An MIT Exploration of Generative AI • From Novel Chemicals to Opera

Implementing Generative AI in U.S. Hospital Systems

Ben Armstrong[1] Kate Kellogg[2] Retsef Levi[3] Julie Shah[4]
Batia Wiesenfeld[5]

[1]MIT Industrial Performance Center, [2]Management and Innovation, Sloan School of Management,
[3]Operations Management, Sloan School of Management,
[4]Department of Aeronautics and Astronautics & CSAIL,
[5]Management, Stern School of Business, NYU

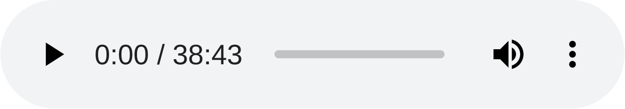
MIT

Published on: Mar 27, 2024

DOI: https://doi.org/10.21428/e4baedd9.1729053f

1. Introduction

There has been widespread optimism that artificial intelligence (AI) applications can transform medical care, improving patient treatment and reducing administrative burdens for hospitals and clinicians.[1],[2],[3] For patients, a healthcare system augmented by AI could mean less wait time due to optimal scheduling and resource allocation and higher-quality diagnostic and treatment decisions due to AI-driven capabilities, such as anomaly detection (e.g., in radiology), risk stratification, and personalized care. For clinicians, AI tools promise to reduce time spent on administrative tasks linked to burnout.



▯ Listen to this article

However, there is also a legacy of challenges when new technologies, and specifically AI-enabled software, are introduced in clinical settings. Past research has underlined barriers to technology adoption when the clinical workforce is insufficiently involved as well as trust and safety challenges when experts are prompted to interact with new technological systems.[4] Another major challenge is the integration with legacy control systems and specifically electronic health record (EHR) systems, which hospitals have invested in making a hub for their IT infrastructure. In fact, past attempts to disseminate new technologies into healthcare systems have often resulted in unintended consequences, such as an increase in the administrative burden on physicians and clinical teams.[5] Moreover, despite several successful proof of concepts and prototype pilots, the number of large-scale field implementations of AI-enabled software within healthcare systems is still relatively low.[6],[7]

As hospitals begin to experiment with generative AI (GAI) tools in clinical settings, it is important to identify known challenges that might stand in the way of achieving better patient care and lower administrative burdens—and, where possible, document practices that can mitigate those challenges. Additionally, it is important to identify areas in which further academic and industry research is needed to better understand the challenges and potential mitigation strategies.

The discussion around the implementation of AI-enabled software within healthcare systems should distinguish between traditional AI models and algorithms, hereafter called Narrow AI (NAI), and the more recent GAI models and algorithms. Although both applications are technically prediction algorithms, the tools have different technical characteristics, which lend themselves to different use cases, different user experiences, and different implications for organizations.

There are at least three major differences between the two. First, NAI models and algorithms are typically built for a specific prediction task (e.g., cancer detection on mammograms).[8] In contrast, GAI tools are typically based on large language models (LLMs) and are capable of performing a wide variety of tasks, such as search, summarization, and text generation tasks (e.g., patient visit note summarization). Second, NAI models and algorithms are typically developed based on a well-defined and labeled dataset specific to the target prediction task. On the other hand, corresponding to their broad functionality, GAI tools require much larger and broader datasets. Third, unlike the output of NAI models, which is typically very structured, the output of GAI models is often complex and unstructured (e.g., newly created text). We examine the challenges facing NAI and GAI applications in healthcare systems from three perspectives: technical, organizational, and cognitive (see summary in Figure 1). Our analysis draws on past studies of NAI applications as well as public reporting, interviews, and early observations of GAI applications in large research hospital settings.

The potential technical, organizational, and cognitive challenges associated with NAI and GAI tools can inform where these tools can be most effective. Organizations exploring new applications of NAI and GAI frequently perform assessments—often scorecards—to determine whether a particular use case is a fit for AI. [9],[10] Hospitals assessing the potential value of AI should grapple with these challenges as they design, adopt, and measure the performance of AI applications.

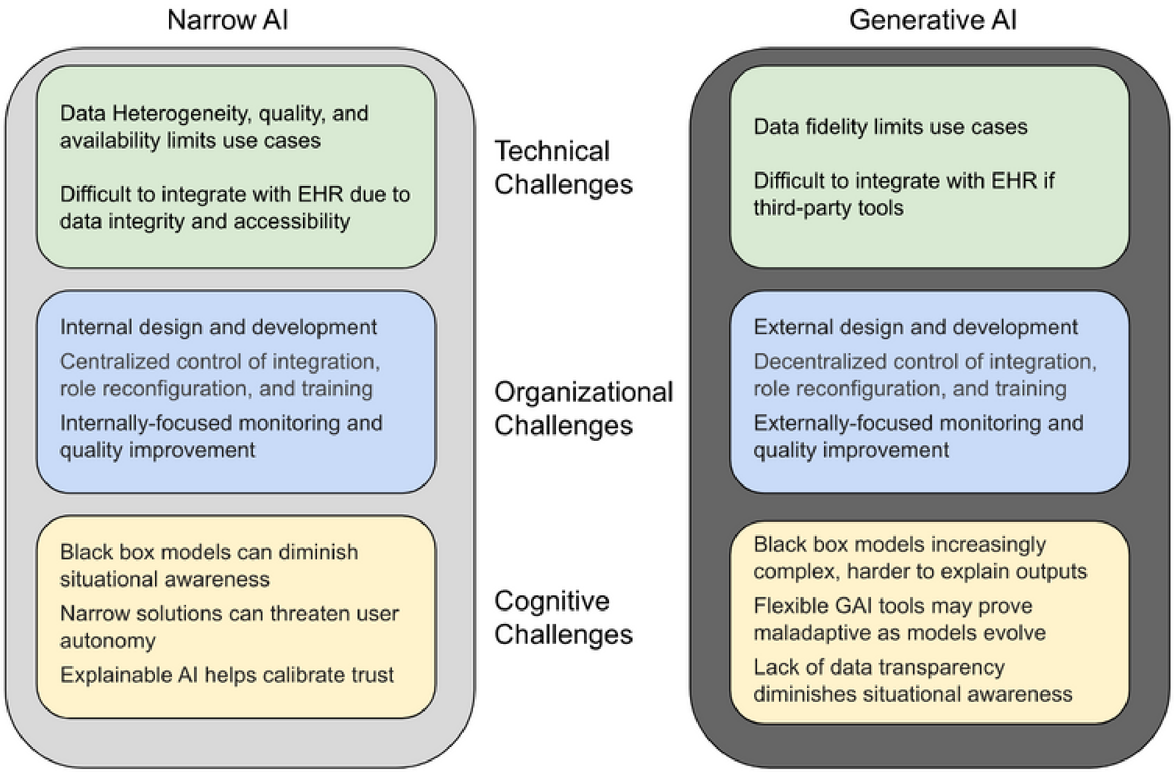

Narrow AI
Generative AI
Data Heterogeneity, quality, and availability limits use cases
Difficult to integrate with EHR due to data integrity and accessibility
Technical Challenges
Data fidelity limits use cases
Difficult to integrate with EHR if third-party tools
Internal design and development
Centralized control of integration, role reconfiguration, and training
Internally-focused monitoring and quality improvement
Organizational Challenges
External design and development
Decentralized control of integration, role reconfiguration, and training
Externally-focused monitoring and quality improvement
Black box models can diminish situational awareness
Narrow solutions can threaten user autonomy
Explainable AI helps calibrate trust
Cognitive Challenges
Black box models increasingly complex, harder to explain outputs
Flexible GAI tools may prove maladaptive as models evolve
Lack of data transparency diminishes situational awareness

Figure 1
Comparing narrow and generative AI: three perspectives.

2. Technical Challenges Related to the Use of NAI versus GAI

Although there have been dramatic leaps in the performance of AI models, it is not guaranteed that these technologies will have a transformative impact on the healthcare system. The potential for new technologies to make a positive difference within the healthcare system depends on their ability to augment specific processes within the broader system and be integrated effectively with legacy IT systems, particularly with EHRs. For example, can new technologies make an existing clinical process faster or less susceptible to error? Might new technologies allow a process to be re-engineered so that it facilitates better performance for clinicians, patients, or other stakeholders? The impact of AI technologies within healthcare systems should be considered in the context of the quadruple aims: enhancing patient experience, improving population health, improving the work life of healthcare providers, and reducing costs.[11]

2.1. The Process Environment in Healthcare

New technologies must work within the healthcare system's process environment. Figure 2 provides a high-level description of the typical process environment of healthcare systems. It consists of system (macro) processes and individual (micro) processes. System processes or protocols capture the way work is supposed to be conducted and is primarily affected by centralized decisions at the practice, department, or hospital level. However, in most scenarios, the system processes do not fully prescribe to the individuals how work should be done, which leaves room for individuals and teams to develop microprocesses at their discretion. EHR systems have become the primary control systems for clinical administrative processes and the hub for clinical data.

Consider the example of a primary care clinic. The appointment system and other related processes are typically decided at the practice level or even at the department level. As part of these processes, MDs and NPs are expected to see patients for specified appointments, treat them, summarize the visit, and submit the appropriate coding for billing. However, the manner by which an MD interacts with a patient within an appointment as well as summarizes the visit is typically left to the discretion of the MD, creating a situation in which there are many personal processes that could vary significantly. These clinical and administrative processes not only use data from the EHR but at the same time create a lot of the data, particularly text data. This leads to an inherent data heterogeneity within and across healthcare systems.

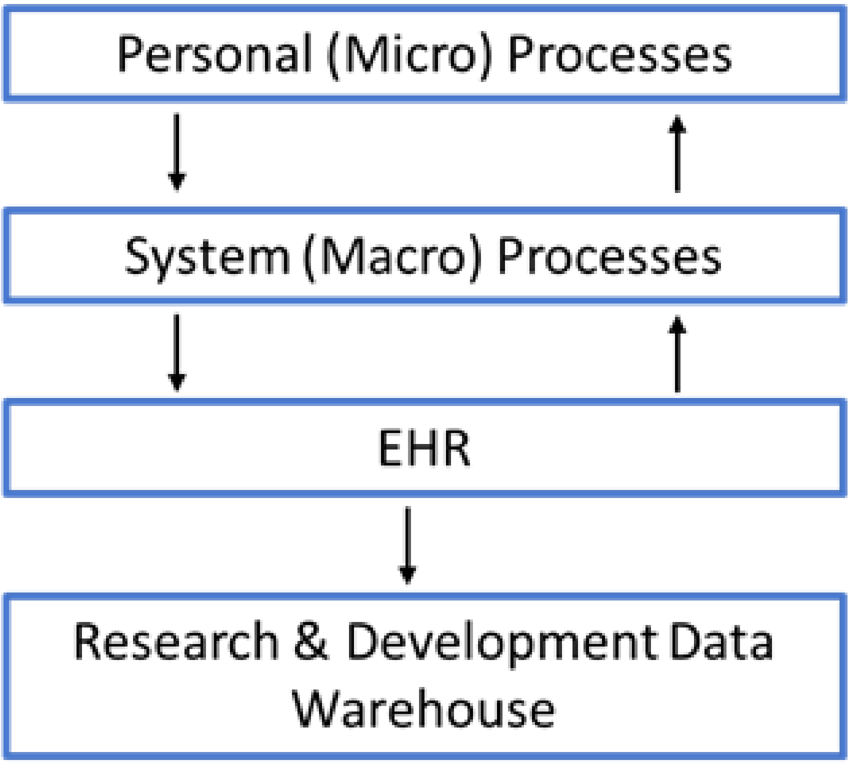

Personal (Micro) Processes
System (Macro) Processes
EHR
Research & Development Data Warehouse

Figure 2
Process and data environment.

2.2. NAI and the Data Heterogeneity Challenge

Over the past decade, experimentation with NAI and machine learning (ML) technologies in clinical healthcare settings has grown as part of efforts to improve clinical efficiency and diagnostic accuracy, among other objectives. The results of these experiments have been mixed: Some studies have shown quality improvements using AI technologies to assist with diagnostic and scheduling tasks,[12] whereas others point out resistance to the new technologies and low uptake of more complex tools.[13]

One of the key technical challenges for introducing and scaling NAI tools is the data infrastructure on which AI applications must be built. For example, if a hospital system sets out to develop even a simple algorithm to predict how long a patient will stay in the hospital given their medical record, there will be several barriers.

First, EHR data is typically not available at scale, which makes it challenging for developers to build AI/ML-enabled applications and tools that require access to training data. To overcome this problem, many large healthcare systems have developed research and development data warehouses that periodically (typically once a day) receive data dumps from the EHR. The purpose of this data warehouse is to enable clinical research, including the development of AI/ML algorithms and models. However, the underlying data scheme of the data warehouse is significantly different from the ones of the real-time operational data system of the EHR, which poses significant challenges to disseminate developed prototype models and algorithms.[14],[15]

Second, given the variability in micro (individual) processes and how clinicians record and store data, a significant portion of the patient's record (e.g., text data) may not be useful for training NAI and ML models. The data heterogeneity and accessibility problems also limit the types of NAI applications that a hospital can build. For example, hospitals may wish to use text-based data from a patient's visit notes to understand how patient sentiments during a visit predict other health outcomes. However, the heterogeneity of how clinicians file their text notes makes it challenging to develop a model with those data.[16]

2.3. GAI and the Data Fidelity Challenge

Hospitals and clinicians have highlighted the promise of GAI to address the data heterogeneity problem and unlock insights from text data and other data sources that have been overlooked. For example, early applications of GAI in clinical settings include inbasket messaging, in which clinicians receive an AI-generated response to patient messages that they can edit if they wish.[3],[17] Another early use case is focused on patient note summarization, in which a doctor's notes on a patient visit are captured by recording, automatically transcribed using NAI tools, and summarized using GAI tools.[18],[19],[20] In both cases, GAI tools employing LLMs have the potential to transform and generate unstructured text data in a more consistent way.

These GAI applications unlock new territory where AI applications can provide significant benefits, but technical challenges remain. For one, GAI models can produce text responses and summaries for heterogeneous data. For example, LLMs could in principle be able to summarize two clinicians' separate patient notes, even if the clinicians' styles were very different. However, it is not immediately clear to what extent GAI tools would be able to maintain differences in style that clinicians find important. GAI suggestions for how to respond to a patient inquiry have scored high in terms of their quality and empathy in one study;[21] however, physicians with a particular style might reject responses that do not match their preferred language. Overall, it seems like these new tools have the flexibility to be useful for processes with high variability at the clinician or office level.

A second, well-documented challenge with GAI models is with their data fidelity. When an LLM summarizes text or responds to a question, it might draw on irrelevant data to generate its response, a problem sometimes referred to as hallucination.[22],[23] The data fidelity problem represents a stark contrast from NAI applications, where the application developers can choose the specific data on which an algorithm is trained and interpret the results based on the training data. For LLM responses, it is far less clear how to evaluate the results the model provides without comparing them to some representation of the results that the model ideally would provide.[24],[25] Due to this challenge, early centralized applications of GAI in hospitals have been deployed in relatively lower-risk settings where accuracy and data fidelity are not critical. In the patient note summarization case, for example, the full patient transcript is available, so even if a note contains inaccuracies, there is a 'ground truth' maintained in the note. In the inbasket messaging case, there are also safeguards. If a clinician opts to use a GAI message, they are prompted to edit the message before sending. Even with these safeguards, there remain organizational and cognitive challenges and tradeoffs associated with GAI applications.

3. Organizational Challenges Related to the Use of NAI versus GAI

Addressing technical challenges associated with the development and implementation of GAI solutions is a necessary, but not sufficient, step in allowing healthcare organizations to achieve the efficiency and productivity improvements promised by GAI. Critically, healthcare leaders will encounter and must address organizational challenges that will likely arise during GAI solution development, implementation, and maintenance.

In particular, three key technical characteristics that differentiate GAI from NAI—(1) large data and computational requirements that make local design and development impractical, (2) flexibility that allows decentralized users to easily implement GAI solutions to perform a wide variety of tasks, and (3) and a rapidly changing software, hardware, and regulatory environment—demand new organizational approaches. These include a shift from internal to external design and development, from centralized to decentralized control, and from internally to externally focused monitoring and process improvement.

3.1. Development: From Internal to External Design and Development

First, while NAI depends on models and algorithms built on well-defined datasets using moderate computing power, GAI depends on LLMs that require significantly larger data and higher computational capabilities that make local design and development by one system likely impractical.[25] Thus, healthcare organization leaders must move from facilitating internal to external solution design and development. This presents an organizational challenge because GAI solution development increasingly depends on expertise in project procurement and adoption that is held by IT department members rather than on clinical and technical expertise held by clinical business unit leaders and AI center of excellence (COE) developers. As a consequence, leaders will need to manage tensions related to cost and standardization that are likely to emerge between these stakeholder groups.

For example, during NAI solution development, healthcare leaders can more easily facilitate the internal development of AI models using a back-and-forth development process between AI COE developers and domain expert champions.16 Developers work closely with a small set of domain experts to identify important use cases; identify, construct, and clean internal historical data; choose the best AI models for each use case; and develop the models.[26]

In contrast, during GAI solution development, because the investment in data and hardware is usually too high for one system to take on, healthcare leaders are more likely to buy access to externally developed solutions. To do this, IT department members must contract with external vendors to procure and adopt third-party technologies, as they did during the acquisition and implementation of EHRs.

For example, at NYU Langone, IT department leaders adopted an inbox messaging solution developed by Epic and Microsoft to draft responses to patient medical advice requests.[21] Epic developers determined the types of patient messages for which draft responses would be generated, the interface through which they were shown to users, and the pricing of the solution. The interface did not allow individual users to engineer their own prompts to elicit messages customized to their needs. Drafts had consistent flaws limiting their promised benefits, and improvements by Epic were gradual.

For example, wordy responses take as much time for healthcare professionals to read and edit as would be required to write the response themselves. COE developers needed to edit Epic's standard prompts to improve them, adding to the cost of generating drafts. These costs would be prohibitive if used for the many millions of patient messages received each year by the hospital system. Factors like these prompted NYU Langone to begin development of its own LLM for "healthcare-specific tasks," using its own proprietary dataset of health records and clinical notes.[27],[28] This example illustrates a more general concern of market consolidation of GAI solutions for healthcare systems, and specifically market control by EHR vendors.

3.2. Implementation: From Centralized to Decentralized Control of Workflow Integration, Role Reconfiguration, and Training

Second, while NAI solutions are developed for a specific purpose, frequently delivered into clinician workflows through the EHR, and often resisted,[29] GAI solutions have the flexibility to perform a wide variety of tasks, can be delivered via software directly to the user's computer, and are remarkably easy to use. On the one hand, this means that GAI solutions allow for greater flexibility to accommodate heterogeneity in individual processes. On the other hand, this presents an organizational control challenge because end users distributed across the organization with little technical training can acquire the ability to use GAI solutions with much less need for or dependence on centrally positioned AI COE technology developers. As a consequence, leaders will need to manage problems related to increasingly heterogeneous processes, including tensions around workflow integration, role reconfiguration, and training.

For example, during NAI solution implementation, internal technology developers work closely with a small set of domain expert champions to map workflows, repair models according to champions' feedback, identify new roles required to most efficiently implement the new technology, and develop and implement training for targeted end users.[30] Typically, the developed solutions enable a new system-level process design but do not allow for the use of heterogenous individual processes.

In contrast, during GAI solution implementation, lead users experiment around where to bring AI solutions into the workflow, share expertise and insights with others around how to use the solutions, and reconfigure roles in interaction with team members to most effectively use them. These solutions typically accommodate individual processes but require oversight to effectively ensure safe and reliable output.[24]

For example, at NYU Langone, doctor lead users are experimenting with GAI to draft letters repealing insurance company denials of payment and to identify the right Current Procedural Terminology codes necessary for billing. They found that these tasks could be done better and faster with the help of GAI and shared their experiences and emergent best practices with colleagues informally. In another example, administrative lead users in the Registration department are developing patient-facing registration chatbots. Instead of implementing systematic workflow integration and structured training, NYU Langone's AI COE has been supporting this decentralized implementation by offering open office hours to support lead users with basic skills and safety guidelines[31] and conducting a series of prompt-a-thons within departments to help facilitate bottom-up solution development and sharing.[32]

3.3. Maintenance: From Internally to Externally Focused Monitoring and Process Improvement

Third, while the external technologies and regulations underpinning NAI have changed incrementally, the pace of change around GAI has been far more rapid. Thus, a third challenge for healthcare organization leaders is the need to change focus from monitoring of internal operations to scanning the external environment to improve model performance and service quality. This presents an organizational challenge because internally focused quality improvement, legal, and Institutional Review Board team members may lose in power to externally focused new product and service scouters. As a consequence, tensions related to innovation and governance are likely to emerge.

For example, during NAI application maintenance, because internal changes in data and system infrastructure more rapidly affect the performance of NAI models than do external changes in software and hardware technologies and regulations, healthcare organizations often maintain AI solutions using an internally focused process to improve current operations.[33] For example, at NYU Langone, internally focused governance units monitor the development and maintenance of AI solutions to ensure that the solutions are in compliance with existing regulations.

In contrast, during GAI solution maintenance, rapidly changing external software and hardware technologies and regulations can affect the performance of AI models. Thus, there is a real risk that healthcare systems will not have timely awareness of important changes that could affect the accuracy and relevance of the GAI solutions they are using. To try to manage this, healthcare organizations may maintain AI solutions using an externally focused environmental scanning process. For example, at NYU Langone, because prompts designed to address edge cases become less effective when the model has been retrained, computer scientists monitor the performance of externally sourced AI solutions to identify when a model has been retrained without disclosure. Langone leaders also partner with government affairs teams to create awareness of government agencies iteratively developing governance policies.

While AI solutions present increasing opportunities to transform medical care, they must be carefully integrated into existing healthcare organizations. At every step of AI development, implementation, and maintenance, key stakeholders make critical choices that shape how and how well AI solutions will work once they move from the laboratory out into the real world. With GAI, control over development shifts toward external vendors and IT department leaders; control over implementation shifts toward decentralized end users; and control over solution maintenance shifts toward technically savvy LLM hardware, software, and regulation scouters. The power shift toward these stakeholders is also reshuffling organizational priorities, creating tensions over costs, standards, and governance.

4. Cognitive Challenges Related to the Use of NAI versus GAI

At the level of the healthcare worker or clinical team, the growing use of NAI—and the early use of GAI—has highlighted known human–automation integration risks with historical parallels, such as aviation's introduction of cockpit automation. These risks include skill atrophy, miscalibration of trust in AI, and information overload. Moreover, the application of these tools in healthcare, where many decisions are high-stakes and ethically complex,[34] raises even further hurdles for deploying these tools effectively and responsibly. This section introduces a framework, The Situation Awareness Framework for Explainable AI (SAFE-AI), that was designed to provide best practices for maintaining situational awareness (SA) for operators of NAI applications. For GAI applications, the SAFE-AI framework can provide some guidance but is more limited considering the complexity and opacity of GAI models.

4.1. The SA Challenge with NAI

When healthcare organizations have adopted AI applications, they often deploy them as clinical decision support systems with a 'human in the loop,' as opposed to deploying the system to perform tasks autonomously.[34],[35] The human-in-the-loop framework combines individual judgment with recommendations from an algorithm to achieve better results than the algorithm or the individual could achieve on their own. This paradigm has been used in other high-stakes domains implementing automation like aviation.[34]

For human operators to be successful in these systems, past research has shown that SA is critical. SA in the canonical model[36] focuses on aligning information systems and processes to provide the human operator the right information to make sound decisions. There are three levels of SA: "the perception of elements in the environment within a volume of time and space (level 1), the comprehension of their meaning (level 2), and the projection of their status in the near future (level 3)."[36]

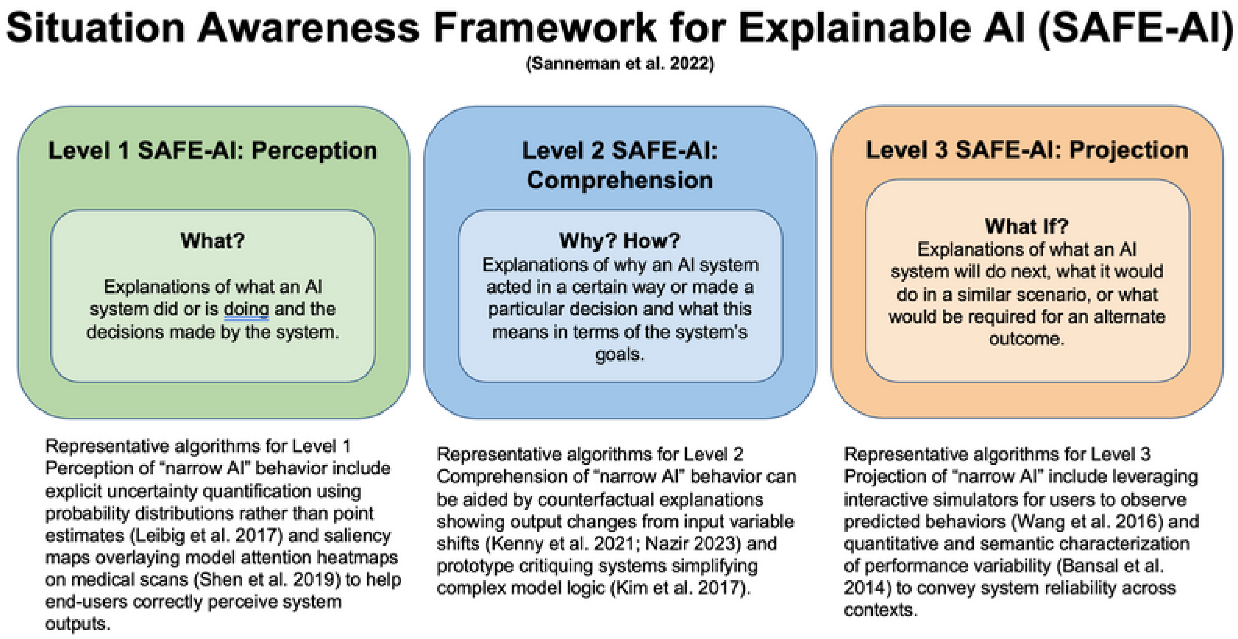
The SA framework originally received attention with the rise of cockpit automation and the need to support pilot awareness of aircraft behavior[37]and has expanded to other domains including healthcare.[38] The SA construct has empirical validation,[39],[40] with connections to performance and errors.[38]

A key cognitive risk of NAI tools is that they will diminish the SA of the healthcare professionals operating them. For example, consider a healthcare worker using an NAI application to manage discharges from the hospital. The worker's SA depends on their ability to explain what the NAI model recommended, why it made its recommendation, and what the model might recommend next as circumstances change. In this context, healthcare workers might require technical expertise on how the model was trained and how it might react in various circumstances. The healthcare worker's need for additional information to maintain their SA has informed the burgeoning field of explainable AI.[35],[36]

There are several ways in which the clinician's lack of SA could lead to costly errors. If the clinician begins over-relying on the model to manage discharges, their skills and intuition for the task could begin to atrophy. The clinician could also miscalibrate their trust in a black box NAI system because they cannot fully understand why the system is providing a given recommendation.[37],[38] Moreover, there are information overload risks[41], as AI-generated alerts can overwhelm the limited processing capacity of time-pressed staff, contributing to poor decision-making and burnout.

4.2. SAFE-AI to Mitigate Cognitive Risks

SAFE-AI[42] provides techniques for maintaining an individual operator's SA in an environment augmented by NAI tools. It builds on design principles integrating automation with human judgment[39] and explainable AI assessments.[36] Explainable AI aims to provide operators insights into how AI tools produce a particular output or recommendation, which is critical for all levels of SA. SAFE-AI can provide design guidance for healthcare systems deploying GAI tools, including the information that an operator might need to maintain their SA (Figure 3).



Situation Awareness Framework for Explainable AI (SAFE-AI)

(Sanneman et al. 2022)

Level 1 SAFE-AI: Perception

What?

Explanations of what an AI system did or is doing and the decisions made by the system.

Representative algorithms for Level 1 Perception of "narrow AI" behavior include explicit uncertainty quantification using probability distributions rather than point estimates (Leibig et al. 2017) and saliency maps overlaying model attention heatmaps on medical scans (Shen et al. 2019) to help end-users correctly perceive system outputs.

Level 2 SAFE-AI: Comprehension

Why? How?

Explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals.

Representative algorithms for Level 2 Comprehension of "narrow AI" behavior can be aided by counterfactual explanations showing output changes from input variable shifts (Kenny et al. 2021; Nazir 2023) and prototype critiquing systems simplifying complex model logic (Kim et al. 2017).

Level 3 SAFE-AI: Projection

What If?

Explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome.

Representative algorithms for Level 3 Projection of "narrow AI" include leveraging interactive simulators for users to observe predicted behaviors (Wang et al. 2016) and quantitative and semantic characterization of performance variability (Bansal et al. 2014) to convey system reliability across contexts.

Figure 3
Summarizing SAFE-AI

SAFE-AI was developed for NAI applications where explainable AI techniques are more readily available. However, these explainability techniques are more complicated to use for GAI applications since the datasets behind the model are vastly larger and more complex. For example, causal probing methods of LLMs like those developed by Tucker et al.[40],[43] allow operators to understand whether factors like syntax might affect model predictions. For example, they found that syntax does play a causal role in predicting masked words in sentences like "I saw the boy and the girl [MASK] tall," with the model prediction changing based on the ordering of the words. These techniques for understanding how a GAI model works at a granular level might not reveal why a model provides a given output each time but can help highlight the factors that the model weighs heavily in making a prediction, supporting goals for Level 1 SAFE-AI. However, there are limitations to the transparency afforded by such methods. Systematically probing many layers, representations, and counterfactuals poses scaling challenges in terms of computation and resources, and the techniques assume access to internal model representations, which may not be accessible.

Another explainability approach called "chain of thought reasoning" aims to support SAFE-AI Level 2 for GAI tools.[44] The method provides prompts to elicit a model's step-by-step 'reasoning.' Its goal is to expose the LLM's implicit inference steps that lead to its final output. This approach has the potential to help operators comprehend gaps and limitations in the LLM's logic.

However, the accuracy and consistency of the reasoning chains degrades rapidly beyond three to four steps and thus provides incomplete information regarding the GAI tool's reasoning. Analysis has also shown deficiencies in this approach when operators aim for the GAI tools to correct errors; LLMs fail to correct simple deductive errors via prompting.[43] The failure to incorporate human guidance on flawed inference chains undermines the potential for using chain of reasoning approaches in supporting SAFE-AI Level 3.

And even if GAI explainability methods did produce results, they might be challenging to interpret. Today's GAI tools commonly leverage neural networks and deep learning that remains inscrutable even to developers—unlike past explicit rules-based systems traceable for human operators. The opacity of GAI models amplifies risks of miscalibrated trust. In low-stakes applications, this might not be a problem. But for diagnostic applications in which a user of GAI must interpret the validity of the output, the black box nature of LLMs can prove challenging.

The flexibility of GAI tools raises an additional challenge. For human operators, it is difficult to understand where GAI tools will produce reliable results—and where they will not.[45] For example, even if an LLM produces a reliable result in early trials to respond to patient messages, it may prove maladaptive over time. AI systems continuously update behaviors via self-supervision on new data, significantly increasing unpredictability from a human operator standpoint.

Taken together, these risks are reason for caution and incremental adoption of GAI tools in high-stakes settings where situational awareness is at a premium. The deployment of traditional automation in high-stakes settings like aviation and manufacturing was deliberately limited in scope and bounded by risk considerations. In contrast, modern techniques have been deployed ambitiously with incentives to scale users quickly. This practice could encourage over-reliance on AI tools in ethically complex settings such as healthcare and end-of-life-care planning. Given pressures for rapid adoption of GAI tools, research can focus on identifying practices for mitigating known risks.

5. Open Questions and New Directions

The different technical, organizational, and cognitive challenges associated with NAI and GAI tools are grounded in recent evidence from early deployments and experiments with AI in hospitals as well as in longstanding evidence on the impact of new technologies such as EHRs. This mix of evidence allows us to propose potential challenges facing new applications of GAI, but it's too early to evaluate how and under what conditions these challenges will manifest in practice. This paper is both a map of potential roadblocks for healthcare organization leaders as well as a set of hypotheses to be tested in future empirical research at hospitals deploying GAI applications. We see new research directions related to the challenges in each section.

5.1. Managing Technical Risks

In past approaches to technology adoption, hospitals have used scorecards to evaluate applications of new software like AI solutions. These scorecards evaluate the complexity of automating a process as well as the potential financial and organizational benefits of the automation. Although we have seen some organizations adopt a scorecard approach to thinking about use cases for GAI, it is not clear how organizations are assessing the risks associated with potential hallucinations and lack of explainability as they are introducing GAI into clinical processes. What tactics will hospital systems use to manage these risks? How will they measure whether the risks are worth taking, particularly when hospitals may be faced with a high probability of productivity gains and a low probability of clinical errors.

5.2. Understanding Winners and Losers

Early laboratory and field evidence from non-healthcare fields has suggested that low-skilled workers have the most to gain when GAI tools are introduced.[46],[47],[48] This is in stark contrast to previous software technologies, which have often been biased in favor of high-skilled individuals, particularly those with college degrees. It is still an open question whether the lowest-skilled workers in healthcare and other fields will benefit most from the introduction of GAI. Even if lowest-skilled individuals improve their performance most when aided by GAI, it is not guaranteed that these individuals will benefit the most in terms of wages, career opportunities, and power within their organizations. It could plausibly be the inverse, where high-skilled individuals gain the most from the introduction of GAI. In this scenario, entrepreneurial individuals with extensive process knowledge and strong networks could use GAI tools to their professional advantage, whereas lower-skilled individuals might experience marginal productivity gains but not see wage or career advantages from GAI.

5.3. Identifying Changes in Performance

Past research raises concerns about the potential for skill atrophy associated with the introduction of high-quality automation. In cases in which the automation is too good—and the cognitive demands on the human operator are low—the human's capability of performing the task on their own and identifying potential errors is diminished. In healthcare settings, there is not clear evidence of skill atrophy or performance challenges linked to the introduction of GAI. However, the skill atrophy problem is a hypothesis to be tested and risk to be managed as organizations measure the impact of new GAI experiments on their workforce.

References

1. Topol, E. J. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. 1st ed. New York: Basic Books, 2019. ↩
2. Lee, P., C. Goldberg, and I. Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. 1st ed. Hoboken: Pearson Education, Limited, 2023. ↩
3. Lin, B. "Generative AI Makes Headway in Healthcare." *Wall Street Journal*, March 21, 2023. Accessed January 1, 2024. https://www.wsj.com/articles/generative-ai-makes-headway-in-healthcare-cb5d4ee2. ↩
4. Carusi, A., Peter D. Winter, Iain Armstrong, Fabio Ciravegna, David G. Kiely, Allan Lawrie, Haiping Lu, Ian Sabroe, and Andy Swift. "Medical Artificial Intelligence Is As Much Social As It Is Technological."

Nature Machine Intelligence 5, no. 2 (Feb. 2023): 2. doi:10.1038/s42256-022-00603-3. ↩

5. Escribe, Célia, Stephanie A Eisenstat, Kerri Palamara, Walter J O'Donnell, Jason H Wasfy, Marcela G Del Carmen, Sara R Lehrhoff, Marjory A Bravard, and Retsef Levi. "Understanding Physician Work and Well-being Through Social Network Modeling Using Electronic Health Record Data: A Cohort Study." *Journal of General Internal Medicine* 37, no. 15 (Nov 2022): 3789–96. doi:10.1007/s11606-021-07351-x. ↩
6. Topol, Eric J. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine* 25, no. 1 (Jan 2019): 44–56. doi:10.1038/s41591-018-0300-7. ↩
7. U.S. Census Bureau. "Annual Business Survey: Extent of Technology Use of Employer Firms by 2-digit NAICS for the United States and States: 2018." https://data.census.gov/table/ABSTCB2018.AB1800TCB01A?q=ab1800tcb%1A&hidePreview=false. ↩
8. Kissinger, H., E. Schmidt, D. P. Huttenlocher, and S. Schouten. *The Age of AI: And Our Human Future*. 1st ed. New York: Little Brown and Company, 2021. ↩
9. Kim, J. Y., et al. "Organizational Governance of Emerging Technologies: AI Adoption in Healthcare," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, in FAccT '23*. New York, NY, USA: Association for Computing Machinery, Jun. 2023, 1396–1417. doi:10.1145/3593013.3594089. ↩
10. Armstrong, B., and B. Berkowitz. "What Two Leading Hospital Systems Can Teach All Companies About Automation." *Sloan Management Review* (Forthcoming). ↩
11. Bodenheimer, Thomas, and Christine Sinsky. "From Triple to Quadruple Aim: Care of the Patient Requires Care of the Provider." *Annals of Family Medicine* 12, no. 6 (Nov-Dec 2014): 573–6. doi:10.1370/afm.1713. ↩
12. Wang, Sabrina M, H D Jeffry Hogg, Devdutta Sangvai, Manesh R Patel, E Hope Weissler, Katherine C Kellogg, William Ratliff, Suresh Balu, and Mark Sendak. "Development and Integration of Machine Learning Algorithm to Identify Peripheral Arterial Disease: Multistakeholder Qualitative Study." *JMIR Formative Research* 7 (2023): e43963–43963. doi:10.2196/43963. ↩
13. Kellogg, K. C. "Opinion: 6 Tactics to Make Artificial Intelligence Work on the Frontlines." *Stat* (Sep. 2022). https://www.statnews.com/2022/09/15/6-tactics-to-make-artificial-intelligence-work-on-the-frontlines/. ↩
14. Sendak, Mark P, Suresh Balu, and Kevin A Schulman. "Barriers to Achieving Economies of Scale in Analysis of EHR Data. A Cautionary Tale." *Applied Clinical Informatics* 8, no. 3 (2017): 826–31. doi:10.4338/ACI-2017-03-CR-0046. ↩

15. He, Jianxing, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. “The Practical Implementation of Artificial Intelligence Technologies in Medicine.” *Nature Medicine* 25, no. 1 (Jan 2019): 30–6. doi:10.1038/s41591-018-0307-0. ↩
16. Singer, Sara J, Katherine C Kellogg, Ari B Galper, and Deborah Viola. “Enhancing the Value to Users of Machine Learning-Based Clinical Decision Support Tools: A Framework for Iterative, Collaborative Development and Implementation.” *Health Care Management Review* 47, no. 2 (Apr-Jun 01, 2022): E21–31. doi:10.1097/HMR.0000000000000324. ↩
17. Diaz, N. “The Hardest Part of Working with Generative AI in Healthcare,” Becker’s Hospital Review. Accessed Jan. 15, 2024. https://www.beckershospitalreview.com/healthcare-information-technology/the-hardest-part-of-working-with-generative-ai-in-healthcare.html. ↩
18. Adams, K. “Epic Is Integrating Abridge’s Generative AI Tool Into Its EHR,” MedCity News. Accessed January 15, 2024. https://medcitynews.com/2023/08/epic-ehr-healthcare-generative-ai/. ↩
19. Capoot, A. “Microsoft Announces New AI Tools to Help Doctors Deliver Better Care,” CNBC. Accessed January 15, 2024. https://www.cnbc.com/2023/10/10/microsoft-announces-microsoft-fabric-and-azure-ai-tools-for-doctors.html. ↩
20. Edwards, B. “GPT-4 Will Hunt for Trends in Medical Records Thanks to Microsoft and Epic,” Ars Technica. Accessed May 5, 2023. https://arstechnica.com/information-technology/2023/04/gpt-4-will-hunt-for-trends-in-medical-records-thanks-to-microsoft-and-epic/. ↩
21. Small, W., et al. “The Potential of AI-Generated Responses to Patients’ In Basket Messages,” NYU Langone working paper, 2024. ↩
22. Engler, A. “Early Thoughts on Regulating Generative AI Like ChatGPT,” Brookings. Accessed April 25, 2023. https://www.brookings.edu/blog/techtank/2023/02/21/early-thoughts-on-regulating-generative-ai-like-chatgpt/. ↩
23. Meskó, Bertalan, and Eric J Topol. “The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare.” *NPJ Digital Medicine* 6, no. 1 (July 6, 2023): 120. doi:10.1038/s41746-023-00873-0. ↩
24. The Lancet Regional Health Europe. “Embracing Generative AI in Health Care.” *Lancet Regional Health Europe* 30 (July 3, 2023): 100677. doi:10.1016/j.lanepe.2023.100677. ↩
25. Wornow, Michael, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. “The Shaky Foundations of Large Language Models and

Foundation Models for Electronic Health Records.” *NPJ Digital Medicine* 6, no. 1 (July 29, 2023): 135. doi:10.1038/s41746-023-00879-8. ↩

26. Sendak, M., et al. “‘The Human Body Is A Black Box’: Supporting Clinical Decision-Making with Deep Learning,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in FAT* ’20*. New York: Association for Computing Machinery, Jan. 2020, 99–109. doi:10.1145/3351095.3372827. ↩
27. Jiang, Lavender Yao, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, et al. “Health System-Scale Language Models Are All-Purpose Prediction Engines.” *Nature* 619, no. 7969 (July 2023): 357–62. doi:10.1038/s41586-023-06160-y. ↩
28. Costa, A. “Large Language Model Predicts Patient Readmission,” NVIDIA Blog. Accessed January 15, 2024. https://blogs.nvidia.com/blog/nyu-large-language-model-patient-readmission-nature/. ↩
29. Winter, P., and A. Carusi. “‘If You’re Going to Trust the Machine, Then That Trust Has Got to Be Based on Something’: Validation and the Co-Constitution of Trust in Developing Artificial Intelligence (AI) for the Early Diagnosis of Pulmonary Hypertension (PH).” *Science & Technology Studies* 35, no. 4 (Dec. 2022): 4. doi:10.23987/sts.102198. ↩
30. Wiens, Jenna, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, et al. “Do No Harm: A Roadmap for Responsible Machine Learning for Health Care.” *Nature Medicine* 25, no. 9 (Sep 2019): 1337–40. doi:10.1038/s41591-019-0548-6. ↩
31. Austrian, J., and Y. Aphinyanaphongs. “Empowering Our Health System with a Private and Secure GPT Service,” Medium. Accessed January 15, 2024. https://nyulangonemcit.medium.com/empowering-our-health-system-with-a-private-and-secure-gpt-service-838fab3fff09. ↩
32. Small, W., et al. “The First Generative AI Prompt-A-Thon in Health Care: A Novel Approach to Workforce Engagement with a Private Instance of ChatGPT,” NYU Langone working paper, 2023. ↩
33. Sendak, M. P., et al., “A Path for Translation of Machine Learning Products into Healthcare Delivery,” *EMJ Innov Innov.* 2020. doi:10.33590/emjinnov/19-00172. ↩
34. Bondi-Kelly, E., et al. “Taking Off with AI: Lessons from Aviation for Healthcare,” in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, in EAAMO ’23*. New York: Association for Computing Machinery, 2023, 1–14. doi:10.1145/3617694.3623224. ↩
35. Nazir, Sajid, Diane M Dickson, and Muhammad Usman Akram. “Survey of Explainable Artificial Intelligence Techniques for Biomedical Imaging with Deep Neural Networks.” *Computers in Biology and Medicine* 156 (April 2023): 106668. doi:10.1016/j.compbiomed.2023.106668. ↩

36. Hoffman, R. R., S. T. Mueller, G. Klein, and J. Litman. “Metrics for Explainable AI: Challenges and Prospects.” Preprint, updated February 1, 2019. https://arxiv.org/abs/1812.04608. ↩

37. Schaefer, Kristin E, Jessie Y C Chen, James L Szalma, and P. A. Hancock. “A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems.” *Human Factors* 58, no. 3 (May 2016): 377–400. doi:10.1177/0018720816634228. ↩

38. Lipton, Z. C. “The Mythos of Model Interpretability.” Preprint, updated March 6, 2017. https://arxiv.org/abs/1606.03490. ↩

39. Endsley, W. Jones. “A Model of Inter- and Intrateam Situation Awareness: Implications for Design, Training and Measurement,” 2001. Accessed January 15, 2024. https://www.semanticscholar.org/paper/A-model-of-inter-and-intrateam-situation-awareness%3A-Endsley-Jones/82e77d0b9927b5c2291c96d40e632e77c6b39e94. ↩

40. Tucker, M., T. Eisape, P. Qian, R. Levy, and J. Shah. “When Does Syntax Mediate Neural Language Model Performance? Evidence from Dropout Probes,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Seattle: Association for Computational Linguistics, July 2022, 5393–5408. doi:10.18653/v1/2022.naacl-main.394. ↩

41. Huang, J., et al. “Large Language Models Cannot Self-Correct Reasoning Yet.” Preprint, submitted October 3, 2023. https://arxiv.org/abs/2310.01798. ↩

42. Sanneman, S., and J. Shah. “The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems.” *International Journal of Human-Computer Interaction* 38, no. 18-20 (2022): 1772–88. doi:10.1080/10447318.2022.2081282. ↩

43. Tucker, M., P. Qian, and R. Levy. “What if This Modified That? Syntactic Interventions with Counterfactual Embeddings,” in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, edited by C. Zong, F. Xia, W. Li, and R. Navigli, Association for Computational Linguistics, 2021, 862–875. doi:10.18653/v1/2021.findings-acl.76. ↩

44. Wei, J., et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” Preprint, updated January 10, 2023. https://arxiv.org/abs/2201.11903. ↩

45. Dell’Acqua, Fabrizio, McFowland, Edward, Mollick, Ethan R., Lifshitz-Assaf, Hila, Kellogg, Katherine, Rajendran, Saran, Krayer, Lisa, Candelon, François, and Karim R. Lakhani. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality,” working paper, Harvard Business School, Technology & Operations Management, 2023. doi:10.2139/ssrn.4573321. ↩

46. Noy, S., and W. Zhang. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence,” MIT, March 2, 2023. https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf. ↩

47. Brynjolfsson, E., J. J. Horton, A. Ozimek, D. Rock, G. Sharma, and H.-Y. TuYe. “COVID-19 and Remote Work: An Early Look at US Data,” National Bureau of Economic Research, Working Paper 27344, June 2020. doi:10.3386/w27344. ↩

48. Eloundou, T., S. Manning, P. Mishkin, and D. Rock. “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.” Preprint, submitted March 17, 2023. https://arxiv.org/abs/2303.10130. ↩