Human-Clinical AI Agent Collaboration

Mason Kadem [1,3,4], **Baraa Al-Khazraji** [2,3,4],
[1]Computing and Software, McMaster University, Canada
[2]Kinesiology, McMaster University, Canada
[3]Research Online, Canada
[4]Vascular AI, Canada
kademm@mcmaster.ca, alkhazrb@mcmaster.ca

Abstract

Balancing automation and accountability is fundamental in any healthcare field, particularly under mandates from the world's first AI act. Yet, the act relies on self-assessment. Here, we draw from a half century of theoretical cognitive neuroscience theories and analyze emerging computer science principles to develop an actionable blueprint to advance beyond self-assessment protocols for responsible Human-Clinical AI Collaboration. Our framework proactively identifies and mitigates risk through four key contributions: (1) interactive healthcare simulations populated by Clinical AI Agents as experimental testbeds to systematically evaluate human-AI collaboration without exposing patients to harm; (2) cognitive-state aware AI that adapts its behaviour based on measured physiological signals indicating cognitive load; and (3) critical safety mechanisms that enable Clinical AI Agents to disengage when detecting insufficient clinician engagement, preventing dangerous over-reliance; (4) emphasizing interpretable models for high-risk decisions and physiologically-adaptive explanations. These innovations address the fundamental mismatch between the dynamic nature of human cognition and the static interaction patterns of current Clinical AI systems, anticipating and mitigating both dangerous over-reliance and disengagement from algorithmic insights.

Introduction

How can we anticipate and optimize human-AI interactions to avoid harm in real world AI deployment? The world's first AI act mandates responsible integration of AI in healthcare (Porsdam Mann, Cohen, and Minssen 2024); yet, it relies on self-assessment (Michel E. van Genderen 2025). As AI continues to innervate clinical decision-making, adequately testing and monitoring human-AI collaboration before and after deployment becomes critical for satisfying mandate requirements.

The rise of large language models (LLMs) enabled computational software to solve problems that humans find difficult. More recently, LLMs have surpassed human experts on medical exams (LLMs: 90.2% (Nori et al. 2023), human experts: 87% (Liévin et al. 2022), passing score: 60%). By extending on LLMs with advanced cognitive architectures, from graph memory and inhibition to goal-directed behaviour, these agents can simulate realistic and responsible clinical behaviour. The human brain encodes memory in connected networks rather than sequential lists. Our implementation of graph-structured memory in Clinical AI Agents reflects this architecture, facilitating both transitive inference across medical concepts and the rapid heuristic reasoning that characterizes expert clinical judgment. This design choice strengthens our first key contribution by enabling more realistic clinical simulations.

We propose leveraging interactive healthcare simulations populated by these intelligent Clinical AI agents to systematically evaluate human-AI interaction without exposing patients to harm. Interactive healthcare simulations of this nature would enhance training of healthcare professionals (Wang and An 2021), particularly for rare or high risk events and difficult to access personnel. Such simulations would also provide testbeds for human-AI interaction theories (Card 2017) to improve decision making, transparency, accountability, reduce cognitive load and blind acceptance (Paleyes, Urma, and Lawrence 2022) of AI recommendations. Furthermore, these simulations could inform future healthcare robots (Yuan et al. 2023) and ease the global healthcare burden (McIntyre and Chow 2020), (i.e., 40% of the global population (WHO 2018). By integrating clinical AI agents for interactive and dynamic natural language interactions, we could reduce the cognitive effort (Shneiderman and Maes 1997) and enhance the realism of clinical simulations. Yet, the deployment of such agents also raises design concerns that must be addressed to ensure safe and effective use in healthcare settings.

Drawing from half a century of theoretical cognitive neuroscience theories (Kahneman and Beatty 1966) and emerging computer science principles, we identify a fundamental mismatch between the dynamic nature of human cognition and the static interaction patterns of current Clinical AI systems. This mismatch creates significant risks: both dangerous over-reliance on AI recommendations and disengagement from potentially valuable algorithmic insights. To address this issue, we theoretically motivate "cognitive-state aware AI" that integrates real-time physiological monitoring (e.g., pupil dilation as a proxy for cognitive load) into clinical AI Agents. Unlike traditional XAI and human-computer interaction approaches that provide static explanations regardless of the human's cortical state, the proposed systems

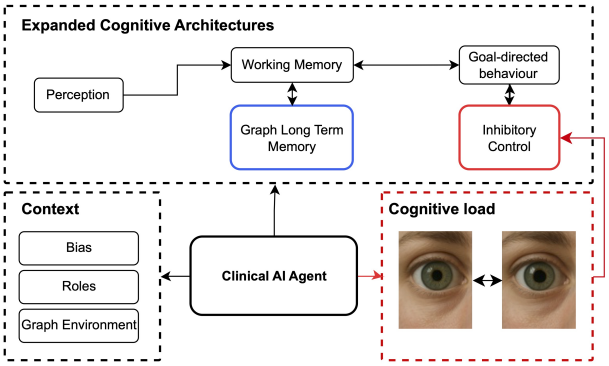

Figure 1: Expanded Cognitive Architecture of Clinical AI Agents.

continuously adapt their behaviour (e.g., explanation) based on the moment-to-moment cognitive state of the human partner, creating a symbiotic paradigm that adapts with detected changes in cognitive capacity, attention, and context.

We propose that leveraging the physiological information to develop critical safety mechanisms will enable Clinical AI Agents to disengage or escalate to additional human oversight when detecting insufficient clinician engagement. By identifying physiological patterns associated with blind acceptance or cognitive disengagement, the system can implement graduated responses: from intensifying cognitive forcing functions to pausing automated processes to disengaging the human-AI Collaboration. This safety layer prevents dangerous over-reliance on AI recommendations in high-stakes clinical contexts. Finally, we emphasize the importance of interpretable models as tools for AI Agents.

Background and Related Work

Recent advances in AI simulation research provide strong foundations for the proposed approaches. Park et al., 2023 leveraged LLMs within an architecture that allowed them to recursively store, prioritize, reflect, react and plan on memories (Park et al. 2023). This architecture within an interactive simulation increases believability. Our conceptual framework builds on this by implementing graph-structured memory representations that better capture the associative nature of clinical knowledge and reasoning patterns. Adapted to healthcare, these simulations can also address several aspects of algorithmic decision-making that are hard to quantify using theoretical models. These include patient variability and complexity, dynamic interactions and feedback loops, modelling human factors (e.g., high stakes, adherence, cognitive load), resource availability (e.g., healthcare professionals, resources), longitudinal outcomes and delayed effects, contextual/environmental factors (e.g., age, sex, socioeconomic factors, geographic location), and ethical and social considerations of AI systems. By capturing these elements, simulations provide a more comprehensive and realistic understanding of healthcare decision-making, leading to better-informed and more effective interventions. Modelling healthcare workers also allows for better human-AI interaction technologies.

While these advanced architectures show promise for healthcare simulations, several limitations must be addressed for practical implementation. The computational demands of LLMs remain significant, and inappropriate responses can break the immersion of the simulation. The simulation described therein might initially seem abstract, but the value lies in its application to certain contexts (e.g., healthcare) and to specific decision-making scenarios. Their practical application helps bridge the gap between theory and practice, ensuring that the insights gained are directly relevant and useful in real-world healthcare settings. Experimental designs may have limited evaluations (e.g., believability).

Additional complexities arise during AI agent implementation in healthcare domains. Current simulations struggle to capture the nuanced factors of clinical settings such as provider-patient trust dynamics, institutional cultures, and emotional responses, which significantly impact decision-making. Park et al., 2022 found that while generative AI can improve design thinking, these systems still face challenges in authentically replicating collaborative decision-making processes between healthcare providers, patients, and families (Park et al. 2022). Furthermore, algorithms that perform well in controlled environments may falter in real-world settings due to unforeseen variables, potentially leading to over-reliance or biased decision-making if training data lacks demographic diversity. These limitations underscore the need for AI systems that account for human cognitive processes rather than relying solely on algorithmic output.

Beyond simulation design, understanding the cognitive interplay between AI systems and human clinicians represents another critical dimension. Buçinca et al., 2021 examined the cognitive system beyond AI explanations (Buçinca, Malaya, and Gajos 2021). However, the examination overlooked common healthcare decision-making approaches like System 1 thinking, heuristics, and experience-based methods. Implementing cognitive forcing functions in AI can be context-dependent. The research by Buçinca et al. did not sufficiently address the long-term effects of these functions on user behaviour and decision-making. While cognitive forcing can reduce over-reliance on AI, it may also increase cognitive load, affecting user experience. Balancing the benefits of reduced over-reliance with the drawbacks of increased cognitive demands should be considered.

The role of explanations in mediating human-AI interaction has emerged as a key factor in addressing over-reliance issues. Vasconcelos et al. provided empirical evidence showing that explanations can effectively mitigate over-reliance on AI (Vasconcelos et al. 2022). The findings have practical implications for the design of AI systems, suggesting that incorporating explanations can enhance user engagement and improve decision quality. Through experiments with varying levels of task difficulty and explanation complexity, the study offers nuanced insights into the effectiveness of explanations. Specifically, it shows that when the task is hard and the explanation is simple, the explanation significantly reduces over-reliance. The study also highlights the impact of rewards, showing that increasing benefits motivates end-

users to avoid mistakes.

However, developing explanations that are both simple and accurate without oversimplifying the underlying complexity of AI decision-making requires a delicate balance. The study by Vasconcelos et al. does not adequately address the potential increase in cognitive load or how simplified explanations can be generated without losing critical information, which is essential for maintaining trust and reliability in AI systems. Additionally, while the study emphasizes the importance of simplifying explanations to reduce overreliance, it primarily relies on financial incentives to motivate participants.

Moving from theoretical frameworks to practical evaluation methods, recent research has attempted to develop comprehensive benchmarks for assessing AI performance in clinical contexts. Schmidgall et al. introduced an interactive multi-modal benchmark specifically designed to evaluate LLMs in their ability to operate as agents for decision-making in simulated clinical scenarios, given the limitations of traditional static medical question-answering benchmarks (Schmidgall et al. 2024). They designed two environments: 1) a multi-modal environment with both images and dialogue, 2) a dialogue-only environment. These environments were populated by language agents: a patient agent that provides symptoms, a doctor agent that diagnoses and requests test from the measurement agent (activate data collection), and a moderator agent which assesses accuracy of diagnosis. Cognitive (e.g., recency bias) and implicit (e.g., racial, sex) bias were included as well, and were found to negatively impact both diagnostic accuracy and patient perceptions. The authors automated the evaluation of diagnostic accuracy of the doctor agent's diagnosis, the compliance (i.e., willingness to follow through with treatment), and confidence (i.e., willingness of the patient to consult the doctor again).

By measuring patient compliance, confidence and willingness to adhere to follow-up consults, the benchmark provides a more realistic/comprehensive evaluation of AI performance in clinical settings. The incorporation of cognitive and implicit biases and their impact on diagnostic accuracy (Schmidgall et al. 2024) and patient perception is an important contribution. The diagnostic evaluation was fully automated and the researchers also evaluated multiple LLMS and found that cross communication between the same model leads to higher accuracy (e.g., patient GPT-4 and doctor GPT-4) and as a function of interactions.

The simulation used by Schmidgall et al. is visually simplistic and only includes patient, doctor, measurement and moderator agents. To enhance believability in the artificial healthcare simulation, cognitive biases should be expanded to include healthcare-specific biases such as anchoring, confirmation, framing, and attribution biases. Incorporating an AI-in-the-loop approach is crucial, as the current fully automated system contradicts the goal of supporting, rather than replacing, clinicians in decision-making. Integrating human oversight allows for real-time adjustments, improves trust, and ensures AI serves as an assistive tool rather than an autonomous decision-maker. The inclusion of GPT-4 and its over-alignment to human values, reduces its ability to represent bias in benchmarks.

Hallucination detection represents a critical safety requirement for Clinical AI Agents in healthcare. Recent work by Farquhar et al. presents a promising approach using semantic entropy to identify confabulations, plausible sounding but factually incorrect outputs that could lead to dangerous clinical decisions (Farquhar et al. 2024). Unlike previous methods focusing on surface level text patterns, this technique analyzes variability in meaning across multiple model generations, with high semantic entropy indicating potential unreliability. While computationally intensive, these safeguards are essential in high stakes medical contexts where factual accuracy directly impacts patient safety.

A recent study presented a sandbox hospital simulation designed to model the full continuum of clinical interactions, from disease onset to treatment in a controlled, dynamic environment (Li et al. 2024). It creates a dynamic hospital environment, providing a realistic platform for evaluating medical agents. The hospital environment consists of 16 distinct areas, including triage stations, consultation rooms, and examination rooms. Patients start as healthy individuals who can develop diseases and seek medical help, mimicking typical patient behaviour. Medical agents perform specific roles and adapt to evolving clinical situations, continuously improving their expertise through interaction and feedback. A strength of this paper is that agents are designed to self-evolve by integrating external knowledge and engaging in reflection processes during task execution. Weaknesses include the lack of AI-in-the loop integration, and scalability (limited scoped scenarios), and limited to diagnostic success rate and would benefit from more detailed performance metrics.

While these existing research efforts provide valuable foundations, they collectively highlight critical gaps in current approaches to Human-Clinical AI collaboration. First, most simulations employ simplified, often list-based memory structures that fail to capture the associative nature of clinical knowledge. Second, they typically lack physiologically-informed adaptation to end-user cortical states. Third, they generally lack robust safety mechanisms for detecting and responding to inadequate human engagement. Fourth, they rarely integrate interpretable models with adaptive explanations tailored to clinical contexts. Our conceptual framework directly addresses these limitations through a comprehensive approach that combines graph-structured memory representations mirroring human memory, interactive simulations, cognitive monitoring and inhibtiory control via physiological signals, and context-aware explanations, creating a novel paradigm for responsible AI deployment that protects patients while enhancing clinical decision-making.

Interactive Simulations as Testbeds

Simulations provide a controlled, safe, and cost-effective environment for healthcare professionals and AI engineers to prototype, test, and evaluate new medical AI technologies, interventions, and treatments (Kadem et al. 2023). These environments, when coupled with clinical AI agents, allow investigation of Human-Clinical AI interactions in ways that

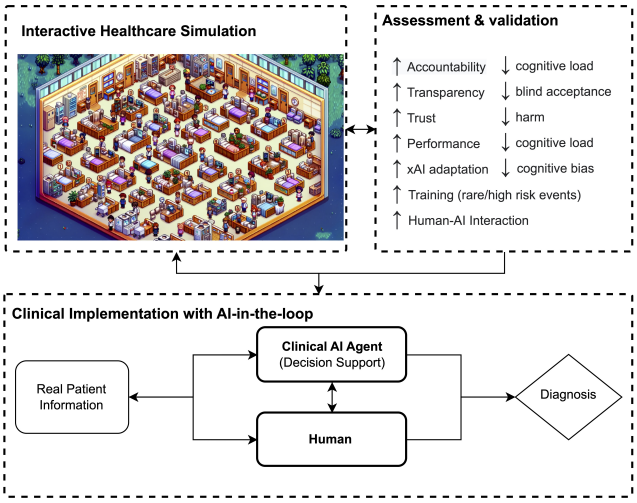

Figure 2: From interactive healthcare simulation to clinical deployment. Enables the anticipation and mitigation of potential harm before clinical implementation while establishing monitoring mechanisms for post-deployment safety. The process enhances accountability, transparency, and performance while reducing cognitive load, blind acceptance, and bias in healthcare decision support.

would be impractical or impossible in real clinical settings.

Simulations enable testing of AI-human interaction theories in difficult-to-access settings (e.g., with astronauts, healthcare workers remote areas), identifying optimal interaction patterns to reduce cognitive load and enhance user experience. They accelerate the deployment time of AI systems in healthcare through rapid prototyping and enhance believability by engaging in dynamic natural language interactions with humans.

This reduces deployment time and addresses concerns about potential harm to patients (Okolo et al. 2024). In simulated healthcare settings, professionals interact with generative health agents to analyze AI-driven decision-making, focusing on effectiveness, reliability, believability, and biases. Developers can also simulate healthcare professionals as AI agents to prototype AI deployment, translating theory into real-world skills without risking patient safety. Simulations replicate healthcare provider interactions, reducing costs and time associated with accessing resources and personnel. They enable interaction with realistic standardized patients, adaptable in emotion and difficulty, facilitating practice in consultations, symptom discussions, and treatment plans.

These environments when coupled with clinical AI agents allow investigation of Human-Clinical AI interactions, coordination of healthcare team responses to emergencies, and safe practice for procedures and decision-making without harming real patients. They offer standardized scenarios with immediate feedback, enhancing skill acquisition, and enable repeated exposure to various situations, aiding skill mastery and confidence building. Simulations also replicate complex, rare, or emergency situations that healthcare providers may not frequently encounter and progressively adapt scenarios to meet end-user needs. Game simulations engage users more effectively than traditional methods, motivating deeper thinking with AI decision support systems and allowing realistic assessments of over-reliance and cognitive interventions.

In algorithmic decision-making, healthcare professionals need to prototype and evaluate the aftermath of their decisions. There is a significant difference between small, homogeneous end-user testing and assessing large-scale AI system deployment. It is urgent to prototype and assess the potential harm of AI systems to ensure their safety and efficacy.

While generative agents hold significant promise for enhancing human-computer interaction in healthcare, they also present notable ethical concerns. Developers must transparently disclose the statistical nature of these clinical AI agents to end-users. The tendency of users to anthropomorphize these agents can lead to over-reliance and high-risk scenarios, necessitating careful design to mitigate these risks in healthcare settings (Abercrombie et al. 2023).

Designing effective AI-human interactions is particularly challenging due to the potential for unpredictable errors, hallucinations (Farquhar et al. 2024) and complex outputs of AI systems (Yang et al. 2020). Following established human-AI design guidelines (Amershi et al. 2019) is crucial for creating user-centric interfaces that improve transparency, and reduce over-reliance and risks. For example, clearly distinguishing the artificial environment from real-world applications (such as game settings) can help mitigate risks by ensuring users remain aware of the artificial nature of the simulation. This awareness prevents over-reliance, potential misuse, and harm from AI errors. Such an approach is beneficial for interactive simulations, providing a safe and controlled environment to test and refine human-AI interactions without real-world consequences.

To mitigate the risk of misinformation, rigorous oversight of AI interactions is essential (Park et al. 2023). Generative healthcare agents should complement, not replace, human input in early-stage prototyping (Park et al. 2022). This approach can reduce deployment time and alleviate the healthcare burden by minimizing the need for professional involvement in testing challenging or high-risk scenarios. Balancing realism with user comfort is essential to maintaining effective and engaging AI-human interactions.

Notably, these interactive healthcare simulations serve a dual purpose: they not only provide valuable training environments for healthcare professionals but also function as controlled experimental testbeds for systematically evaluating AI-human interactions prior to deployment in high-risk clinical settings (Fig. 2). This approach aligns with the growing regulatory emphasis on robust pre-implementation testing for high-risk AI systems in healthcare.

Blind Acceptance

AI decision support tools can lead to blind acceptance (Buçinca, Malaya, and Gajos 2021). Blind acceptance, a common error in AI-human interactions, is especially concerning in high-stakes decision-making as it forgoes human accountability. Instead of integrating their own insights

with AI suggestions, users tend to accept the AI's decisions, even when they are incorrect. Conflicting data exist on whether explanations reduce blind acceptance (Vasconcelos et al. 2022; Buçinca, Malaya, and Gajos 2021), since the mere presence of an explanation increases AI trust (Zhang, Liao, and Bellamy 2020). Dual-process theory suggests humans rely more often on rapid, cognitively biased (System 1) thinking than on System 2 deliberate cognition (Wason and Evans 1974).

It might be better to motivate end users to engage in higher-level cognition with AI (Kaur et al. 2020), rather than rely solely on AI explanations (Buçinca, Malaya, and Gajos 2021). To reduce the impact of cognitive biases on decision-making and reduce human errors, interventions can disrupt autopilot cognition and encourage higher-level thinking (Lambe et al. 2016; Graber et al. 2012). If we aim to better engage humans with AI, especially in high stakes decision making, optimizing the cognitive effort required for tasks is important, as both low and high cognitive load disengages humans.

Believability and blind trust in human-AI interactions are related but distinct concepts. Believability ensures AI recommendations are perceived as credible and trustworthy, fostering confidence in their use. Blind trust, however, occurs when users place excessive trust in AI, potentially neglecting their own judgment and critical thinking. In healthcare simulations, maintaining this balance is crucial. While believable AI can enhance prototyping and decision-making, preventing blind trust ensures that healthcare professionals continue to apply their expertise alongside AI insights for the best patient outcomes. Interactive healthcare simulations offer a powerful way to achieve this balance.

Cognitive Load

Cognitive load in AI-human interaction involves the mental effort required to engage with AI systems effectively. It encompasses various executive functions, including working memory, attention, and problem-solving abilities, all of which are managed by complex neural networks in the brain.

The prefrontal cortex plays a critical role in managing tasks such as decision-making, attention, and working memory. Within the prefrontal cortex, the dorsolateral prefrontal cortex handles working memory and planning, the ventromedial prefrontal cortex aids in decision-making and emotional regulation, and the anterior cingulate cortex is vital for error detection and behaviour adjustment. Neurotransmitters also play a significant role in cognitive load. Dopamine is essential for attention (inspiration for the attention mechanism in transformers), learning, and maintaining focus. Norepinephrine enhances alertness and response to stress, while acetylcholine is important for attention and memory formation.

High cognitive load can lead to mental fatigue, reducing efficiency in processing information and making decisions. Cognitive engagement follows an inverted U-shaped curve known as the Yerkes-Dodson law, where performance initially improves with increasing mental activation but then declines as these levels become too high. Performance is poor when activation is too low because attention is insufficient, and poor when activation is too high because mental resources become overwhelmed. Healthcare workers using AI systems need to stay in this middle zone of optimal performance. Too little mental engagement leads to over-reliance on automation and loss of critical thinking. Too much mental demand from managing complex technology while caring for patients overloads their cognitive capacity, leading to worse decisions. System design must balance AI support, enough to prevent cognitive overload but not so much that healthcare workers lose their skills and vigilance toward important clinical details.

AI systems that adapt to the user's context and provide personalized support can manage intrinsic cognitive load. Context-aware systems ensure relevant information is available when needed. Interactive tutorials and immediate feedback help users build mental models and improve performance. Providing hands-on experience with guided instructions enhances learning and retention. AI systems can support users by reducing distractions, offering flexible problem-solving paths, and providing adaptive learning experiences. These features enhance cognitive flexibility, focus, and memory retention. Understanding and managing cognitive load in AI-human interaction is essential for creating effective, user-friendly AI systems. By addressing the cognitive demands placed on users, we can design AI technologies that enhance performance and user satisfaction. This theoretical understanding of cognitive dynamics necessitates practical implementation strategies for real-time assessment and response to clinician cognitive states.

Physiological Monitoring of Cognitive Load

The dynamic nature of clinical decision-making necessitates real-time assessment of cognitive load to ensure optimal AI-human collaboration. Unlike traditional approaches that rely on retrospective self-reporting, technologies that enable continuous, non-intrusive monitoring of clinicians' cognitive states, which allows clinical AI systems to adapt explanations and support in real-time based on the detected cognitive load level.

Fifty years of research demonstrates that pupil dilation reliably indicates cognitive load when controlling for luminance and accommodation. Kahneman and Beatty (Kahneman and Beatty 1966) established that "anything that increases the brain's processing load will dilate the pupil." Research has confirmed this relationship across multiple cognitive domains including attention, memory, processing load, and executive functions. Physiologically, pupil size depends on the balance between two autonomic nervous system branches: the sympathetic system activates the iris dilator muscle causing dilation, while the parasympathetic system activates the iris sphincter muscle causing constriction. The locus coeruleus, a key arousal center, influences both pathways by activating the sympathetic nervous system while inhibiting the parasympathetic nervous system via the Edinger-Westphal nucleus.

Multiple lines of evidence establish pupil size as a reliable index of locus coeruleus activity. Electrical stimulation of the locus coeruleus in various animal models produces rapid pupil dilation, while pharmacological manipulations

of arousal via locus coeruleus modulation similarly affect pupil size. The locus coeruleus is the brain's primary source of norepinephrine and innervates much of the neocortex including the fronto-parietal network associated with executive functions like working memory and goal-directed behaviour. This locus coeruleus-norepinephrine system plays a major role in attention regulation and arousal states, making pupillometry an ideal non-invasive method for continuously monitoring cognitive states during clinical decision-making.

Our prior work demonstrates that pupillometry can detect increased processing demands from subtle linguistic challenges during comprehension. For example, pupil dilation increases in response to lexical ambiguities in text, even when participants do not consciously perceive these ambiguities (Kadem et al. 2020). This sensitivity to unconscious cognitive processing highlights pupillometry's potential for detecting nuanced mental workload in real-world clinical environments. Modern eye-tracking technologies and even standard webcams (Kadem and Cusack 2017) can now capture these physiological signals in affordable and non-intrusive ways.

The integration of these assessment techniques enables what we term "cognitive-state aware AI systems" (Fig. 1). Clinical decision support tools that continuously monitor the clinician's cognitive capacity and adjust their interaction style accordingly. In high-stakes scenarios where cognitive resources are already taxed, these systems can perhaps automatically shift toward more aligned explanations. This symbiotic design represents a fundamental shift from existing explanation approaches that offer one-size-fits-all justifications regardless of the user's cognitive state. By continuously adapting explanations based on physiological feedback, the system creates a responsive, cognitively-aware partnership that maintains appropriate human engagement while providing decision support.

A critical component of our framework is the ability to detect dangerous disengagement patterns and implement appropriate safety responses. While adaptive explanations help maintain optimal cognitive engagement, certain situations may still lead to blind acceptance or insufficient scrutiny of AI recommendations. Physiological indicators can provide early warning signs of these problematic interaction patterns.

AI-in-the-Loop

We shift from a Human-in-the-Loop model to an AI-in-the-Loop approach that centers human cognition. In high-risk medical settings, human expertise remains essential because large language models (LLMs) are inherently error-prone and cannot be fully controlled. While fields like optometry may soon be automated (Zhou et al. 2023; Zekavat et al. 2022), medicine involves complex, high risk, and ambiguous cases that demand human oversight. A human-centered approach ensures AI supports, rather than replaces, high-risk clinical decision-making, preserving the cognitive flexibility needed in healthcare environments.

Designing effective human-clinical AI agent collaboration requires integrating our framework's key innovations into practical clinical workflows. First, context-aware functionality is critical. AI systems must adapt recommendations based on patient-specific information and clinical context, similar to how our graph-structured memory approach enables associative relationships between clinical concepts. The system should connect symptoms, medical history, and current state with appropriate care pathways, rather than matching isolated data points to generic recommendations.

Second, physiologically-informed adaptivity must be embedded in every interaction. Our cognitive-state monitoring enables the system to detect when a clinician's cognitive load increases, for example, during complex cases or emergencies, and dynamically adjust both information presentation and decision support. This might include simplifying explanations, highlighting critical information, or adjusting the threshold for safety interventions based on detected engagement levels.

Human oversight identifies nuances that AI systems miss. Clinicians recognize atypical presentations, cultural factors, and emotional responses that current AI cannot adequately capture. By incorporating human intelligence into the process, the system leverages prior experience while extending rather than replacing human judgment. This collaborative approach is crucial for addressing cases that AI alone might struggle with. Human clinicians recognize patterns from experience in rare or atypical presentations, consider psychosomatic factors, detect olfactory cues (which AI cannot perceive at all) and dynamically adjust their diagnostic approach as symptoms evolve, capabilities current AI systems lack. Additionally, human doctors provide critical ethical judgment in sensitive situations and gather information through physical examination and tactile feedback.

Third, graduated safety mechanisms should operate across all interactions. When pupillometry or other physiological signals indicate potential disengagement or blind acceptance, the system should implement escalating interventions: from subtle emphasis of key information, to explicit "cognitive forcing" techniques that require active clinician engagement, to complete disengagement and escalation to additional oversight for high-risk decisions.

Fourth, communication design must balance transparency with cognitive efficiency. Natural language capabilities allow healthcare professionals to query the system with questions like "What additional tests are recommended?" while receiving explanations matched to their expertise level and current cognitive state. The interface should support bidirectional feedback, allowing clinicians to override AI recommendations and provide rationales that improve system performance over time.

This human-AI collaboration leverages complementary strengths: AI's consistency and pattern recognition with human clinicians' contextual understanding, ethical reasoning, and adaptability. By maintaining this balance, the system can handle the complexities of real-world medical scenarios while reducing the burden of creating theoretically flawless algorithms. Implementation of these principles requires careful evaluation in simulated environments before clinical deployment. Our approach enables systematic testing across various clinical scenarios, measuring not only diagnostic accuracy but also team performance, cognitive engagement,

and appropriate reliance.

Use Case: Spaceflight Medicine

Spaceflight medicine exemplifies our framework's value, where access to trained astronauts, aerospace physicians, and microgravity environments would be prohibitively expensive and logistically impossible, making comprehensive evaluation of AI systems financially unfeasible. Our simulations overcome these barriers by populating scenarios with intelligent agents that embody diverse astronaut profiles that vary in physiological responses, medical histories, and communication styles.

Critically, these simulations can be run by developers and AI engineers early in the development process to anticipate and mitigate issues before involving actual astronauts or medical personnel, creating a safer, iterative development pathway for high-risk medical AI systems. Moreover, our architecture supports simulation-within-simulation analysis, where meta-agents observe interaction with the simulation itself, revealing subtle biases in how users perceive and trust simulated environments versus real-world scenarios.

This approach not only reduces costs dramatically but also enables testing hundreds of rare emergency scenarios that would otherwise require decades of actual spaceflight to encounter naturally. The simulation captures complex relationships between microgravity exposure and physiological responses via our graph-structured memory architecture. Pupillometry sensors monitor physicians' cognitive load, triggering adaptive responses: detailed explanations under normal conditions shift to streamlined guidance during high cognitive load. When pupillary data indicate disengagement, safety protocols activate from highlighting critical values to complete system disengagement, similar to autonomous vehicle safeguards. Performance evaluation compares AI-alone, human-alone, and collaborative approaches across various scenarios, enabling optimization before actual deployment. By systematically identifying potentially dangerous interaction patterns in this controlled environment, our approach enables risk mitigation before deployment to actual space missions, where a single error could cost lives and mission resources.

Explainability vs. Interpretability

While cognitive-state aware AI and safety mechanisms address the dynamic aspects of human-AI collaboration, the underlying AI models themselves require careful consideration. The distinction between explainability and interpretability is crucial for designing effective clinical AI agents. While these terms are often used interchangeably, they represent fundamentally different approaches to AI transparency that impact decision support in healthcare (Kadem, Noseworthy, and Doyle 2023; Kadem 2023).

Interpretability refers to the inherent transparency of an AI system, where humans can directly understand how inputs lead to outputs by examining the model's structure. Interpretable models like decision trees, linear regression, or rule-based systems allow clinicians to trace the exact reasoning path, facilitating trust through structural clarity. For example, a decision tree might show that if a patient's systolic blood pressure exceeds 140 mmHg and they have a family history of cardiovascular disease, their risk of stroke increases by 30%. This transparency is particularly valuable in high-stakes medical scenarios where accountability is important.

Explainability, in contrast, refers to post-hoc methods that attempt to clarify how complex, often opaque AI systems (like deep neural networks) arrive at their conclusions. Rather than offering inherent transparency, explainability techniques generate approximations of the model's reasoning. The critical distinction for healthcare applications is that explainability does not guarantee true understanding of the model's internal processes. It provides justifications that may approximate but not fully capture the complex interactions within the model. This limitation becomes particularly relevant in scenarios like our spaceflight use case, where multiple physiological systems interact in complex ways under microgravity conditions.

Our framework's contribution lies in integrating cognitive monitoring with explanation generation. Unlike traditional approaches that provide static explanations regardless of the clinician's cortical state, we propose physiologically-adaptive explanations that adjust in complexity and format based on measured cognitive load. The same underlying clinical information can be presented differently depending on the clinician's momentary cognitive capacity, from detailed explanations during normal cognitive load to streamlined, action-oriented guidance when cognitive resources are taxed.

For safety-critical decisions in healthcare, we recommend using inherently interpretable models where possible (Kadem, Noseworthy, and Doyle 2023). Decision trees (Quinlan 1986) use hierarchical if-then rules with clear decision paths. Linear models offer coefficients directly quantifying feature importance, making the impact of each clinical variable transparent. Bayesian networks encode probabilistic relationships between variables, making inference steps explicit.

For complex pattern recognition tasks where interpretable models may sacrifice performance, post-hoc explanation methods become necessary. Feature importance techniques (Breiman 2001) rank input features by their contribution to predictions. For image-based diagnostics, visual methods like saliency maps highlight influential image regions. Attention mechanisms reveal which input elements the model prioritized. To understand how specific clinical variables affect outcomes, partial dependence plots (Friedman 2001) show how individual features affect outcomes. Counterfactual explanations (Wachter, Mittelstadt, and Russell 2017) identify minimal input changes that would alter predictions.

More sophisticated approaches include LIME (Ribeiro, Singh, and Guestrin 2016), which approximates complex models locally with interpretable ones, and SHAP (Lundberg and Lee 2017), which assigns feature importance using game theory principles. When clinicians need to relate new cases to familiar ones, example-based approaches (Arjovsky and Bottou 2017) can explain by referencing similar cases from the model's training data.

The effectiveness of explanations depends not only on their content but critically on the cognitive state of the recipient. The same explanation can either enhance decision quality or be disregarded depending on the clinician's momentary cognitive capacity, attention resources, and engagement level. This insight motivates our approach to physiologically-adaptive explanations that respond to measured cognitive states.

In our clinical AI agent framework, we implement different explanation strategies based on the clinical context and detected cognitive load. For example, in the spaceflight medicine scenario, when monitoring shows normal cognitive load, the system might provide detailed explanations of how multiple physiological parameters contribute to deconditioning risk assessment. However, when high cognitive load is detected (e.g., during an emergency), the system automatically shifts to simplified, action-oriented explanations highlighting only the most critical factors and recommended actions.

This adaptive approach addresses the challenges of cognitive load discussed earlier by providing appropriate levels of complexity of explanation based on the clinical context, the level of expertise and the momentary cognitive state. It also helps mitigate blind acceptance by reinforcing human judgment when physiological indicators suggest potential disengagement or undue reliance on automation.

Even well-explained AI recommendations may lead to blind acceptance if the explanation increases perceived reliability without improving actual understanding (Schmidgall et al. 2024). This highlights why AI-in-the-loop oversight remains essential; experienced clinicians can detect inconsistencies between explanations and medical knowledge that might not be apparent from the explanation alone.

Our framework suggests that the effectiveness of explanations depends not only on their content but critically on the cognitive state of the recipient. The same explanation can improve decision quality or be disregarded depending on the clinician's momentary cognitive capacity, attention resources, and engagement level. This insight requires a more nuanced approach to explainability that considers the dynamic cognitive needs of healthcare professionals.

Various explanation approaches exist for clinical AI systems. For feature attribution, techniques like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016) help quantify how specific inputs influence predictions, while Conditional GANs (Mirza and Osindero 2014) and latent space visualization (Radford, Metz, and Chintala 2015) help interpret generative models. Visual methods such as saliency maps and Grad-CAM highlight regions of interest in medical images, while counterfactual explanations (Wachter, Mittelstadt, and Russell 2017) identify which changes would alter predictions, particularly valuable for clinical decision support. Decision trees (Quinlan 1986), gradient-based explanations (Baehrens et al. 2010), and partial dependence plots (Friedman 2001) visualize decision boundaries and feature relationships, offering clinicians transparent insights into model reasoning. For complex clinical scenarios, example-based approaches (Arjovsky and Bottou 2017) and perturbation analysis (Fong and Vedaldi 2017) provide contextual understanding, while anchor explanations (Ribeiro, Singh, and Guestrin 2018) and global surrogate models (Craven and Shavlik 1995) offer simplified rule-based interpretations. These techniques vary in complexity and application context, reinforcing the need for cognitive-state adaptive approaches that match explanation sophistication to the clinician's momentary cognitive capacity.

Conclusion and Future Work

Clinical AI agents represent a transformative shift in healthcare, enabling clinicians to simulate, predict, and refine decisions in high-stakes scenarios. Our conceptual framework for cognitive-state aware AI addresses the fundamental mismatch between static AI systems and dynamic human cognition by continuously adapting explanations and safety mechanisms based on measured physiological indicators of cognitive load. By integrating real-time pupillometry and other physiological signals, these systems can detect moments of over-reliance or disengagement and adjust their behaviour accordingly. This approach maintains meaningful human engagement throughout the decision process while providing valuable decision support.

Our four key innovations, interactive healthcare simulations with graph-structured memory, physiologically adaptive AI, safety disengagement mechanisms, and context-aware explanations, collectively create a new paradigm for human-AI collaboration that preserves human agency while leveraging the analytical power of AI systems. The graph-based memory structures integrated into our agents enable the formation of meaningful associations between clinical concepts and past experiences, reflecting how expert clinicians actually think rather than how algorithms traditionally process information.

We must remain realistic about current limitations in these agents. All systems have inherent error probabilities requiring human verification, especially in high-stakes healthcare scenarios. By framing AI not as a replacement but as a cognitive partner that adapts to human states, we can mitigate risks like algorithmic complacency and ensure these tools enhance healthcare decision-making while maintaining appropriate human oversight.

This conceptual framework opens several promising research directions, including implementation across diverse healthcare domains, development of standardized physiological monitoring protocols suitable for clinical environments, and creation of regulatory frameworks that incorporate cognitive-state awareness into safety certification requirements. By addressing the fundamental mismatch between static AI systems and dynamic human cognition, this framework lays the groundwork for more effective, safer, and more responsible human-AI collaboration in healthcare and beyond.

References

Abercrombie, G.; Curry, A. C.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages: On Anthropomorphism in Dialogue Systems. *arXiv*.

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; Teevan, J.; Kikin-Gil, R.; and Horvitz, E. 2019. Guidelines for Human-AI Interaction. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45035970-2.

Arjovsky, M.; and Bottou, L. 2017. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv*.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and Müller, K.-R. 2010. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.*, 11: 1803–1831.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1): 1–21.

Card, S. K. 2017. *The Psychology of Human-Computer Interaction*. Andover, England, UK: Taylor & Francis. ISBN 978-0-20373616-6.

Craven, M.; and Shavlik, J. 1995. Extracting Tree-Structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 8.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630: 625–630.

Fong, R.; and Vedaldi, A. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *arXiv*.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29(5): 1189–1232.

Graber, M. L.; Kissam, S.; Payne, V. L.; Meyer, A. N. D.; Sorensen, A.; Lenfestey, N.; Tant, E.; Henriksen, K.; Labresh, K.; and Singh, H. 2012. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ Qual. Saf.*, 21(7): 535–557.

Kadem, M. 2023. *Interpretable Machine Learning in Alzheimer's Disease Dementia*. Master's thesis, McMaster University, Hamilton, Ontario, Canada.

Kadem, M.; and Cusack, R. 2017. Pearls and Perils of Pupillometry Using a Webcam.

Kadem, M.; Garber, L.; Abdelkhalek, M.; Al-Khazraji, B. K.; and Keshavarz-Motamed, Z. 2023. Hemodynamic Modeling, Medical Imaging, and Machine Learning and Their Applications to Cardiovascular Interventions. *IEEE Reviews in Biomedical Engineering*, 16: 403–423.

Kadem, M.; Herrmann, B.; Rodd, J. M.; and Johnsrude, I. S. 2020. Pupil Dilation Is Sensitive to Semantic Ambiguity and Acoustic Degradation. *Trends in Hearing*, 24.

Kadem, M.; Noseworthy, M.; and Doyle, T. 2023. XGBoost for Interpretable Alzheimer's Decision Support. *Proceedings of the AAAI Symposium Series*, 1(1): 135–141.

Kahneman, D.; and Beatty, J. 1966. Pupil Diameter and Load on Memory. *Science*, 154(3756): 1583–1585.

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45036708-0.

Lambe, K. A.; O'Reilly, G.; Kelly, B. D.; and Curristan, S. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Qual. Saf.*, 25(10): 808–820.

Li, J.; Wang, S.; Zhang, M.; Li, W.; Lai, Y.; Kang, X.; Ma, W.; and Liu, Y. 2024. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv*.

Liévin, V.; Hother, C. E.; Motzfeldt, A. G.; and Winther, O. 2022. Can large language models reason about medical questions? *arXiv*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.

McIntyre, D.; and Chow, C. K. 2020. Waiting Time as an Indicator for Health Services Under Strain: A. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 57.

Michel E. van Genderen, P., MD. 2025. Moving Toward Implementation of Responsible Artificial Intelligence in Health Care: The European TRAIN Initiative — Artificial Intelligence — JAMA — JAMA Network.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *arXiv*.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; Luo, R.; McKinney, S. M.; Ness, R. O.; Poon, H.; Qin, T.; Usuyama, N.; White, C.; and Horvitz, E. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv*.

Okolo, C. T.; Agarwal, D.; Dell, N.; and Vashistha, A. 2024. \. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1): 1–28.

Paleyes, A.; Urma, R.-G.; and Lawrence, N. D. 2022. Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Comput. Surv.*, 55(6): 1–29.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. New York, NY, USA: Association for Computing Machinery. ISBN 979-840070132-0.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *UIST '22: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45039320-1.

Porsdam Mann, S.; Cohen, I. G.; and Minssen, T. 2024. The EU AI Act: Implications for U.S. Health Care. *NEJM AI*, 1(11).

Quinlan, J. R. 1986. Induction of decision trees. *Mach. Learn.*, 1(1): 81–106.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. \. In *ACM Conferences*, 1135–1144. New York, NY, USA: Association for Computing Machinery.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. *AAAI*, 32(1).

Schmidgall, S.; Harris, C.; Essien, I.; Olshvang, D.; Rahman, T.; Kim, J. W.; Ziaei, R.; Eshraghian, J.; Abadir, P.; and Chellappa, R. 2024. Addressing cognitive bias in medical language models. *arXiv*.

Shneiderman, B.; and Maes, P. 1997. Direct manipulation vs. interface agents. *interactions*, 4(6): 42–61.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M.; and Krishna, R. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *arXiv*.

Wang, C.; and An, P. 2021. Explainability via Interactivity? Supporting Nonexperts' Sensemaking of pre-trained CNN by Interacting with Their Daily Surroundings. In *CHI PLAY '21: Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, 274–279. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45038356-1.

Wason, P.; and Evans, J. 1974. Dual processes in reasoning? *Cognition*, 3(2): 141–154.

WHO. 2018. Health workforce requirements for universal health coverage and the Sustainable Development Goals. *World Health Organization*, 1–58.

Yang, Q.; Steinfeld, A.; Rosé, C. P.; and Zimmerman, J. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '20)*.

Yuan, K.; Sajid, N.; Friston, K.; and Li, Z. 2023. Hierarchical generative modelling for autonomous robots. *Nat. Mach. Intell.*, 5: 1402–1414.

Zekavat, S. M.; Raghu, V. K.; Trinder, M.; Ye, Y.; Koyama, S.; Honigberg, M. C.; Yu, Z.; Pampana, A.; Urbut, S.; Haidermota, S.; O'Regan, D. P.; Zhao, H.; Ellinor, P. T.; Segrè, A. V.; Elze, T.; Wiggs, J. L.; Martone, J.; Adelman, R. A.; Zebardast, N.; Del Priore, L.; Wang, J. C.; and Natarajan, P. 2022. Deep Learning of the Retina Enables Phenome- and Genome-Wide Analyses of the Microvasculature. *Circulation*, 145(2): 134–150.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAT∗ '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45036936-7.

Zhou, Y.; Chia, M. A.; Wagner, S. K.; Ayhan, M. S.; Williamson, D. J.; Struyven, R. R.; Liu, T.; Xu, M.; Lozano, M. G.; Woodward-Court, P.; Kihara, Y.; Allen, N.; Gallacher, J. E. J.; Littlejohns, T.; Aslam, T.; Bishop, P.; Black, G.; Sergouniotis, P.; Atan, D.; Dick, A. D.; Williams, C.; Barman, S.; Barrett, J. H.; Mackie, S.; Braithwaite, T.; Carare, R. O.; Ennis, S.; Gibson, J.; Lotery, A. J.; Self, J.; Chakravarthy, U.; Hogg, R. E.; Paterson, E.; Woodside, J.; Peto, T.; Mckay, G.; Mcguinness, B.; Foster, P. J.; Balaskas, K.; Khawaja, A. P.; Pontikos, N.; Rahi, J. S.; Lascaratos, G.; Patel, P. J.; Chan, M.; Chua, S. Y. L.; Day, A.; Desai, P.; Egan, C.; Fruttiger, M.; Garway-Heath, D. F.; Hardcastle, A.; Khaw, S. P. T.; Moore, T.; Sivaprasad, S.; Strouthidis, N.; Thomas, D.; Tufail, A.; Viswanathan, A. C.; Dhillon, B.; Macgillivray, T.; Sudlow, C.; Vitart, V.; Doney, A.; Trucco, E.; Guggeinheim, J. A.; Morgan, J. E.; Hammond, C. J.; Williams, K.; Hysi, P.; Harding, S. P.; Zheng, Y.; Luben, R.; Luthert, P.; Sun, Z.; McKibbin, M.; O'Sullivan, E.; Oram, R.; Weedon, M.; Owen, C. G.; Rudnicka, A. R.; Sattar, N.; Steel, D.; Stratton, I.; Tapp, R.; Yates, M. M.; Petzold, A.; Madhusudhan, S.; Altmann, A.; Lee, A. Y.; Topol, E. J.; Denniston, A. K.; Alexander, D. C.; and Keane, P. A. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163.