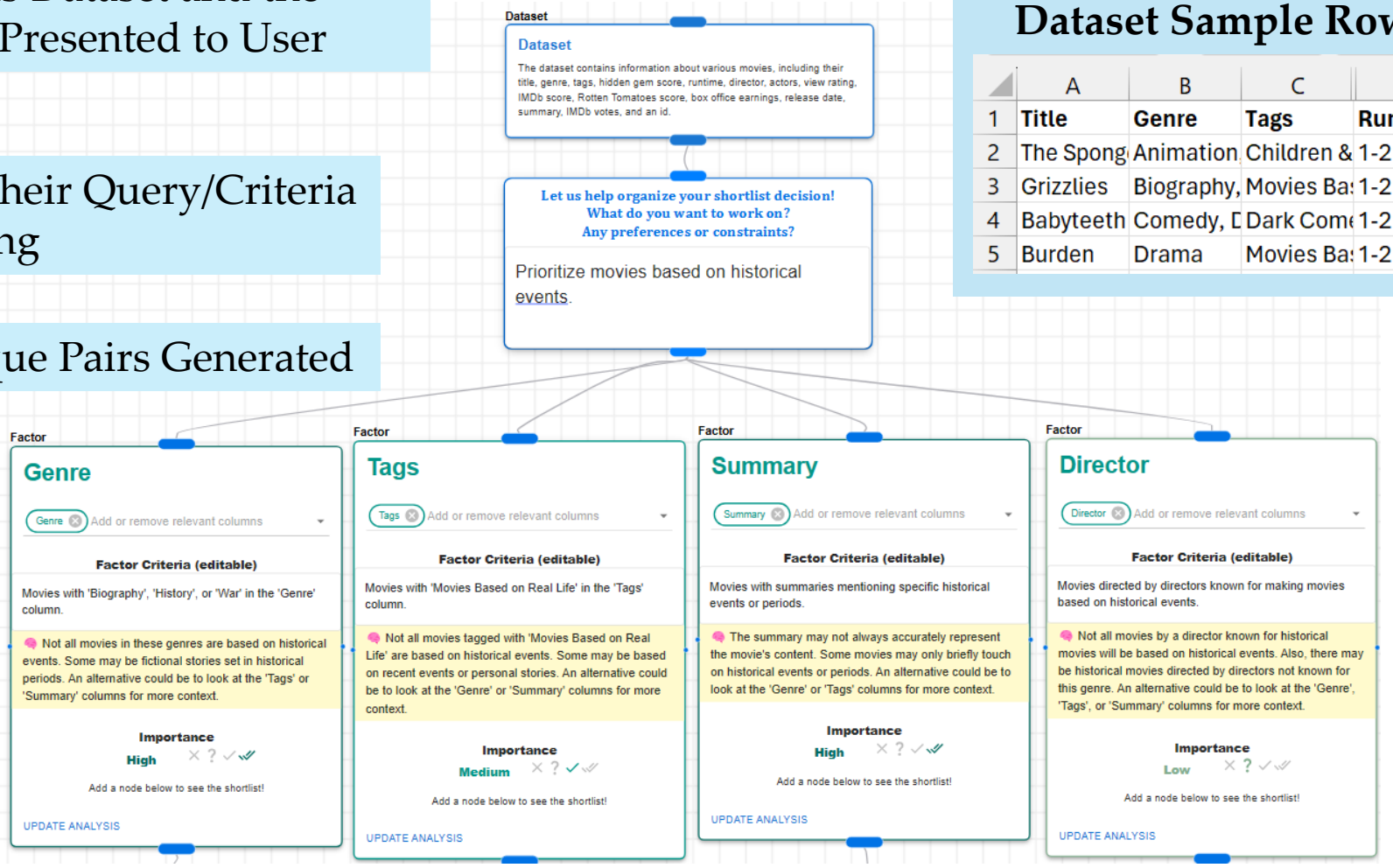
✅ Support the user in evaluating and updating output

Sarkar, A. et al., (2024, *forthcoming*). When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution. In *Proceedings of the Annual Conference of the European Spreadsheet Risks Interest Group*.

Slide credit: Ian Drosos

Critical Shortlisting Prototype

1. User Uploads Dataset and the Summary is Presented to User
2. User enters their Query/Criteria for shortlisting
3. Factor-Critique Pairs Generated



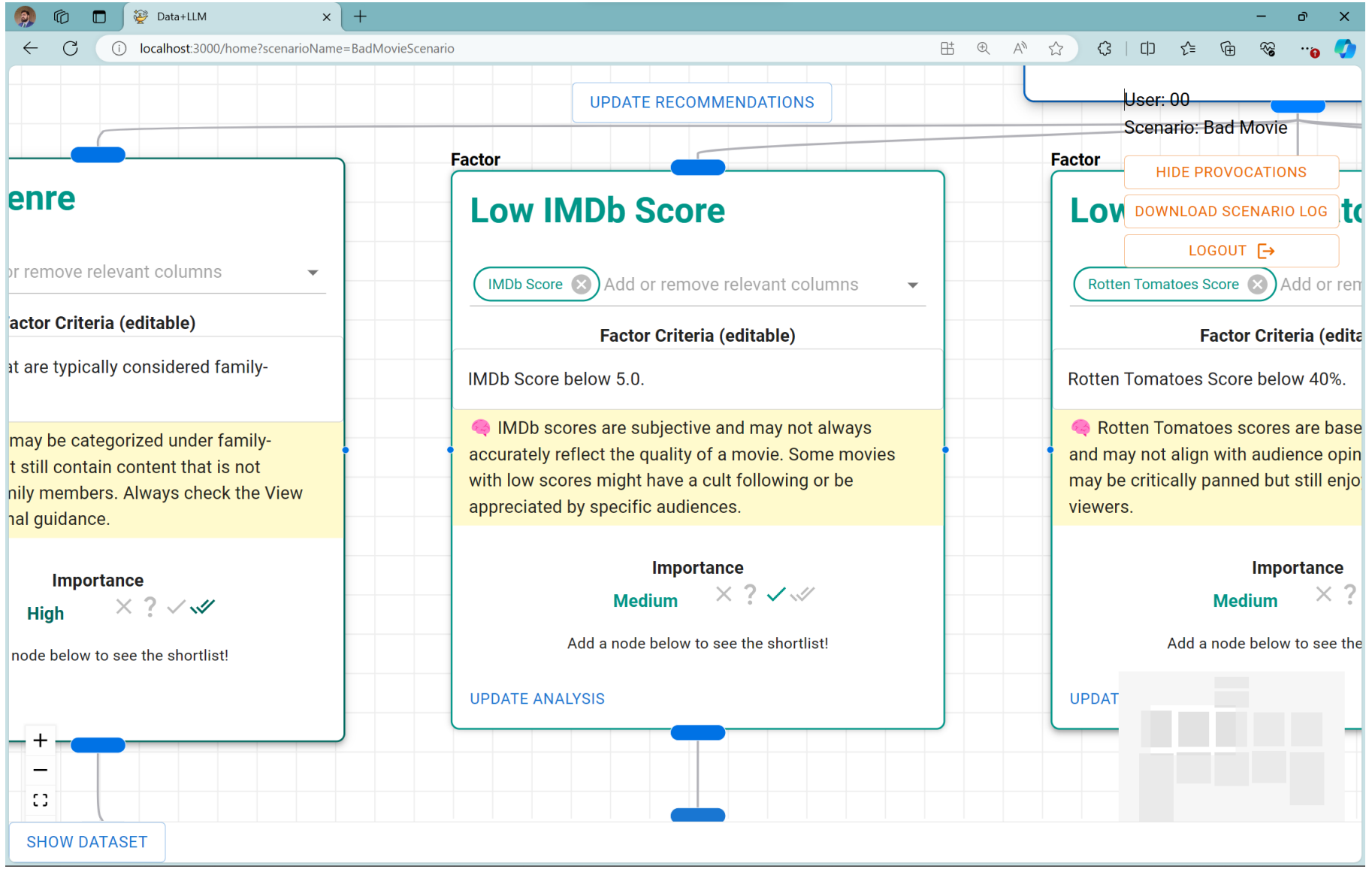
Dataset Sample Rows: Netflix Movies

	A	B	C	E	M	N
1	**Title**	**Genre**	**Tags**	**Runtime**	**Summary**	**IMDb Votes**
2	The Spong	Animation	Children &	1-2 hour	When his bes	10485
3	Grizzlies	Biography,	Movies Ba:	1-2 hour	A group of dis	1314
4	Babyteeth	Comedy, D	Dark Come	1-2 hour	Seriously ill t	9757
5	Burden	Drama	Movies Ba:	1-2 hour	A young man	1814

Sarkar, A. et al., (2024, *forthcoming*). When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution. In *Proceedings of the Annual Conference of the European Spreadsheet Risks Interest Group*.

Slide credit: Ian Drosos

Critical Shortlisting Prototype

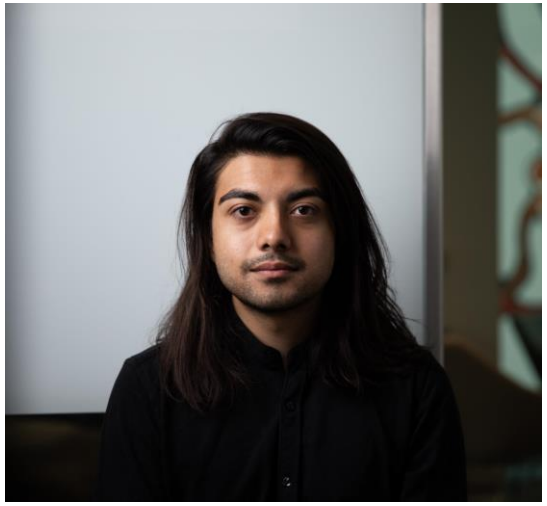


Generative AI *shifts knowledge work* from material production to critical integration.

But the tendency for *mechanised convergence* shows a deterioration of critical thinking.

In-context *provocations*, grounded in the theory of critical thinking, help.

When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution



Advait Sarkar

🌐 https://advait.org

🔗 https://www.linkedin.com/in/advaitsarkar/

🐥 https://x.com/AdvaitSarkar/

advait@microsoft.com

Microsoft · University of Cambridge · UCL

Collaborators: Xiaotong (Tone) Xu, Ian Drosos, Carina Negreanu, Christian Poelitz, Andy Gordon, Nick Wilson, Neil Toronto, Sean Rintel, Lev Tankelevitch, Richard Banks