Microsoft

Microsoft New Future of Work Report 2023

A summary of recent research from Microsoft and around the world that can help us create a new and better future of work with AI.

Editors and Authors

- **Editors**: Jenna Butler (Principal Applied Research Scientist), Sonia Jaffe (Principal Researcher), Nancy Baym (Senior Principal Research Manager), Mary Czerwinski (Partner Research Manager), Shamsi Iqbal (Principal Applied & Data Scientist), Kate Nowak (Principal Applied Scientist), Sean Rintel (Senior Principal Researcher), Abigail Sellen (VP Distinguished Scientist), Mihaela Vorvoreanu (Director Aether UX Research & EDU), Brent Hecht (Partner Director of Applied Science), and Jaime Teevan (Chief Scientist and Technical Fellow)
- **Authors:** *Najeeb Abdulhamid, Judith Amores, Reid Andersen, Kagonya Awori, Maxamed Axmed, danah boyd, James Brand, Georg Buscher, Dean Carignan, Martin Chan, Adam Coleman, Scott Counts, Madeleine Daepp, Adam Fourney, Dan Goldstein, Andy Gordon, Aaron Halfaker, Javier Hernandez, Jake Hofman, Jenny Lay-Flurrie, Vera Liao, Siân Lindley, Sathish Manivannan, Charlton Mcilwain, Subigya Nepal, Jennifer Neville, Stephanie Nyairo, Jacki O'Neill, Victor Poznanski, Gonzalo Ramos, Nagu Rangan, Lacey Rosedale, David Rothschild, Tara Safavi, Advait Sarkar, Ava Scott, Chirag Shah, Neha Shah, Teny Shapiro, Ryland Shaw, Auste Simkute, Jina Suh, Siddharth Suri, Ioana Tanase, Lev Tankelevitch, Mengting Wan, Ryen White, Longqi Yang*

Referencing this report:

- On social media, please include the report URL (https://aka.ms/nfw2023).
- In academic publications, please cite as: *Butler, J., Jaffe, S., Baym, N., Czerwinski, M., Iqbal, S., Nowak, K., Rintel, R., Sellen, A., Vorvoreanu, M., Hecht, B., and Teevan, J. (Eds.). Microsoft New Future of Work Report 2023. Microsoft Research Tech Report MSR-TR-2023-34 (https://aka.ms/nfw2023), 2023.*

Welcome to the 2023 Microsoft New Future of Work Report!

In the past three years, there have been not one but two generational shifts in how work gets done, both of which were only possible because of decades of research and development. The first shift occurred when COVID made us realize how powerful remote and hybrid work technologies had become, as well as how much science was available to guide us in how to (and how not to) use these technologies. The second arrived this year, as it became clear that, at long last, generative AI had advanced to the point where it could be valuable to huge swaths of the work people do every day.

We began the New Future of Work Report series in 2021, at the height of the shift to remote work. The goal of that report was to provide a synthesis of new – and newly relevant – research to anyone interested in reimagining work for the better as a decades-old approach to work was challenged. The second New Future of Work Report, published in 2022, focused on hybrid work and what research could teach us about intentionally re-introducing co-location into people's work practices. This year's edition, the third in the series, continues with the same goal, but centers on research related to integrating LLMs into work.

Throughout 2023, AI and the future of work have frequently been on the metaphorical – and often literal – front page around the world. There have been many excellent articles about the ways in which work may change as LLMs are increasingly integrated into our lives. As such, in this year's report we focus specifically on areas that we think deserve additional attention or where there is research that has been done at Microsoft that offers a unique perspective. This is a report that should be read as a complement to the existing literature, rather than as a synthesis of all of it.

This is a rare time, one in which research will play a particularly important role in defining what the future of work looks like. At this special moment, scientists can't just be passive observers of what is happening. Rather, we have the responsibility to shape work for the better. We hope this report can help our colleagues around world make progress towards this goal.

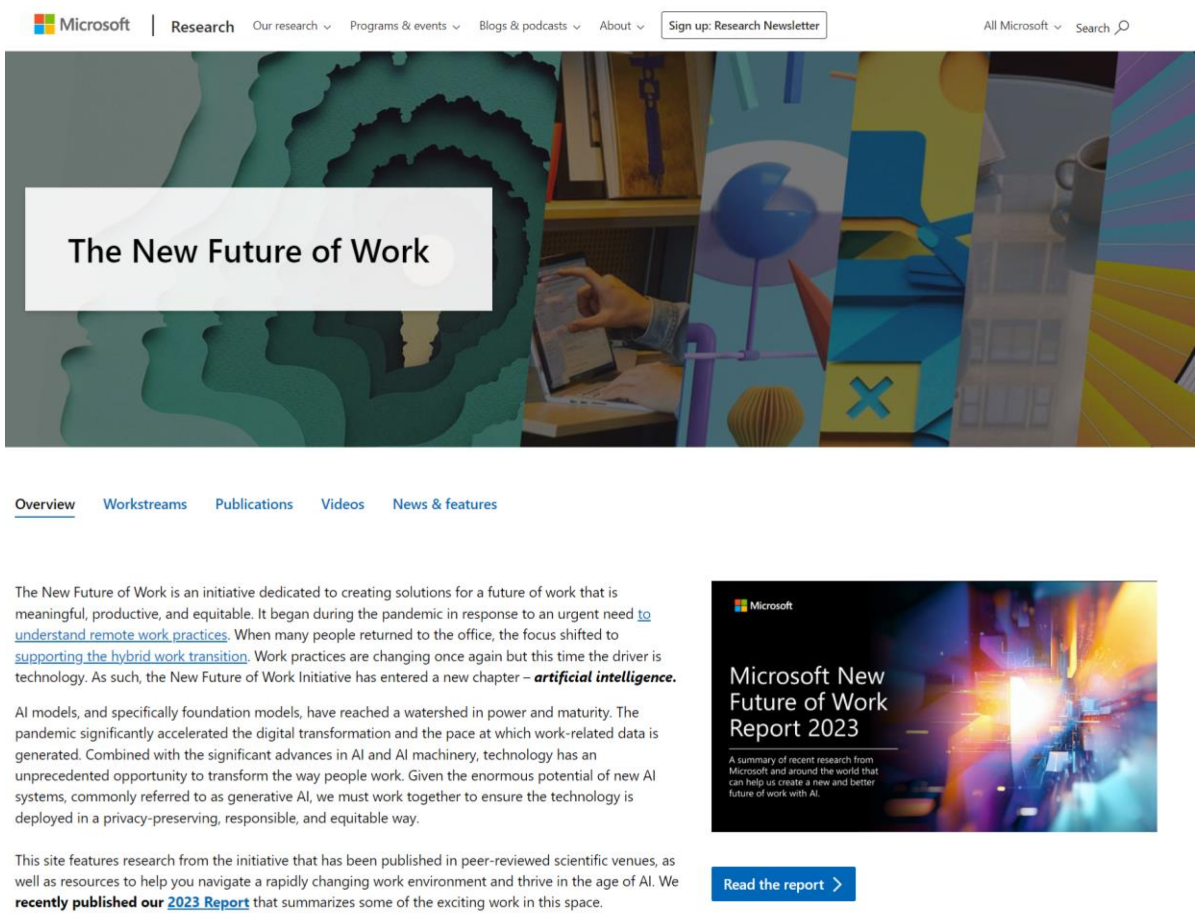
- Jaime Teevan, Chief Scientist and Technical Fellow

This report emerges from Microsoft's New Future of Work initiative

Microsoft has helped shape information work since its founding. However, a confluence of recent circumstances – remote work, hybrid work, LLMs – have created an unprecedented opportunity for the company to reimagine how AI and other digital technologies can make work better for everyone.

Since its inception, the *New Future of Work* (NFW) initiative has brought together researchers from a broad range of organizations and disciplines across Microsoft to focus on the most important technologies shaping how people work. The initiative is working to create the new future of work – one that is equitable, inclusive, meaningful, and productive – instead of predicting or waiting for it. It does this by conducting primary research and synthesizing existing research to share with the research community. This report is one of the many public resources it has produced.

The reader can find the New Future of Work initiative's many other research papers, practical guides, reports and whitepapers at the initiative's website: https://aka.ms/nfw.



https://aka.ms/nfw

Report overview

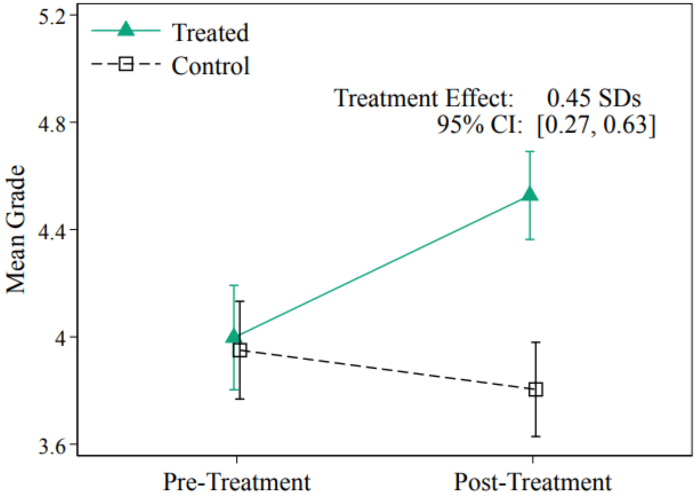
This report provides insight into AI and work practices. In it you will find content related to:

- **LLMs for Information Work:** How do LLMs affect the speed and quality of common information work tasks? LLMs can boost productivity for information workers, but they also require careful evaluation and adaptation.
- **LLMs for Critical Thinking:** How can LLMs help us break down and build up complex tasks? LLMs can help us tackle complex tasks by provoking critical thinking, enabling microproductivity, and shifting the balance of skills.
- **Human-AI Collaboration:** How can we collaborate effectively with LLMs? Effective collaboration with LLMs depends on how we prompt, complement, rely on, and audit them.
- **LLMs for Complex and Creative Tasks:** How can LLMs tackle tasks that go beyond simple information retrieval or generation? LLMs can support complex and creative tasks by, for instance, enhancing metacognition.
- **Domain-Specific Applications of LLMs:** How are LLMs being used and affecting different domains of work? We focus specifically on software engineering, medicine, social science, and education.
- **LLMs for Team Collaboration and Communication:** How can LLMs help teams work and communicate better? LLMs can help teams improve interaction, coordination, and workflows by providing real-time, retrospective feedback and leveraging holistic frameworks.
- **Knowledge Management and Organizational Changes:** How is AI changing the nature and distribution of knowledge in organizations? LLMs might, for instance, finally eliminate knowledge silos in large companies.
- **Implications for Future Work and Society:** What implications will AI have for the future of work and society? We can shape AI’s impact by addressing adoption disparities, fostering innovation, leading like scientists, and remembering that the future of work is in our control.

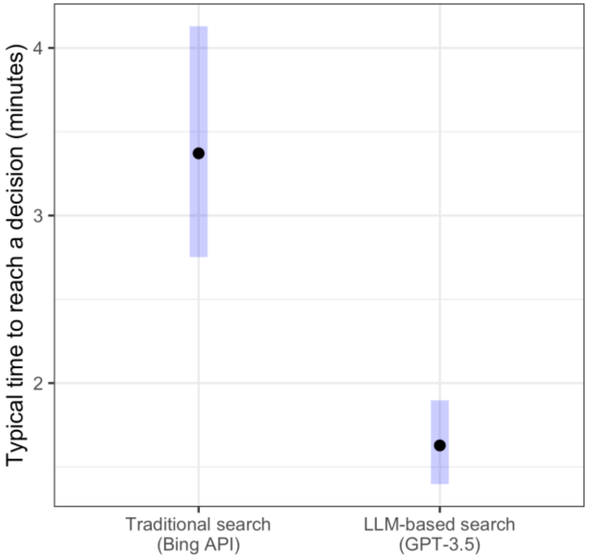
Lab experiments show LLMs can substantially improve productivity on common information work tasks, although there are some qualifiers

LLM-based tools can help workers complete a variety of tasks more quickly and increase output quality

- Studies have found that people complete simulated information work tasks much faster and with a higher quality of output when using generative AI-based tools:
    - People took 37% less time on common writing tasks (Noy & Zhang 2023).
    - BCG consultants produced >40% higher quality on one simulated consulting project (Dell'Acqua et al. 2023).
    - Users were also 2x faster at solving simulated decision-making problems when using LLM-based search over traditional search (Spatharioti et al. 2023).
- For some tasks, increased speed can come with moderately lower correctness.
    - When the LLM made mistakes, BCG consultants with access to the tool were 19 percentage points more likely to produce incorrect solutions (Dell'Acqua et al. 2023).
    - Spatharioti et al. (2023) developed a simple UX-based interventions that can work well at helping people navigate these tradeoffs.
- Users may need help negotiating the tradeoffs involved to maximize productivity gains.
- How task-level gains translate to job-level gains will depend on whether gains extend to other tasks and how the tools are integrated into workflows.



Quality of output (Treated = using ChatGPT) (Noy & Zhang 2023)



Estimates and confidence intervals for average log(time) by condition, (Spatharioti et al. 2023)

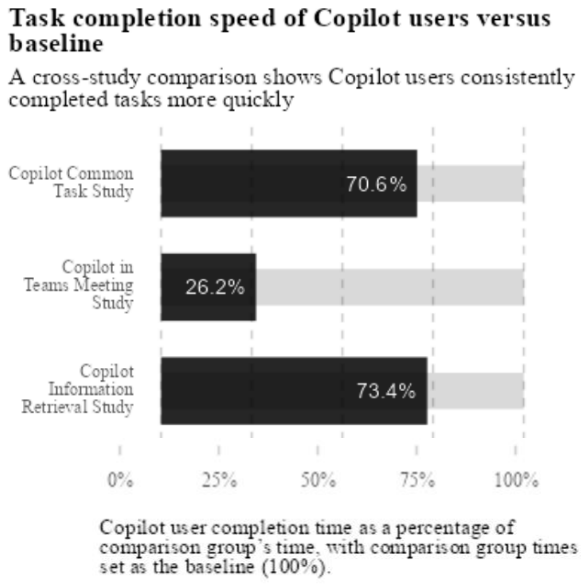
Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. SSRN preprint.
Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. SSRN Working Paper 4573321.
Microsoft Study: Spatharioti, S. E., et al. (2023). Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv preprint.

Copilot for M365 saves time for a variety of tasks in lab studies and surveys

Users also report Copilot reduces the effort required, effects on quality are mostly neutral

Microsoft's AI and Productivity Report synthesizes results from 8 early studies, most focused on the use of M365 Copilot for information worker tasks for which LLMs are most likely to provide significant value (Cambon et al. 2023).

- Tasks included meeting summarization, information retrieval, and content creation.
- Study participants with Copilot completed experimenter-designed tasks in 26-73% as much time as those without it.
- A survey of enterprise users with access to Copilot also showed substantial perceived time savings.
  - 73% agreed that Copilot helped them complete tasks faster, and 85% said it would help them get to a good first draft faster.
- Many studies found no statistically significant or meaningful effect on quality.
  - However, in the meeting summarization study where Copilot users took much less time, their summaries included 11.1 out of 15 specific pieces of information in the assessment rubric versus the 12.4 of 15 for users who did not have access to Copilot.
  - In the other direction, the study of M365 Defender Security Copilot found security novices with Copilot were 44% more accurate in answering questions about the security incidents they examined.
  - A study of the Outlook "Sound like me" feature found Copilot users like many aspects of the emails it generated more than human-written ones but could sometimes tell the difference between Copilot writing versus human writing.
  - Of enterprise Copilot users, 68% of respondents agreed that Copilot actually improved quality of their work.
- Users also reported tasks required less effort with Copilot.
  - In the Teams Meeting Study, participants with access to Copilot found the task to be 58% less draining than participants without access.
  - Among enterprise Copilot users, 72% agreed that Copilot helped them spend less mental effort on mundane or repetitive tasks.



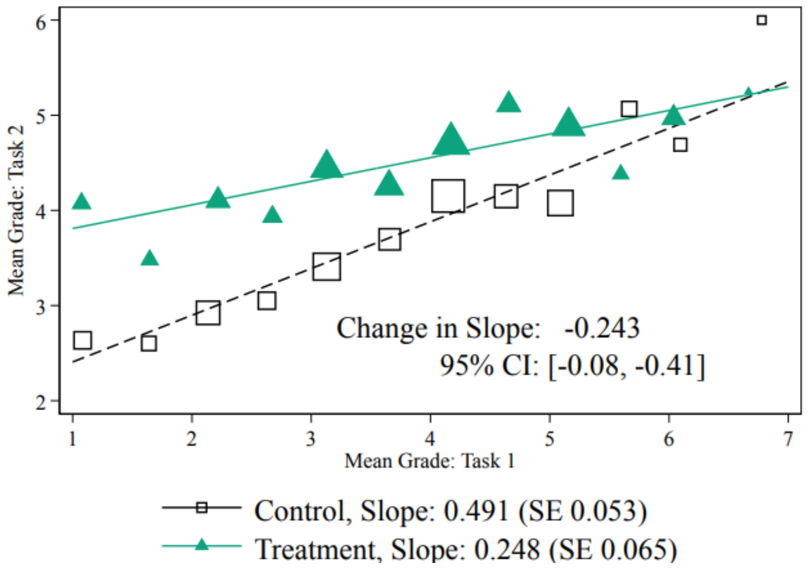
Task completion times for lab studies of Copilot for M365 (Cambon et al. 2023)

Microsoft Study: Cambon, A., et al. (2023), Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. MSFT Technical Report.

The evidence points to LLMs helping the least experienced the most

Mostly early studies have found that new or low-skilled workers benefit the most from LLMs

- In studying the staggered rollout of a generative AI-based conversational assistant, Brynjolfsson et al. (2023) found that the tool helped novice and low-skilled workers the most.
    - They found suggestive evidence that the tool helped disseminate tacit knowledge that the experienced and high-skilled workers already had.
- In a lab experiment, participants who scored poorly on their first writing task improved more when given access to ChatGPT than those with high scores on the initial task (see graph, Noy & Zhang 2023).
- Peng et al. (2023) also found suggestive evidence that GitHub Copilot was more helpful to developers with less experience.
- In an experiment with BCG employees completing a consulting task, the bottom-half of subjects in terms of skills benefited the most, showing a 43% improvement in performance, compared to the top half whose performance increased by 17% (Dell'Acqua et al. 2023).
- Recent work by Haslberger et al. (2023) highlights some complexities and nuance in these trends, including cases in which LLMs might increase performance disparities.



Green triangles represent those who got access to ChatGPT for the second task. Their scores across the two tasks are less correlated. (Noy & Zhang 2023)

Brynjolfsson, E., et al. (2023). Generative AI at Work. NBER Working Paper 31161.
Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. SSRN Working Paper 4375283.
Microsoft Study: Peng, S., et al. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint 2302.06590.
Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. SSRN Working Paper 4573321.
Haslberger, M., et al. (2023) No Great Equalizer: Experimental Evidence on AI in the UK Labor Market. SSRN Working Paper 4594466,

Critical thinking: LLM-based tools can be useful provocateurs

Reconceptualizing AI systems as “provocateurs” in addition to “assistants” can promote critical thinking in knowledge work

- As AI is applied to more generative tasks, human work is shifting to “critical integration” of AI output, requiring expertise and judgement (Sarkar 2023).
- Moving beyond just error correction, AI provocateurs would challenge assumptions, encourage evaluation, and offer counterarguments.
- Interaction design of provocative AI needs to strike a balance between useful criticism and overwhelming people.
- Frameworks that structure critical thinking objectives (e.g., Bloom’s taxonomy) and Toulmin’s model operationalize argument analysis, which could inform provocative AI design (Kneupper 1978).
- Interactive technologies that spark discussion and engage users contribute to critical thinking development (Sun et al. 2017; Lee et al. 2023).

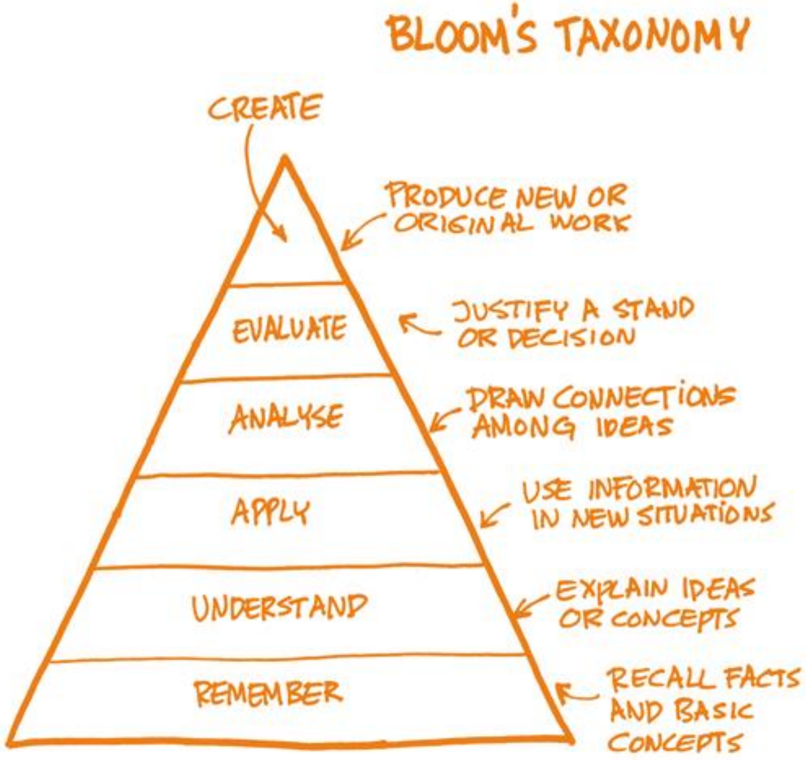


Image of Bloom’s Taxonomy (Bezjak et al. 2018)

Microsoft Study: Sarkar, A. (2023). Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots Proceedings of the ACM Symposium on Human-Computer Interaction for Work (CHIWORK 2023).
Kneupper, C. W. (1978). Teaching argument: An introduction to the Toulmin model. *College Composition and Communication* 29, 3..
Sun, N., et al. (2017). Critical thinking in collaboration: Talk less, perceive more. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*.
Lee, S., et al. (2023). Fostering Youth’s Critical Thinking Competency About AI through Exhibition. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.
Bezjak, S. et al, (2018). Open Science Training Handbook

AI can enhance microproductivity practices

AI can be harnessed to augment human capabilities through novel task management strategies

- The concept of "microproductivity", in which complex tasks are decomposed into smaller subtasks and performed in "micromoments" by the person most skilled to do so, can be enhanced through automation (Teevan 2016).
    - For example, Kokkalis et al. (2013) demonstrated that high level tasks broken into multistep action plans through crowdsourcing result in people completing significantly more tasks (47.1% task completion) compared to the control condition of no plans (37.8%). These benefits were scaled by applying NLP algorithms to automatically create action plans for a larger variety of tasks based on a training set of similar tasks, and the plans were further refined through human intervention.
    - Kaur et al. (2018) showed that using a fixed vocabulary to break down comments in a document into a series of subtasks resulted in a 28% increase in subtasks that can be handed off to crowdsourcing or automation, leaving a smaller percentage of subtasks left for the document author.
- AI can help with automatic identification of micromoments and microtasks, improving overall quality and efficiency.
    - Contextual identification of micromoments based on preceding activities and location can yield up to 80.7% precision (Kang et al. 2017); such micromoments can be used for learning (Cai et al. 2017), creation of audiobooks (Kang et al. 2017), editing documents (August et al. 2020), and coding (Williams et al. 2018).
    - White et al. (2021) demonstrated how machine learning can be leveraged to automatically detect microtasks from user-generated task lists resulting in a positive precision of 75%, and forecast duration, with the best classifier performance for tasks with duration of 5 minutes.

John enters into his task list:

- Exercise more frequently

TaskGenies responds with the action plan:

- Find a workout buddy to keep you accountable
- Get a gym membership
- Create a weekly exercise schedule
- Start working out this Monday and stick to the schedule

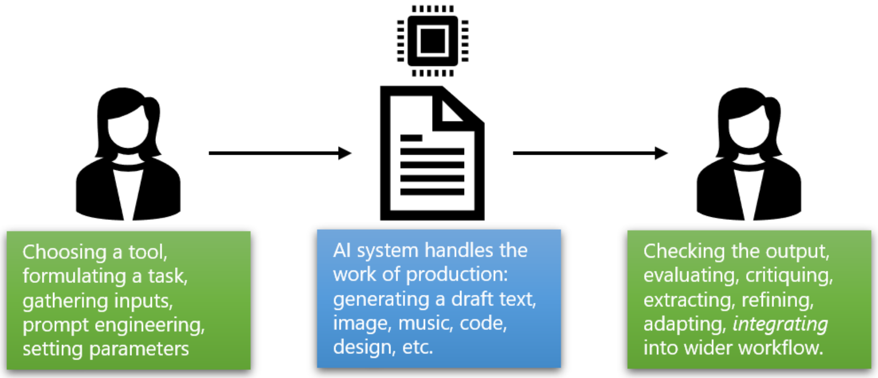
Decomposing high level tasks into concrete steps (plans) makes them more actionable resulting in higher task completion rates. Online crowds do the decomposition, algorithms identify and reuse existing plans. (Kokkalis 2013)

Microsoft Study: Teevan, J. (2016). The future of microwork. XRDS 23, 2.
Kokkalis, N., et al. 2013. TaskGenies: Automatically Providing Action Plans Helps People Complete Tasks. ACM Transactions on Computer-Human Interaction 20, 5.
Kaur, H. et al. 2018. Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning. Proceedings of the ACM on Human-Computer Interaction. 2, CSCW.
Kang, B. et al. (2017). Zaturi: We Put Together the 25th Hour for You. Create a Book for Your Baby. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).
Cai, C. J., Ren, A., & Miller, R. C. (2017). WaitSuite: Productive Use of Diverse Waiting Moments. ACM Transactions on Computer Human Interaction 24, 1.
Microsoft Study: August, T., et al. (2020). Characterizing the Mobile Microtask Writing Process. 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20).
Microsoft Study: Williams, A., (2019). Mercury: Empowering Programmers' Mobile Work Practices with Microproductivity. Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology
Microsoft Study: White, R. W., et al. (2021). Microtask Detection. ACM Trans. Inf. Syst. 39, 2.

Analyzing and integrating may become more important skills than searching and creating

With content being generated by AI, knowledge work may shift towards more analysis and critical integration

- Information search as well as content production (manually typing, writing code, designing images) is greatly enhanced by AI, so general information work may shift to integrating and critically analyzing retrieved information.
- Writing with AI is shown to increase the amount of text produced as well as to increase writing efficiency (Biermann et al. 2022; Lee et al. 2022).
- With more generated text available, the skills of research, conceptualization, planning, prompting and editing may take on more importance as LLMs do the first round of production (e.g., Mollick 2023).
- Skills not directly related to content production, such as leading, dealing with critical social situations, navigating interpersonal trust issues, and demonstrating emotional intelligence, may all be more valued in the workplace (LinkedIn 2023).



The critical integration “sandwich”: when AI handles production, human critical thinking is applied at either end of the process to complete knowledge workflows (Sarkar 2023).

Biermann, O. C., et al. (2022). From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22).
Lee, M., et al. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).
Mollick, E. (2023). My class required AI. Here's what I've learned so far. One Useful Thing
LinkedIn (2023). Future of Work Report: AI at Work.

Constructing optimal prompts is difficult

Prompts are the primary interface for both users and developers to interact with large language models, but consistently developing effective prompts is a challenge

- Precise prompt composition is critical in achieving the desired LLM output, with semantically similar prompts yielding significantly different, sometimes incorrect, outputs (Jiang et al. 2020).
- Writing effective prompts can require significant effort, including multiple iterations of modification and testing (Jiang et al. 2022).
- Prompt behavior can be brittle and non-intuitive.
  - Seemingly minor changes, including capitalization and spacing can result in dramatically different LLM outputs (Holtzman 2021; Arora et al. 2023)
  - The order of prompt elements, such as sections, few-shot examples or even words can significantly impact accuracy, in some cases varying from near random chance to state-of-the-art (Zhao et al. 2021; Kaddour et al. 2023).
  - The same prompt can result in significantly different performance across model families, even with models of similar parameter size (Sanh et al. 2022).
  - While many prompting techniques have been developed, there is little theoretical understanding for why any particular technique is suited to any particular task (Zhao et al. 2021).
- End users of prompt-based applications struggle more than prompt engineers to formulate effective prompts (Zamfirescu-Pereira et al. 2023).

Jiang, Z., et al. (2020). How Can We Know What Language Models Know? Transactions of the Association for Computational Linguistics, 8.
Jiang, E., et al. (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems
Holtzman, A., et al. (2021). Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. EMNLP.
Arora, S., et al. (2023). Ask me anything: A simple strategy for prompting language models. The Eleventh International Conference on Learning Representations.
Zhao, Z., et al. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. Proceedings of the 38th International Conference on Machine Learning.
Kaddour, J., et al. (2023). Challenges and Applications of Large Language Models. arXiv preprint.
Sanh, V., et al. (2022) Multitask Prompted Training Enables Zero-Shot Task Generalization. International Conference on Learning Representations.
Zamfirescu-Pereira, J.D., et al. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. (CHI '23).

But constructing effective prompts is becoming easier

Base model training, tools, and LLMs themselves are helping improve prompt performance

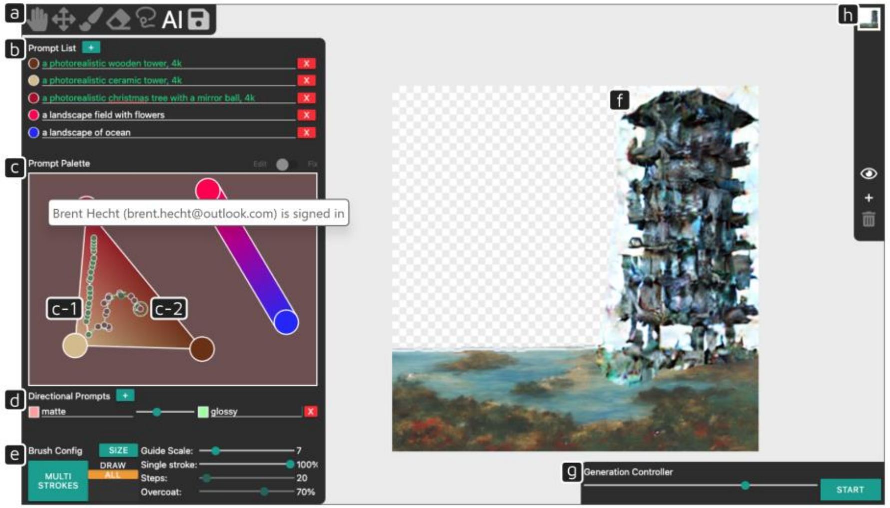
- Significant research is devoted to improving model instruction following.
    - Fine-tuning with human feedback can dramatically improve LLMs ability to follow prompt instructions, even when compared to models with 100x parameters (Ouyang et al. 2022).
    - Utilizing multi-task and chain-of-thought training data significantly improved instruction-following capabilities (Chung et al. 2022).
- LLMs have been shown to be effective prompt optimizers.
    - Prompt optimization techniques that utilize an LLM to iteratively provide feedback and produce new versions of a hand-crafted seed prompt can significantly improve performance (Pryzant et al. 2023).
    - Multi-step optimization with natural language task descriptions and scored optimization examples can induce an LLM to generate new, higher performing prompt variations (Yang et al. 2023).
    - Inspired by evolutionary algorithms, an LLM can be used to generate new prompt candidates by mutating prompts from a population and evaluating their fitness against a test set over multiple generations (Fernando et al. 2023).
    - Recent work suggests optimized prompts can outperform specifically fine-tuned models in a number of important domains, especially medicine (Nori et al. 2023).

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35.
Chung, H. W., et al. (2023) Scaling instruction-finetuned language models. arXiv preprint.
Pryzant, R., et al. (2023). Automatic Prompt Optimization with Gradient Descent and Beam Search. arXiv preprint.
Yang, C., et al. (2023). Large language models as optimizers. arXiv preprint.
Fernando, C., et al. (2023). Promptbreeder: Self-referential self-improvement via prompt evolution. arXiv preprint.
Nori, Harsha, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine arXiv preprint.

People are also learning to prompt more effectively

As people get better at communicating with LLMs, they are getting better results

- Prompt guidance is commonly used as a way for people to learn to prompt better.
    - Research suggests that training on how to prompt can lead to greater productivity gains from LLM tools (Dell'Acqua et al. 2023).
    - Using a lens informed by the psycholinguistic concept of grounding (Clark 1996), Teevan (2023) argues in HBR that effective communication with generative AI requires providing contextual information, specifying the desired output, and verifying the accuracy of the generated content.
    - Many other guides and reference materials are also available, including a recent WorkLab article (Microsoft 2023) and OpenAI's documentation on prompt engineering (OpenAI 2023).
- Tools can help users develop more effective prompts.
    - Researchers are building interactive tools that can help people iteratively refine their prompts (e.g., Brade et al. 2023; Chung & Adar 2023).
    - Human-in-the-loop LLM-based optimization was shown to enable non-experts to improve prompt performance for medical note generation (Yao et al. 2023).
    - Copilot Lab is one Microsoft effort to help people learn how to effectively interact with LLMs, e.g. by providing a collection of suggested prompts.



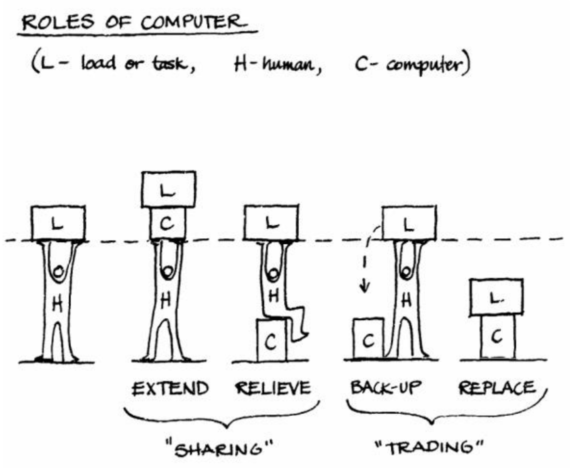
The PromptPaint interface, which uses non-textual affordances to help people refine image generation (Chung & Adar 2023)

Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. SSRN working paper.
Clark, H. H., 1996. Using Language (1st edition ed.). Cambridge University Press.
Microsoft Study: Teevan, J. (2023). To work well with GenAI, you need to learn how to talk to it. Harvard Business Review.
Microsoft Study: Microsoft WorkLab (2023). The art and science of working with AI.
OpenAI (2023) Prompt Engineering.
Brade, S., et al. (2023). Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. arXiv preprint.
Chung, J. J. Y., & Adar, E. (2023) PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.
Yao, Z., et al. (2023) Do Physicians Know How to Prompt? The Need for Automatic Prompt Optimization Help in Clinical Note Generation. arXiv preprint.

Complementarity is a human-centered approach to AI collaboration

Humans and AI can “collaborate” in many ways: from each party acting as a collaborative team member, to a person overseeing an AI automation loop, to AI simulating a human

- Sheridan & Verplank (1978) introduced the Level of Automation (LOA) framework, to classify how responsibility can be divided between human and automation (see figure). It has been widely applied, e.g., in self-driving vehicles and process control.
  - Computers share load with humans by extending human capabilities or relieving the human to make their job easier, or
  - Computers trade load with humans by through being a back-up in case the human falters, or completely replacing the human.
- Based on the idea of LOAs, Parasuraman & Wickens (2000) outlined a model to determine what should be automated and to what extent. It has been applied in the analysis of contemporary systems (Mackeprang et al. 2019).
- A human-centered approach takes a complementary perspective, in which human and AI are partners that balance out each other’s weaknesses (Lubars & Tan 2019). Examples include mixed initiative-interaction (Horvitz 1999), collaborative control where human and machines are involved in the same activity (Fong et al. 2001) and coactive design that focuses on supporting interdependency between the human and AI (Johnson et al. 2011).



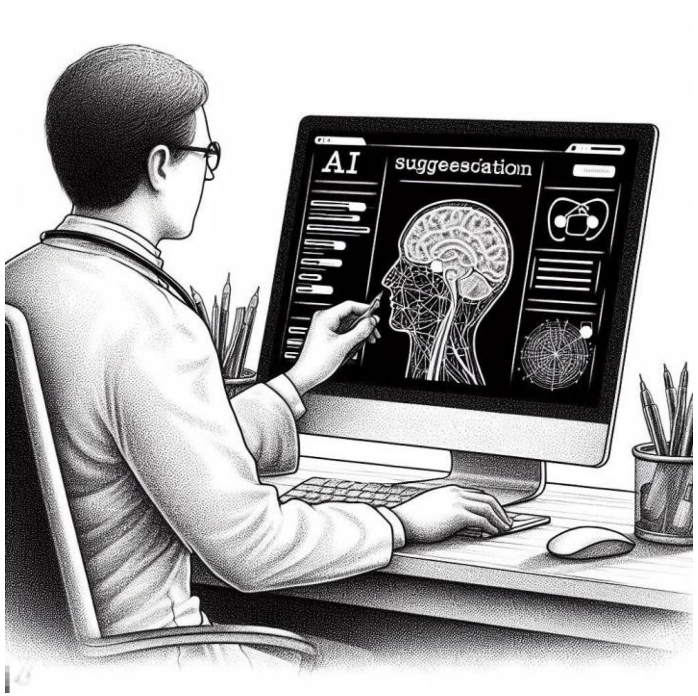
Distribution of task-load between humans and computers/automation (Sheridan & Verplank 1978)

Sheridan, T. B., & Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. Technical Report.
Parasuraman, R., & Wickens, C. D. (2008). Humans: Still Vital After All These Years of Automation. Human Factors, 50(3).
Mackeprang, M., et al. (2019). Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. Proceedings of the ACM Human-Computer Interaction. 3, CSCW.
Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: towards a framework of task delegability. Proceedings of the 33rd International Conference on Neural Information Processing Systems.
Microsoft Study: Horvitz, E. (1999). Uncertainty, Action, and Interaction: In Pursuit of Mixed-Initiative Computing. Intelligent Systems, 6.
Fong, T., et al. (2001). Collaborative control: A robot-centric model for vehicle teleoperation. The Robotics Institute
Johnson, M., et al. (2011). Beyond Cooperative Robotics- The Central Role of Interdependence in Coactive Design. IEEE Intelligent Systems 26, 3.

Appropriate reliance on AI is a key challenge in human-AI interaction

For many reasons, people often over-rely on AI, but careful design can create appropriate reliance

- Overreliance on AI happens when people accept incorrect AI outputs. Many things can affect overreliance, such as familiarity with task, AI literacy, automation bias, and confirmation bias (Passi & Vorvoreanu 2022).
- Overreliance on AI leads to poorer performance than either the human or the AI acting alone (Agarwal et al. 2023; Passi & Vorvoreanu 2022), so it's important to keep in mind when designing AI systems with which people interact.
- Many techniques exist for reducing overreliance, including effective onboarding, transparency techniques (Danry et al. 2023), uncertainty visualizations (next slide), cognitive forcing functions, and more. However, mitigation techniques, particularly explanations, can backfire and increase rather than reduce overreliance, so careful design and evaluation are needed to create appropriate reliance (Passi & Vorvoreanu 2022)
- Passi & Vorvoreanu (2022) provides a review of research about antecedents, consequences, and mitigations of overreliance on AI.



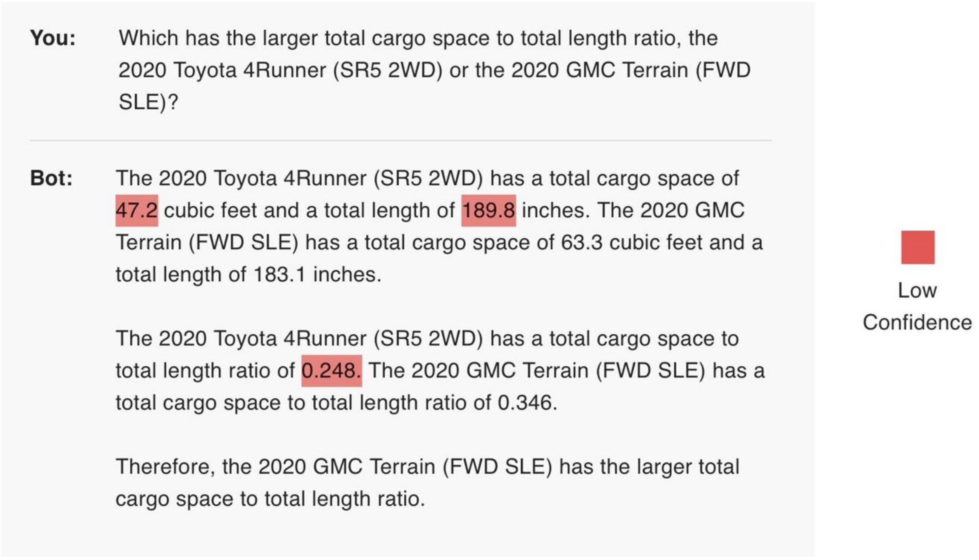
In a study about medical decision making, clinicians with low AI literacy were 7 times more likely to select medical treatments aligned with AI recommendations (Jacobs et al. 2021; Image credit: Bing Image Creator)

Passi, S., & Vorvoreanu, M. (2022). Overreliance on AI Literature Review. Microsoft Research preprint.
Agarwal, N., et al. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper 31422.
Danry, V., et al. (2023). Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).

Uncertainty visualization can help create appropriate reliance

Highlighting uncertain content in LLM-enabled search engine answers improved humans' decision accuracy

- Spatharioti at al. (2023) created a confidence-based scheme that highlighted uncertain parts of an LLM-enabled search engine's response (see image).
- For challenging tasks in which the LLM tended to err, highlighting uncertain content improved decision accuracy compared to unannotated output.
- Highlighting uncertain content can build awareness that AI-generated content may be wrong.
- Similarly, in a study with software developers, highlighting uncertain code suggestions increased task accuracy (Vasconcelos et al. 2023)



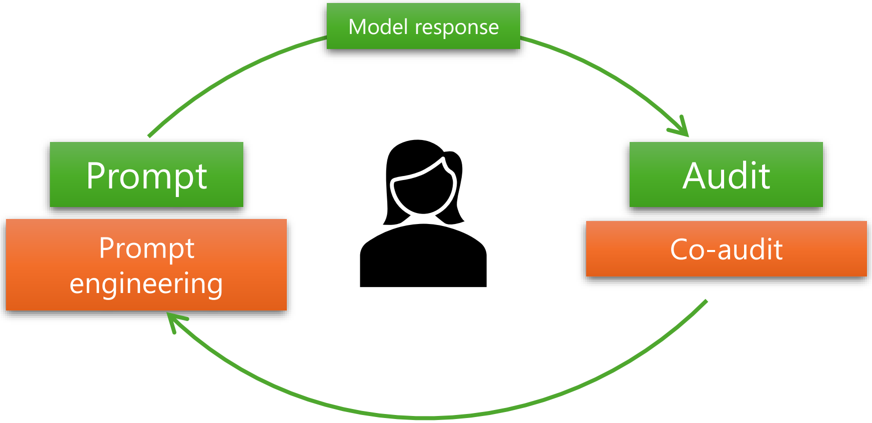
UX showing uncertainty in results to improve reliance (Spatharioti et al 2023)

Microsoft study: Spatharioti, S., et al. (2023). Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv preprint.
Vasconcelos, H., et al (2023). Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. arXiv preprint.

Co-audit tools help users check LLM outputs

Prompt engineering and co-audit are complementary aspects of human-AI dialog

- Co-audit (Gordon et al. 2023) is the opposite of prompt engineering: Co-audit tools aim to help users to audit or evaluate AI outputs for mistakes. Co-audit tools aim to help with abstraction matching, correctness checking, and repair decisions for AI content.
- Examples include tools for AI-generated spreadsheet computations (Liu et al. 2023; Ferdowsi et al. 2023), which help users understand how their words are matched to a computation and inspect how the computation behaves.
- ChatProtect (Mündler et al. 2023) is an AI-based co-audit tool that itself is based on AI. It is a chat experience with features to detect and remove hallucinated content from generated text. The co-audit experience lets the user inspect different sentences to detect hallucinations via sampling multiple times from the LLM.
- Co-audit may help low-confidence users, who may over-rely on or be intimidated by AI-generated outputs. (Gordon et al. 2023)
- Microsoft has proposed principles for co-audit (Gordon et al. 2023).



The relationship between co-audit and prompt engineering: one helps construct the input prompt, while the other helps double-check the output response (Gordon et al. 2023)

Microsoft Study: Gordon, A., et al. (2023). Co-audit: tools to help humans double-check AI-generated content. Microsoft Research preprint.
Liu, M. X., et al. (2023). "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.*
Ferdowsi, K., et al. (2023). ColDeco: An End User Spreadsheet Inspection Tool for AI-Generated Code. *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC).*
Mündler, N., et al. (2023). Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *arXiv preprint,*

Generative AI demands greater metacognition from users but also has potential to support it

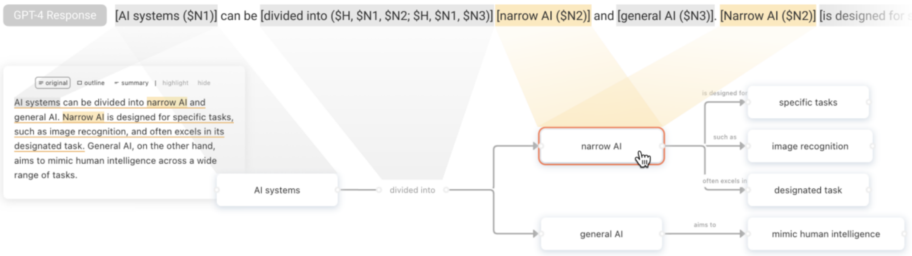
Users of Generative AI require self-awareness and well-calibrated confidence for effective interactions

- Working with generative AI tools like Copilot has implications for users' *metacognition* – the ability to analyze, understand, and control one's own thought processes, including aspects like self-awareness, well-calibrated confidence, and flexibility (Norman et al. 2019).
- Generative AI *demands greater metacognition* from users, for example:
  - Users of AI systems must be self-aware of, and explicit about, their goals, translating them into precisely specified prompts (Zamfirescu-Pereira et al. 2023; Chen et al. 2023). Ready-made prompts are helpful, but nevertheless require adaptation and evaluation based on users' goals and intentions.
  - Generative AI's ability to rapidly produce entire documents makes evaluating these outputs for quality far more important and effortful than word or phrase suggestions with "auto-complete". Users need to maintain a well-calibrated level of confidence in their own evaluation ability and in the AI system (Chong et al. 2022; Steyvers & Kumar 2023).
- Generative AI can also *support users' metacognition*, for example:
  - Systems can support users' self-awareness by proactively identifying and organizing ideas. *Graphologue* is a system that creates interactive, graphical node-link diagrams out of lengthy LLM responses to facilitate information exploration, organization, and comprehension (Jiang et al. 2023).
  - Similar to how human experts can guide end-users in co-creating with AI, generative AI systems can provide proactive self-reflective prompts to help end-users calibrate their confidence when working with them – e.g., *"How confident are you in understanding this output? Does anything require explanation?"* (Gmeiner et al. 2023).

Peoples' confidence in AI and in themselves: The evolution and impact of confidence on adoption of AI advice

	Regression coefficient against the probability of accepting AI suggestion		
	Low (Poor)	Mid (Fair)	High (Good)
Confidence in AI	−0.0737 *(P=0.9)*	0.0820 *(P=0.7)*	−0.691 *(P=0.2)*
Self-confidence	−1.36 *(P=0.09)*	−0.813 *(P<0.05)*	1.78 *(P<0.05)*

For decision-makers in a chess game, self-confidence is related to acceptance of AI suggestions, while confidence in AI is not. Good decision-makers effectively translate their self-confidence into appropriate reliance on AI. Adapted from Chong et al. (2022).



Graphologue creates node-link diagrams out of LLM responses to help end-users make sense of outputs (Jiang et al. 2023)

Norman, E., et al. (2019). Metacognition in psychology. Review of General Psychology, 23(4).
Zamfirescu-Pereira, J. D., et al. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. (CHI '23).
Chen, X. A., et al. (2023). Next Steps for Human-Centered Generative AI: A Technical Perspective. arXiv preprint.
Chong, L., et al (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. Computers in Human Behavior, 127, 107018.
Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. Perspectives on Psychological Science.
Gmeiner, F., et al (2023). Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.
Jiang, P., et al. (2023). Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. arXiv preprint arXiv:2305.11473.

LLMs have made giant steps forward in multilingual performance, but there is still much to be done

Performance drops for languages other than English where training data and context are lacking

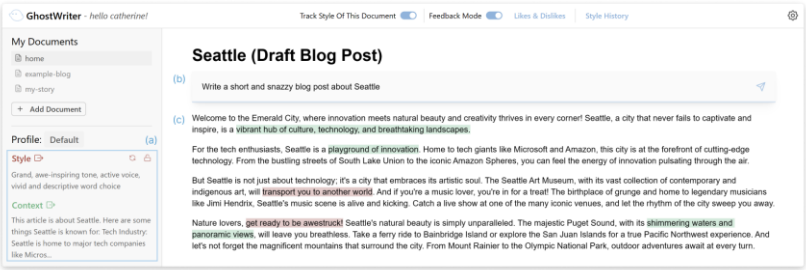
- Multilingual LLMs will reduce the barriers to information access (Nicholas et al. 2023) and help realize transformative applications at scale (Nori et al. 2023).
- The impact of this can be much higher in low and middle socioeconomic regions where resources are scarce.
- However, many problems still remain. For instance, GPT4 performance is still best on English, and performance drops substantially as we move to mid- and low-resource languages (Ahuja et al. 2023).
  - Many language families don't have enough data for adequate training (Patra et al. 2023).
- Non-Latin scripts are under-represented on the web, so LLMs perform worse on non-Latin text even in high resource languages, such as Japanese (Ahuja et al. 2023).
- Lack of relevant linguistic and societal context in languages and cultures will impact task level performance for LLMs, for example in handling dialects within the same language family (Hada et al. 2023).
- There is still little investigation into the multilingual performance of applications built on LLM derived artifacts, for example knowledge-bases built on low quality embeddings will not perform as well.

Nicholas, G., et.al. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis. arXiv pre-print
Nori, H., el. al. (2023). Capabilities of GPT-4 on Medical Challenge Problems. arXiv preprint.
Ahuja, K., et.al. MEGA: Multilingual Evaluation of Generative AI. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing
Patra, B., et.al. Everything you need to know about Multilingual LLMs. ACL 2023 Tutorial
Hada, R., et.al. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? arXiv preprint.
OpenAI (2023). GPT-4 Technical Report. arXiv preprint.

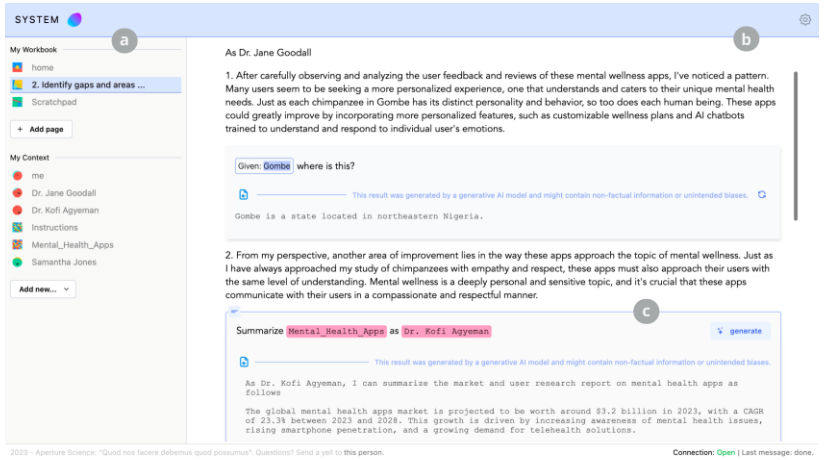
New systems point to how LLMs can aid creative activities

A few in-progress projects at Microsoft Research investigate how LLMs can assist in creative tasks

- Preserving people's agency over the creative process (orchestration) is fundamental for successful, meaningful augmentation (Palani et al. 2023). MSR researchers developed a system called "Ghostwriter" that the explores new ideas to champion agency in controlling the output style of LLM-based writing, and novel ways to express personalization.
- One in-progress study highlights how important personalization is to preserving a creator's authenticity and improves the sense of authorship (Hwang et al. 2023).
- Another study highlighted how creativity is not a discrete event that is served in a lightning bolt moment. Supporting creativity is primarily about also supporting the creative's process as well as providing generative tools. MSR has developed a system that probes ideas to support the creative process (Palani et al. 2023).
- An MSR project gathered feedback from participants interacting with two writing-enhancing prototypes (Ghostwriter and Amethyst) and their feedback indicates that their mental models about their relationship with the systems varies between being a tool, to an assistant to a collaborator. This is due to tasks not being monolithic in their demands.
- There is potential in applying LLMs to the acceleration of game narratives creation (Brockett et al. 2023). Ongoing work aims at exploring how LLMs can augment the development and testing of games.
- People can teach about writing style besides prompts and chats by working directly on style description documents and directly annotating/marking the writing document. The process can also give them literacy about how to understand and talk about style (Yeh et al. 2023).



Screenshot of GHOSTWRITER (Yeh, et al 2023)



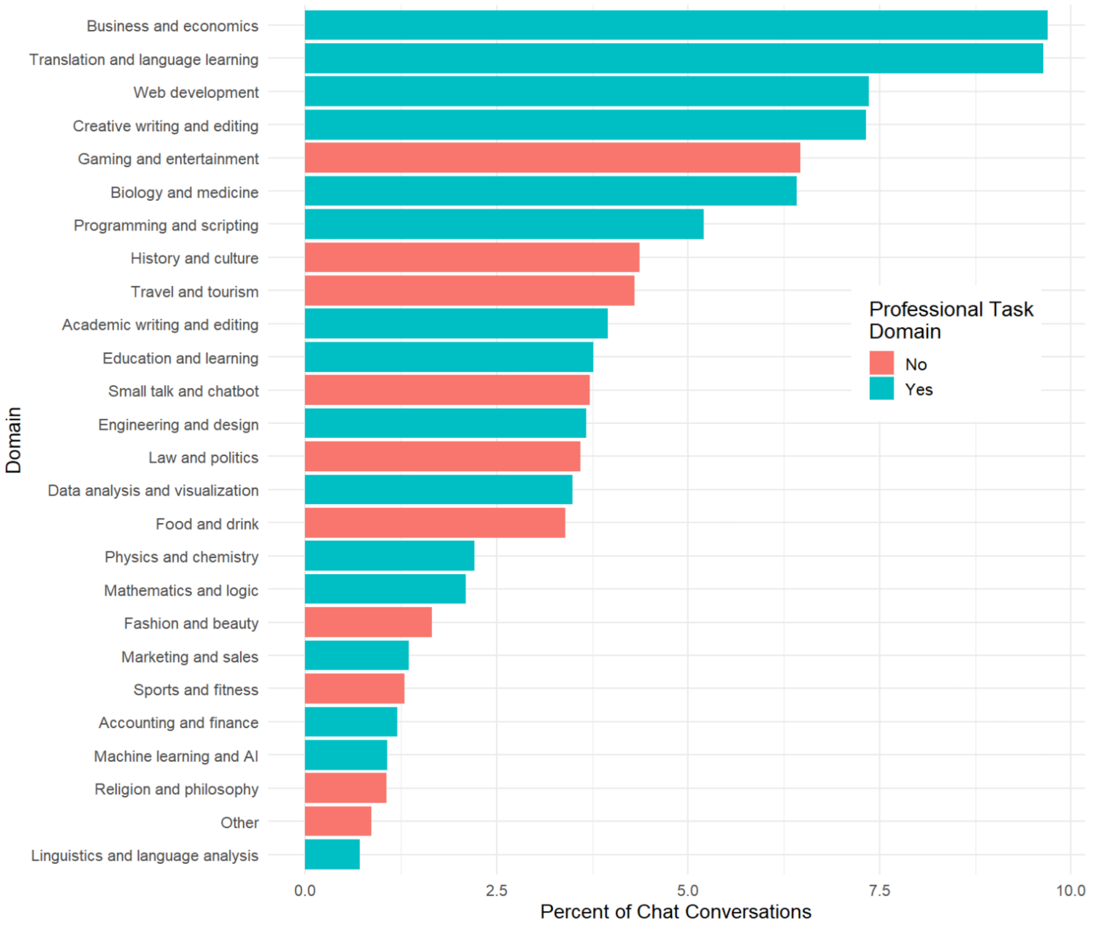
Screenshot of Amethyst (Palani et al. 2023)

Microsoft Study: Palani, S., et al. (2023). Amethyst: A Creative Process-Focused Notebook That Leverages Large Language Models. (under review)
Microsoft Study: Hwang, A., et al (2023). Seeking authenticity in creative writing with LLMs. In preparation
Microsoft Study: Brockett, C., et al. (2023) Project Emergence
Microsoft Study: Yeh, C., et al (2023). GhostWriter: Augmenting Human-AI Writing Experiences Through Personalization and Agency. (under review)

Bing Chat is frequently used for professional and more complex tasks

Compared to traditional search, consumers use (LLM-based) Bing Chat for more topics in professional domains and for more complex tasks

- Counts et al. (2023) analyze a sample of fully-anonymized, consumer-facing Bing Chat conversations and Bing searches from May-June 2023.
- Using GPT-4 to group these conversations and searches by topics, they find (see graph):
  - 69% of Bing Chat conversations are in domains oriented toward professional tasks.
  - 39% of Bing Search sessions are in professional task domains.
- Counts et al. also categorize the complexity of the chats and searches sessions according to Anderson & Krathwohl's et al.'s (2001) taxonomy of "Remember", "Understand", "Apply", "Analyze", and "Create".
  - In Bing Chat 36% of conversations are high complexity (Apply, Analyze, or Create).
  - But in Bing Search, only 13% are high complexity.



Domains of Bing chat conversations (Counts et al. 2023)

Microsoft Study: Counts, S., et al. (2023). Completing Knowledge Work and Complex Tasks with a Generative Search Engine. In preparation.
Anderson, L., & Krathwohl, D. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman

"Fast AI" and "Slow AI": Different LLM experiences require different latencies

Many interactions with LLMs require rapid iteration, however some don't, and the "slow search" literature points to ways systems can use that extra time to deliver better results to end users

- One well-known challenge with LLM systems is latency between issuing a prompt and receiving a response (e.g., Lee et al. 2023) and a great deal of research is happening to reduce this latency (e.g., Kaddour et al. 2023).
- For many use cases, low latency is essential: we know from traditional search that even small increases in latency can substantially affect the user experience (e.g., Shurman & Brutlag 2009).
- However, the literature on "slow search" (Teevan et al. 2014) highlights how some use cases do not need fast responses, and this additional time can open up a whole new design space for AI applications.
- People are willing to wait hours and days for responses to many types of high-importance questions, such as in forums like StackOverflow (Bhat et al. 2014) and in social media (Hecht et al. 2012).
- With more time to return a response, LLMs can issue multiple prompts, search over more documents using retrieval-augmented generation approaches, do additional refining of answers, and much more that probably has not been considered yet. Researchers might want to ask, "If I had minutes and not milliseconds, what new types of experiences could I create?"
- The "Slow AI" user experience needs to be different than the "fast AI" experience, clearly communicating the system's status, helping people understand the benefits of delayed response, and providing ways to interrupt or redirect if it appears things are off-track (Teevan et al. 2013).
- Bing's Deep Search experience provides a real-world example of how a "fast AI" experience (standard Bing Chat) can be complemented by a "slow AI" one (Microsoft 2023).

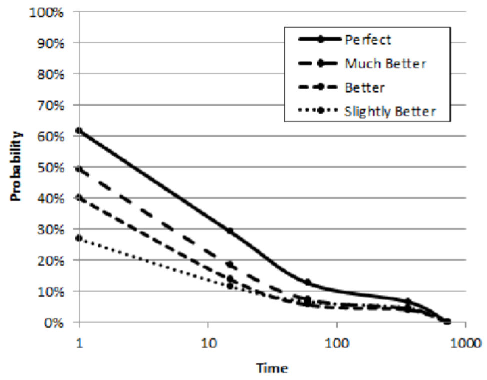


Figure 4. The probability participants were willing to wait at least *T* minutes for their search results for different answer quality levels. (Time on log scale.)

The observed relationship in one study between willingness-to-wait and wait time for different levels of search result quality in traditional search (Teevan et al. 2013)

Lee, M., et al. (2023) Evaluating Human-Language Model Interaction. arXiv preprint.
Kaddour, Jean, J.H., et al. (2023). "Challenges and Applications of Large Language Models." arXiv preprint.
Shurman, E., & Brutlag, J. (2009). Performance related changes and their searcher impact. Velocity.
Microsoft study: Teevan, J., et al. (2014) Slow Search. Communications of the ACM 57, 8.
Bhat, V., et al. (2014). Min(e)d your tags: Analysis of question response time in stackoverflow. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
Microsoft study: Hecht, B., et al. (2012). SearchBuddies: Bringing Search Engines into the Conversation. Proceedings of the International AAAI Conference on Web and Social Media, 6, 1.
Microsoft study: Teevan, J., et al. (2013) "Slow Search: Information Retrieval without Time Constraints." HCIR '13.
Microsoft Bing Blog (2023). Introducing Deep Search,

For software engineering, benefits of LLMs depend on the task

LLM coding tools are still nascent, and both lab studies and experience reports show varying levels of assistance, often depending on task and developer skill level

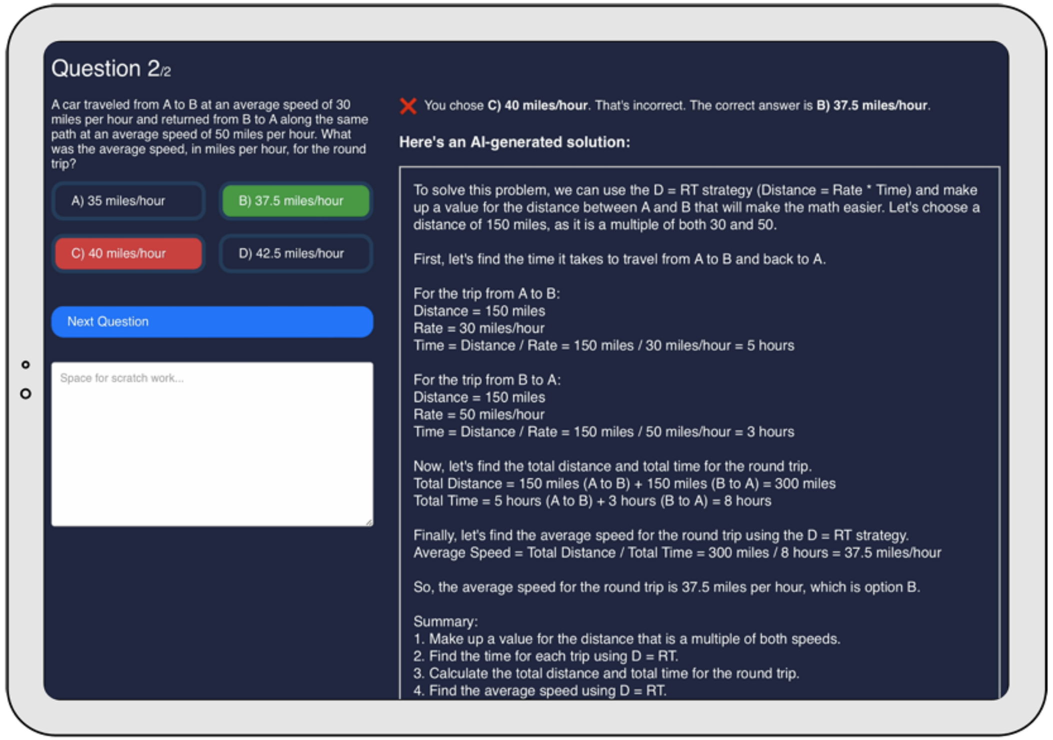
- LLM-based tools like GitHub Copilot can generate code from natural language prompts and code snippets, going beyond traditional syntax-directed autocomplete (Chen et al. 2021). Despite similarities, these new tools also differ from compilation, pair programming, and search/reuse metaphors, exhibiting distinct interaction patterns (Sarkar et al. 2022).
- In a lab study, those with GitHub CoPilot implemented an HTTP server in JavaScript 56% faster than those without (Peng et al. 2022).
- While some lab studies found no effect of AI programming assistance on completion time or correctness (Vaithilingam et al. 2022; Xu et al. 2022), developers nevertheless appreciated the capabilities of AI programming assistance and find it a positive asset (Vaithilingam et al. 2022; Xu et al. 2022).
- Experience reports show AI programming assistance reduces task time for repetitive tasks, boilerplate code, and discovering APIs (Sarkar et al. 2022).
- In a study of 69 students, the use of Codex boosted their performance on self-paced Python training. Importantly, this did not impact their manual code-modification abilities (Kazemitabaar et al. 2023).
- However, issues can arise with misinterpreted prompts and subtle bugs in generated code; debugging generated code can be challenging (Sarkar et al. 2022).
- Applying LLMs to end-user programming introduces issues like intent specification, code correctness, comprehension, behavior change, and target language mismatch (Srinivasa Ragavan et al. 2022).

Chen, M., et al. (2021). "Evaluating large language models trained on code." *arXiv preprint arXiv:2107.03374*.
Microsoft Study: Sarkar, A., et al. (2022). What is it like to program with artificial intelligence?. In Proceedings of the 33rd Annual Conference of the Psychology of Programming Interest Group (PPIG 2022).
Microsoft Study: Peng, S., et al. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint.
Vaithilingam, P., et al. (2022). "Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models." In *Chi conference on human factors in computing systems extended abstracts*.
Xu, F. F., et al, (2022). "In-ide code generation from natural language: Promise and challenges." *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2.
Kazemitabaar, M., et al. (2023). Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.
Microsoft Study: Srinivasa Ragavan, S., et al. (2022). "Gridbook: Natural language formulas for the spreadsheet grid." In *27th international conference on intelligent user interfaces*.

New research highlights some of the benefits of LLMs in education

There has been much important coverage of the challenges that LLMs introduce in education, but recent evidence also suggests the significant promise LLMs have in education as well

- In one of the first randomized experiments on LLMs and education, LLM-based explanations positively impacted learning relative to seeing only correct answers, regardless of whether students consulted them before or after attempting practice problems (Kumar et al. 2023).
    - The study also found that pre-prompting LLMs to act as tutors with customized instructions also showed promise.
- Recent work leverages a sports analogy to understand the spectrum of human-AI relationships that are possible in educational contexts (Hofman et al. 2023).
    - On one extreme, there is the concern that LLMs might act as "steroids", with students using them as substitutes for studying or doing their own work at the cost of learning skills and concepts themselves.
    - On the other extreme, there is the hope that LLMs will instead serve as "coaches", providing personalized, low-cost tutoring to a wide range of students, helping them improve their own capabilities. This could also democratize access to education and provide students normally “without” access to more 1-1 education.



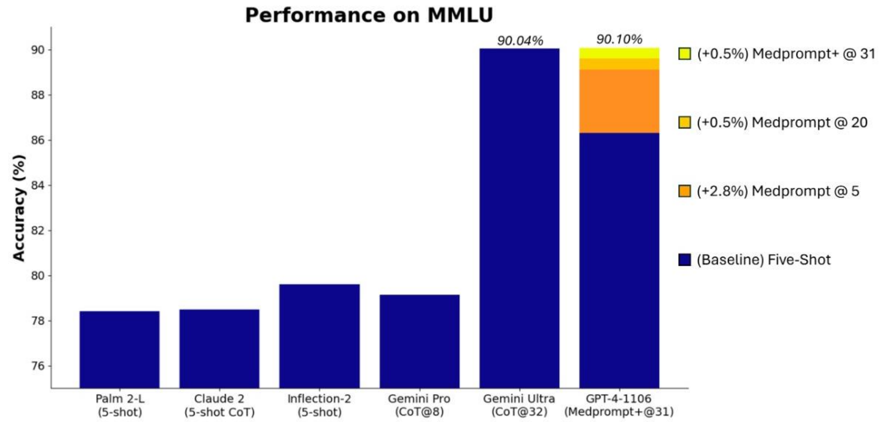
Example questions from the practice phase of Kumar et al. 2023. In this example, the student tried first and then received the answer along with tutoring help from a customized LLM (Kumar et al. 2023)

Microsoft Study: Kumar, H., et al. (2023). Math Education with LLMs: Peril or Promise? (Work in progress.)
Microsoft Study: Hofman, J. M., et al. (2023). A Sports Analogy for Understanding Different Ways to Use AI. *Harvard Business Review*.

GPT-4 excels at core examinations for medical licensure and practice

Performance improves even more when using novel prompting strategies

- Microsoft's Office of the Chief Scientific Officer (OCSO) – in conjunction with OpenAI – tested GPT-4's performance on the USMLE medical exam required to practice medicine in the US (Nori et al. 2023a,b).
- GPT-4 achieved 80% accuracy (+20 pts higher than the average passing human score) with no finetuning or advanced prompting techniques, comparable to Google's Med-PaLM2 despite the latter model being significantly finetuned (Nori et al. 2023a).
- Incorporating advanced prompting strategies boosted GPT-4's performance to 90%, far exceeding Med-PaLM2 (Nori et al. 2023b).
- GPT-4 is shown to be significantly better calibrated than GPT-3.5, demonstrating a much-improved ability to predict the likelihood that its answers are correct.
    - The model also show impressive capabilities to explain medical reasoning, personalize explanations to students, and interactively craft new counterfactual scenarios around a medical case.
- The prompting strategies in Nori et al. (2023b) generalize beyond medicine, enabling GPT-4 to outperform Google Gemini on the broad-based MMLU reasoning benchmark.



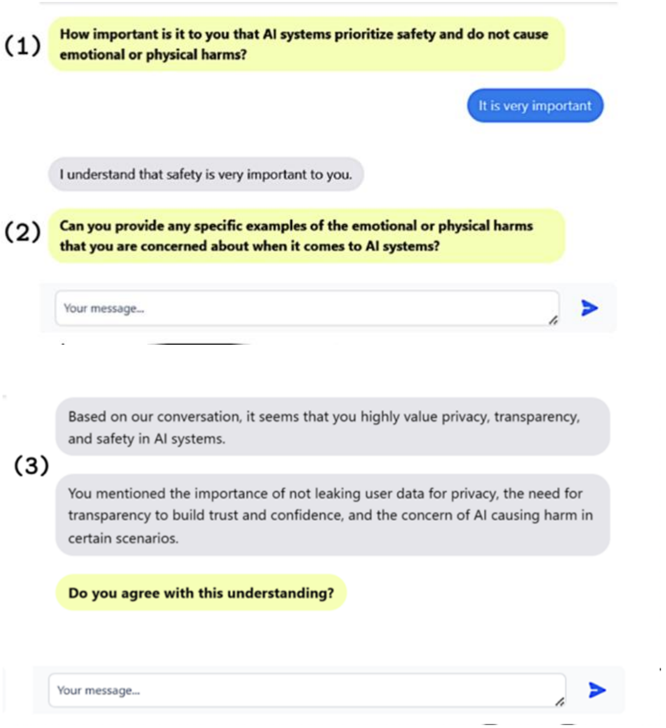
Reported performance of multiple models and methods on the MMLU benchmark (Nori et al. 2023b)

Microsoft study: Nori, H., et al. (2023a) Capabilities of GPT-4 on Medical Challenge Problems. arXiv preprint.
Microsoft study: Nori, H., et al. (2023b) Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Microsoft blog.

LLMs will change the way social science research is done

LLMs can rapidly analyze data from humans and generate synthetic data to accelerate science in new ways

- Early work suggests that LLMs respond to surveys and economic games similarly to humans, directionally and sometimes in magnitude (Argyle et al. 2022, Horton 2023, Brand et al. 2023).
  - These findings open new opportunities to test hypotheses on simulated data prior to experimenting with humans.
  - They also raise new questions about the meaning of LLM-generated survey data: How to conduct statistical analysis? How to validate the of analysis on such synthetic data? How to combine data from humans with data from LLMs?
- LLMs may accelerate the collection and analysis of non-quantitative data from human subjects through expanded text processing capabilities that facilitate near-real-time sensemaking or even interacting directly with human participants as an interviewer or other conversational aid (Chopra & Haaland 2023; Vilalba et al. 2023).
- LLM-based Code Interpreter from OpenAI makes preliminary data analysis accessible even to people without data science or statistical training.



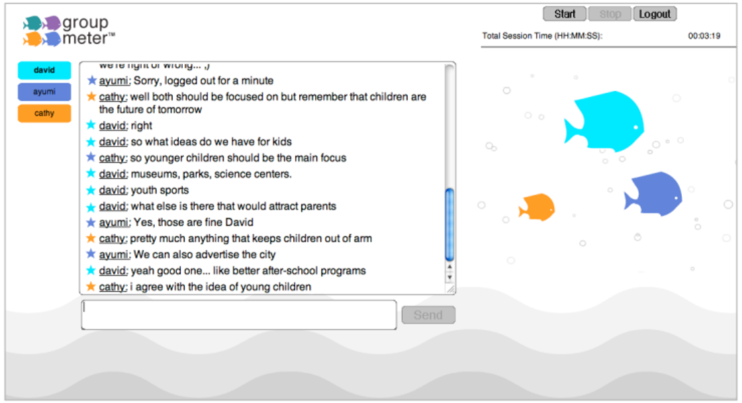
A screenshot of automated alignment conversations: A multi-agent system enables adaptive surveys in which an LLM is used to generate follow-up questions and a conversation summary for participant review (Villalba et al. 2023)

Argyle, L.P., et al. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3).
Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER preprint.
Brand, J., et al. (2023). Using GPT for market research. SSRN preprint.
Chopra, F., & Haaland, I. (2023). Conducting Qualitative Interviews with AI. SSRN preprint.
Villalba, A.C., et al. (2023).Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models. arXiv preprint

Instant AI feedback may improve real-time interactions in meetings

LLMs might be able to solve endemic problems with real-time interactions at work – e.g., encouraging more equal participation in meetings when doing so is valuable – but more research is needed to figure out how to minimize cognitive load and fit to team dynamics

- Monitoring and displaying participation and agreement rates during meetings can encourage more equal participation and higher agreement, respectively (DiMicco et al. 2007; Samrose et al. 2017; Leshed et al. 2009).
    - However, equal participation isn't always optimal; if an expert is present, it may be preferable to let them contribute more. Similarly, more agreement isn't always more productive, and could attenuate engagement with critical and creative tasks.
- Researchers have developed prototypes that delivered feedback on the level of engagement and information exchange in a meeting (Tausczik & Pennebaker 2013).
    - Only teams with low levels of information exchange objectively benefited from the feedback. This suggests that feedback should be tailored to specific teams' meeting dynamics.
    - Displaying both types of feedback resulted in worse outcomes, suggesting cognitive overload. With limited capacity to digest instantaneous feedback, the system must be precise in both the content and quantity of feedback.



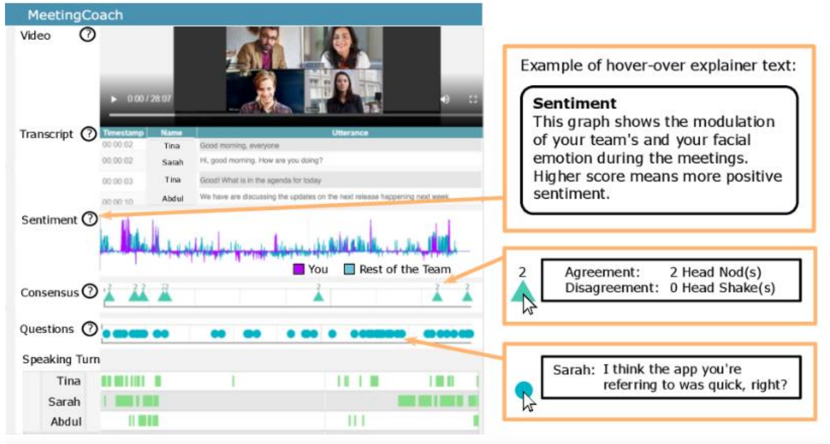
Real-time interface for displaying feedback on agreement and participation (Leshed et al. 2009)

DiMicco, J. M., et al. (2007) The Impact of Increased Awareness While Face-to-Face, Human–Computer Interaction, 22:1-2.
Samrose, S., et al. (2020). Immediate or Reflective?: Effects of Real-time Feedback on Group Discussions over Videochat. arXiv preprint.
Leshed, G., et al. (2009). Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09).
Tausczik, Y. R., & Pennebaker, J. W. (2013). Improving teamwork using real-time language feedback. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).

Retrospective AI feedback may improve long-term meeting interactions

Retrospective AI feedback on meetings must be delivered in a way that is actionable, engaging and personalised – it should seek to reduce the burden of reviewing meetings and may need to be incorporated into training

- After a meeting, team members may benefit from reviewing the information shared. Kim & Shah (2016) created a system that detected topic areas with poor shared understanding and recommended these areas for review. However, while the system increased shared understanding, participants did not perceive it to be helpful.
- Samrose et al. (2021) provided study participants with transcripts as well as measures of variables like consensus, questions, and time speaking. Users perceived the feedback as important for the team, suggesting feedback should be provided alongside actionable changes.
- In a busy work schedule, reviewing meetings is a time burden. A conversational interface could be more engaging, asking users about their teamwork, and making specific recommendations (Webber et al. 2019). Generative AI could deliver highly personalised feedback, in both content and delivery, enriched with pictures, videos, and music to support its message.



A post-meeting dashboard where participants can review their behaviours and those of others in the meeting (Samrose et al. 2021).

Kim, J., & Shah, J. A. (2016). Improving Team's Consistency of Understanding in Meetings. IEEE Transactions on Human-Machine Systems 46.5.
Samrose, S., et al. (2021). MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).
Webber, S., et al. (2019). Team challenges: Is artificial intelligence the solution? *Business Horizons*, *62*(6).
Ellwart, T., et al (2015). Managing information overload in virtual teams: Effects of a structured online team adaptation on cognition and performance. European Journal of Work and Organizational Psychology, 24:5.

AI may help leaders and teams plan and iterate on workflows

Workflow planning will benefit from AI’s ability to track task interdependence

- AI can help to allocate team member roles based on their present work schedules and their skill sets, attitudes, and actions (Sowa et al. 2020).
- AI can track how well task interdependence status is synchronized, measuring workload and redistributing the workload of individual team members to ensure that a team acts in a coherent manner (Khakurer & Blomqvist 2022).
    - Case 1: Train traffic control. An AI assistant could effectively measure and inform team members about their own and other team members’ workload, and effectively automate task delegation (Harbers & Neerincx 2017).
    - Case 2: Construction. ChatGPT generated a logical sequence of tasks, breaking down steps needed and handling dependencies among the proposed tasks (Prieto et al. 2023). Results suggested that AI-enabled tools could generate or enhance agendas based on project details, such as the scope of work a user provides. Not all the proposed tasks agreed with the scope of work, but ChatGPT showed promising performance and received positive user feedback (Prieto et al. 2023).
    - Case 3: Urban planning. With enough information about the project scope and the team, AI could effectively plan the workflow. However, collaborative planning platforms should integrate human feedback in the loop to refine workflow suggestions, offer alternatives, and balance multiple perspectives and considerations (Wang et al. 2023).
- AI help in delegating management responsibilities can be an effective form of human-AI collaboration (Hemmer et al. 2023), freeing management to focus on team vision.
- As AI becomes more prominent in workflow planning, it is critical to consider the possible externalities and challenges raised in the “algorithmic management” literature (e.g., Lee 2018).

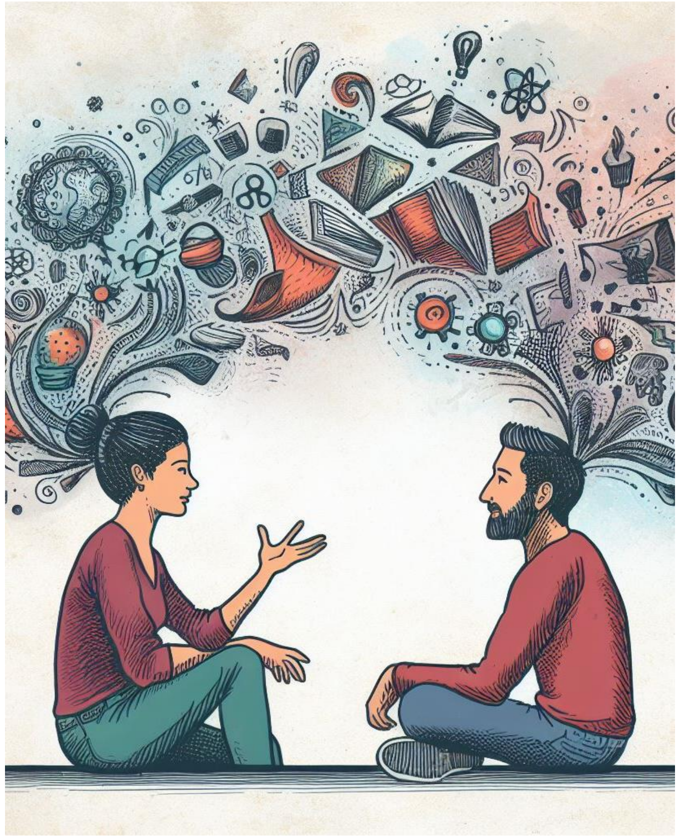
Workflow planning can benefit from AI’s ability to track task interdependence (Image credit: Bing Image Creator)

Sowa, K. ,(2021). Cobots in knowledge work: Human–AI collaboration in managerial professions. Journal of Business Research, 125.
Khakurel, J., & Blomqvist, K. (2022). Artificial Intelligence Augmenting Human Teams. A Systematic Literature Review on the Opportunities and Concerns. International Conference on Human-Computer Interaction.
Harbers, M., & Neerincx, M. A. (2017). Value sensitive design of a virtual assistant for workload harmonization in teams." Cognition, Technology & Work 19.
Prieto, S., et al (2023). Investigating the use of ChatGPT for the scheduling of construction projects. Buildings 13,. 4.
Wang, D., (2023). Towards automated urban planning: When generative and chatgpt-like ai meets urban planning. arXiv preprint.
Hemmer, P., et al (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. IUI 2023.
Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society. 5, 1.

Digital knowledge is moving from documents to dialogues

Knowledge is no longer only embedded in documents, spreadsheets, and text – it is now embedded in conversation and can be served up dynamically through that same medium

- Digital content historically has existed in the form of documents, but is increasingly captured in the form of conversations, be it via digitally mediated conversations between people or between people and an LLM.
- The knowledge embedded in these conversations can be leveraged by LLMs.
  - Facts from previous conversations may be directly surfaced at contextually appropriate times.
  - Past conversations can also be used for personalization.
  - Successful conversations can provide patterns for prompt engineering.
- *Grounding* is the process by which participants in a conversation come to a mutual understanding (Clark 1996).
  - Grounding conversations can lead to *grounded* content. For example, a brainstorming conversation may lead to the creation of a slide deck once everyone is on the same page.
  - Traditionally, grounded content is what people turn to for knowledge re-use. But with LLMs the grounding conversation itself can be re-used.
- Given how important conversations are for knowledge creation, additional research is needed on how to help people have great conversations, externalizing what they know and generating interesting new ideas.



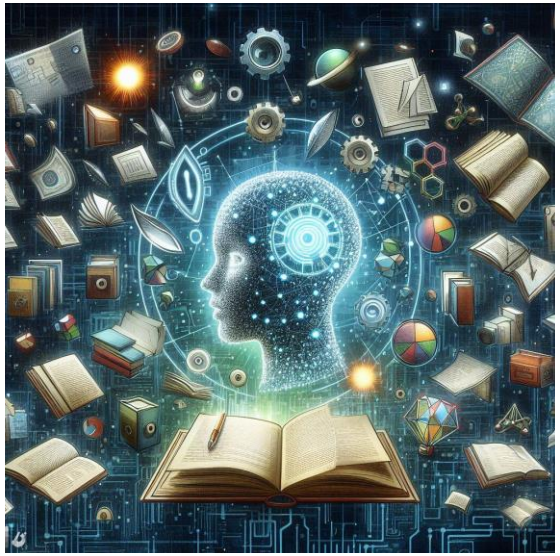
With LLMs mining transcriptions of conversations, conversations become shared and searchable knowledge (Image credit: Bing Image Creator)

Clark, H. H. (1996). *Using Language*. Cambridge University Press.
Microsoft Study: Teevan, J. (2023) From Documents to Dialogues. Generative AI: Hackathon Closing Ceremony, Carnegie Melon University.

LLMs may help address one of the greatest problems facing organizations: knowledge fragmentation

Organizational knowledge is fragmented across documents, conversations, apps and devices, but LLMs hold the potential to gather and synthesize this information in ways that were previously impossible

- Knowledge fragmentation is a key issue for organizations. Organizational knowledge is distributed across files, notes, emails (Whittaker & Sidner 1992), chat messages, and more. Actions taken to generate, verify, and deliver knowledge often take place outside of knowledge 'deliverables', such as reports, occurring instead in team spaces and inboxes (Lindley & Wilkins 2023).
- LLMs can draw on knowledge generated through, and stored within, different tools and formats, as and when the user needs it. Such interactions may tackle key challenges associated with fragmentation, by enabling users to focus on their activity rather than having to navigate tools and file stores, a behavior that can easily introduce distractions (see e.g., Bardram et al. 2019).
- However, extracting knowledge from communications raises implications for how organization members are made aware of what is being accessed, how it is being surfaced, and to whom. Additionally, people will need support in understanding how insights that are not explicitly shared with others could be inferred by ML systems (Lindley & Wilkins 2023). For instance, inferences about social networks or the workflow associated with a process could be made. People will need to learn how to interpret and evaluate such inferences.



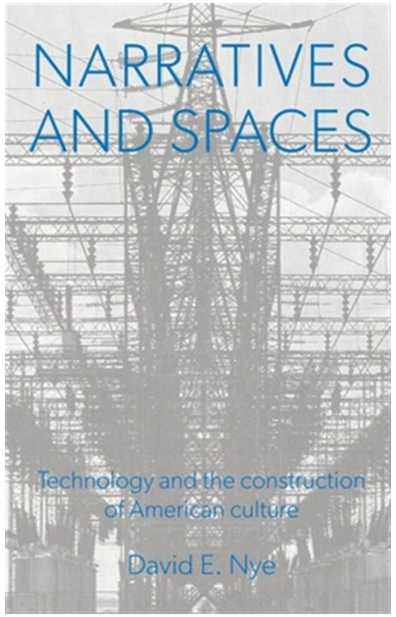
Fragmented knowledge could be pulled together with AI (Image credit: Bing Image Creator)

Whittaker, S., & Sidner, C. (1996). Email overload: exploring personal information management of email. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96).
Bardram, J., et al. (2019). Activity-centric computing systems. Communications of the ACM, 62, 8.
Lindley, S., & Wilkins, D. J. (2023). Building Knowledge through Action: Considerations for Machine Learning in the Workplace. ACM Transactions on Computer-Human Interaction 30, 5.

The introduction of AI into any organization is an inherently *sociotechnical* process

It's a two-way street – people influence technology just as technology influences people

- New technologies always land in contexts that are filled with meaning and expectation that shape whether and how technologies are adapted and with what consequences (Baym & Ellison 2023).
- David Nye's (1997) classic study of how Americans responded to the invention of electricity argues that interpretations fell on a spectrum from utopian hopes (ranging from world peace to modest life improvements) to dystopian fears (ranging from global destruction to more daily inconveniences). Contemporary discourses of AI dramatically increasing productivity or leading to human extinction can reflect the same sociotechnical interpretive dynamics.
- People in organizations do not always accept technologies that on the face of it seem to be improvements. Action research in British coal mines in the 1950s (Trist & Bamforth 1951) showed that understanding this resistance required understanding people, organizations, and technologies as part of a single *sociotechnical* system: "a web-like arrangement of the technological artefacts, people, and the social norms, practices, and rules" (Sawyer & Tyworth 2006, p. 51).
- An important implication is that new technologies, such as applications powered by LLMs, should be developed through participation with people in the contexts in which they will be deployed. "The rationale for adopting socio-technical approaches to systems design is that failure to do so can increase the risks that systems will not make their expected contribution to the goals of the organization" (Baxter & Sommerville 2011, p. 4)



Nye's classic text on technology and American culture

Baym, N., & Ellison, N. B. (2023). Toward work's new futures: Editor's Introduction to Technology and Future of Work special issue. Journal of Computer-Mediated Communication 28(4).
Nye, D. E. (1997) Narratives and Spaces: Technology and the Development of American Culture, New York: Columbia University Press
Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human relations* 4.1.
Sawyer, S., & Tyworth, M. (2006) Social informatics: Principles, theory, and practice. *Social Informatics: An Information Society for all? In Remembrance of Rob Kling: Proceedings of the Seventh International Conference on Human Choice and Computers (HCC7), IFIP TC 9.*
Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with computers* 23.1.

Knowledge worker perceptions of AI influence adoption

Effective adoption can also be influenced by how well AI tools fit workflows

- Perceptions of new technologies and knowledge workers' willingness to adopt them can be influenced by how they are used and discussed in workplaces. For example, early work in the Social Influence Model of Technology Use found that initially, perceptions of email's usefulness were influenced by how co-workers used and talked about the technology (e.g., Schmitz & Fulk 1991).
- Knowledge workers' ability to effectively adopt new technologies can also be influenced by how well the tools fit their workflows. Poor contextual fit means they might feel limited and lack the means or time to make an informed decision (Yang et al. 2019; Khairat et al. 2018). Human Factors research shows that disrupting domain experts' workflows can also limit their ability to apply their expertise (Elwyn et al. 2013; Klein, 2006) and decision-making strategies learned with experience (Sterman & Sweeney 2004).
- Knowledge workers form perceptions of AI systems and anticipate related workflow changes before using them. For example, Rezazade Mehrizi's (2023) ethnographic study of how radiologists interpret AI shows that even though most had not worked with technology, they co-constructed frames for understanding how it would shape their work, ranging from expectations that it would automate them away, to envisioning AI as likely to enhance or rearrange their work, to expecting that their work would become increasingly about communicating to the AI to make it work more effectively.

How AI tools are perceived by knowledge workers and whether they fit their work context can determine if they will be effectively adopted (Image credit: Microsoft stock image)

Schmitz J., & Fulk J. (1991). Organizational colleagues, media richness, and electronic mail: A test of the social influence model of technology use. Communication Research, 18(4).
Yang, Q., et al. (2019). Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems
Khairat, S., et al. (2018). Reasons for physicians not adopting clinical decision support systems: Critical analysis. JMIR medical informatics, 6, 2.
Elwyn, G., et al. (2013). Many miles to go…": A systematic review of the implementation of patient decision support interventions into routine clinical practice. BMC medical informatics and decision making, 13(2).
Klein, G., et al. (2006). Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems, 21(4).
Sterman, J. D., & Sweeney, L. B. (2004). Managing complex dynamic systems: challenge and opportunity for. In Henry Montgomery, Raanan Lipshitz, & Berndt Brehmer (Eds.), How professionals make decisions. CRC Press.
Rezazade Mehrizi, M. H. (2023). Pre-framing an emerging technology before it is deployed at work: the case of artificial intelligence and radiology, Journal of Computer-Mediated Communication, 28, 4.

Human-AI working: Monitoring and takeover challenges

Many jobs might increasingly require individuals to oversee what intelligent systems are doing and intervene when needed, however automation studies reveal potential challenges

- Monitoring requires vigilance, but people struggle to maintain attention on monitoring tasks for more than half an hour, even if they are highly motivated (Mackworth 1950). Studies with air traffic controllers show that vigilance requiring jobs can also lead to stress (Loura et al. 2013).
- An increase in automation can result in deterioration of cognitive skills that are crucial when automation fails, and human needs to take control (Bainbridge 1983). Automation also limits opportunities to develop problem-solving skills needed to critically evaluate the output of the system (Bainbridge 1983; Weiner & Curry 1980).
- Humans struggle to shift attention between manual and automated tasks (Wickens et al. 2007; Metzger & Parasuraman 2005), especially under high workload conditions (Janssen et al. 2019). This can interfere with their ability to effectively monitor and take control in cases of failure.
- When passively monitoring automation, humans have not historically used the freed-up time effectively. In semi-automated driving tasks, participants' attention shifted to unrelated activities, e.g., reading, which led to a delayed response if the vehicle failed (de Winter et al. 2014). Passive monitoring might also lead to increased distractedness and mind-wandering (Yoon & Ji 2019).

People struggle to maintain attention on monitoring tasks for more than half an hour, even when highly motivated (Image Credit: Bing Image Creator)

Mackworth, N. H. (1950). Researches on the measurement of human performance. *Medical Research Council Special Report*, No. 2680.
Loura, J., et al. (2013). Job stress in air traffic controllers: A review. *IJMSSR*, 2(6).
Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19.
Weiner, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, 23.
Wickens, C. D., et al (2006). Imperfect diagnostic automation: An experimental examination of priorities and threshold setting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 3.
Metzger, U., & Parasuraman, R., (2017). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. In *Decision Making in Aviation.*
Janssen, C. P., et al. (2019). History and future of human-automation interaction. *International journal of human-computer studies*, 131.
De Winter, J. C., et al. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation research part F: traffic psychology and behaviour*, 27.
Yoon, S. H., & Ji, Y. G., (2019). Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. *Transportation research part F: traffic psychology and behaviour*, 60.

We need to work to mitigate increased risk of “moral crumple zones"

Studies of past automations teach us that when new technologies are poorly integrated within work/organizational arrangements, workers can unfairly take the blame when a crisis or disaster unfolds

- Elish (2019) examined the history of autopilot in aviation. Some of her key observations were:
  - AI-supported autopilot systems were deemed "safer" than pilot-flown airplanes, but policymakers mandated pilots/copilots to be available "just in case" the machine failed.
  - Pilots were not trained for this new role and sometimes were ill-equipped to handle sudden hand-off when things went wrong.
  - Pilots became a “moral crumple zone”: Since pilots had to take over at the worst possible moments and struggled, they were often blamed for crashes.
- Elish’s work and others highlights the importance of building technologies that deeply engage with actual human capacity and of ensuring that an entire sociotechnical system works well in the context in which it is operated.
- As Elish writes, these findings highlight the importance of focusing on the true “value and potential of humans…in the context of human-machine teams”.

Engaging Science, Technology, and Society 5 (2019), 40-60 DOI:10.17351/ests2019.260

Moral Crumple Zones:
Cautionary Tales in Human-Robot Interaction

MADELEINE CLARE ELISH[1]
DATA & SOCIETY RESEARCH INSTITUTE

Abstract
As debates about the policy and ethical implications of AI systems grow, it will be increasingly important to accurately locate who is responsible when agency is distributed in a system and control over an action is mediated through time and space. Analyzing several high-profile accidents involving complex and automated socio-technical systems and the media coverage that surrounded them, I introduce the concept of a *moral crumple zone* to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system. Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions. While the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system, at the expense of the nearest human operator. The concept is both a challenge to and an opportunity for the design and regulation of human-robot systems. At stake in articulating moral crumple zones is not only the misattribution of responsibility but also the ways in which new forms of consumer and worker harm may develop in new complex, automated, or purported autonomous technologies.

Keywords
autonomous vehicles; responsibility; machine learning; human factors; accidents; social perceptions of technology; self-driving cars, robot; human-in-the-loop; human-robot interaction

Introduction
On March 18, 2018, a self-driving Uber car struck and killed a pedestrian crossing her bike in the middle of an Arizona roadway. At the steering wheel of the putative “autonomous vehicle,” a safety driver sat. Her job was to monitor the car’s systems and take over in the event of an emergency. The safety driver now may face criminal charges of vehicular manslaughter (Somerville and Shepardon 2018). A devastating accident has forced the question that had been

[1] Madeleine Clare Elish, Email: mcelish@datasociety.net

Copyright © 2019 (M.C. Elish). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at estsjournal.org.

Moral crumple zone paper (Elish 2019)

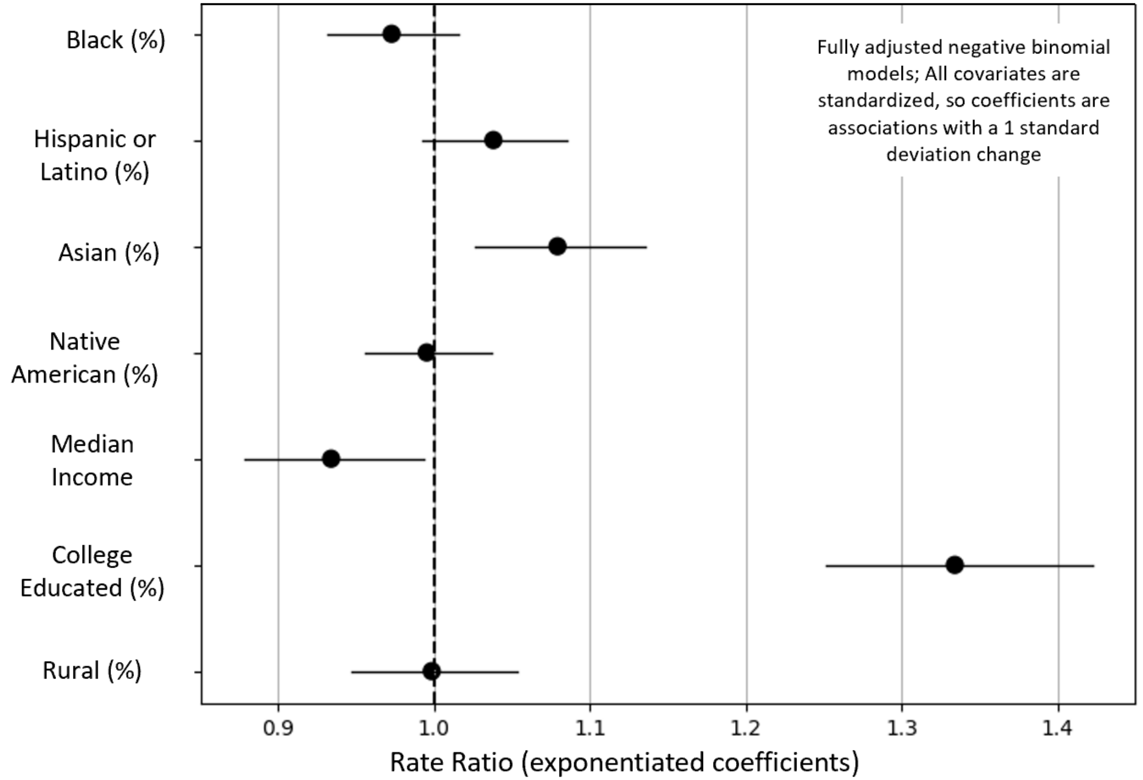
Elish, M. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society* 5.

Early evidence shows disparities in adoption follow traditional digital divide

Looking at searches in traditional Bing for "ChatGPT" or "Chat GPT" can show which counties have higher rates of interest

- Daepp (2023) looks at searches in traditional Bing for "ChatGPT" or "Chat GPT" and matches it with county-level demographic data.
    - Many more people are searching for these terms in counties where a higher share of people are college educated.
    - Such searches are also slightly more common in places with a higher percentage of Asians.
    - Perhaps surprisingly, the rate of searching is slightly negatively correlated with the county's median income.

(This analysis can't measure actual usage of ChatGPT, just the interest in it from people searching for it.)



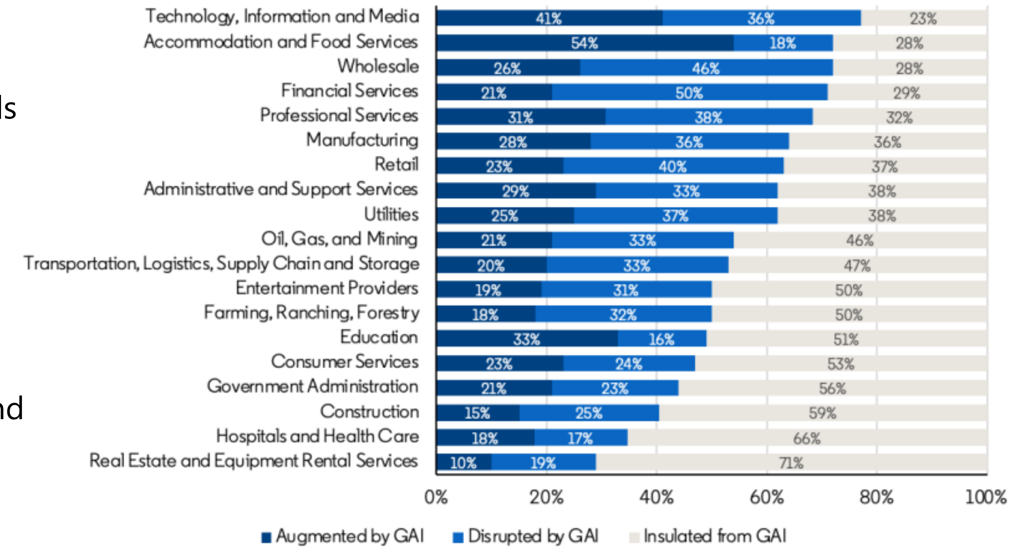
Association between rates of search for Chat GPT and a one standard deviation difference in county-level variables (Daepp 2023)

Microsoft study: Daepp, M. (2023). The Emerging AI divide in the United States. In progress.

Most jobs will likely have at least some of their tasks affected by LLMs

Many studies have used AI's current capabilities to try to measure where AI will have the most impact – either by making some people more productive or by replacing some roles

- A study by OpenAI found that approximately 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of GPTs (Eloundou et al. 2023).
  - Around 19% of workers may see at least 50% of their tasks impacted.
- A study by LinkedIn researchers categorized each job category by whether few of its associated skills will be impacted by AI (Insulated) or, if many of its skills will be impacted, whether it also has many skills that are complementary (Augmented) or does not have complementary skills (Disrupted). (Kimbrough & Carpanelli 2023, see graph).
  - Augmented jobs are particularly likely to see a shift in the composition of tasks workers do and the skills they rely on most.
- Research by Goldman Sachs suggests that organizations in Developed Markets may have more tasks exposed to AI than in Emerging Markets.
- However, the ultimate effects of new technologies on jobs are very hard to predict because they depend on how the technology is adopted. Historical examples show a wide range of possible effects:
  - Direct Distance Dialing technology almost entirely replaced the profession of switchboard operation in the 1930s. (Carmi 2015).
  - ATMs did not replace bank tellers, despite fears that they would. Instead, the jobs evolved - less time spent on basic tasks like counting bills, and more on complex customer issues (Bessen 2015).
  - Similarly, the introduction of basic chatbots in the early 2010's generated changes to jobs in the customer service industry, but did not eliminate them (CFPB 2022).



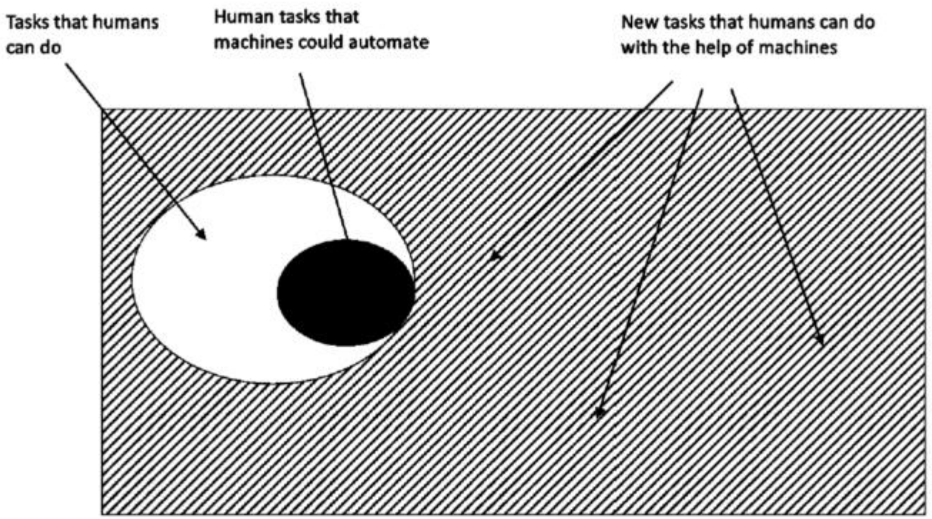
Share of LinkedIn members in occupations likely to be augmented, disrupted or insulated, by industry as calculated by the Linkedin Economic Graph Research Institute (Kimbrough & Carpanelli 2023)

Eloundou et al. (2023) GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint.
Kimbrough, K., & Carpanelli, M. (2023) Preparing the Workforce for Generative AI Insights and Implications. *Linkedin Economic Graph Research Institute*
Goldman Sachs (2023) The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)
Carmi, E. (2015). Taming Noisy Women: Bell Telephone's female switchboard operators as a noise source. *Media History*, *21*(3).
Bessen, J. (2015). Learning by Doing: The Real Connection Between Innovation, Wages, and Wealth. Yale University Press.
Consumer Financial Protection Bureau (2022). Chatbots in consumer finance.

Innovation is the secret sauce to job creation with new technologies

"Innovation vs. automation" is often a better framework than "augmentation vs. substitution"

- Over time, new technologies have helped create billions of new jobs and new types of jobs (e.g., train conductors, switchboard operators, computer programmers).
  - This is a mechanism by which technology has raised living standards (Acemoglu 2023; Koyama & Rubin 2022).
- While the net effect has been positive thus far, new technologies have also substituted for many types of human labor (e.g., stable hands, switchboard operators, human calculators).
- A technology that only substitutes for existing labor can only increase productivity by so much. To paraphrase Brynjolfsson (2023), if the ancient Greeks had invented something that automated all of the labor that existed in their time, no one would have to work, but everyone would still be using latrines and they wouldn't have vaccines.
- A key factor to ensuring that a new technology creates more jobs than it costs and can unlock massive productivity gains is *innovation*: what new things can the new technology allow us to do that we couldn't do before? What new, more productive uses of human labor does it create?
- In this respect, "innovation vs. automation" is often a better framework to use than "substitution vs. augmentation"
  - Augmentation will still substitute for human labor if there is not enough demand in the market for a lot more output of an existing task. If there is a lot of unmet demand, a technology that makes people more productive at an existing task can help meet that demand. If there isn't, it can mean fewer people are needed working on that task.
- While harder to measure, it is important to try to track whether and where human labor is being used in innovative new ways.



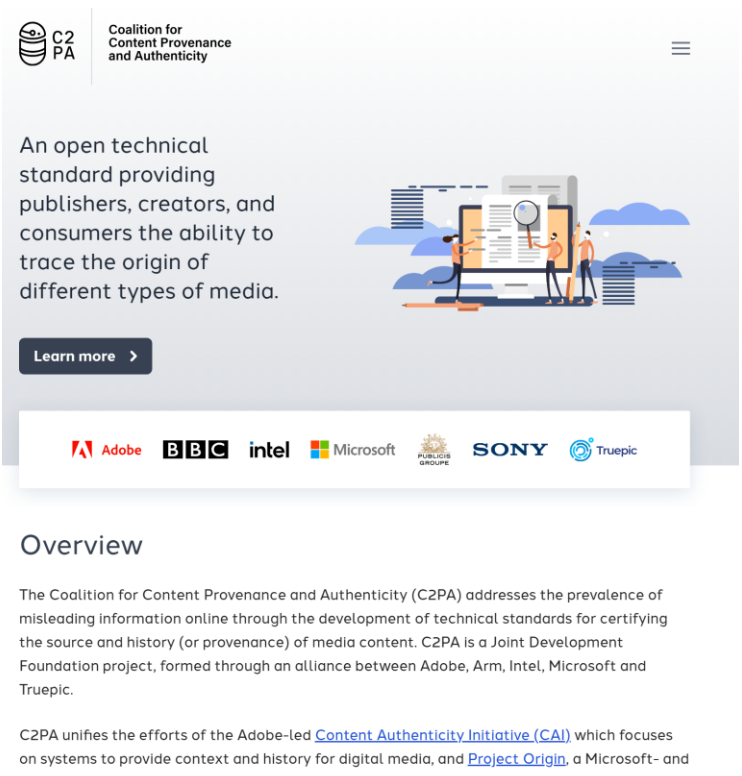
A graphic depicting some of the themes on this slide from Brynjolfsson (2023)

Acemoglu, D., & Johnson, S. (2023) Power and Progress: Our Thousand-year Struggle Over Technology and Prosperity. PublicAffairs
Koyama, M., & Rubin, J. (2022) How the World Became Rich: The Historical Origins of Economic Growth. John Wiley & Sons.
Brynjolfsson, E. (2022) The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. Daedalus.

The future of work is a choice, not a predetermined destiny

Instead of "How *will* AI affect work?", the question should be "How *do we want* AI to affect work?"

- Despite the way people sometimes talk about innovation, it is not a natural force; it is largely the product of societal factors, all of which are within human control (e.g., Bijker et al. 2012).
- As was the case for hybrid work, it is often important to reframe predictive questions about AI's relationship to work into questions about values and strategic goals (e.g., Weyl 2022). Rather than "What *will* the future of work look like?", we should ask "What *do we want* it to look like?".
- Several major actors in AI have stated what they think the future of work should look like, including in OpenAI's charter and Microsoft's Copilot vision.
- The scientific literature suggests that achieving many goals regarding the future of work and AI will require joint action across and within model builders, people who use models, and people who create content that is used by models (e.g., Vincent & Hecht 2023).
- If we anticipate problems emerging at the intersection of technology, work and who they benefit, it is almost always within the ability of humans – collaborating together – to fix those problems (Hecht et al. 2018).
- Some examples of coalitions in which Microsoft is involved that are tackling key problems include the Coalition for Content Provenance and Authenticity, the Biden-Harris administration's voluntary AI commitments, and Microsoft partnership with the AFL-CIO.



The C2PA is one coalition Microsoft is involved in to help address key challenges raised by LLMs.

Bijker, W. E., et al. (2012). *The Social Construction of Technological Systems, anniversary edition: New Directions in the Sociology and History of Technology*. MIT Press
Weyl, E. G. (2022). Sovereign Nonsense. *RadicalxChange*.
Vincent, N., & Hecht, B. (2023). Sharing the Winnings of AI with Data Dividends: Challenges with "Meritocratic" Data Valuation. *EAAMO '23* (2023).
Hecht, B., et al. (2018). It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. *ACM Future of Computing Blog*.

Call to action: Lead like a scientist

Science can provide insight about how to lead in this time of significant change

- We are all going through a period of rapid learning and growth. Fortunately, there's a model for that: Science. Leaders can take insight from the scientific process.
- This means developing a hypothesis and metrics, then doing the experimentation to test the hypothesis.
- It also means learning from existing knowledge. While LLMs appear very new, as demonstrated in this report there is great deal that is already know about them. We must build on the state-of-the-art to keep pushing forward.
- Sharing what we learn gives others something to build on and creates the opportunity to validate results. We must be open to debate about the best way forward.
- Science can also help us consider the externalities we create as we develop new norms, embed new tools, and change how we work.

Using scientific principles on building on current knowledge, testing a hypothesis and validating results, we can build a new equitable, productive and inclusive future of work with AI (Image Credit: Bing Image Creator)

Teevan, J. (2023) From Documents to Dialogues. Generative AI: Hackathon Closing Ceremony, Carnegie Melon University.