COPENHAGEN BUSINESS SCHOOL

Solbjerg Plads 3
DK-2000 Frederiksberg
Danmark

www.cbs.dk

ISSN 0906-6934

Print ISBN: 978-87-7568-327-7
Online ISBN: 978-87-7568-328-4

COGNITIVE CHALLENGES IN HUMAN-AI COLLABORATION

PhD Series 04-2025

CBS

CBS PhD School
Department of Accounting

PhD Series 04.2025

RIKARD ROSENBACKE

COGNITIVE CHALLENGES IN HUMAN-AI COLLABORATION

A Study on Trust, Errors, and Heuristics in Clinical Decision-Making

CBS

Cognitive Challenges in Human-AI Collaboration

A Study on Trust, Errors, and Heuristics in Clinical Decision-Making

Rikard Rosenbacke

Department of Accounting, Center for Corporate Governance

Supervisors: Ioanna Constantiou, Tom Kirchmaier, Åsa Melhus

CBS PhD School
Copenhagen Business School

Rikard Rosenbacke
Cognitive Challenges in Human-AI Collaboration
A Study on Trust, Errors, and Heuristics in Clinical
Decision-Making

First edition 2025
Ph.D. Serie 04.2025

ISSN 0906-6934

Print ISBN: 978-87-7568-327-7
Online ISBN: 978-87-7568-328-4

DOI: https://doi.org/10.22439/phd.04.2025

FOREWORD / ACKNOWLEDGEMENT

After 25 years immersed in the high-octane world of investment banking and entrepreneurship, I found myself seeking the kind of cognitive spark that a rigorous academic pursuit could provide. Insights from cognitive psychology and theories on human decision-making illuminated a path for me to better understand the cognitive challenges in human-AI collaboration at a time when discussions about AI oscillated wildly between doomsday and utopian extremes.

The field of medicine, with its life-or-death stakes and evidence-based rigor, called out to me as a sphere where I could apply my newfound scholarly passion to make a meaningful difference. It is here, in the nuanced interstices of clinical decision-making, that I have sought to address the critical gaps in our understanding of AI—a quest that potentially has implications beyond healthcare.

First of all, I would like to express my deepest gratitude to my supervisors, Ioanna Constantiou, Tom Kirchmaier, and Åsa Melhus, for their invaluable guidance, wisdom, and encouragement throughout this journey.

My heartfelt thanks also go to my co-authors, who contributed richly to the collaborative research process. I am especially grateful to David Stuckler for teaching me the intricacies of systematic literature reviews and the art of academic writing.

A special appreciation goes to the memory of Daniel Kahneman, who graciously took the time to discuss my research and whose work has profoundly inspired my intellectual curiosity. My gratitude extends to Erik Brynjolfsson for his generous time and for sharing a visionary perspective on the future of AI at MIT and Stanford.

I would also like to acknowledge my friends and colleagues, Erik Urnes, Carl-Johan Hagman, and John Abrahamson, whose insights and support were instrumental as I considered stepping into the world of research.

Finally, to my family—my wife, Katarina, our daughter, Philippa, and our sons, Carl and Victor—my deepest love and gratitude for your boundless patience and unwavering support. This accomplishment would have remained a distant dream without each of you by my side.

Rikard Rosenbacke

ABSTRACT

English abstract

Artificial Intelligence (AI) have the potential to transform healthcare. Applications are far-reaching, from diagnosing and detecting disease through to implementing treatments and surgeries. Yet curiously, while AI is now being integrated into diverse economic sectors like finance, retail, and automotive, healthcare institutions have been slow to adopt AI. The reasons for this slow uptake are multiple but relate to low trust in AI among clinicians and a conflict with the prevailing culture of evidence-based medicine, whereby physicians critically engage in diagnostic discourse to reach clinical decisions. There is a growing shift toward explainable AI (XAI) systems that promise to transform the opaque "black-box" into a more interpretable "glass-box."

This thesis aims to develop and test a framework for understanding how clinicians collaborate with AI and XAI. In so doing, I aim to move beyond common characterizations of "AI aversion" or "AI appreciation," which have been used to describe when clinicians engage with AI or not, to understand the cognitive underpinnings of clinicians' engagement with AI. I further seek to understand when AI collaboration is effective, leading to more accurate medical decisions or worsening performance, leading to more or new errors. To do so, I perform a mixed-methods study of clinician-AI collaboration dynamics, with a focus on trust, errors, and heuristics.

Paper I (published in the *Journal of Medical Internet Research AI*) is a systematic literature review of empirical studies on clinicians' trust in different types of AI, including both AI, which is a 'black box,' where AI provides diagnostic advice, and XAI, in which AI provides advice accompanied by clinical explanations. The comprehensive review found that clinicians had greater trust in XAI, likely reflecting better coherence with evidence-based medicine, but that they could also create an overreliance on XAI, risking potential new errors. This review revealed a number of critical gaps, including how to optimize trust so that clinicians' reliance on AI corresponded to an actual improvement in clinical accuracy.

In Paper II, I drew upon these insights to create a novel decision-making framework for clinician-AI collaboration, mapping the potential errors that could occur. Specifically, these included: *True Conflict errors* (when the clinician is incorrect but does not heed the correct AI advice), *False Conflict errors* (where the clinician is correct but is persuaded by an incorrect AI to adopt an incorrect diagnosis), and *False Confirmation errors* (where an incorrect clinician is falsely confirmed by an incorrect AI). Prior research had extensively focused on the first error, but largely ignored the latter two possibilities, in part reflecting dubious assumptions about infallibility of AI, as I show in the thesis. This novel framework then laid the foundation for successive mixed-methods investigations, integrating quantitative data on errors and qualitative data to provide "thick" descriptions of the cognitive challenges associated with them.

To test this framework, I recruited eleven physicians, asking them to diagnose recurrent ear infections in children, based on data from previous medical studies. The physicians made an

initial diagnosis without AI, which formed a "baseline" for comparison. Subsequently, they were given the opportunity to revise their diagnosis with the aid of AI, followed by XAI. In total, they made 330 diagnoses, which enabled me to track how the physicians engaged with AI and whether and when decision errors were made. At each step, I interviewed physicians to describe their reasoning process (so-called "think-aloud" method) in deciding whether or not to trust the AI diagnosis.

With this experimental design, I was able, for the first time to my knowledge, to generate several key insights that significantly advance our understanding in this area. (Paper II). Quantitatively, I found that two major errors identified in my framework, False Conflict and False Confirmation, were responsible for the majority of errors made in clinician-AI collaboration. Although AI tended to improve overall diagnostic performance slightly, in several cases, it induced additional errors when it persuaded physicians to switch to an incorrect diagnosis. Qualitatively, the interviews revealed that physicians had considerably more trust and engagement with XAI than AI overall. In cases of False Confirmation, physicians appeared to blindly trust both AI and XAI, creating risks of undetected medical errors.

Building on these insights, Paper III (*Proceedings in ECIS 2024*) sought to characterize the underlying psychological heuristics that could account for these decision-making patterns. While I identified evidence consistent with multiple heuristics, the main two that correlated with errors were i) commitment bias, when physicians stubbornly clung to their initial, incorrect decision; and ii) confirmation bias, when physicians failed to seek additional information when AI and XAI confirmed an incorrect decision.

Having revealed the prominence of False Confirmation and False Conflict errors, I then revisited two contemporary seminal papers that had neglected their importance. Drawing on my novel framework, in Paper IV (published *in British Medical Journal - Medical Ethics),* I was able to respond critically to authors who had argued that AI could serve as a "second medical opinion", an important institution in the practice of evidence-based medicine. They argued that no action would be needed when AI confirmed physicians' judgment. My data in Paper II revealed that this scenario corresponded to over two-thirds of all errors, posing major risks to acting upon this advice. Instead, I laid out an alternative framework for when and how AI could serve as a second opinion, arguing that a management framework should be calibrated to the risks to patients and the underlying accuracy of the AI instrument.

In Paper V (published in *Nature Communications*), I tested my finding's reproducibility; I reanalyzed a dataset of dermatologist-AI collaboration to diagnose melanoma (109 clinicians made 4,512 diagnoses). Similar to prior studies, it focused on True Conflict errors. Revisiting their data, I revealed a significant increase in new False Conflict errors with AI collaboration, which ultimately negated most improvements in True Conflict cases. This issue was even more pronounced among the most experienced clinicians. In this study, there was also a large proportion of undetected False Confirmation errors. Importantly, these papers externally validated and reproduced the findings from my Paper II study.

Through the development of a novel, systematic framework to identify and measure diagnostic errors in clinician-AI collaborations, this research provides a significant contribution to understanding the complexities of AI and XAI in healthcare. While this work addresses a critical gap, it also lays an important foundation for future studies aiming to reduce errors and improve the clinician-AI partnership. Hopefully, these findings offer a practical approach to enhancing the reliability and safety of AI integration in medicine, supporting the broader goal of optimizing AI's role in clinical decision-making

These contributions are distinct and, to my knowledge, represent the first thorough exploration of the intersection between cognitive bias, trust dynamics, and diagnostic errors in clinician-AI collaboration. Each paper presents a unique contribution, offering valuable insights into different aspects of clinician-AI interactions. What makes this thesis particularly significant is how the papers build upon each other. Together, the thesis creates a framework that provides a more holistic and synergistic understanding than each single study could offer alone, making a novel contribution to the scientific literature and offering practical implications for optimizing AI integration in healthcare.

While the contributions of this thesis mark an important step in understanding cognitive challenges in clinician-AI collaboration, it is only the beginning. AI has considerable promise for improving healthcare, but to realize this potential, future research will be needed to identify interventions that can effectively mitigate False Conflict and False Confirmation errors while optimizing AI systems to enhance reasoned clinician judgment without overreliance or distrust.

Svensk sammanfattning

Artificiell intelligens (AI) har potentialen att förändra hälso- och sjukvården. Användningsområdena är omfattande, från att diagnostisera och upptäcka sjukdomar till att implementera behandlingar och operationer. Trots detta har hälso- och sjukvårdsinstitutioner varit långsammare med att ta till sig AI jämfört med andra sektorer som finans, detaljhandel och fordonsindustrin. Skälen till denna långsamma anpassning är flera, men relaterar ofta till låg tillit till AI bland kliniker och en konflikt med den rådande kulturen av evidensbaserad medicin, där läkare engagerar sig kritiskt i diagnostiska beslut. Därför ser vi ett ökat intresse för förklarande AI (XAI), som syftar till att förvandla den "svarta lådan" till modeller som är mer transparenta.

Denna avhandling syftar till att utveckla och testa ett ramverk för att förstå hur kliniker samarbetar med AI och XAI. Jag försöker gå bortom de rådande beskrivningar som "AI-aversion" eller "AI-appreciation", vilka ofta används för att beskriva om kliniker litar på AI eller inte, för att istället studera de kognitiva utmaningarna som uppstår när kliniker interagerar med AI. Jag försöker även att förstå när AI-samarbete leder till bättre medicinska beslut och när AI leder till fler eller nya fel.

Jag gör både kvalitativa och kvantitativa studier för att undersöka dynamiken i kliniker-AI-samarbete, med fokus på tillit, fel och heuristik. Artikel I (publicerad i *JMIR AI*) är en systematisk litteraturöversikt av empiriska studier rörande klinikers tillit till olika typer av AI, inklusive både AI som fungerar som en "svart låda", där AI ger diagnostiska råd, och XAI, där AI ger råd tillsammans med kliniska förklaringar. Denna omfattande översikt visade att kliniker hade större tillit till AI med förklaringar, vilket sannolikt speglar en samstämmighet med evidensbaserad medicin. Men det fanns också en risk för överdriven tillit till XAI, vilket kan skapa nya potentiella fel.

I artikel II använde jag dessa insikter för att skapa ett ramverk för kliniker-AI-samarbete där jag kartlägger de potentiella fel som kunde uppstå. Specifikt inkluderade dessa: sanna konfliktsfel (när klinikern har fel men ignorerar det korrekta AI-rådet), falska konfliktsfel (när klinikern har rätt men övertalas av ett felaktig AI till att anta en felaktig diagnos) och falska bekräftelsefel (när en felaktig kliniker falskt bekräftas av felaktig AI). Tidigare forskning hade främst fokuserat på det första felet men i stort sett ignorerat de andra två möjligheterna, delvis på grund av tvivelaktiga antaganden att AI har rätt.

För att testa detta ramverk rekryterade jag elva läkare och bad dem att diagnostisera återkommande öroninfektioner hos barn, baserat på data från tidigare medicinska studier. Läkare gjorde en initial diagnos utan AI, vilket utgjorde en "baslinje" för jämförelse. Därefter fick de möjlighet att revidera sin diagnos med hjälp av AI, följt av XAI. Totalt gjorde de 330 diagnoser, vilket gjorde det möjligt för mig att se hur läkarna interagerade med AI och om och när diagnosfel gjordes. Vid varje steg intervjuade jag läkarna för att beskriva deras resonemangsprocess (så kallad "think-aloud"-metod) för att se huruvida de skulle lita på AI eller XAI-diagnosen.

Med hjälp av denna experimentella design kunde jag göra flera viktiga insikter (Artikel II). Kvantitativt fann jag att de två stora felen som identifierades i mitt ramverk, falska konflikter och falska bekräftelser, var orsaken till majoriteten av de fel som gjordes i samarbetet mellan kliniker och AI. Även om AI tenderade att förbättra den övergripande diagnostiska prestandan, orsakade det i flera fall ytterligare fel när AI övertygade läkare att byta till en felaktig diagnos. Kvalitativt visade intervjuerna att läkare hade avsevärt större tillit till XAI än för AI. I fall av falsk bekräftelse verkade läkarna blint lita på både AI och XAI, vilket ökade risken för oupptäckta medicinska fel.

Utifrån dessa insikter syftade artikel III (konferenspublikation i *ECIS 2024*) till att karakterisera de underliggande psykologiska heuristiker som kunde förklara dessa beslutsmönster. Medan jag identifierade bevis som överensstämmer med flera heuristiker, var de två främsta i) ”commitment bias”, när läkare envist höll fast vid sin initiala, felaktiga diagnos; och ii) bekräftelsebias, när läkare slutat att söka ytterligare information när AI och XAI bekräftade en initial kliniskt felaktig diagnos.

Efter att ha visat frekvensen av falska bekräftelser och falska konflikter undersökte jag två banbrytande artiklar som hade negligerat dessa fels betydelse. Med utgångspunkt i mitt ramverk, i artikel IV (publicerad i *British Medical Journal - Medical Ethics*), kunde jag kritiskt svara författarna som hade hävdat att AI kunde fungera som en "second opinion", en viktig del i evidensbaserad medicin. De argumenterade för att inga åtgärder skulle behövas när AI bekräftade läkarnas bedömning. Mina data i artikel II visade att detta scenario motsvarade över två tredjedelar av alla fel, vilket utgör stora risker om man följer författarnas råd. Istället lade jag fram ett alternativt ramverk för när och hur AI kunde fungera som en ”second opinion”, och argumenterade för att ett ramverk borde kalibreras med hänsyn till riskerna för patienterna och AIs träffsäkerhet.

I artikel V (publicerad i *Nature Communications*) testade jag reproducerbarheten av mina fynd; jag alyserade ett data set där dermatologer samarbetade med AI för att diagnostisera melanom (109 kliniker gjorde 4 512 diagnoser). Liknande tidigare studier fokuserade de på sanna konfliktsfel men när jag analyserar deras data avslöjade jag en betydande ökning av nya falska konfliktsfel i AI-samarbete, vilket reducerade de flesta förbättringar i de sanna konfliktsfallen. Detta problem var ännu mer uttalat bland de mest erfarna klinikerna. I denna studie fanns också en stor andel oupptäckta falska bekräftelsefel. Dessa artiklar validerade och reproducerade i huvudsak resultaten från min studie i Artikel II.

Genom att utveckla ett nytt, systematiskt ramverk för att identifiera och mäta diagnostiska fel i samarbeten mellan kliniker och AI, bidrar denna forskning med ett viktigt perspektiv på komplexiteten kring AI och XAI inom hälso- och sjukvården. Samtidigt som arbetet fyller en kritisk lucka, utgör det också en värdefull grund för framtida studier med målet att minska felaktigheter och förbättra samarbetet mellan kliniker och AI. Förhoppningen är att dessa resultat ska erbjuda ett praktiskt tillvägagångssätt för att öka tillförlitligheten och säkerheten vid integrering av AI i medicin, vilket stöder det bredare målet att optimera AI:s roll i kliniska beslutsprocesser.

Dessa bidrag är unika och, så vitt jag vet, representerar de den första grundliga utforskningen av samspelet mellan kognitiva biaser, förtroendedynamik och diagnostiska fel i samarbete mellan människa och AI. Varje artikel erbjuder ett unikt bidrag och ger värdefulla insikter i olika aspekter av interaktionen mellan kliniker och AI. Det som gör detta arbete särskilt betydelsefullt är hur artiklarna bygger vidare på varandra. Tillsammans skapar de en ram som ger en mer holistisk och synergistisk förståelse än vad någon enskild studie skulle kunna erbjuda. Summan av resultaten i avhandling är större än delarna, vilket gör det till ett nyskapande bidrag till den vetenskapliga litteraturen och erbjuder praktiska implikationer för optimering av AI-integrationen inom hälso- och sjukvården.

Medan bidragen i denna avhandling markerar ett viktigt steg i förståelsen av de kognitiva utmaningarna i kliniker-AI samarbete, är detta bara i sin linda. AI har stor potential att förbättra hälso- och sjukvården, men för att realisera denna potential behövs framtida forskning som identifierar interventioner som effektivt kan minska falska konflikt- och bekräftelsefel, samtidigt som AI-system behöver optimeras för att förbättra välgrundade kliniska beslut och undvika både blind tillit och blind misstro.

Original Papers

Throughout my PhD studies, I have authored nine papers, eight of which have been peer-reviewed, accepted, and published. As per the thesis guidelines, a maximum of five papers are included in this thesis, each referenced by its respective Roman numeral.

Thesis papers

I: Rosenbacke, R.; Melhus, Å.; Mckee, M.; Stuckler, D (2024) 'How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Healthcare: Systematic Review', *Journal of Internet Medical Research AI.*

II: Rosenbacke, R. (2024) 'Errors in Physician-AI Collaboration: Insights From a Mixed-methods Study of Explainable AI and Trust in Clinical Decision-making', *SSRN Electronic Journal.* doi: 10.2139/SSRN.4773350.

III: Rosenbacke, R. (2024) 'Heuristics and Errors in XAI-Augmented Clinical Decision-Making: Moving Beyond Algorithmic Appreciation and Aversion', *In Proceedings of the 32th European Conference on Information Systems (ECIS) Association for Information Systems*. AIS Electronic Library (AISeL).

IV: Rosenbacke, R.; Melhus, Å.; Mckee, M.; Stuckler, D (2024) 'The AI and XAI Second Opinion: The Danger of False Confirmation in human-AI Collaboration' *British Medical Journal - J. Med. Ethics* (2024). doi:10.1136/jme-2024-110074

V: Rosenbacke, R.; Melhus, Å.; Stuckler, (2024) 'False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making ' *Nature Communications. 2024 151* 15, 1–2 (2024)

Other relevant papers

McKee, M., Rosenbacke, R. & Stuckler, D. 'The power of artificial intelligence for managing pandemics: A primer for public health professionals' *Int. J. Health Plann. Manage.* (2024). doi:10.1002/HPM.3864

Nyberg, J., Rosenbacke, R. & Ben-Menachem, E. 'Digital clinics for diagnosing and treating migraine' *Curr. Opin. Support. Palliat. Care* 18, (2024).

Rosenbacke, R., Tajhizi, N., Constantiou, I. & Melhus, Å. 'Designing a Digital Artifact for Data Collection to Improve Daily ADHD Medication' *ECIS 2022 Res. Pap.* (2022).

Weis, C.; Hauser, K.; Rosenbacke, R.; Brinker, T., (2024) 'Investigating Interaction Errors in Clinical Decision-Making: Implications for Risk Understanding and XAI Assistance in Melanoma Diagnostics' *In Proceedings of the Cancer Prevention Research Conference, American Cancer Society.*

Wies, C., Hauser, K. & Brinker, T. J. 'Reply to: False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making' *Nature Communications. 2024 151* 15, 1–3 (2024).

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AI	Artificial Intelligence
BMJ	British Medical Journal
DARPA	Defense Advanced Research Projects Agency
DNN	Deep Neural Network
ECIS	European Conference on Information Systems
EBM	Evidence-Based Medicine
FN	False Negative
FP	False Positive
HCI	Human-Computer Interaction
IS	Information Systems
JMIR	Journal of Medical Internet Research
ML	Machine Learning
NPV	Negative Predictive Value
rAOM	Recurrent Acute Otitis Media or middle ear infection
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RCT	Randomized Controlled Trial
RQ	Research Question
SHAP	SHapley Additive exPlanations
SPSS	Statistical Package for the Social Sciences
TN	True Negative
TP	True Positive
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

TABLE OF CONTENTS

Cognitive Challenges in Human-AI Collaboration

Introduction

Background

Artificial Intelligence (AI) has the potential to revolutionize healthcare, with applications ranging from early diagnostics and personalized treatments to complex surgical procedures. (Brynjolfsson and Mcafee, 2017; Schwalbe and Wahl, 2020; Rajpurkar *et al.*, 2022; Moor *et al.*, 2023). While industries like finance, retail, and e-commerce have embraced AI-driven technologies, healthcare has been notably slower to adopt these advancements (Bruce, 2024). This reluctance largely stems from unique challenges in the medical field, where trust, ethics, and safety are paramount concerns (Reddy *et al.*, 2020; Petersson *et al.*, 2022). In healthcare, clinicians are often skeptical of AI because of its "black-box" nature, which makes it difficult for them to interpret or justify AI recommendations (Fazal *et al.*, 2018; Rajpurkar *et al.*, 2022). Given the responsibility clinicians bear for patient outcomes, this lack of transparency raises ethical concerns, especially regarding patient consent and informed decision-making (Reddy, 2022).

Explainable AI (XAI) has emerged as a potential solution to this problem, promising to make AI decision processes more transparent, thereby fostering trust among healthcare professionals (Gunning and Aha, 2019; Loh *et al.*, 2022). However, existing literature often assumes that XAI will automatically increase trust and adoption without sufficient empirical evidence to confirm this assumption (Rosenbacke *et al.*, 2024b). Furthermore, while XAI might enhance trust, there is a risk of overreliance on its explanations, especially when the AI's recommendations are incorrect, potentially leading to new types of diagnostic errors (Kiani *et al.*, 2020). Thus, understanding the nuanced role of XAI in clinical decision-making is essential to safely and effectively integrate these technologies into healthcare.

Before presenting the aim of this thesis and the primary research question, *'What are the cognitive challenges for clinicians in XAI collaboration?'* I will begin with an introduction to artificial intelligence, its applications in healthcare, and the potential role of explainable AI within clinical settings.

Artificial intelligence

Artificial Intelligence is sometimes misleadingly characterized as a single entity, when it actually refers to an expansive field within computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. AI is an interdisciplinary science with multiple approaches, but advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector and industry (Brynjolfsson and Mcafee, 2017; Brynjolfsson and Mitchell, 2017; Rajpurkar *et al.*, 2022; Moor *et al.*, 2023).

Machine Learning (ML), a subset of AI, involves the development of algorithms that can learn and make predictions or decisions based on data. These learning algorithms can identify patterns and features in the data they process, allowing them to make informed decisions without being explicitly programmed for each specific task. ML is often divided into supervised, unsupervised, and reinforcement learning, each differing in how the machine uses data to learn (Lecun, Bengio and Hinton, 2015; Goodfellow I, Bengio Y, 2016).

Deep Neural Networks (DNN), which fall under the umbrella of ML, are inspired by the structure and function of the human brain, Figure 1. They are composed of layers of interconnected nodes or "neurons" that can learn to recognize patterns of input data by adjusting the weights of the connections through a process known as backpropagation. DNNs are particularly well-suited for handling large and complex data sets, which has led to significant breakthroughs in fields such as image and speech recognition, natural language processing, and autonomous vehicles (Lecun, Bengio and Hinton, 2015; Goodfellow I, Bengio Y, 2016).

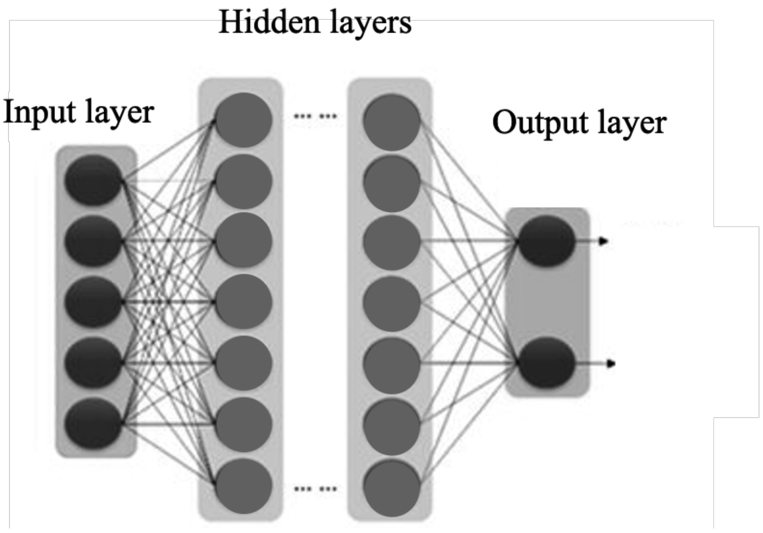


Figure 1: Illustration of a Deep Neural Network

Each node in a DNN acts like a tiny processor for performing computations. The basic operation involves receiving inputs from previous nodes (or the initial input data), which are then multiplied by weights—a form of parameter in neural networks that determines the influence of inputs. These weighted inputs are summed together, and often another parameter is added to this sum. The resulting value is then passed through a function known as an activation function, which determines the output of the node. This output becomes the input to the next layer of nodes. The activation function's purpose is crucial; it introduces non-linear properties to the network, enabling it to learn more complex patterns than just straight-line correlations among data. By adjusting the weights and parameters during the training process (using techniques like backpropagation and gradient descent), the network learns to make increasingly accurate predictions or decisions based on input data (Lecun, Bengio and Hinton, 2015; Goodfellow I, Bengio Y, 2016). Next, a transition to discussing AI applications in healthcare, along with the challenges and barriers to broader implementation.

AI technologies in healthcare

AI technologies are rapidly emerging as potentially important instruments for decision-making in modern healthcare (Schwalbe and Wahl, 2020; Rajpurkar *et al.*, 2022). There is hope and potential for AI to support clinical decision-making in a number of domains, including to improve clinical diagnostic assessments, reduce medical errors, streamline management processes like triage, and improve overall patient outcomes (Sutton *et al.*, 2020).

Yet perhaps curiously, in contrast to other business domains, the impact of AI on clinical practice is relatively modest (Rajpurkar *et al.*, 2022; Kolasa *et al.*, 2023). Uptake in medicine lags behind other sectors, such as media, defense, e-commerce, and software development. Where it has been applied, progress has been uneven. One recent umbrella meta-analysis (Kolasa *et al.*, 2023) of 220 systematic literature reviews covering over 7,000 papers mapped machine learning AI applications in healthcare over the past decade. It revealed that AI was primarily being applied to clinical prediction and prognosis of diseases, particularly for imaging data in the clinical specialties of oncology and neurology. One recent survey found that less than 5% of healthcare organizations are using AI tools (Dai and Ching, 2022), and only 9% of healthcare employees say they feel at an 'advanced' level of AI fluency, the least of all industries surveyed (Bruce, 2024). This slow and variable uptake across fields likely reflects a combination of ethical, technical, institutional, and practitioner-related hurdles (Petersson *et al.*, 2022). Overcoming these challenges will be crucial for the safe and successful application of AI in healthcare (Rajpurkar *et al.*, 2022).

Challenges with AI in healthcare

Multiple reasons have been posited for the slow uptake of AI in healthcare. Ethical concerns have been raised, including by physicians themselves, about the ability of AI to respect confidentiality and patient privacy (Rajpurkar *et al.*, 2022). AI may also have important risks to patient safety, even if it can outperform doctors' own accuracy, analogous to those identified with automatic driverless vehicles. Clinicians may also face perceived existential threats, as with other labor market sectors, of redundancy and deskilling as AI tools begin to replace their roles (Chew and Achananuparp, 2022). Further, if clinicians cannot fully comprehend AI decisions and so articulate them to patients, it could infringe upon patients' rights to informed consent and autonomy (Reddy, 2022).

There are additionally a series of technical and institutional challenges (Rajpurkar *et al.*, 2022). One pertains to evidence on quality improvement of AI applications. For example, the umbrella systematic review (Kolasa *et al.*, 2023) found suboptimal and inconsistent quality in reporting on the development and modification of machine learning algorithms intended for clinical application. At least one-third of published studies failed to evaluate the accuracy of the AI systems. Additionally, AI applications in healthcare will likely need to undergo regulatory approval processes similar to those for other medical devices and drugs, which can create long lag periods between innovation and successful integration into practice (Reddy *et al.*, 2020). Institutionally, barriers arise from investments made by hospitals and clinics into technologies which do not interface with the newer generation of AI tools, so creating effective barriers to modernization.

Notwithstanding these limitations, in general, clinicians have demonstrated a reluctance to change patterns of practice so as to implement AI tools into their clinical routines (Gupta, Boland and Aron, 2017). However, recent research highlights that despite this reluctance to AI, there is a growing momentum toward digitalization in healthcare, driven by the need for personalized care and precision medicine (Constantinides, 2023). These trends underscore that while some clinicians may be slow to adopt AI, others are embracing digital transformation in response to evolving patient care needs. A recent scoping review of research on healthcare providers' perceptions begins to explain why there is a general reluctance to implement AI in healthcare (Chew and Achananuparp, 2022). It revealed that, curiously, although healthcare providers had an overall favorable view of AI, recognizing its potential for enhancing service efficiency and reducing costs due to its accessibility and ease of use, they hesitated to embrace it. This apprehension stemmed from ethical and technical concerns, including about trustworthiness (pertaining to accuracy), data confidentiality, patient safety, the current state of technology, and the possibility of AI leading to full automation, which could render physicians obsolete. Another challenge that healthcare providers have highlighted pertains to AI systems' complexity which could not only hinder their uptake but also effective use in clinical settings (Beede *et al.*, 2020), and unexpected challenges might emerge from the interactions between humans and AI. Kiani and colleagues found that their AI algorithm improved physicians' accuracy as long it was correct. However, when the AI advice was incorrect, overall accuracy decreased significantly, irrespective of pathologist experience or case difficulty levels (Kiani *et al.*, 2020). In an additional review, tracking weekly updates in medical AI over the last two years, the authors (Rajpurkar *et al.*, 2022) concluded that the adoption of AI systems in daily clinical operations presents a substantial yet underutilized opportunity.

Another major barrier to the integration of AI in healthcare is its conflict with current cultural prevailing practices of evidence-based medicine (EBM). EBM calls for clear and transparent decision-making processes (Amann *et al.*, 2020; Kundu, 2021). Even if AI can make accurate diagnoses, physicians may not be able to interpret or engage with them. There is a common perception among doctors that AI operates as a “black box”, without providing clear justification for its health-related advice (Fazal *et al.*, 2018; Wadden, 2021; Reddy, 2022). When healthcare providers do not understand clinical advice, they are much less likely to use it (Cui and Zhang, 2021). Unlike rule-based diagnostic systems, draw upon intricate statistical frameworks, such as deep learning neural networks, that are inherently difficult for humans to interpret (Castelvecchi, 2016). This can make AI platforms are less transparent, making both their decision-making improvement and errors harder to identify (Jussupow *et al.*, 2021). Clinicians could also be forgiven for some degree of skepticism, as less than 1% of AI healthcare algorithms have been externally validated (that is tested and confirmed for accuracy beyond the scope of their initial training datasets) (Kolasa *et al.*, 2023).

This lack of transparency is especially important to physicians-in-charge, who maintain the responsibility, including legal responsibility in several jurisdictions, for the final decision-making (Reddy, 2022). They also may need to be able to articulate the clinical rationale to patients or their families. Thus human oversight is arguably indispensable in healthcare (Jongsma and Sand, 2022). Rajpurkar and colleagues argue that collaborative arrangements warrant further investigation, as they have the potential to independently outperform the capabilities of either AI or humans and are likely to be more representative of genuine medical practice. Rajpurkar and colleagues argue that they "*would like to see collaborative setups receive more study*" (Rajpurkar *et al.*, 2022, p. 6).

The role of Explainable AI in healthcare

Recently, explainable AI has been developed to overcome transparency limitations, increasing its uptake in diverse management domains as well as healthcare (Nazar *et al.*, 2021; Loh *et al.*, 2022). The opacity of AI is a primary barrier to its practical applications, particularly in healthcare settings (Loh *et al.*, 2022). Consequently, XAI has emerged as a methodology aimed at bolstering decision confidence (Kepecs *et al.*, 2008) or trust (Chanda *et al.*, 2024) in AI predictions by elucidating the processes through which these predictions are made. This transparency is intended to foster greater utilization of AI systems in healthcare (Constantiou, Joshi and Stelmaszak, 2024).

Several researchers have proposed that XAI will foster greater AI uptake in clinics. They argue that XAI is essential for securing the safety, approval, and adoption of AI systems among both providers themselves and the institutions where they work (Antoniadi *et al.*, 2021; Evans *et al.*, 2022; Reddy, 2022; Haque, Islam and Mikalef, 2023; Chanda *et al.*, 2024). The central goal of XAI, according to computer scientists and developers, is building trust and doing so through greater transparency (Gerlings, Shollo and Constantiou, 2021). For example, the US defense XAI program Defense Advanced Research Projects Agency (DARPA) underscores the necessity of XAI for understanding, trusting, and effectively managing the next wave of AI technologies (Gunning and Aha, 2019).

Not everyone agrees, however. Ghassemi and colleagues argue in *Lancet Digital Health,* that current applications of explainable AI are flawed, offering only incomplete insights into the inner mechanics of AI algorithms (Ghassemi, Oakden-Rayner and Beam, 2021). They argue that AI systems are fundamentally simplified models of reality, and the explanations provided by XAI are merely additional layers of simplification. This can lead to oversimplifications that may misrepresent the underlying complexity of the AI's decision-making process, potentially leading to misunderstandings or misplaced trust. They advocate for stakeholders to shift their focus from insisting on explainability and to seek alternative approaches, such as better validation processes of AI algorithms through randomized controlled trials, as an alternative pathway for integrating AI insights into practice. This would synchronize the contribution of AI with the clinical systems for evaluating accuracy and, as a result, trustworthiness. Thus, Ghassemi and colleagues argue that the potential of XAI for improving healthcare is limited at best.

Reddy questions this pessimistic view. He argues that notwithstanding the known limitations that Ghassemi and colleagues cite (such as how XAI mostly approximates the underlying machine learning mechanisms to explain decision-making), XAI adds considerable value to the prevailing culture of evidence-based medicine (Kundu, 2021; Reddy, 2022). Reddy argues that, without XAI, the AI "black-box" medical models prevent clinicians from evaluating their quality, potentially breaching patient consent and autonomy. Only when physicians can interpret the clinical logic behind the AI diagnosis can they meaningfully collaborate with AI in a way consistent with evidence-based medicine. If, as Ghassemi and colleagues suggest, we neglect explainable AI, we inevitably hamper AI integration into healthcare, even if randomized controlled clinical trials prove its efficacy because the lack of explanations limits accountability, trust, and compliance, so making it difficult for doctors to, for example, explain diagnoses to patients and fulfill their ethical responsibilities. Reddy argues that explainable AI fosters trust and transparency, better aligning AI performance with clinical standards and evidence based medicine (Reddy, 2022).

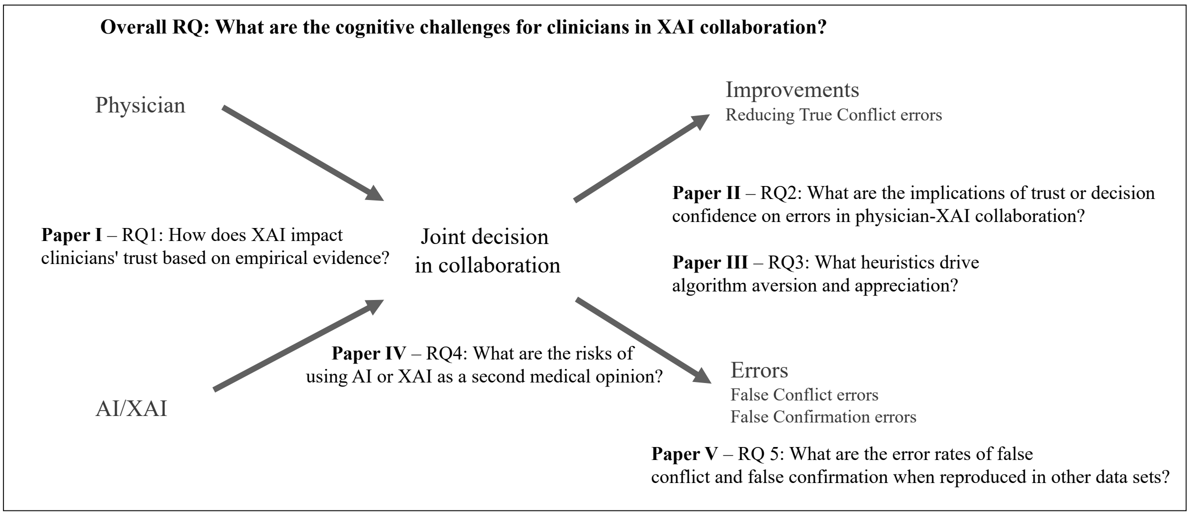
In summary, while it is difficult to pinpoint a specific singular factor, the lack of transparency, and resulting lack of trust, appear to be major forces underlying the relatively low acceptance of AI among healthcare practitioners (Lee *et al.*, 2017; Rajpurkar *et al.*, 2022; Reddy, 2022). In this thesis, I explore the cognitive challenges of human-AI collaboration, with a particular emphasis on the impact of incorporating explanations into AI-generated advice.

Aims of the thesis

This thesis will begin to respond to these collaborative challenges through a series of investigations of how clinicians engage with AI versus XAI to make decisions and go beyond past analyses to understand the psychological underpinnings and cognitive challenges of effective or ineffective human-AI collaboration. It will specifically investigate the potential for AI innovations like explainable AI to better align with evidence-based medicine and, in so doing, facilitate not only uptake but also effective integration of AI improvements into clinical practice. Hence, the overarching research question of the thesis is as follows:

Overall RQ: What are the cognitive challenges for clinicians in XAI collaboration?

The series of empirical studies included in the thesis, the research questions, and their links are depicted in Figure 2 below.



Overall RQ: What are the cognitive challenges for clinicians in XAI collaboration?

Physician
AI/XAI
Joint decision in collaboration
Improvements
Reducing True Conflict errors
Errors
False Conflict errors
False Confirmation errors

Paper I – RQ1: How does XAI impact clinicians' trust based on empirical evidence?

Paper II – RQ2: What are the implications of trust or decision confidence on errors in physician-XAI collaboration?

Paper III – RQ3: What heuristics drive algorithm aversion and appreciation?

Paper IV – RQ4: What are the risks of using AI or XAI as a second medical opinion?

Paper V – RQ 5: What are the error rates of false conflict and false confirmation when reproduced in other data sets?

Paper I: How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Healthcare: a Systematic Review

Paper II: Errors in Physician-AI Collaboration: Insights from a Mixed-Methods Study of Explainable AI and Trust in Clinical Decision-Making

Paper III: Heuristics and Errors in XAI-Augmented Clinical Decision-Making: Moving Beyond Algorithmic Appreciation and Aversion

Paper IV: The AI and XAI Second Opinion: The Danger of False Confirmation in human-AI Collaboration

Paper V: False Conflict and False Confirmation Errors are Crucial Components of AI Accuracy in Medical-Decision Making

Figure 2: Framework of the interrelation of research papers

Throughout my PhD studies, I have authored nine papers, of which eight have been peer-reviewed, accepted, and published at the time of writing (Rosenbacke *et al.*, 2022, 2024b, 2024a; McKee, Rosenbacke and Stuckler, 2024; Nyberg, Rosenbacke and Ben-Menachem, 2024; Rosenbacke, 2024a, 2024b; Rosenbacke, Melhus and Stuckler, 2024; Weis *et al.*, 2024) In line with the thesis guidelines, a maximum of five papers are included here. Each included paper has addressed distinct yet interconnected aims, which I synthesize within this thesis according to the framework illustrated in Figure 2, depicting the interrelation of my research contributions.

Current literature indicates that clinicians often struggle to fully trust AI due to its "black-box" nature (Rosenbacke *et al.*, 2024b). Although XAI aims to increase transparency, actual evidence of its impact on clinician trust is sparse. The first aim is to systematically examine empirical evidence to determine whether XAI genuinely builds trust or, conversely, fosters overreliance, potentially leading to diagnostic errors.

Additionally, while much of the existing research focuses on how clinicians respond to accurate AI advice, there is a lack of studies exploring how XAI affects decisions when the AI guidance is incorrect (Rosenbacke, 2024a). This research aims to fill this gap by investigating the types of errors that may arise in such scenarios, offering new insights into how XAI influences decision-making accuracy in complex clinical settings.

Furthermore, I aim to explore the underlying cognitive heuristics and biases that contribute to errors and clinicians' tendencies to either over-rely on or avoid AI advice (Rosenbacke, 2024b). This work aims to deepen our understanding of the psychological factors influencing clinician-AI collaboration and investigates whether XAI exacerbates or mitigates these biases.

A critical aspect of this thesis is the examination of XAI's role as a "second opinion" in clinical settings. This research aims to investigate the potential risk of False Confirmation errors, where clinicians and AI mutually reinforce each other's incorrect conclusions, which can jeopardize diagnostic accuracy (Rosenbacke *et al.*, 2024a).

Finally, I aim to assess the generalizability and robustness of my findings across different datasets and medical contexts. By examining the consistency of these cognitive errors, particularly False Confirmation errors, in varied clinical environments, the research aims to establish a broader, more reliable framework for clinician-AI interactions (Rosenbacke, Melhus and Stuckler, 2024).

The aim of the thesis kappa is to synthesize the key findings across the five individual papers and highlight how their collective insights generate synergies that go beyond what each paper could achieve in isolation. Without the integration provided in the kappa, the connections and interplay between the papers might remain obscured. The kappa brings these threads together, making the sum of the thesis more valuable and impactful than the individual parts. It draws on the distinct contributions of each paper to build a comprehensive framework for understanding the cognitive challenges and dynamics in human-AI collaboration, emphasizing how these insights collectively advance our understanding of optimizing AI in clinical settings.

Before presenting each of the five interlinked empirical studies and corresponding research questions that form the core of this thesis, I will outline the interdisciplinary approach, discuss my research philosophy, and review the relevant theoretical models and methodologies that guide this work.

Interdisciplinary approach

Drawing on different disciplines, this thesis adopts a multi-faceted approach to explore how clinicians collaborate with AI systems. The study utilizes both qualitative and quantitative methods, including a systematic literature review, field study data sets from Sweden and Germany, statistical analyses, and qualitative interviews. Through this diverse methodology, the thesis seeks to provide a comprehensive understanding of the opportunities and risks associated with human-AI collaboration in clinical practice while also addressing the broader governance and ethical challenges.

Approach and Positioning

This thesis is positioned at the crossroads of several key disciplines, reflecting the multi-disciplinary nature of research at the Center for Corporate Governance and the Department of Digitalization at Copenhagen Business School. By integrating perspectives from corporate governance, digitalization, healthcare, computer science, and cognitive psychology, the research addresses the complex interplay between AI systems and human decision-making within healthcare (Figure 3).

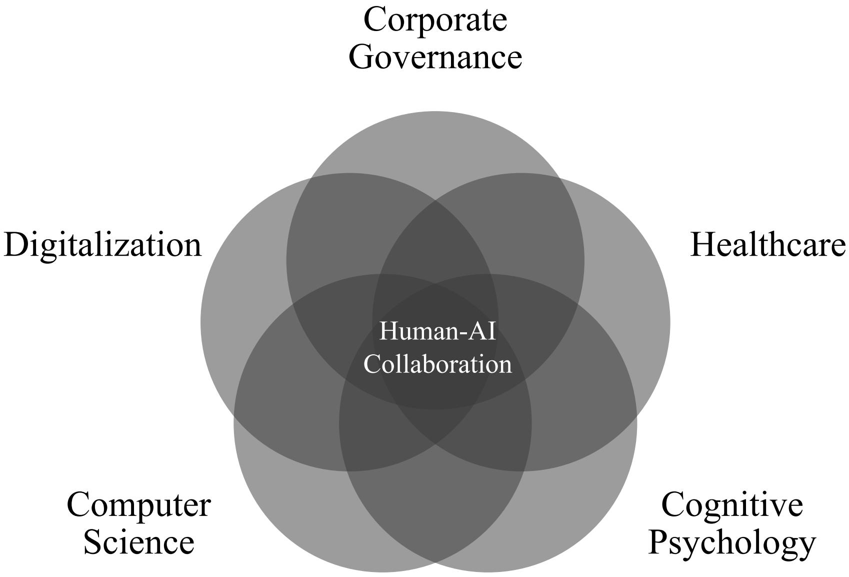


Figure 3: Interdisciplinary approach

The transformation of healthcare through digitalization and AI technologies forms the core of this study. Concepts like digital transformation are vital, reflecting the integration of advanced technologies into healthcare processes, with a particular focus on clinical decision-making. In addition to this thesis, I have also published papers related to digital transformation in the care of migraine patients (Nyberg, Rosenbacke and Ben-Menachem, 2024) and ADHD patients (Rosenbacke *et al.*, 2022).

At its foundation, the study investigates the corporate governance and oversight of AI integration into healthcare institutions, assessing the ethical, regulatory, policy and accountability frameworks necessary when utilizing AI systems (McKee, Rosenbacke and Stuckler, 2024; Rosenbacke *et al.*, 2024a). This includes understanding how AI reshapes decision-making authority and responsibility within healthcare organizations.

The cognitive underpinnings of decision-making are explored through the lens of cognitive psychology, particularly how clinicians trust, interpret, and respond to AI recommendations (Rosenbacke, 2024a; Rosenbacke *et al.*, 2024b; Rosenbacke, Melhus and Stuckler, 2024). This research draws on cognitive theories to explain biases, heuristics (Rosenbacke, 2024b), and mental models (Rosenbacke, 2024a) that influence how clinicians interact with AI in clinical settings.

The study delves into how clinicians interact with AI systems in real-world settings (Rosenbacke, 2024a) through principles of human-computer interaction (HCI). This approach helps explore usability and overall user experience, which are critical to understanding how AI tools can facilitate trust, ease of use, and efficient collaboration between humans and machines.

Information systems (IS) play a vital role in this research, providing the infrastructure for how data-driven technologies, including AI systems, collect, process, and present information to clinicians. IS is key to understanding how AI integrates into healthcare workflows (Rosenbacke *et al.*, 2024a), ensuring data security, interoperability, and decision support.

IT and e-health provide the technological foundation for AI systems, covering the development of tools like XAI and their practical implementation in clinical environments (McKee, Rosenbacke and Stuckler, 2024). This research explores how the operationalization of AI and the digital healthcare platforms are transforming patient care.

The combination of these diverse disciplines creates a unique vantage point, enabling the thesis to bridge gaps between corporate governance (responsibility and accountability frameworks), digitalization and IS (technical and operational capabilities), and the human-centered focus of clinical practice (psychology, HCI, and healthcare). This interdisciplinary approach allows for a comprehensive understanding of both the opportunities and cognitive challenges involved in integrating AI into healthcare workflows. This synthesis is closely aligned with the pragmatic philosophy, emphasizing a balance between theory, evidence based medicine and practical outcomes in real-world contexts.

Philosophy of science:

In this thesis, the integration of artificial intelligence within healthcare is examined from a pragmatic perspective, emphasizing practical relevance and real-world problem-solving (Alvesson and Sandberg, 2011). Pragmatism, as rooted in the works of philosophers like John Dewey, prioritizes actionable knowledge and focuses on resolving tangible issues (Morgan, 2014; Kaushik and Walsh, 2019). This approach insists not only on theoretical exploration but also on the empirical validation of AI benefits to clinical decision-making and patient outcomes (Alvesson and Sköldberg, 2000). Explainable AI takes a central role in this research, reflecting a commitment to demystifying AI decision-making processes, thereby enhancing transparency and balancing trust among clinicians (Alvesson and Kärreman, 2007).

This pragmatic stance is particularly suited to the complex interactions between clinicians and AI, where theoretical constructs must be empirically tested to yield concrete benefits in clinical decision-making and patient outcomes. By prioritizing actionable insights, pragmatism provides a flexible and outcome-oriented approach that aligns well with healthcare needs, especially as clinicians integrate AI systems into diagnostic and therapeutic processes.

A key aspect of pragmatism is its flexibility in methodology, particularly in mixed-methods research, as employed in this thesis. Creswell et al. argue that mixed methods are inherently pragmatic, allowing researchers to employ both qualitative and quantitative techniques to answer research questions effectively (Creswell and Plano, 2017). In studying clinician-AI collaboration, this research combines experimental data with qualitative insights, providing a comprehensive view of the cognitive and practical challenges clinicians encounter. This mixed-methods approach underscores the pragmatic focus on utility, which Dewey described as a measure of truth in terms of practical application (Kaushik and Walsh, 2019). Pragmatism's inherent flexibility supports methodological pluralism, allowing the researcher to adapt the methods to the problem at hand rather than adhere rigidly to a single epistemological stance (Morgan, 2014).

Epistemologically, which refers to the study of what we can know and how we can know it, this research emphasizes "epistemic utility," which refers to the practical value of knowledge in a real-world context. In simpler terms, this means that the research values knowledge that clinicians can use directly to make better decisions. In this study, it's not enough for AI to produce information—it must also provide insights that genuinely help clinicians diagnose more accurately and improve patient care. This aligns with philosopher John Dewey's view that knowledge should be a tool for action, emphasizing that useful, applicable knowledge is more valuable than abstract theories alone (Kaushik and Walsh, 2019).

Ontologically, which refers to how we see or define what is real or true, pragmatism permits a functional view of AI, regarding it not as a "self-thinking" autonomous entity but as a collaborative tool that complements clinical expertise. Pragmatism's rejection of rigid categories supports this perspective, focusing on the roles AI assumes in healthcare based on its practical applications rather than on strict definitions (Morgan, 2014). The trust that clinicians place in AI is scrutinized in light of AI's ontological nature—how XAI transforms the "black box" into a "glass box" of understandable and actionable information (Alvesson and Sandberg, 2011). Thus, AI is viewed as a flexible partner in the clinical environment, adapting to the needs of clinicians rather than operating in isolation.

Ethically, pragmatism, with its focus on the consequences of actions, provides a guiding framework for the responsible integration of AI in healthcare. In clinical settings, this pragmatic approach emphasizes transparency, accountability, and practical utility, which are essential for the ethical deployment of AI. This framework is particularly relevant for upholding patient autonomy, ensuring informed decision-making, and preserving clinicians' roles in AI-supported environments, where human oversight is indispensable for ethical and safe practice.

While evidence-based medicine (EBM) traditionally relies on empiricist and positivist principles, emphasizing objective observation and reproducibility in clinical practice (Timmermans and Mauck, 2005; Straus et al., 2019), the practical challenges of clinical settings have increasingly led to the integration of pragmatism as a complementary philosophy. Pragmatism is compatible with EBM's focus on real-world applicability, particularly in complex cases that may not fit neatly within the framework of randomized controlled trials (Tonelli, 2006; Greenhalgh, 2019). In practice, medical decision-making often involves balancing empirical evidence with clinician expertise, patient preferences, and situational factors—elements that align well with a pragmatic approach.

Pragmatism's adaptability allows for "what works" in specific contexts, a concept foundational to "implementation science" (Bauer and Kirchner, 2020) and "patient-centered care" (Epstein and Street, 2011) approaches. Unlike pure empiricism, pragmatism supports the integration of qualitative insights, subjective experiences, and clinician expertise, allowing evidence to be adapted to individual patient cases. Clinicians frequently encounter unique cases that deviate from the typical conditions studied in randomized controlled trials. Pragmatism, therefore, encourages a flexible approach to applying evidence, one that respects the situational nuances and ethical considerations inherent in patient care (Creswell and Plano, 2017).

In summary, pragmatism provides a flexible, problem-centered approach to clinician-AI collaboration, supporting methodological adaptability and a focus on outcomes. By adopting a pragmatic stance, this thesis seeks to yield findings that are both theoretically rigorous and practically meaningful, advancing our understanding of AI's role in healthcare and contributing to the development of safer and more effective AI systems.

Next, by building on the pragmatic approach's focus on practical outcomes, I elaborate on established decision-making models to interpret the cognitive processes clinicians experience when collaborating with AI.

Models for decision-making

To interpret the cognitive challenges and the decision-making process for the clinicians in the studies, I mainly draw upon the dual process model (Wason and Evans, 1974). Wason and Evans argue that there is a difference between unconscious behavior and conscious thought. The dual process model was later popularized by Daniel Kahneman as two systems of thinking—System 1, which is fast, automatic, and intuitive, and System 2, which is slower, more deliberate, and more reason-based (Kahneman, 2011).

The most common critiques against Kahneman's dual-process theory, particularly from scholars like Gary Klein (Klein, 2015) and Gerd Gigerenzer (Gigerenzer and Gaissmaier, 2011), focus on the overemphasis on cognitive biases and the underestimation of the power of intuition in expert decision-making. Klein, an advocate for naturalistic decision-making, suggests that Kahneman's model underrepresents how experts can make accurate decisions rapidly through pattern recognition and experience without the slow deliberation that Kahneman's System 2 suggests. Gigerenzer, meanwhile, criticizes Kahneman for not giving enough credit to heuristics' adaptive value. He sees heuristics as fast, frugal, and often accurate tools that can guide decision-making, especially in an uncertain world.

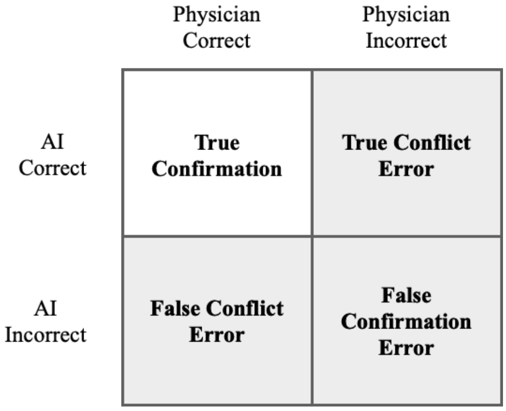
Psychologists hold differing opinions on whether heuristic errors stemming from System 1 can be corrected. One dominant perspective asserts that these errors cannot be rectified; at best, decision-makers can become aware of common decision-making pitfalls and attempt to avoid them. As the creator of the dual-process theory, Kahneman states, "*It's false to hope that if you become more aware of your errors you will make better decisions*" (Matias, 2017). Conversely, Klein's naturalistic decision-making approach offers a more optimistic view. This approach does not separate heuristic and systematic processes but rather explores the comprehensive cognitive processes that allow decision-makers to manage and regulate their reasoning (System 2) alongside their intuition (System 1) (Klein, 2015). Similar to Klein, Ackerman and Thompson argue that decision-makers can balance intuitive, heuristic, and deliberate reasoning activities by employing metacognition (Ackerman and Thompson, 2017).

In addition, Bandura's social cognitive theory emphasizes the role of self-efficacy in decision-making. According to Bandura, enhancing one's belief in their ability to make effective decisions can improve the use of both intuitive and deliberate reasoning, thereby potentially reducing errors (Bandura, 1986). This perspective adds the dimension of self-confidence and social learning to the discussion, suggesting that cognitive and emotional factors play a crucial role in managing decision-making processes. By fostering a stronger sense of self-efficacy, individuals may better leverage their cognitive abilities, aligning with Klein's and Ackerman and Thompson's views while also highlighting the importance of belief in one's capabilities.

Previous studies, (Buçinca, Malaya and Gajos, 2021; Naiseh, Cemiloglu, *et al.*, 2021; Bertrand *et al.*, 2022), tend to draw more on Kahneman's work to better understand the potential cognitive challenges in physician-AI collaboration while others (Jussupow *et al.*, 2021) use Klein's and Ackerman's approach that decision-makers can control and monitor their reasoning process to avoid being influenced by heuristics. Following this, I will elaborate on the categorization of decision-making outcomes, providing a structured framework for analyzing and interpreting these results.

Model for categorizing decision-making outcomes

When it comes to a model for categorizing decision-making outcomes, I draw on the work of Jussupow *et al.*, see Figure 4, which categorizes the possible missteps that can arise during human-AI interaction (Jussupow *et al.*, 2021). Building upon their 2x2 matrix, I introduce distinct labels for each quadrant, elucidating the nature of four potential errors: True Confirmation, True Conflict error, False Conflict error, and False Confirmation error. This contrasts with Jussupow et al.'s original taxonomy of "confirmation I and II" and "disconfirmation I and II." I contend that distinguishing between these errors is crucial, both conceptually and in terms of their cognitive implications.



	Physician Correct	Physician Incorrect
AI Correct	**True Confirmation**	**True Conflict Error**
AI Incorrect	**False Conflict Error**	**False Confirmation Error**

Figure 4: Model for categorizing decision-making outcomes adopted from Jussupow et al.

However, this 2x2 cross-table has its constraints. It resembles of a confusion matrix used in machine learning (Zhang *et al.*, 2021) since it offers a crisp classification system for decision-making outcomes. However, its binary structure is less effective in the medical decision-making context, where uncertainty prevails, and the 'maybe' answers cannot be ignored. These intermediate cases, where neither a clear positive nor negative can be asserted, challenge the matrix's dichotomous approach. Furthermore, the reliance on a gold standard for diagnostic accuracy is problematic in the medical field, where the truth is a shifting concept. The expectation of 100% certainty in diagnosis from AI or clinicians is unattainable; the ground truth is not always clear, as often evidenced when a patient's cause of death is not definitively known (De Koning *et al.*, 2003). This issue is compounded in patients with comorbidities, who present with multiple interlinked diseases that can obscure and complicate the diagnostic process. In essence, the utility of the 2x2 matrix is constrained by the complex realities of medical practice, which frequently exist in the gray areas between the binary opposites of traditional classification systems.

Methods and materials

This study utilizes a mixed-methods approach to analyze AI impact on clinical decision-making, integrating qualitative and quantitative data for a more holistic perspective. Pragmatism, which values methodological flexibility, supports this combination of methods to address my complex research questions effectively (Creswell and Plano, 2017). This approach aligns with the view that the "truth" of a method is in its practical application, focusing on insights that enhance real-world utility (Kaushik and Walsh, 2019). By embracing pragmatism, this research adapts methods to the research problem itself, embodying methodological pluralism (Morgan, 2014).

My research employs a variety of methodological approaches to examine the role of AI in clinical decision-making. The methodology draws from established frameworks in both medical research and information systems, ensuring a rigorous and systematic approach to data collection and analysis. Through the use of systematic literature reviews, field studies, thematic analyses, and statistical analyses, this thesis explores the cognitive, technical, and social dimensions of human-AI collaboration in healthcare settings.

Systemic literature review methodology

In the medical field, the hierarchy of evidence pyramid (Yetley *et al.*, 2017) is often used to describe the quality of evidence as shown in Figure 5. The pyramidal shape qualitatively integrates the amount of evidence generally available from each type of study design and the strength of evidence expected from indicated designs. In each lower level, the amount of available evidence generally declines. Study designs in higher levels of the pyramid generally exhibit increased quality of evidence and reduced risk of bias. Confidence in causal relations increases at the upper levels. Meta-analyses have the highest level of evidence followed by systematic reviews.

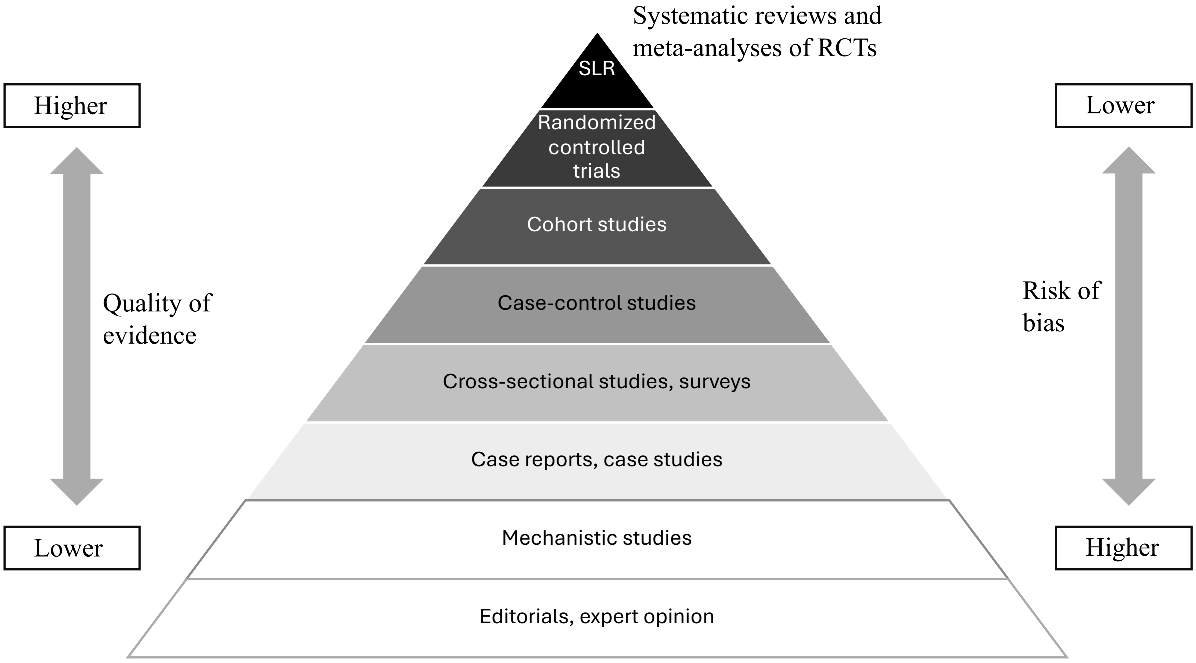

Systematic reviews and meta-analyses of RCTs
SLR
Randomized controlled trials
Cohort studies
Case-control studies
Cross-sectional studies, surveys
Case reports, case studies
Mechanistic studies
Editorials, expert opinion
Higher
Quality of evidence
Lower
Lower
Risk of bias
Higher

Figure 5: Hierarchy of evidence pyramid (Yetley *et al.*, 2017).

A systematic review differs from a conventional narrative literature review in that it involves a structured and replicable process summarizing the current research. Systematic reviews developed from the health sciences and other fields is now making its way into Information Systems (IS) research (Okoli and Schabram, 2012; Boell and Cecez-Kecmanovic, 2015).

In Paper I, we searched two different databases (PubMed and Web of Science), following the PRISMA guidelines (Page *et al.*, 2021). Studies were included if they empirically measured the impact of XAI on clinicians' trust, using either cognition- or affect-based measures. A total of 778 articles were screened. Ten of them fulfilled the inclusion and the exclusion criteria and were further analyzed. For further details, refer to Paper I (Rosenbacke *et al.*, 2024b).

Swedish field study data set

The impact of AI advice, as compared with XAI advice, on physicians' decision-making processes was investigated in a field study (Rosenbacke, 2024a). The algorithms were applied to data extracted from a previous vaccination trial conducted at the Department of Otorhinolaryngology, Head and Neck Surgery, Lund University Hospital, Lund, Sweden (Gisselsson-Solén *et al.*, 2014, 2015) concerning risk predictors of recurrent middle ear infections (rAOM) in young children. The data set presented a near-ideal scenario. The data was meticulously close to what one might consider 'ground truth'. The patient cases of otitis media, characterized by recurrent ear infections, were documented with a high degree of precision—affirmed by the definitions and diagnoses established by medical professionals. This level of detail and accuracy in the data set not only bolstered the credibility of the study but also provided a strong foundation for evaluating the algorithm, firmly anchoring it in the realities of clinical practice.

Physicians were presented with both correct and incorrect AI advices (60% accuracy), and their decision-making processes were followed in three steps: i) an initial part where physicians diagnose patients' risk of recurrent middle ear infections; ii) a second part where physicians have the opportunity to update their judgment when provided with AI advice; iii) finally, the last part where physicians had a second opportunity to update their judgment when provided with XAI. In total, the physicians made 330 judgments.

Eleven physicians (nine males and two females) with a range of medical specialties and from three Swedish hospitals participated in the study. Five were Ph.Ds. The study did not mandate prior AI experience. For further details, refer to Paper II (Rosenbacke, 2024a).

German field study data set

The research by Chanda et al. (2024) on malignant melanoma presents notable parallels to the investigation I conducted in my Paper II.

For this data set, AI accuracy was 80% and followed in three steps: i) an initial part where clinicians diagnose malignant melanoma; ii) a second part where clinicians were provided with AI advice; iii) finally, the last part where clinicians were provided with XAI (Chanda *et al.*, 2024). For further details, refer to Paper V and related studies (Rosenbacke, Melhus and Stuckler, 2024; Weis *et al.*, 2024; Wies, Hauser and Brinker, 2024), and the original study XAI (Chanda *et al.*, 2024).

Leveraging their extensive dataset, I was able to re-apply my framework and method from Paper II and test the reproducibility of the outcomes in my study on a significantly larger scale. Their work provided important data for quantifying the instances of False Confirmations as well as True and False Conflicts that occur when clinicians engage with AI in decision-making processes. The dataset included decision-making instances from 109 clinicians, encompassing a total of 4,512 decisions. (Chanda *et al.*, 2024).

Quantitative data methods

The focus was initially on quantifying physicians' switching decisions—instances where doctors either altered or maintained their initial clinical judgment upon exposure to AI or XAI. A heat map was generated to visualize patterns, which were subsequently labelled and tallied. The dataset underwent multifaceted analysis, examining scenarios where AI was accurate or erroneous, as well as breaking down the data per patient, per physician, and per type of decision. For further details, refer to Paper II (Rosenbacke, 2024a).

I tested the statistical significance ($p<0.05$) of physician switches with AI and XAI in different ways. Following prior papers as a validation exercise (Chanda *et al.*, 2024), we tested whether AI led to improvements in decision-making accuracy, using t-tests to compare the accuracy of decisions with and without AI/XAI. I also applied a chi-squared test to observe whether departures from original decisions were beyond what could be expected through random decision-making. I also performed multivariate regression to quantify the added benefit of AI and XAI on overall diagnostic accuracy, adjusting for potential confounding factors, such as individual patient effects. For further details, refer to Paper II.

Qualitative data methods

In Paper II and III (Rosenbacke, 2024a, 2024b), investigating the decision-making processes of clinicians, the study adopted qualitative data collection methods that offered insight into the cognitive underpinnings of medical diagnosis. I used semi-structured interviews complemented by "think-aloud" protocols, (Van Someren, Barnard and Sandberg, 1994), a technique that enables participants to verbalize their thoughts in real-time as they engage in diagnostic tasks. This method is particularly valuable for capturing the sequence and structure when studying cognitive challenges.

Compared to structured interviews or surveys, the semi-structured nature of these interviews afforded a degree of flexibility, allowing for deeper exploration of topics as they arose naturally during the dialogue. The "think-aloud" component provided a live commentary of the physicians' reasoning, giving me a window the mental operations that guided the physicians' judgment.

Other qualitative methods, such as focus groups or narrative analysis, may reveal broader patterns and shared experiences but can lack the specificity and detail that "think-aloud" protocols can elicit when examining individual cognitive processes.

The "think-aloud" method is not without limitations. It relies on participants' ability to articulate their thought processes, which can be challenging under the cognitive load of complex tasks. Moreover, physicians might have been influenced by the observer's paradox (Gordon, 2013)—

the presence of me, as a researcher with open-ended questions, may have altered the behavior of the physicians.

A thematic analysis was conducted on the qualitative dataset following Braun and Clarke's methodology (Braun and Clarke, 2006, 2012). It involves 6 stages: 1) Familiarization with the data, 2) Generating codes, 3) Searching for themes, 4) Reviewing themes, 5) Defining and naming themes, and 6) Writing (Paper II and III).

Software

The software used in the different studies are shown in Table 1.

Software	**Paper**
Mendeley reference manager	II-V
Microsoft Excel	I-V
Microsoft Word	I-V
Random forest: Python package (*API Reference — scikit-learn 1.1.3 documentation*)	II, III
R software for statistical computing and graphics	V
SPSS	II, V
XAI: SHapley Additive exPlanations (SHAP)	II, III
Zotero reference manager	I

Table 1: Software.

To conclude, this research employed a mix of systematic review, quantitative, and qualitative methods to investigate human-AI collaboration in healthcare settings. With quantitative data methods offering a macro perspective on decision accuracy and qualitative approaches capturing the nuances of human cognition, this mixed-methods approach allowed for a multidimensional exploration.

The following chapter will build upon these methodologies by diving into the problematization of current literature and a summary of the key findings and contributions of each of the five included papers in the thesis, with a focus on decision-making errors, trust in AI, and the cognitive mechanisms shaping human-AI interaction.

Cognitive challenges in Human-AI collaboration

In this chapter, I explore the cognitive challenges in human-AI collaboration, focusing on how clinicians engage with AI systems and the errors that arise in this interaction, as presented in five interconnected papers.

In the first paper, I conducted a systematic literature review (published in the *Journal of Internet Medical Research AI*) to determine if explainable AI increases trust among clinicians (Rosenbacke *et al.*, 2024b). While most studies suggested that XAI enhances trust, I found that overreliance on AI due to increased trust can lead to new errors when the AI advice is incorrect. This finding called for a deeper exploration of the balance between trust and accuracy, leading to the second paper.

Most existing literature focuses on whether and to what extent physicians adjust their decision-making when AI models are correct, but there is little research on how XAI impacts diagnostic accuracy. In the second paper, I developed a framework categorizing different types of errors in AI-human collaboration. It highlighted how clinicians navigate incorrect and correct AI advice, revealing that while XAI can improve accuracy by convincing incorrect clinicians to change, it can also mislead them when the AI advice is wrong (Rosenbacke, 2024a). This paper demonstrates the importance of understanding the cognitive biases behind these errors, which are further explored in the third paper.

In the third paper (published in *Proceedings of ECIS 2024*), I analyzed the cognitive biases and heuristics—such as commitment bias and false confirmation bias—that drive decision-making in AI-human collaboration (Rosenbacke, 2024b). These biases explain why clinicians may either over-rely on or reject AI advice, leading to errors. These insights form the basis for the discussion in the fourth paper, which addresses the role of AI as a second medical opinion.

The fourth paper contributes to a debate in the *British Medical Journal of Ethics* on whether AI can act as a reliable second opinion in clinical settings (Rosenbacke *et al.*, 2024a). My findings emphasize the dangers of False Confirmation errors, where clinicians stop investigating after being incorrectly confirmed by AI advice. These errors can reduce accuracy by up to 30 percentage points, highlighting the risks of using AI and XAI without proper oversight. This prompted me to test the reproducibility of these findings.

In my fifth paper (published in *Nature Communications*), I applied my framework to a larger dataset from a previous study published in the same journal (Chanda *et al.*, 2024; Rosenbacke, Melhus and Stuckler, 2024). This replication confirmed my previous findings, showing that the same errors occur in different clinical contexts, demonstrating the robustness of my framework.

Together, these papers provide a comprehensive understanding of the cognitive challenges in human-AI collaboration and offer valuable guidance for improving AI integration in healthcare. These contributions are distinct and represent, to my knowledge, the first thorough exploration of how cognitive bias, trust dynamics, and diagnostic errors in human-AI collaboration intersect. What makes this body of work particularly powerful is that each paper adds a unique layer to the understanding of these interactions, while the combined insights form a synergistic whole. The sum of the findings in the kappa offers a more comprehensive understanding than what each single paper could provide on its own, creating a framework that is not only novel but greater than the sum of its parts.

A systematic literature review to understand trust in XAI (Paper I)

Problematization

Recent developments in explainable AI seek to address the transparency problem, particularly in sensitive domains like healthcare, where opaque AI systems often struggle to gain user trust (Nazar *et al.*, 2021; Loh *et al.*, 2022). XAI promises to bolster decision confidence by providing clearer insights into how AI reaches its conclusions, thereby fostering greater clinician engagement and improving safety and adoption rates (Antoniadi *et al.*, 2021; Evans *et al.*, 2022; Chanda *et al.*, 2024). However, critics argue that current XAI systems may oversimplify complex AI processes, risking misunderstandings and misplaced trust (Ghassemi, Oakden-Rayner and Beam, 2021). Despite these limitations, proponents like Reddy (2022) emphasize XAI's alignment with evidence-based medicine, advocating for its role in enhancing clinical accountability and patient autonomy.

Ultimately one way to resolve this debate is through data; by evaluating whether and how clinicians engage with XAI, and whether or not they outperform AI. However, partly owing to the novelty of XAI, there is a lack of research on trust and uptake of XAI, and whether it could improve the decision-making and ultimate patient outcomes over and above AI alone.

Although it is highly plausible that XAI could enhance trust in AI systems, as suggested by Reddy, recent systematic reviews indicate a lack of concrete evidence to support this assumption, most papers just assume that explanations will increase trust. One systematic review found that most studies neglect trust, instead evaluating user satisfaction as an indicator of whether or not explanations are effective (Jung *et al.*, 2023). Only two studies out of the 882 screened articles actually discussed the impact of XAI on clinicians trust (Jung *et al.*, 2023). Another systematic review focused on explanations to the end-using clinician to create a trustworthy environment. However, they only assume that transparency and explanations go hand-in-hand with clinicians trust in the algorithm (Nazar *et al.*, 2021); another review speculated that XAI could enhance decision confidence and trust for clinicians (Antoniadi *et al.*, 2021), while another forcefully argued that XAI could instill trust in the users, and assist clinicians in decision-making (Giuste *et al.*, 2023). The systematic reviews suggest that providing explanations for AI algorithms promote transparency, which in turn can enhance the perceived trustworthiness of the algorithm. Knowing the inner workings of the algorithms is believed to foster a sense of reliability in the technology. However, this perceived trustworthiness is not synonymous with the actual behavior of trust by clinicians. The choice of clinicians to trust and follow AI advice is distinct from the notion that an algorithm is trustworthy simply because it has higher accuracy than human clinicians.

Curiously, these researchers, despite conducting extensive reviews, appear to presuppose that XAI will boost clinicians' trust and their likelihood of using AI recommendations. Yet empirical support for this is scarce, which leads to my first research question:

RQ1: How does XAI impact clinicians' trust based on empirical evidence?

To address this question, I conducted a systematic literature review, Paper I (Rosenbacke *et al.*, 2024b), published in the *Journal of Internet Medical Research AI*. I evaluated empirical evidence of the impact of XAI on trust and uptake in physician-AI collaboration. As discussed in the methods section, systematic reviews are frequently employed in health sciences and, increasingly, in Information Systems research. This method differs from a conventional narrative literature review in that it involves a structured, transparent, and replicable process for gathering, appraising, and summarizing the current research and professional contributions (Okoli and Schabram, 2012; Boell and Cecez-Kecmanovic, 2015).

Key findings and contribution

Many argue that XAI can build trust by making AI decisions more transparent. My research provides, to the best of my knowledge, the first systematic empirical evidence supporting this claim; however, it also reveals an important caveat: the increased trust in XAI can sometimes lead to overreliance, especially when AI-generated advice is incorrect.

Briefly, I found that the vast majority of studies report greater clinician trust in XAI over AI without explanations. Noteworthy, two studies found explanations could decrease trust, especially when physicians could not understand the explanation provided. In general, the included studies paid scant attention to the phenomenon of "too much trust", or overreliance, which can manifest as blind trust. A few studies also highlighted the difference between affect-based trust /System 1) and cognitive-based trust (System 2) (Naiseh, Al-Thani, *et al.*, 2021; Naiseh *et al.*, 2023). This points to the cognitive challenges associated with how explanations are presented, as well as searching for means to better optimize trust.

It also emerged from the review that studies had yet to investigate how trust related to AI performance. Specifically, doctors could trust XAI, but in so doing, make more errors if AI was actually incorrect. In a study where pathologists used AI to detect cancer, the authors concluded that when the AI advice was incorrect, overall accuracy decreased significantly, irrespective of clinicians' experience or how difficult the case was (Kiani *et al.*, 2020). Hence, more detailed studies are needed to identify the dynamics of trust and decision-making in relation to the actual accuracy achieved in AI-clinician collaborations, especially when the AI advice is incorrect, which leads us to my second paper.

Errors in human-AI collaboration and associated cognitive challenges (Paper II)

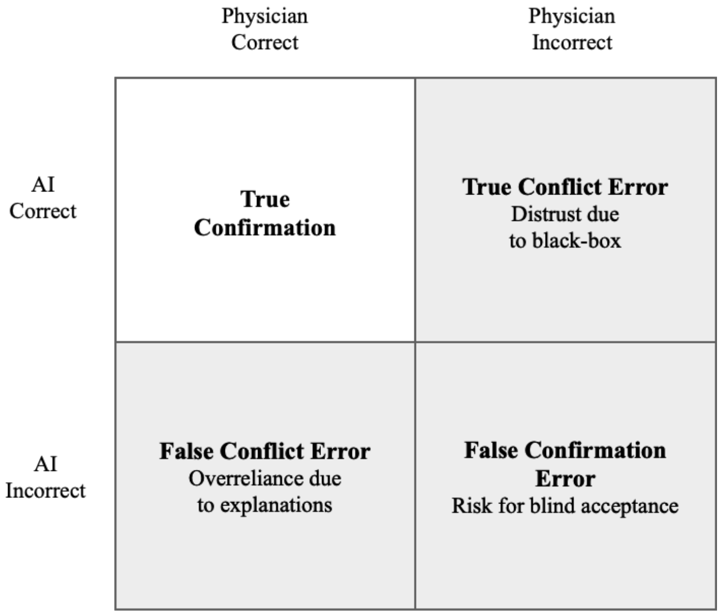
Problematization

Prior studies have tended to focus on whether and to what extent physicians adjust their decision-making when AI models are correct or perform significantly better than clinicians. To the best of my knowledge, the only study to examine the cognitive difficulties faced by clinicians when AI systems provide erroneous advice is by Jussupow and colleagues (2021).

This study focuses solely on AI, but I extend the research by incorporating XAI to provide a more comprehensive understanding. In their seminal work, they note that "*most prior work has assumed that provided [AI] system advice is correct and beneficial. In doing so, it has largely neglected the cognitive challenges entailed in incorrect system advice*"(Jussupow *et al.*, 2021).

It is worth noting that, even with best efforts and significant algorithmic improvements, AI will not achieve 100% accuracy in the foreseeable future. As long as AI remains imperfect, explanations will be an important instrument to identify potential errors, for example, since medical data sets are naturally imperfect (Amann *et al.*, 2020). Yet even if perfect accuracy were theoretically attained, there is no assurance that the AI system would be devoid of biases, particularly when trained with diverse and intricate datasets typical in medical contexts that differ from clinical practice (Reddy, 2022).

Since AI is imperfect, like any model, it is critical to investigate the cognitive processes that doctors implement to evaluate it when AI models provide results that are incorrect (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023). Drawing on the work of Jussupow et al., Figure 6 categorizes the possible missteps that can arise during human-AI interaction (Jussupow *et al.*, 2021). I extend their 2x2 matrix to label each quadrant to highlight the four possible errors, which I label: True Confirmation, True Conflict error, False Conflict error, and False Confirmation error. In the original, Jussupow classified only "confirmation I and II" and "disconfirmation I and II", yet, as I will argue and demonstrate, it is fundamental to differentiate these errors and their associated cognitive underpinnings.



	Physician Correct	Physician Incorrect
AI Correct	**True Confirmation**	**True Conflict Error** Distrust due to black-box
AI Incorrect	**False Conflict Error** Overreliance due to explanations	**False Confirmation Error** Risk for blind acceptance

Figure 6: Potential errors in Human-AI collaboration and decision-making, adapted from Jussupow et al., (2021).

Taking each of the four decision-making outcomes where errors can arise (either by the physician or the AI or both) in turn:

i) *True Confirmation*: The first confirmation occurs when both the clinical diagnosis and the AI advice are correct. In normal cases this is not a source of error unless the physicians suddenly change mind and override both their initial clinical diagnosis and the AI advice.

ii) *True Conflict errors*: The first conflict occurs when there is a discrepancy between a physician's incorrect clinical diagnosis and the correct diagnosis made by AI. This situation creates a dilemma for the physician: should they adhere to their own clinical assessment or override their initial judgment, trusting the "black-box" AI prediction? Adding explanations to the AI advice is an intervention that can potentially reduce these errors.

iii) *False Conflict errors*: A second conflict emerges when physicians are correct, but AI makes an incorrect judgment. Using and trusting a correct algorithm is intuitively a correct judgment; however, algorithms can err, and in this conflict scenario, high trust by physicians can potentially be counterproductive. Explanations for AI advice can potentially lead to physician overreliance by creating a false sense of trust in the AI's decision-making process.

iv) *False Confirmation errors*: The third main error arises when both the physician and AI system are incorrect. In this case, the AI falsely confirms a physician's erroneous judgment, which could create a false degree of confidence. When conflict occurs between clinical diagnosis and AI, it may seem natural for physicians to probe the underlying reasons for this divergence. However, this error has been described as a clinical "worst-case scenario" as clinicians potentially will fail to detect the problem (Jussupow *et al.*, 2021).

The prevalence and impact of different types of errors on diagnostic accuracy, particularly how they differ between AI and XAI systems, and how they vary with physicians' experience or specialization, are not yet well understood. Clarifying these variations is essential to improve diagnostic reliability and AI-human collaboration in clinical settings.

In True Conflict cases, there is evidence that clinicians are quite 'sticky' and unlikely to adjust their decision-making. This 'stubbornness' in the face of new evidence is not only seen with AI but also in reluctance to adopt new technologies and diagnostic tools into their practice (Petersson *et al.*, 2022). Qualitative studies have suggested that physicians, like humans in general, are "resistant to change" and "creatures of habit" (Gupta, Boland and Aron, 2017). Extensive research has proposed that XAI, with its explanations, is an intervention that will help clinicians to understand the AI rationale and foster greater AI uptake in clinics (Antoniadi *et al.*, 2021; Evans *et al.*, 2022; Reddy, 2022; Haque, Islam and Mikalef, 2023; Chanda *et al.*, 2024).

Turning to False Conflicts, the main threat is that AI convinces a physician with low decision confidence to switch from a correct to an incorrect diagnosis. While on the surface, this may seem unlikely, there is evidence, partly summarized above, that this indeed does happen, especially with XAI (Ribeiro, Singh and Guestrin, 2018; Lucic, Haned and de Rijke, 2020). For instance, explanations might engender undue confidence in AI recommendations; researchers have found that simply providing explanations boosts trust in the AI prediction, a "mere exposure effect" (Kliegr, Bahník and Fürnkranz, 2021), (note that this definition differs from the term "mere exposure effect" originally coined by Zajonc (Bornstein and D'Agostino, 1992). Eiband and colleagues demonstrate that placebic (false) explanations can engender a level of trust comparable to genuine explanations (Eiband *et al.*, 2019). Additional research (Fürnkranz, Kliegr and Paulheim, 2020; Chromik *et al.*, 2021; Nourani *et al.*, 2021) has revealed that explanations can lead to cognitive errors, such as backward reasoning, a cognitive process where individuals begin with a conclusion and work backward to find supporting evidence, often overlooking contrary evidence.

A much more devious cognitive challenge occurs in cases of False Confirmation, where the AI and its explanations confirm the clinician's initial incorrect clinical diagnosis. An undue trust in an explainable AI system can emerge from confirmation bias (Wang *et al.*, 2019), a psychological tendency where humans are more likely to trust an AI system that consistently produces outputs aligning with their pre-existing beliefs or initial hypotheses (Naiseh *et al.*, 2023), and a reluctance to seek disconfirmatory evidence. This over-reliance on XAI can pose significant risks, particularly when the outputs of the systems are erroneous but reaffirm the user's prior convictions (Naiseh *et al.*, 2023). These findings underscore the risks of implementing AI system explanations in critical situations without confirming their congruence with the cognitive mechanisms of users (Bertrand *et al.*, 2022). However, previous research has yet to investigate whether and to what extent False Confirmation errors occur in XAI collaboration, and how they can potentially be mitigated. The only prior study, to my knowledge, reported False Confirmation in cases of human-AI collaboration as the "*Worst-case scenario as decision makers do not detect problem,*" noting "*participants felt confirmed by incorrect [AI] advice*" (Jussupow *et al.*, 2021).

It is clear that more research is needed to identify these specific errors and how they are improved or exacerbated by differing features of AI and XAI. For example, Jussupow and colleagues, who found physicians' metacognitive challenges when aided by AI hindered accuracy gains, called for research on whether explainable AI may help overcome these challenges (Jussupow *et al.*, 2021). Evans and colleagues call for "*empirical studies of user interaction with explainability elements embedded into more true-to-life workflow would provide further valuable insights*." (Evans *et al.*, 2022). Furthermore, Naiseh and colleagues call for "*future work to explore XAI design modalities and principles to mitigate potential over-reliance risk when explanations are provided*" (Naiseh *et al.*, 2023). In response to these calls for research, I elaborate a series of study designs to investigate the cognitive challenges, trust or decision confidence implications, and potential errors introduced by XAI with the following second research question:

RQ2: What are the implications of trust or decision confidence on errors in physician-XAI collaboration?

In my second paper (Rosenbacke, 2024a), I conducted a mixed-methods field study to differentiate better trust (or decision confidence) and errors that emerge when AI is correct or incorrect and either confirms or conflicts with doctors' diagnoses. I intentionally designed an AI setup where, a significant portion of the time, the AI system was incorrect (40%). This enabled me to hold constant 'trustworthiness', by which I commonly refer to the objective aspects of AI system (here, accuracy), and 'trust', referring to subjective perception of the AI systems' reliability, credibility, and the degree to which individuals are willing to rely on AI systems.

The design involved a series of decisions, with incrementally provided AI and XAI information. Initially, physicians made a clinical diagnosis, then received AI advice, and finally, they got explanations for the AI advice, with opportunities to revise their diagnosis after each new piece of additional advice. Data were collected both quantitatively and qualitatively at every stage. Employing "think-aloud" protocols enabled a deeper exploration of the physicians' reasoning and decision-making processes.

Key findings and contribution

This paper addresses a critical, previously unexplored gap in the literature: while most studies examine how AI impacts decision-making when its advice is correct, little attention has been paid to the role of XAI in scenarios where AI advice is incorrect and the implications for diagnostic accuracy. My research systematically investigates these error types and introduces a novel framework for understanding them, offering groundbreaking insights into how XAI shapes clinician decision-making and diagnostic accuracy in these challenging contexts.

In terms of uptake, physicians exhibited "stickiness" in their diagnostic decisions in about two-thirds of all cases, consistent with AI distrust and a potential commitment bias (Dolan *et al.*, 2012). Adding explanations with XAI did persuade more physicians to use it, but nonetheless about half of the doctors remained unchanged with the aid of XAI. In cases of conflicts, drawing on the qualitative data, I could clearly identify the physicians' hard cognitive work (System 2) where they tried to understand why there was a conflicting view from the XAI. However, virtually none of the physicians altered their decisions when AI confirmed their incorrect diagnosis and increased the physicians' decision confidence (a "False Confirmation"), which accounted for two-thirds of all errors identified in my study. The qualitative analysis showed that physicians neglected the possibility of AI error in cases of confirmation reminiscent of a confirmation bias (Nickerson, 1998) or a System 1 error.

Physicians' commitment to their own clinical diagnosis is in line with previous research that has focused on True Conflict errors where physicians override correct AI advice. However, the findings highlight a critical oversight in previous research: There has been a lack of investigation into cognitive challenges cases where AI or XAI provides erroneous advice that is accepted by the physician.

Both the commitment to their clinical diagnosis and the acceptance of a False Confirmation seem to be intuitive System 1 errors. They are consistent with doctors employing a bias rather than a reasoned approach to decision-making, which leads to the next topic in my third paper: what cognitive processes are behind these errors?

Beyond AI algorithm aversion or appreciation (Paper III)

Problematization

Much research in information systems has reported that decision-makers are generally more likely to trust and incorporate advice from humans than AI algorithms. A recent systematic literature review investigated 80 empirical studies on algorithm aversion and found that, in general, "*People tend to rely less on algorithms even when algorithms provide better decisions*" (Mahmud *et al.*, 2022, p. 17). This phenomenon is typically defined as "algorithm aversion" (Dietvorst, Simmons and Massey, 2015) (when rejecting advice), or "algorithm appreciation" (Logg Jennifer, 2018) (when incorporating it).

While studies have recognized that people tend to exhibit algorithm aversion much more than appreciation, the reasons are not fully understood. The aforementioned systematic review found that these studies of algorithmic aversion have tended to be conducted in artificial laboratory settings, often with students or crowd-sourced workers (like Mechanical Turk), which may not reflect the actual performance of AI systems in real-world settings. Mahmud and colleagues call for more qualitative studies with practitioners, noting that "*scholars should undertake more qualitative research on this area [algorithm aversion], involving practitioners*" (Mahmud *et al.*, 2022, p. 15).

However, these terms—aversion and appreciation—might have been useful descriptions in early research but are too simplistic for understanding the nuanced interactions in physician-AI/XAI collaboration. I argue that this simplistic dichotomy is analytically unhelpful or, even worse, inaccurate. Aversion can be useful when the AI advice is incorrect, and appreciation can be useful when the AI advice is correct and vice versa.

There is a lack of psychological depth in the analysis of human trust in AI and XAI, and the analysis is often driven by medical or computer science in this space. There are risks of implementing AI explanations in critical situations without confirming their congruence with the cognitive mechanisms of users (Bertrand *et al.*, 2022). Researchers have started employing insights into human cognition and behavior, such as the dual-process theory (Kahneman, 2011), to understand better XAI and cognitive challenges (Miller, 2019; Wang *et al.*, 2019). Buçinca and colleagues argue that trust is not as rational as many assume, "*Informed by the dual-process theory of cognition, we posit that people rarely engage analytically with each individual AI recommendation and explanation, and instead develop general heuristics about whether and when to follow the AI suggestions.*" (Buçinca, Malaya and Gajos, 2021).

In line with previous research, I draw on the dual-process theory to better understand the cognitive challenges physicians face in AI/XAI-collaboration. This theory posits that human cognition operates through two distinct processes: intuition (fast thinking System 1) and reasoning (slow thinking System 2) (Kahneman, 2011). When acquiring new skills, such as reading, the slower-thinking System 2 engages in intensive cognitive processing to identify patterns within each letter. With prolonged practice, when System 1 is effectively trained, it becomes capable of swiftly recognizing letter patterns, enabling effortless recognition of words and even entire sentences without exerting conscious effort. System 1 functions efficiently when individuals have undergone extensive training and have become adept in a particular domain. However, these mental shortcuts, or heuristics, may not be well-suited for new contexts, highlighting the potential for misjudgments and mistakes when intuitive thinking is applied outside its accustomed domain (Kahneman, 2011). While the term "bias" or "heuristic" often suggests judgment errors, in line with previous research on XAI and heuristics (Bertrand *et al.*, 2022), I frame it as cognitive or mental shortcuts. These shortcuts can sometimes lead to mistakes, but as my findings (Rosenbacke, 2024a) highlight, they can also serve as beneficial heuristics.

It is most likely the case that algorithm aversion or appreciation is fueled by System 1 and its associated heuristics pertaining to trust. A systematic review of how cognitive biases affect XAI-assisted decision-making argues that heuristics like AI algorithm aversion and appreciation are trust-related heuristics that arise from System 1 (Bertrand *et al.*, 2022). To trust AI, or have the intention to use AI, can be based on cognition-based trust (System 2), where trust is derived from the perceived understandability, reliability, and technical competence of AI, rooted in reasoning. However, trust can also be intuitive or affect-based (System 1), involving emotional attachment and faith (Madsen and Gregor, 2000; Lewicki and Brinsfield, 2011). Independent of which one of these facets of trust that is engaged, trust can serve as a System 1 decision-making shortcut, enabling the decision-maker to select information while ignoring other information to simplify a complex decision.

Surprisingly, little effort is spent on understanding the cognitive challenges of decision augmentation with AI-based systems in healthcare, although these systems make it more difficult for decision-makers to evaluate the correctness of system advice and to decide whether to reject or accept it. Furthermore, there seems to be even less research on what happens when adding explanations to the AI-advice. As little is known about the cognitive mechanisms that underlie such evaluations, leading to my third research question:

RQ3: What heuristics drive algorithm aversion and appreciation?

In Paper III, a qualitative thematic analysis study, I examined 330 clinical decisions using "think aloud" protocols to identify heuristics employed with AI and explainable AI (Rosenbacke, 2024b). The paper has been published in the proceedings of the *European Conference on Information Systems* (ECIS).

Key findings and contribution

Current literature falls short in thoroughly examining the nuanced psychological factors that lead clinicians to either over-rely on or avoid AI and XAI in clinical decision-making. Terms like 'algorithm aversion' and 'algorithm appreciation' are overly simplistic, failing to capture the complex cognitive mechanisms and biases at play. This research challenges these limited frameworks, offering a deeper investigation into the cognitive processes that shape clinician interactions with AI and XAI. By filling this critical gap, it introduces a more sophisticated understanding of the cognitive dynamics that influence these behaviors, paving the way for new approaches to optimizing human-AI collaboration in healthcare.

I found that algorithm aversion or appreciation arises from underlying decision-making heuristics such as pro-innovation bias, ambiguity aversion, commitment bias, mere exposure effect, and false confirmation bias. The "mere exposure effect" occurred commonly with XAI, when physicians, feeling uncertain about their diagnoses, altered their decision to an incorrect AI diagnosis.

However, in nearly 50% of cases, physicians demonstrated a strong commitment to their initial clinical diagnosis, which suggests the presence of commitment bias. However, this adherence was somewhat mitigated when explanations accompanied the AI advice, indicating that explainability may partially reduce such biases by encouraging a more open-minded approach to alternative perspectives.

False Confirmation errors were observed in instances where the AI confirmed an incorrect diagnosis, which led clinicians to refrain from seeking further information in virtually all cases. This behavior reflects a reliance on System 1 thinking—driven by cognitive shortcuts and confirmation bias—rather than the more deliberate, analytical processes of System 2 reasoning. As a result, clinicians may inadvertently accept AI-confirmed errors without critically evaluating alternative diagnoses.

In the paper I also discuss how cognitive interventions could redress these heuristics in decision-making to better optimize clinical accuracy. For further elaboration of AI/XAI and trust related heuristics or biases, please refer to Appendix VI.

Having developed this rich framework for identifying errors, and understanding the potential cognitive challenges and associated biases which underpin them, I then was able to apply these insights to contemporary academic discourse to break new ground. I contribute to emerging academic debates on AI in healthcare, including Paper IV (Rosenbacke *et al.*, 2024a), published in *British Medical Journal-Journal of Medical Ethics* (BMJ Medical Ethics), and Paper V (Rosenbacke, Melhus and Stuckler, 2024), published in *Nature Communications*, and finally an additional paper (not included in this thesis) in the proceedings of the *Cancer Prevention Research Conference*, *American Cancer Society* (Weis *et al.*, 2024).

Can AI serve as a 'Second Opinion' in medicine? (Paper IV)

Problematization

AI has been argued to have the potential to 'disrupt medicine.' One way it could potentially alter practice is through the institution of a 'second opinion'. This can occur when AI algorithms analyze medical data, such as imaging or tabular information, to support or challenge diagnoses and treatment plans proposed by human clinicians. This leads to my fourth research question:

RQ 4: What are the risks of using AI or XAI as a second medical opinion?

Whether or not AI can be a real second opinion is, at the time writing, a source of considerable debate in *BMJ Medical Ethics*. Kempt and Nagel argue for a "rule of disagreement", by which AI can provide a second opinion (Kempt and Nagel, 2022). When AI diagnosis concurs with the initial physician assessment, no further action is required, but when it differs substantially, another human opinion is imperative. However, responding to Kempt and Nagel, Jongsma and Sand disagree, arguing that there is "symmetry in the burden of proof" in both agreement and disagreement (Jongsma and Sand, 2022). They emphasize the inherent fallibility of both human and AI judgements, advocating a second human opinion regardless of the concurrence or dissent of AI.

Key findings and contribution

I draw upon my physician-AI collaboration framework to argue that the crux of this debate hinges on the prevalence and impact of "False Confirmation", a scenario where AI erroneously validates an incorrect human decision (Rosenbacke *et al.*, 2024a). These errors seem exceedingly difficult to detect, reminiscent of bias akin to confirmation bias. Furthermore, this debate has yet to engage with the emergence of explainable AI, which elaborates on why the AI tool reaches its diagnosis.

To progress this debate, I draw upon my framework for conceptualizing decision-making errors in physician-AI collaborations. The False Confirmation error rate was 26% in the Swedish dataset and 9% (Rosenbacke, 2024a) in the German study (Rosenbacke, Melhus and Stuckler, 2024). Through a series of simulations, I show that, even in the most accurate AI systems, False Confirmations are likely to be pervasive in clinical practice, decreasing overall diagnostic accuracy to between 5% and 30%. For further elaboration, please also refer to Appendix VII.

Based on these insights, I develop a pragmatic approach to employing AI as a second opinion, with three main recommendations: i) emphasizing the need for physicians to make clinical decisions before consulting AI; ii) employing nudges to increase awareness of False Confirmations and reduce blind trust; and iii) critically engaging cognitively to understand the clinical rationale associated with XAI explanations to avoid the mere exposure effect (Rosenbacke *et al.*, 2024a). This approach underscores the necessity for a cautious, evidence-based methodology when integrating AI into clinical decision-making.

Reproducing the findings – the role of False Confirmation in recent published studies (Papers V)

Problematization

Research reproducibility is a cornerstone in science. It confirms the validity of research findings and exposes potential errors or biases, thereby reinforcing the credibility of scientific knowledge. It ensures that results are not just isolated occurrences but reliable indicators of broader truths (Ioannidis, 2018).

To test if my findings were reproducible, I used my novel framework (Rosenbacke, 2024a) and spotted papers published in leading journals that completely omitted False Confirmation errors (as well as False Conflict errors), leading to the fifth research question:

RQ5: What are the error rates of False Conflict and False Confirmation when reproduced in other data sets?

I revisited the most recent and largest study conducted in which dermatologists draw upon explainable AI to diagnose melanoma (Chanda *et al.*, 2024). This original study reports that XAI enhances both trust and confidence among physicians. However, the paper considers only overall accuracy, which masks key decision-making errors and overlooks the specific user groups who stand to gain the most from AI applications.

My work was featured in the 'Matters Arising' section in *Nature Communications* (Rosenbacke, Melhus and Stuckler, 2024), paper V. According to the journal, 'Matters Arising' papers are exceptionally interesting and timely scientific comments and clarifications on original research papers published in Nature.

Next, given the constrained format of Nature's 'Matters Arising' papers, I will expand on my findings in this section, delving further into the details and extending the insights beyond those presented in my original paper. Additionally, I will incorporate relevant findings from a co-authored paper in the proceedings of the *Cancer Prevention Research Conference, American Cancer Society* (Weis *et al.*, 2024).

Key findings and contribution

Accuracy for best performing versus worst performing physicians

The original data includes 4,512 diagnoses made by 109 clinicians (Chanda *et al.*, 2024). I found a rate of False Confirmation of 6.7%, rising to about one in ten for the worst performing physicians. As shown in Figure 7, for the best-performing physicians, I also found that performance worsened even more with XAI than with AI. This echoes the points of my first paper (Rosenbacke *et al.*, 2024b) that trust in AI can be 'double-edged'- here, worsening accuracy (inducing False Conflict errors) for the best performing physicians.

Consistent with my hypotheses, I found that AI, for the lowest-performing clinicians, helped stamp out True Conflict errors. For these clinicians, accuracy could have increased to 76.1% had they fully trusted the AI. It's also worth noting that AI accuracy for cases shown to the worst performers was 76.1%, slightly lower than the overall average of 80.4%, while for the best performers, it was 80.9%. Thus, the lower performance among the worst-performing clinicians seems not only to be due to lower performance, but it may also partly reflect the increased difficulty of cases they faced.

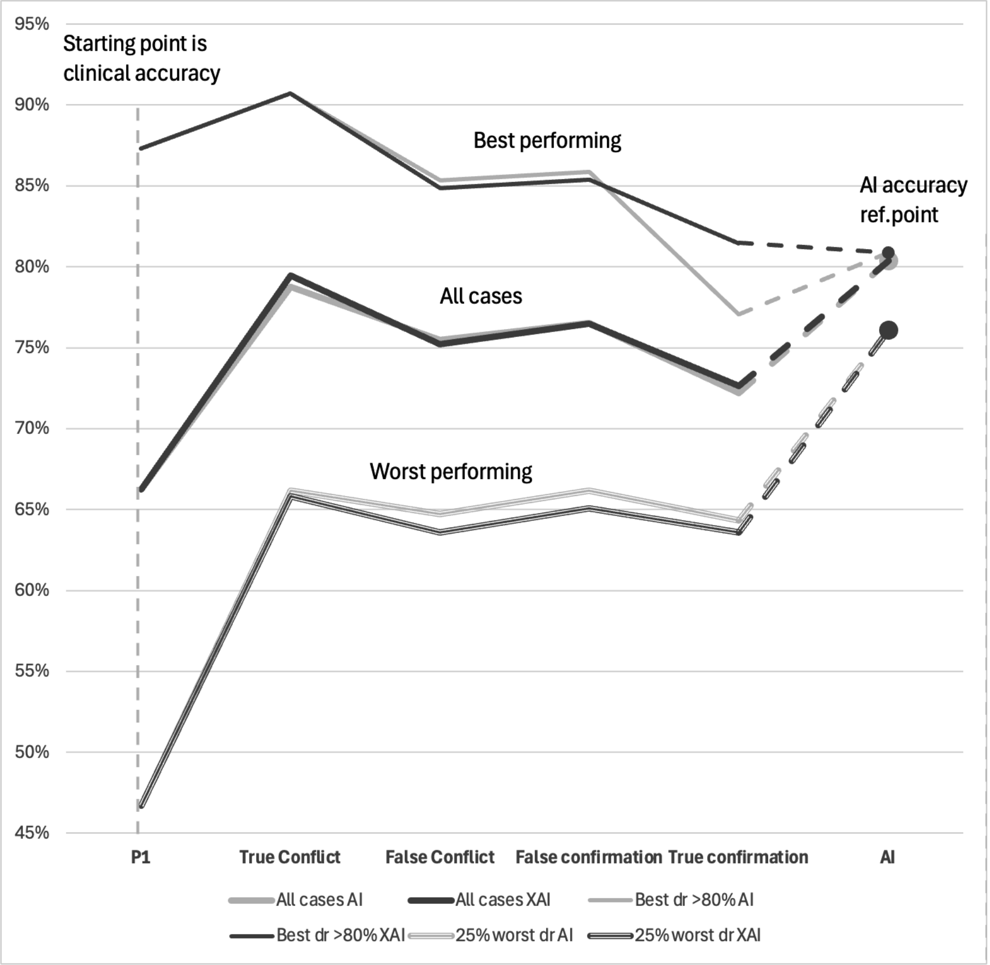


Figure 7: Clinical accuracy for best and worst performers across different error types.

Note: P1 is the initial clinical accuracy for the physicians.
True Conflict: Accuracy improves after accepting correct AI advice.
False Conflict: Accuracy decreases after accepting incorrect AI advice
False Confirmation: Slight improvement after rejecting incorrect AI and changing diagnosis.
True Confirmation: Accuracy decreases after rejecting correct AI advice and changing an initially correct diagnosis.

AI accuracy reference point is the AI accuracy for the cases presented to the respective group. Light grey lines represent diagnosis with AI advice and dark grey XAI advice.

Diagnostic accuracy in cases with high trust versus low trust in XAI

In the original study conducted by Chanda and colleagues, physicians were asked to rate their trust in the XAI advice for each diagnosis on a Likert scale from 1 to 10. The total number of diagnoses analyzed was 1,508. A key distinction to highlight is that this trust rating is based on individual cases, not on physicians' overall trust in XAI. This means that even the same physician could report high trust in some cases and low trust in others. When analyzing the impact of XAI on diagnostic accuracy by focusing on cases with low trust (trust < 4, n=164) and high trust (trust > 8, n=369), several important findings emerged. This analysis provides a more granular understanding of how trust in XAI influences clinical decision-making on a case-by-case basis.

I found that for the high-trust cases, clinical accuracy in Phase 1—before receiving any AI or XAI advice—was noticeably higher than the average, whereas for low-trust cases, clinical accuracy was clearly lower, as seen in Figure 8. Specifically, for cases where physicians reported trust in XAI to be >8, the clinical accuracy in Phase 1 was 79.3% (compared to the average of 66.3% for all cases). For low-trust cases (trust <4), the clinical accuracy dropped to 52.6%. This suggests that high-trust cases are relatively easier for both clinicians and the AI, while low-trust cases are more challenging.

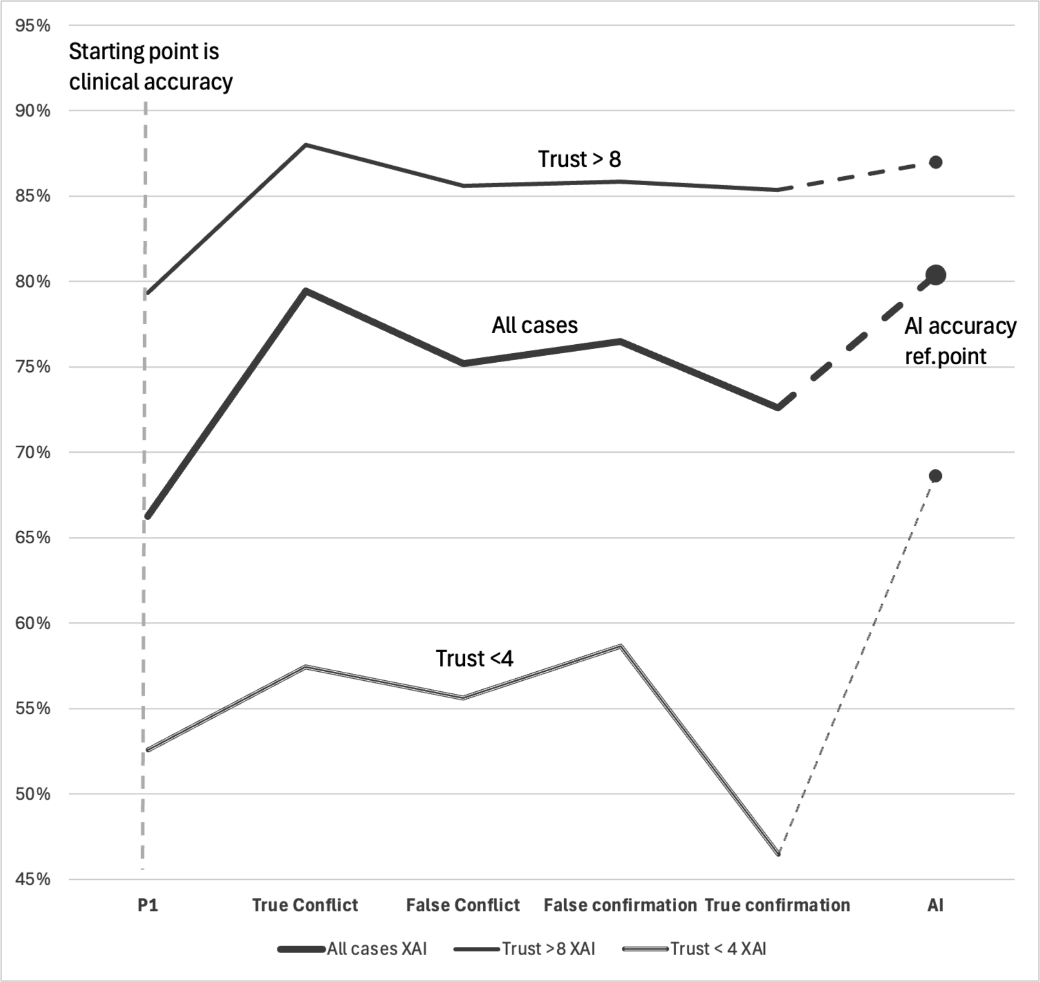


Figure 8: Clinical accuracy for high XAI trust cases (trust >8), low XAI trust cases (trust <4), and all cases measured on a 1-10 Likert scale.

The change rate has significant implications for different error types. In cases of True Conflict and False Confirmation, a high change rate is beneficial since the physician is initially incorrect. As shown in Figure 9, the change rate for True Conflict cases was 85.7% for high-trust cases, whereas, for low-trust cases, the change rate was only 15.8%, indicating missed opportunities for improvement. In these cases, high trust and frequent changes lead to better outcomes, as physicians adopt correct AI advice. Conversely, in False Confirmation cases, low trust and higher change rates are advantageous. Low-trust cases had a change rate of 50.0%, compared to 2.3% in high-trust cases, where clinicians failed to realize that both they and the AI were incorrect, demonstrating a clear confirmation bias.

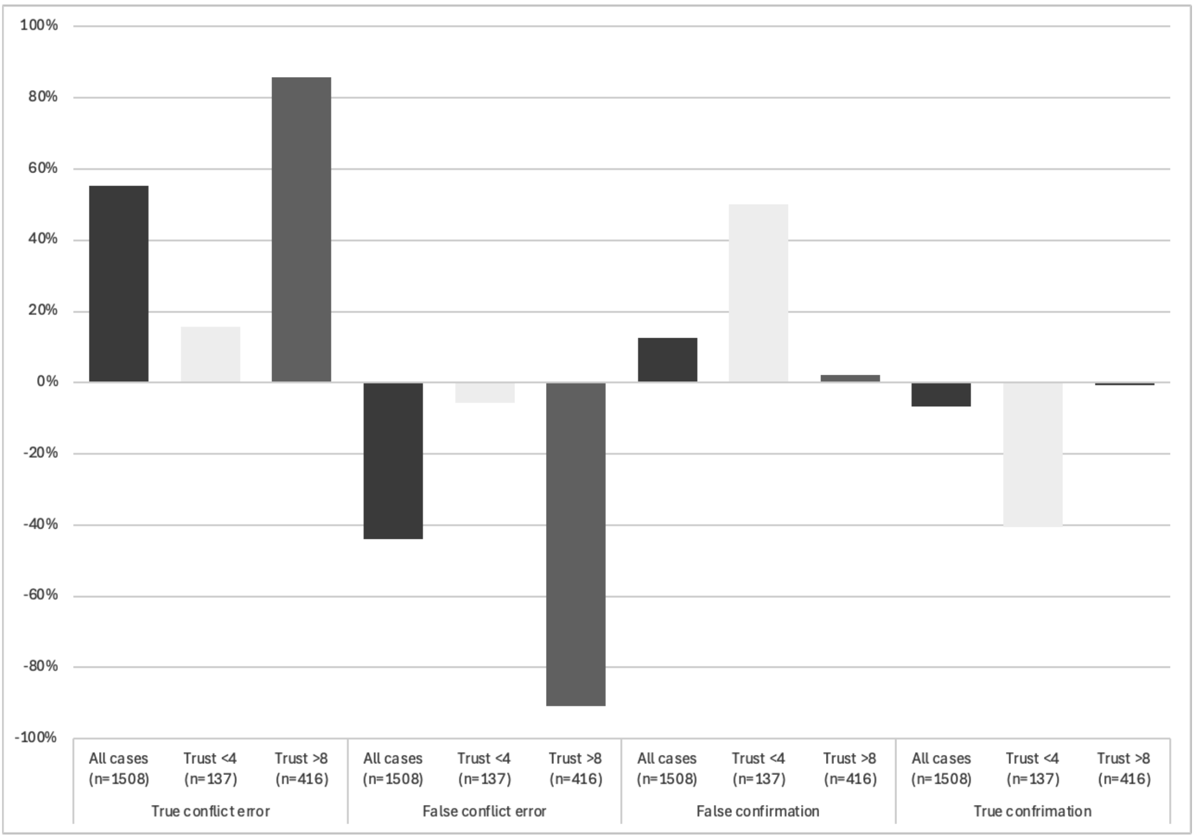


Figure 9: Change rate with XAI in Phase 3.

True Conflict: The physician should change to correct errors and improve accuracy.
False Conflict: The physician is correct and should not change.
False Confirmation: Physicians should change to recognize that both their diagnosis and the AI's advice are incorrect.
True Confirmation: When both the physician and AI are correct, distrust in AI may decrease accuracy.
Dark grey bars represent all cases; light grey bars represent cases with trust <4, and mid-grey bars represent cases with trust >8.
A positive bar indicates accuracy improvement, while a negative bar indicates accuracy loss

In cases of False Conflict and True Confirmation, where the physician is already correct, a low change rate is beneficial. In False Conflict cases, high trust in AI has a detrimental effect, with a change rate of 90.9% in high trust cases, leading to new errors. In contrast, the change rate for low trust cases is only 5.7%, avoiding errors to a larger extent. In True Confirmation cases, high trust cases show a negligible change rate of 0.6%, while low trust cases exhibit a much larger change rate of 40.5%, indicating more frequent and unnecessary changes from correct to incorrect diagnoses. This highlights that in low-trust cases, the physicians are more likely to second-guess themselves, even when their initial diagnosis is correct. However, it is important to note that the low trust cases are more difficult.

These findings underscore the importance of carefully managing trust in AI within healthcare settings. High trust can lead to beneficial changes when the physician is wrong, but in cases where the physician is correct, too much trust can result in unnecessary changes, introducing errors. These insights can guide healthcare institutions in refining AI adoption strategies to optimize outcomes based on the balance between trust and decision accuracy.

Robustness of findings

These findings have substantial implications for guiding how healthcare institutions implement AI in practice.

The observed trends in False Conflict and False Confirmation errors in my study have shown consistency when applied to new conditions. Notably, Chanda et al.'s research differs from mine: i) in its application of XAI to image-based data as opposed to the tabular data in my study, ii) the use of a different XAI technique, with them employing an unspecified method and my use of SHAP, and iii) the focus on a distinct medical specialty, with their work involving dermatology professionals and mine targeting ear infection diagnoses. Despite these variations, the robustness of these findings is evident across larger datasets, diverse XAI technologies, and multiple medical disciplines.

Together with the original authors (Chanda *et al.*, 2024), we replicate my framework (Rosenbacke, 2024a) and apply it to their much larger data set. We also co-authored a conference paper for the *Cancer Prevention Research Conference in Boston, American Cancer Society* (Weis *et al.*, 2024).

Next, the general discussion will encompass key themes not previously addressed in the individual papers, followed by a concluding section for the thesis.

General discussion

Artificial Intelligence has the potential to revolutionize healthcare, with broad-ranging applications that span from early diagnosis and disease detection to personalized treatments and even complex surgical procedures. However, despite significant advancements and the integration of AI into sectors such as finance and retail, healthcare has been notably slow to embrace this technology. This hesitation is attributed to several factors, with the most significant being clinicians' lack of trust in AI systems. Additionally, the nature of evidence-based medicine, which encourages clinicians to engage in in-depth diagnostic discussions and critically assess their decisions, conflicts with the use of opaque AI "black-box" systems. As a result, there has been a growing movement toward the development of explainable AI, which aims to shift these black-box models into more transparent systems, referred to as "glass-boxes," with the goal of enhancing trust and improving adoption among healthcare professionals.

However, as explored in my first paper, there is a significant assumption in the current literature that XAI automatically leads to increased clinician trust and adoption of AI recommendations (Rosenbacke et al., 2024b). My systematic literature review addresses this oversight, revealing empirical evidence that, while XAI generally enhances trust, it may also lead to overreliance, especially when AI advice is incorrect. This highlights a new gap in the current literature; there is a need for more detailed research into the dynamics of trust and its impact on diagnostic accuracy in clinician-AI/XAI collaborations.

Building on this, the second paper addresses an additional gap in the literature: the focus on AI's influence when its recommendations are correct, neglecting the impact when AI or XAI is wrong (Rosenbacke, 2024a). I developed a novel framework to classify different errors that can arise in human-AI collaboration, particularly focusing on scenarios where physicians are misled by incorrect AI advice. Though XAI can improve diagnostic accuracy in some cases, it can also lead to significant errors, particularly when incorrect AI explanations convince clinicians to change a correct diagnosis or falsely confirm an error. This leads to a new gap, what are the reasons for these errors?

The third paper dives deeper into the cognitive biases that drive these errors. Previous research has focused narrowly on concepts like "algorithm aversion" or "algorithm appreciation," but my work looks at underlying heuristics such as commitment bias, confirmation bias, and others (Rosenbacke, 2024b). These psychological mechanisms help explain why clinicians might either over-rely on or avoid AI advice, highlighting the complexities of trust in AI.

Fourthly, my findings fed into ongoing debates in the field, such as whether AI could serve as a second medical opinion (Rosenbacke *et al.*, 2024a). I explore how False Confirmation errors—where both AI and clinicians are wrong but confirm each other's mistakes—can severely impact clinical accuracy.

Finally, I was able to replicate these findings by using my novel framework across different datasets, demonstrating the robustness of my research and providing critical insights for optimizing AI integration in healthcare (Rosenbacke, Melhus and Stuckler, 2024; Weis *et al.*, 2024; Wies, Hauser and Brinker, 2024).

This discussion integrates findings from all five papers, offering a holistic framework that not only identifies these cognitive challenges but also proposes theory-building and practical directions for improving clinician-AI collaboration.

Contribution to theory and practice

This thesis provides empirical evidence of clinicians' behavior and cognitive challenges in collaboration with AI and XAI. To the best of my knowledge, the thesis, for the first time:

i) Empirically demonstrates that XAI can improve trust in AI compared to "black-box" models. However, it also reveals a potential downside: convincing XAI can lead to overreliance on algorithmic advice, which may cause new errors when the AI is incorrect.

ii) The research quantifies the types of errors that occur in both human-AI as well as human-XAI collaboration and demonstrates how these errors impact diagnostic accuracy.

iii) Identifies the cognitive challenges and how they change with the introduction of XAI versus “black-box” AI.

iv) Describe the underlying reasons for the cognitive challenges and explain why biases and heuristics are likely responsible for the errors.

This discussion highlights the contributions of the thesis, both theoretical and practical, in the context of current literature on human-AI and XAI collaboration. It demonstrates how each aspect of this research pushes the field forward by addressing specific cognitive challenges and decision-making processes involved in AI-assisted clinical environments. Additionally, the discussion provides pathways for future research to continue exploring and refining these insights, ensuring ongoing advancements in both academic and real-world applications of AI in healthcare.

I will begin by discussing how XAI, compared to traditional AI, can enhance clinicians' trust and their willingness to engage with the technology. However, this increased trust may also lead to overreliance, which introduces new types of errors, indicating a complex dynamic between trust, accuracy, and decision-making.

XAI improves clinicians’ trust but can lead to overreliance

Previous literature have highlighted that transparency is essential for building trust in AI among decision-makers (Glikson and Woolley, 2020). To achieve this, it has been suggested that additional mechanisms should be incorporated to make AI models more explainable and transparent to users, thereby enhancing trust (Barredo Arrieta *et al.*, 2020).

Current literature suggests that XAI could improve trust in AI, but this assumption lacks concrete empirical evidence (Rosenbacke *et al.*, 2024b). Reviews have focused on XAI's potential to enhance trust and decision confidence, but the actual impact on clinician trust has been largely overlooked. In a recent review, only two studies out of the 882 screened articles actually discussed the impact of XAI on clinicians' trust (Jung *et al.*, 2023). Several reviews have assumed a strong connection between transparency and clinicians' trust in AI algorithms. For example, Nazar et al. (2021) emphasized the role of explanations in fostering trust, while Antoniadi et al. (2021) suggested that XAI could enhance decision confidence and trust in clinical settings. Furthermore, Giuste et al. (2023) argued that explainability not only instills trust but also assists in clinical decision-making. However, despite these claims, there remains a gap in empirical evidence to substantiate the direct link between transparency and increased trust.

My systematic review extends previous research by providing empirical evidence that, to the best of my knowledge, shows for the first time that clinicians demonstrate higher levels of trust in XAI compared to AI without explanations (Rosenbacke *et al.*, 2024b). Nonetheless, some studies have identified an issue of "over-trust" or overreliance, where clinicians may develop an excessive dependence on AI recommendations, even when the AI is incorrect (Naiseh, Al-Thani, *et al.*, 2021; Naiseh *et al.*, 2023).

It also emerged from the review that studies had yet to investigate how trust related to AI performance. Only one study found no difference in trust between experts and non-experts but reported that the performance of non-experts who drew upon XAI was superior in clinical practice (Gaube *et al.*, 2023). This indicates a gap in understanding how XAI affects diagnostic accuracy. Addressing this, the papers in this thesis delve deeper into these dynamics, investigating how XAI influences diagnostic performance across various clinical scenarios (Rosenbacke, 2024a; Rosenbacke, Melhus and Stuckler, 2024).

The relatively low number of studies included in my systematic review (n=10) poses a limitation in terms of generalizability to other populations and settings. Additionally, there was inconsistent or weak reporting on the trust measurement instruments used across these studies, as well as variability in the number of respondents, which affects the robustness of conclusions drawn. Another critical gap identified was that few studies provided information on the accuracy of the underlying XAI algorithms, which could also influence the level of trust clinicians have in XAI. Future research should aim to improve reporting standards, particularly regarding trust metrics and algorithmic accuracy.

In summary, my systematic review uncovered an unexplored area in current literature: while clinicians may trust XAI, there is little research exploring how this trust relates to the actual accuracy of AI recommendations. This reveals the need for further detailed studies on how trust in XAI affects diagnostic outcomes, which is the focus of my second paper.

Quantifying XAI diagnostic errors and their impact on clinical accuracy

Only a few prior studies have discussed errors when AI is correct or incorrect, such as by Jussupow *et al.* (2021) and Naiseh *et al.* (2023). Jussupow and colleagues specifically explored how AI influences different diagnostic errors. Importantly, they called for more research to determine if XAI impacts these errors differently compared to regular AI.

Furthermore, previous studies have largely neglected the distinct effects of AI versus XAI on diagnostic accuracy, leaving a notable gap in the literature. This thesis addresses this oversight by exploring how the transparency offered by XAI impacts clinical decision-making accuracy in comparison to traditional 'black-box' AI models."

In both the Swedish study (Rosenbacke, 2024a) and the German (Rosenbacke, Melhus and Stuckler, 2024; Weis *et al.*, 2024), diagnostic accuracy improved with both AI and XAI advice compared to pure clinical diagnosis. In the Swedish dataset, clinical accuracy improved from 50% to 53% with AI advice and 57% with XAI advice. In the German dataset, accuracy improved from 66.2% to 72.3% with AI and 73.2% with XAI.

In a previous study, Gaube et al. (2023) found that non-experts who utilized XAI performed better in clinical practice. However, my findings from the German dataset take this further and offer a more nuanced view. The worst performers initially improved their clinical diagnostic accuracy from 46.7% to 63.6% with XAI. On the other hand, for the best performers, XAI reduced their clinical accuracy from 87.3% to 81.5%. These findings highlight the crucial observation that AI and XAI benefit different groups differently. The worst performers should rely more on XAI, as full adherence to XAI advice would have brought their accuracy up to 80.4% (the same accuracy as the XAI). Conversely, the best performers who outperform XAI should avoid relying on it, as it reduces their diagnostic accuracy.

This distinction between worst and best performers introduces a new dimension to the literature on who should and should not use AI or XAI. While the findings are new and insightful, the practical challenge of distinguishing between worst and best performers in real-time clinical settings remains. Future research is needed to explore how these insights could be applied in practice, including whether worst and best performers can be identified ahead of time and how AI systems could be tailored to different user groups.

Importantly, I go beyond previous research that merely focuses on overall accuracy improvements that mask underlying False Conflict and False Confirmation errors. In the Swedish study, the net accuracy gains in terms of reducing True Conflict errors was 16%, but this was offset by a net accuracy loss of -9% due to False Conflict errors. When I reanalyzed the German data, I found that the original authors in the journal *Nature Communications* (Chanda *et al.*, 2024) omitted that the net accuracy gain (by reducing True Conflict errors) was 13%, but this was offset by a -4% loss due to new False Conflict errors. The Swedish study had a higher rate of False Conflict errors because the AI accuracy was set to 60%, compared to 80% for the German AI. As elaborated earlier, most prior studies have focused on situations where AI advice is correct, and the incorrect physicians override the AI advice. XAI can potentially help physicians understand and accept the advice, but prior recent studies have highlighted that explanations can induce overreliance and thus introduce new False Conflict errors (Naiseh *et al.*, 2023; Rosenbacke *et al.*, 2024b). However, previous studies have not developed a systematic framework for identifying, measuring, and quantifying these error rates—an oversight that my novel framework addresses for the first time.

The most devious errors are the False Confirmation errors, where incorrect physicians are confirmed by incorrect AI/XAI advice. In the Swedish study, False Confirmation errors contributed to a net accuracy loss of -26%, and in the German study, a net accuracy loss of -9%. Again, the higher error rate in the Swedish study is related to the lower AI accuracy of 60% compared to the German 80%. With a higher rate of errors by the AI, there will be more False Confirmation cases. Only a few prior studies have identified this error, but none have investigated its frequency or its reasons (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023). Again, the original authors in the German study (Chanda *et al.*, 2024) failed to identify these errors that had a significant impact on diagnostic accuracy.

The most important limitations of these findings are rooted in the fact that both the Swedish and German studies focused on specific AI accuracy rates (60% and 80%, respectively), which influenced the occurrence of False Conflict and False Confirmation errors. This might not represent scenarios where AI accuracy is higher or lower. The findings, therefore, may not generalize to all AI systems with varying accuracy levels. Additionally, the sample sizes and specific medical contexts (ear infections and melanoma diagnoses) could constrain the broader applicability of these results to other clinical settings.

Future research should investigate AI accuracy across diverse medical datasets to evaluate the generalizability of these error rates and explore additional clinical contexts to assess how widespread these cognitive errors are. Additionally, the reasons behind False Confirmation errors need further examination, and future studies could explore whether specific XAI designs could reduce these error types by enhancing critical thinking or skepticism in clinical decision-making, rather than overreliance on XAI recommendations. Further studies should also aim to identify more robust strategies for detecting and mitigating these errors in practice, such as tailoring XAI systems to different levels of clinician expertise. In the next section, I will theorize the underlying cognitive causes of these errors and the associated challenges clinicians face in AI and XAI collaboration.

Theorizing underlying reasons for errors and cognitive challenges in human-AI/XAI collaboration

My empirical studies have demonstrated various errors and cognitive challenges in physician-AI/XAI collaboration, highlighting the interplay between reasoning (System 2) and intuitive System 1) decisions. I am moving beyond Jussupow and colleagues (2021) by i) investigating if other theories than metacognition can explain these phenomena and, ii) examining the differences in collaboration with AI versus XAI and, iii) quantifying the different errors and their impact on diagnostic accuracy

Psychologists differ on whether System 1 heuristic or biased errors can be corrected. One view suggests they cannot be rectified, only recognized, as Kahneman states, "*It's false to hope that if you become more aware of your errors you will make better decisions*" (Matias, 2017). Conversely, Klein's approach integrates heuristic and systematic processes, focusing on managing and regulating reasoning (System 2) alongside intuition (System 1) (Klein, 2015).

Similarly, Ackerman and Thompson argue that metacognition allows for balancing intuitive, heuristic, and deliberate reasoning (Ackerman and Thompson, 2017). In addition, Bandura's social cognitive theory emphasizes self-efficacy in decision-making. Enhancing one's belief in their ability to make effective decisions can improve both intuitive and deliberate reasoning, potentially reducing errors (Bandura, 1986). This perspective adds self-confidence and social learning to the discussion, highlighting the importance of belief in one's capabilities alongside cognitive strategies.

Jussupow and colleagues (2021) investigated the cognitive challenges physicians face when collaborating with "black-box" AI systems. They argue that physicians employ metacognition (Ackerman and Thompson, 2017) by balancing intuitive, heuristic, and deliberate reasoning activities. These metacognitive processes are crucial for physicians to fully benefit from AI assistance. However, Jussupow *et al.* argue that diagnostic errors often arise from deficiencies in utilizing metacognition, leading physicians to make decisions based on beliefs (System 1) rather than actual data (System 2) or to conduct overly superficial information searches. Furthermore, Jussupow and collegues argue that without explanations, physicians have limited ability to recognize when AI advice is incorrect. They call for research into how physicians collaborate with XAI and its explanations. With XAI, physicians might have the opportunity to understand the basis of the AI's recommendations, potentially improving their diagnostic decisions. I respond to this call by, for the first time, investigating and comparing how physicians interact with both AI and XAI, pushing the literature forward in understanding these dynamics.

Conflict errors

My empirical studies show, in line with prior studies (Naiseh, Al-Mansoori, *et al.*, 2021; Naiseh, Cemiloglu, *et al.*, 2021; Naiseh *et al.*, 2023), that XAI can reduce True Conflict errors, though this improvement is largely offset by an increase in False Conflict errors. Interventions like XAI involve balancing the reduction of True Conflict errors while minimizing the introduction of new False Conflict errors. Naiseh et al. (2023) argue that when introducing XAI, calibrating trust is crucial.

My empirical data show that XAI only slightly reduces True Conflict errors more effectively than standard AI, resulting in a net positive effect on accuracy. However, almost 50% of clinicians in the studies (Rosenbacke, 2024a; Rosenbacke, Melhus and Stuckler, 2024) did not change their initial diagnosis despite being guided by explanations, indicating a stubbornness reminiscent of a commitment bias. This commitment to the initial clinical diagnosis was illustrated in my qualitative data where one of the physicians stated, ”*I don't believe I'm better [then the AI], but I can't see a reason why I should change*.” (J, Step 2) (Rosenbacke, 2024b).

Using Kahneman's view that biases, such as commitment bias, are hard to remedy is overly simplistic. I have shown that XAI helps reduce overall conflict errors. Instead, drawing on Klein’s and Ackerman’s work, it seems that explanations help physicians combine both intuitive (System 1) and reasoned (System 2) decision-making. As noted by Jussupow and colleagues, physicians in their study failed to fully use metacognitive processes to reveal all cases where AI was correct or not. This pattern is similar in my studies with AI, but I take this further and have shown that XAI more effectively helps physicians use metacognitive traits and reduce the total number of conflict errors. The physicians' commitment to their initial clinical diagnosis could also be explained by Bandura's self-efficacy theory. Trust in an algorithm, or decision confidence, also relates to self-efficacy. Decision confidence or belief in one's own view seems not only to differ between physicians (inter-personal) but also for the same physicians across patients (intra-personal). For example, for two patients in the Swedish study, the AI advice was correct, and cases of True Conflict errors. For patient 2, accuracy increased from 27% to 100% with XAI, as all doctors took up its suggestions. However, for patient 9, the accuracy similarly starts at 18%, and only one doctor updated the diagnosis with XAI (Rosenbacke, 2024b). With strong self-efficacy, trust in AI and XAI seems to be very low. In summary, while Kahneman is correct to some extent that commitment bias is difficult to eliminate, Klein and Ackerman's view that decision-makers can reduce intuitive errors (System 1) through deliberate reasoning (System 2) is also valid to some extent, and Bandura's self-efficacy also adds an additional layer of understanding this phenomenon.

Furthermore, it is important to note that physicians' commitment bias to their own clinical diagnosis is not always a source of error. In cases where AI is incorrect, it is beneficial to override the advice; here, self-efficacy reduces errors. This is particularly important for the best physicians. In the German data, the best performers' clinical diagnosis exceeded AI accuracy, but they performed worse in AI and XAI collaboration. The top performers would have benefited from a higher level of commitment, decision confidence, or self-efficacy in their initial diagnosis. In other words, self-efficacy proves advantageous. Next, I turn to confirmation errors.

Confirmation errors

Previous studies have only emphasized the errors that arise when AI or XAI incorrectly confirm the clinician's mistaken diagnosis (Naiseh, Al-Mansoori, *et al.*, 2021; Naiseh, Cemiloglu, *et al.*, 2021; Naiseh *et al.*, 2023). My research goes beyond this by not only quantifying these errors (Rosenbacke, 2024a; Rosenbacke, Melhus and Stuckler, 2024) but also qualitatively investigating the underlying cognitive and decision-making processes that lead to such mistakes (Rosenbacke, 2024a, 2024b). For the first time, this approach provides a deeper understanding of the reasons behind these errors, offering valuable insights into how overreliance on AI/XAI recommendations can develop and its impact on clinical accuracy.

In False Confirmation cases, virtually all physicians did not consider that they could be wrong. I found a strong Pearson correlation (r=0.91, p-value < 0.01) between the initial clinical diagnosis and the final diagnosis with XAI advice in scenarios where both the XAI and the physicians' assessments were incorrect (Rosenbacke, 2024a). This correlation underscores the tendency of physicians to adhere to their original diagnoses when falsely confirmed, indicating a substantial confirmation bias. To the best of my knowledge, these studies are the first to quantify False Confirmation errors, and it seems that XAI does not help physicians detect these errors.

Jussupow and colleagues argued that XAI might help physicians detect this error: "*physicians may use provided explanations to learn from CAID[Computer Aided Intelligent Diagnosis] systems and gain new insights derived from the systems' pattern analysis*" (Jussupow et al., 2021), but in my studies, XAI had no impact. My qualitative data showed that this error is hard to detect. As one physician put it, "*I don't change anything where I was right from the beginning. That seems foolish*" (J, Step 3) (Rosenbacke, 2024b). Unfortunately, being confirmed is not the same as being correct.

Theorizing why this erroneous confirmation happens, neither Klein's nor Ackerman's view that decision-makers can reduce intuitive errors (System 1) through deliberate reasoning (System 2) seems applicable. This confirmation bias aligns more closely with Kahneman's view that System 1 errors are hard to rectify. Bandura's self-efficacy theory is also not applicable. While this perspective adds self-confidence and social learning to the discussion, highlighting the importance of belief in one's capabilities alongside cognitive strategies, physicians did not show any positive signs of this. However, drawing on Bandura's learning perspective from a negative point of view, when False Confirmation errors are undetected, there is a high risk that physicians learn from the incorrect AI and its incorrect explanations. This potentially creates a vicious circle where False Confirmation further increases overreliance and hence increases False Conflict errors. False Confirmation errors are likely to be between 5% to 30% of the total diagnoses (Rosenbacke *et al.*, 2024a) and hence likely to have a significant impact on what Bandura calls self-confidence and social learning but with a negative outcome.

Current research commonly addresses variations in trust toward AI, categorizing responses as either "algorithm aversion" or "algorithm appreciation." Algorithm aversion, a term introduced by Dietvorst, Simmons, and Massey (2015), refers to the tendency to reject AI advice, often in favor of human judgment, even when the algorithm has been demonstrated to provide better outcomes. In contrast, "algorithm appreciation," coined by Logg Jennifer (2018), describes the opposite effect, where users place greater trust in algorithmic advice than in human recommendations.

I argue that the current literature's simplistic dichotomy of algorithmic aversion or appreciation is analytically unhelpful (Rosenbacke, 2024b). Aversion can be useful when the AI advice is incorrect, as it prompts physicians to rely on their clinical judgment rather than blindly following erroneous AI recommendations. Conversely, appreciation can be useful when the AI advice is correct, as it allows physicians to benefit from accurate AI insights. However, these terms—aversion, and appreciation—might have been useful descriptions in early research but are too simplistic for understanding the nuanced interactions in physician-AI/XAI collaboration (Rosenbacke, 2024b).

Despite the theoretical contributions of this thesis, it is possible that my theory of diagnostic errors in physician-AI/XAI collaboration will also be viewed as a simplification over time. As physicians and AI systems evolve and learn from their interactions, the patterns of errors and the dynamics of collaboration are likely to change. This thesis represents just an entry point in investigating these phenomena. As the famous quote by statistician George E.P. Box goes, "*All* *models are wrong, but some are useful*" (Box, 1976). This underscores the idea that while current models and theories may not capture every detail perfectly, however, they provide a valuable framework for understanding and improving physician-AI interactions.

In this section, I explored the interplay between reasoning and intuition (System 1 vs. System 2 thinking) in clinical decision-making with AI and XAI, building on and extending prior theories like metacognition (Jussupow *et al.*, 2021). One of the key limitations of this work is the generalizability of findings. The errors I have identified, particularly True Conflict, False Conflict and False Confirmation errors, are specific to the datasets used in this thesis. The medical contexts and AI models studied here might differ in other settings or specialties, limiting broader applicability. Additionally, the focus on cognitive biases such as commitment bias and confirmation bias provides only a partial explanation for decision-making processes; other factors like institutional culture, team dynamics, or technology acceptance may play important roles but have not been fully considered here.

Future research could build upon these findings by applying my novel framework across various clinical contexts, AI models, and specialties to assess the generalizability of these cognitive challenges. Also, exploring how different types of XAI designs or interventions can mitigate False Confirmation and Conflict errors will be vital. Lastly, interdisciplinary studies involving computer science, psychology, sociology, and medicine could better address how social factors, team dynamics, and system-wide trust issues influence human-AI collaboration in healthcare. This will lead to a more robust understanding of errors and how to manage them in real-world settings.

This discussion on cognitive biases and decision-making in AI-human collaboration transitions to a broader but critical concept: the distinction between "trustworthy AI" and "trust in AI." These terms are often used interchangeably, but they represent distinct aspects of AI-human interaction that must be clearly delineated to ensure the successful integration of AI systems into healthcare.

Trustworthy AI versus trust in AI

In this thesis, I focus on the clinician’s trust behavior and intention to use AI or XAI algorithmic advice. If the algorithms are trustworthy, is a different topic. Trustworthiness (Carter, 2023) typically pertains to the system's reliability in terms of accuracy, sensitivity, and specificity. These are quantifiable metrics in the natural sciences domain, serving to evaluate the AI performance against objective standards.

The lack of externally validated healthcare AI algorithms is concerning. A comprehensive umbrella meta-analysis conducted by Kolasa et al. in 2023, which synthesized data from 220 systematic literature reviews encompassing more than 7,000 academic articles, reveals a critical oversight in the field of healthcare AI. A mere fraction of 1% of AI algorithms in 7,000 studies had undergone external validation (Kolasa *et al.*, 2023). This lack of trustworthiness due to the lack of external validation signifies a deficit in the rigorous evaluation of AI algorithms, testing them for reliability, and verifying their accuracy outside the confines of their original training datasets.

Conversely, human trust (Carruthers; Carter, 2023) in AI is more nuanced and multifaceted, rooted in social sciences and cognitive psychology. It involves the subjective perception of the AI systems reliability, credibility, and the degree to which individuals are willing to rely on AI systems. This perception is influenced by, for example, past experiences, social context, individual differences in risk tolerance, and the ability to understand and predict AI behavior. Although AI systems, no matter how advanced, function based on algorithms and data, they lack the ethical decision-making framework inherent to humans. This raises significant concerns about trust, particularly in scenarios requiring moral or ethical judgment. Users may feel uneasy relying on AI for decisions in complex, high-stakes environments like healthcare, where human integrity and ethical reasoning play a crucial role. The key question remains: can AI ever be truly "trusted" in the same way as humans?

While there's a correlation between the natural science aspect of AI trustworthiness and the social science aspect of human trust in AI, they don't necessarily cause each other. High accuracy in an AI system does not automatically engender trust among users, and conversely, a person's trust in AI doesn't always imply that the AI outputs will be accurate. This issue is examined in Paper V, wherein I found that for the top-performing clinicians, clinical accuracy exceeded that of the AI system (Rosenbacke, Melhus and Stuckler, 2024). However, as they trusted the AI advice and integrated several AI-induced errors, their diagnostic accuracy decreased when collaborating with AI, compared to their solo clinical judgments. The intersection of trust versus trustworthy is an area of rich exploration, with implications for AI design, regulation, and user education (Reinhardt, 2023).

The concept of trust is complicated. In my systematic literature review (Rosenbacke *et al.*, 2024b), we found that in the articles reviewed, there was considerable heterogeneity in the use of the term ‘trust’ and how it is operationalized in healthcare research. To avoid potentially missing important studies in our search, we adopted a conservative search strategy in which we did not specify trust as a keyword but rather manually searched for all papers, including a broad set of trust-related outcomes. Generally, the study designs widely varied, from qualitative investigations to experimental quantitative studies, making it difficult to draw direct comparisons. Another limitation present in several studies was the weak reporting of trust measurement instruments, as well as the number of respondents, particularly in qualitative studies. Few studies reported the accuracy (trustworthiness) of the underlying XAI algorithm, which could also alter the healthcare providers’ engagement and trust in XAI technologies.

In my studies (Rosenbacke, 2024a, 2024b) I established the foundation for the participating physicians by asserting the trustworthiness of the algorithm in use. This trust was instilled by informing the physicians that they should assume that the algorithm had been subject to rigorous validation processes—a statement framed specifically for the context of the experimental setup. This set-up was deliberate to avoid discussions of trust versus trustworthiness. However, I deliberately calibrated the AI algorithms at 60% accuracy to be able to study the clinician’s trust and the cognitive challenges when the AI advice was incorrect. Next, I will examine how the accuracy of AI systems influences the error rate.

Frequency of False Confirmation errors

To estimate the rate of False Confirmation errors, we must consider the product of the clinician's error rate and the AI error rate. Specifically, this is expressed by the formula:

False Confirmation = (1 - AI Accuracy) x (1 - Physician Accuracy).

However, an essential preliminary step involves discerning the respective error rates of both the AI system and the clinicians. In recent systematic reviews and meta-analyses, the authors compared the performance of AI versus physicians (Shen *et al.*, 2019; Nazarian *et al.*, 2021). They found that AI accuracy varies from 60% to 99%, with sensitivity and specificity from 60% to 99%. Physicians' clinical performance was 48-99% when measuring accuracy, sensitivity, and specificity.

The selection of a 60% accuracy rate for the AI in Paper II was a deliberate methodological choice. This figure falls within the lower spectrum of the 60-99% accuracy range for AI as reported in recent systematic reviews and meta-analyses (Shen *et al.*, 2019; Nazarian *et al.*, 2021). In Paper II, the clinicians' accuracy for the lowest performers was 41% and 65% for the best performers. In Paper V, the accuracy for the lowest performers was 47%, and for the best, 87%. This aligns with the reported variability in the clinical performance of physicians, ranging from 48-99% in terms of accuracy, sensitivity, and specificity. By choosing a lower-performing AI model, Paper II aimed to illuminate the cognitive challenges faced by clinicians when the AI guidance is both correct and incorrect.

In paper IV I simulate the theoretical likelihood of False Confirmation errors. With an AI accuracy of 60% and a clinical accuracy of 50%—particularly relevant for lower-performing clinicians that could have even lower accuracy—the estimated rate of False Confirmation errors could reach as high as 20% to 30% (Rosenbacke *et al.*, 2024a). However, it is also posited that these lower-performing clinicians might experience a marked decrease in True Conflict errors, as evidenced by the improved overall accuracy in Papers II and V. For further details, also refer to Appendix VII.

As AI systems become more accurate, the rate of diagnostic errors, including False Confirmations, is expected to decrease, leading to more effective collaboration between clinicians and AI and enhancing patient outcomes and diagnostic reliability. However, AI will never be 100% infallible. Medical data is highly variable and complex, making it challenging for AI to generalize across all cases. AI relies on the quality of its training data, which can be incomplete or biased, and requires constant updates as new medical conditions and treatments emerge. AI may struggle with nuanced interpretations of medical images and tests that require contextual and experiential understanding, and it cannot fully grasp the ethical considerations often involved in medical decisions. Additionally, AI systems can be prone to technical issues, such as software bugs and cyber threats, affecting their reliability. Therefore, I argue that AI should be viewed as a tool to augment human decision-making, combining AI's data-processing capabilities with the expertise and judgment of human clinicians for the best outcomes.

Societal contribution

My work on trust and cognitive challenges when introducing explainable AI in clinical decision-making allowed me to contribute to a study on trust for the World Health Organization (WHO), the policy brief "*Trust: The foundation of health systems*" published by the European Observatory on Health Systems and Policies (McKee, Greenley and Permanand, 2023). I advised the WHO team of researchers on trust and artificial intelligence in health systems. In summary, this study found that "*those involved in implementing AI solutions must consider how they will be received by those who must use them and, especially, whether they will engender the appropriate level of trust, neither too much nor too little. There is some evidence that XAI can help but much more research is needed to understand how it can be most effective and in what circumstances and therefore when it should be seen as trustworthy (and trusted to the extent that is appropriate).*"(McKee, Greenley and Permanand, 2023, p. 26).

The WHO policy brief highlights the importance of not only ensuring that AI and XAI systems are trustworthy but also that clinicians trust these systems in practice while managing cognitive challenges. In my research, I contributed to the WHO study and policy brief and also responded to their call for greater clarity on how to optimize trust in AI for healthcare applications. The WHO emphasizes fostering the "right" level of trust—not too much to avoid overreliance and not too little to ensure clinicians benefit from AI's potential. This aligns with the findings in my systematic review (Rosenbacke *et al.*, 2024b), where I demonstrate that while XAI may enhance clinician trust, it can also create risks of overreliance, especially when AI is incorrect. This dual aspect—trustworthiness and appropriate levels of trust—must be carefully managed, as highlighted in my research, to ensure effective AI integration into clinical decision-making.

Further, in my other papers, I delve into the errors that occur when clinicians trust AI too much, including False Confirmation errors (Rosenbacke, 2024a; Rosenbacke, Melhus and Stuckler, 2024), and the underlying cognitive biases, such as commitment and confirmation bias (Rosenbacke, 2024b), that contribute to these issues. This deepens the WHO's concern regarding the fine balance between ensuring AI is trusted enough to be used but not blindly followed.

I have also contributed to the ongoing call for further analysis in *The International Journal of Health Planning and Management* (Lopes, Martins and Correia, 2024) regarding the applications and implications of artificial intelligence in policy, planning, and management (McKee, Rosenbacke and Stuckler, 2024). It is important to note that although I refer to this contribution, it is not included within the scope of this thesis. However, this additional work broadens the impact of my research, highlighting its relevance beyond clinical AI integration and into health system management and policy frameworks.

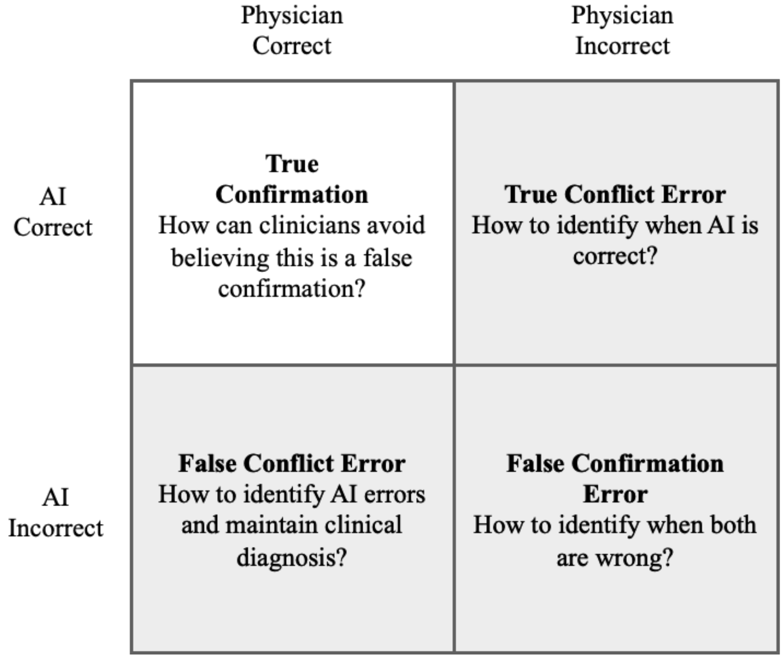
In my work, I provide empirical evidence and a theoretical framework for understanding these dynamics. This makes a meaningful contribution to the global conversation about trust in AI in healthcare systems, offering insights into how these systems can be designed and implemented to maximize their benefits while mitigating potential harms.

The subsequent section will discuss potential strategies for how to improve True Conflict errors while avoiding False Conflict and Confirmation errors in human-AI collaborative settings, which presents an expansive field for future research.

Future research on how to get the most out of a clinical AI collaboration

In an environment like healthcare, where human-AI interaction is pivotal, a central inquiry is whether the synergy between humans and artificial intelligence can surpass the capabilities of either part independently. The concept hinges on the effectiveness of each entity in recognizing and augmenting the other's strengths, potentially leading to an entity that is greater than the sum of its parts (Bansal *et al.*, 2019; Bansal, Wu and Zhou, 2021; Fügener *et al.*, 2021; Hemmer *et al.*, 2023).

Utilizing my framework, in Figure 10, to identify True and False Conflicts, as well as confirmation errors, can potentially provide insights into optimizing these interactions. The integration of AI in clinical settings should ideally complement clinical expertise, thereby increasing diagnostic accuracy. The goal is to correct clinical errors, but introducing AI can also bring about new types of errors, such as False Conflicts, which may neutralize some of the improvements achieved by resolving True Conflicts. This reflects the challenges seen in my studies, where the theoretical benefits of AI assistance were not always realized in practice (Rosenbacke, 2024a; Rosenbacke *et al.*, 2024b; Rosenbacke, Melhus and Stuckler, 2024; Weis *et al.*, 2024).



	Physician Correct	Physician Incorrect
AI Correct	**True Confirmation** How can clinicians avoid believing this is a false confirmation?	**True Conflict Error** How to identify when AI is correct?
AI Incorrect	**False Conflict Error** How to identify AI errors and maintain clinical diagnosis?	**False Confirmation Error** How to identify when both are wrong?

Figure 10: How can we create synergies where the human-AI collaboration is better than the parts?

Without AI, the False Confirmation errors could likely have happened with only a clinical diagnosis. However, there is a risk that with AI support, the clinician will not do the same diligent and reason-based thought process as soon as confirmed. It seems that in my studies, clinicians have stopped any further diligence. It might be that AI increases these errors. Further studies are needed.

In both my work (Rosenbacke, 2024a) and the work of Chanda et al. (2024), an enhancement in diagnostic precision was observed, signifying a net positive effect from AI integration. Nonetheless, these studies did not adequately successfully address the phenomenon of False Confirmation errors. Specifically, even with the presence of explanatory frameworks and risk factor rankings that conflicted with evidence-based medicine or the clinician's perspective, no reduction in the incidence of False Confirmation errors was achieved. Interventions such as cognitive forcing strategies appeared ineffective in mitigating these errors despite the indications that the AI conclusions were inconsistent or incorrect (Rosenbacke, 2024a). This suggests that the implementation of AI in clinical settings may require additional methods to manage and minimize the occurrence of False Confirmation errors effectively.

To optimize human-AI collaboration so that it is more effective than its individual components, a multi-faceted approach is necessary. Here are several considerations and potential strategies that need further research to investigate if they could contribute to such synergy:

- Regularly engage clinicians in simulated environments where they encounter controlled False Conflict and False Confirmation scenarios. This training can sharpen their skills in distinguishing between AI errors and correct assertions.
- Educate healthcare professionals about common cognitive biases that may lead to False Confirmations, such as confirmation bias and overconfidence in AI. Training could include strategies to mitigate these biases.
- Maintain comprehensive audit trails for both AI and clinician decisions to facilitate retrospective analysis and learn from instances of False Conflicts and False Confirmations.
- Foster collaboration between clinicians, AI developers, data scientists, and cognitive psychologists to create a holistic approach to decision-making that considers diverse perspectives and expertise.
- Implementing nudges within AI systems can prompt critical evaluation by physicians. These nudges could act as reminders, encouraging doctors to question consensus and consider the possibility of errors from both AI and their initial diagnosis.
- Promoting a comparative analysis of AI explanations (including ranking and weights of risk factors) compared to physician clinical judgment. Examining whether they rely on the same clinical factors and assign similar weights can uncover discrepancies, particularly in cases of False Confirmation.

∞ Finally, in medical diagnostics, 'ground truth' can be elusive due to the ambiguous nature of some conditions. These are the gray areas—not strictly positive or negative—where symptoms may partially match several different diagnoses. If an AI system, designed to aid in diagnostics, operates only in a binary framework, it could erroneously reinforce a physician's initial, potentially incorrect judgment, discouraging further investigation. To address this, AI could adopt a tri-class system to account for ambiguity, identifying cases as positive, negative, or uncertain. This could potentially encourage further scrutiny and reduce False Confirmation errors and the risk of premature diagnostic closure, acknowledging the nuanced reality of medical practice. Further research is needed.

I argue that studies of human-AI collaboration demand an interdisciplinary approach, one that marries the precision of natural sciences and the nuance of social sciences. Clinicians, who primarily rely on evidence-based medicine, operate within a culture accustomed to the complexities of human conditions, which often present in shades of gray rather than the black-and-white outcomes typical in computer science. Furthermore, to harness the full potential of AI in clinical settings, it's crucial to integrate insights from cognitive psychology, which aims to understand the intricacies of human decision-making.

My contribution lies at this intersection — serving as an interpreter between the definitive stances of computer science, the flexibility of social science, and the variable, patient-centered approach of medicine. By facilitating this dialogue, my research aligns with the World Health Organization's calls to optimize AI for healthcare, ensuring it complements the nuanced decision-making processes of clinicians (McKee, Greenley and Permanand, 2023). Bringing together the predictive power of AI with the context-driven acumen of clinicians can create a synergistic partnership. This partnership not only improves diagnostic accuracy but also enhances the efficacy of healthcare delivery, advancing both the science and practice of medicine.

Nevertheless, this thesis merely scratches the surface of a vast and complex field. There remains an extensive scope for further research. As we move forward, it's essential to pursue a more granular understanding of when and how clinicians should place their trust in AI, balancing between AI system validation and user-centered design. Recognizing this, my research invites a broader conversation and more comprehensive studies to fully unravel the potential and perils of AI in healthcare. The journey toward seamless and effective human-AI partnerships in clinical settings is ongoing, and much work remains to ensure these partnerships are as productive and safe as possible.

Conclusion

In conclusion, this research has unearthed the intricate cognitive challenges that arise from the collaboration between clinicians and AI and XAI, particularly highlighting the need for a nuanced understanding of errors in the clinician-AI interface. Most of the current literature has concentrated on how and to what extent physicians adjust their decision-making when AI models provide accurate recommendations. However, there has been a significant oversight regarding the impact of AI errors, especially when these are accompanied by explanations. This gap leaves unanswered questions about the influence of incorrect AI guidance on clinical accuracy and the potential risks of overreliance on AI explanations in such scenarios.

My novel framework identifies that diagnostic errors are manifold: True Conflict errors, where clinicians dismiss correct AI advice; False Conflict errors, where clinicians are misguided by incorrect AI; and False Confirmation errors that masquerade as consensus, but the clinician and the AI are wrong in their diagnosis, and the clinicians halt further investigation. The investigation into these errors is pivotal for the successful integration of AI in clinical practice.

This research provides a pioneering empirical analysis across diverse datasets and thousands of clinical decisions, establishing these error patterns as both replicable and significant. This thesis is the first, to the best of my knowledge, to quantify how XAI, despite its potential to enhance trust, can also inadvertently increase False Conflict errors due to overreliance on explanations. Moreover, it identifies and quantifies False Confirmation errors—a particularly insidious form of error where clinicians, misled by incorrect AI advice, prematurely halt further investigation. This behavior reflects a tendency toward confirmation bias rather than a deliberate decision. To the best of my knowledge, this research is the first to demonstrate that XAI explanations fail to effectively mitigate these specific types of errors. The research further demonstrates that even experienced clinicians are vulnerable to these errors, especially when falsely confirmed by AI, revealing a complex relationship between trust, expertise, and diagnostic accuracy in clinician-AI collaboration

By developing a novel systematic framework to measure and understand these errors, my work advances the discourse on AI and XAI as a second opinion and provides a practical approach to refining AI and XAI applications in healthcare. This research does not just fill a gap; it potentially sets a new foundation for future studies aimed at mitigating these errors and optimizing the clinician-AI partnership, ultimately contributing to safer and more reliable AI integration in medicine.

Despite that I have been able to reproduce my findings, still the main limitations of this thesis lie in its generalizability and scope. The studies are still based on specific medical contexts (ear infections and melanoma) and AI accuracy rates, which limit the broader applicability of the findings to other clinical settings or AI systems. Additionally, in the systematic review, the small sample sizes and inconsistent reporting on trust measures and algorithm accuracy across studies constrain the robustness of the conclusions. Future research should broaden the scope, exploring different clinical contexts, higher accuracy AI systems, and improving consistency in measuring trust and algorithm performance.

Nonetheless, it is clear from the findings that we stand at the cusp of a nascent field that warrants extensive further investigation. There is a significant gap in understanding the long-term impacts of AI and XAI on clinical decision-making. This thesis stresses the imperative for continued empirical research to mitigate the risks of False Conflict and Confirmation errors and to fine-tune AI systems to support rather than undermine critical clinician judgment.

The insights garnered from this work have laid a foundational understanding of the cognitive interplay between clinicians and AI. My research calls for a continuation of this inquiry, underscoring the need for systematic and thorough research to navigate the complexities of AI integration in healthcare. Only through such dedicated scholarly endeavors can the scope of the AI potential be harnessed and its pitfalls adequately addressed to foster better clinical outcomes.

Echoing the words of the cognitive psychologist Steven Pinker: "*The beauty of reason is that it can always be applied to understand failures of reason*" (Pinker, 2018), this thesis illustrates the need to refine and elevate our rational endeavors in healthcare through continued investigation and reflection.

References

Ackerman, R. and Thompson, V. A. (2017) 'Meta-Reasoning: Monitoring and Control of Thinking and Reasoning', *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2017.05.004.

Alvesson, M. and Kärreman, D. (2007) 'Constructing mystery: Empirical matters in theory development', *Academy of Management Review*. doi: 10.5465/AMR.2007.26586822.

Alvesson, M. and Sandberg, J. (2011) 'Generating research questions through problematization', *Academy of Management Review*. doi: 10.5465/amr.2009.0188.

Alvesson, M. and Sköldberg, K. (2000) *Reflexive Methodology: new vistas for qualitative research (second edition)*, *Sage*.

Amann, J. *et al.* (2020) 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective', *BMC Medical Informatics and Decision Making*. BioMed Central Ltd, 20(1), pp. 1–9. doi: 10.1186/S12911-020-01332-6/PEER-REVIEW.

Antoniadi, A. M. *et al.* (2021) 'Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review', *Applied Sciences (Switzerland)*. MDPI AG, 11(11). doi: 10.3390/app11115088.

'API Reference' (2006) in *The Definitive Guide to MySQL5*, pp. 693–720. doi: 10.1007/978-1-4302-0071-0_23.

Bandura, A. (1986) 'Social foundations of thought and action : a social cognitive theory / Albert Bandura.', *New Jersey: Prentice-Hall, 1986*, 16(1).

Bansal, G. *et al.* (2019) 'Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance', in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 2–11. doi: 10.1609/hcomp.v7i1.5285.

Bansal, G., Wu, T. and Zhou, J. (2021) 'Does the whole exceed its parts? The efect of ai explanations on complementary team performance', in *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3411764.3445717.

Barredo Arrieta, A. *et al.* (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58. doi: 10.1016/j.inffus.2019.12.012.

Bauer, M. S. and Kirchner, J. A. (2020) 'Implementation science: What is it and why should I care?', *Psychiatry Research*. Elsevier, 283, p. 112376. doi: 10.1016/J.PSYCHRES.2019.04.025.

Beede, E. *et al.* (2020) 'A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy', in *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3313831.3376718.

Bertrand, A. *et al.* (2022) 'How cognitive biases affect XAI-Assisted decision-making: A systematic review', in *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91. doi: 10.1145/3514094.3534164.

Boell, S. K. and Cecez-Kecmanovic, D. (2015) 'Debating systematic literature reviews (SLR) and their ramifications for IS: A rejoinder to Mike Chiasson, Briony Oates, Ulrike Schultze, and Richard Watson', *Journal of Information Technology*. Palgrave Macmillan, 30(2), pp. 188–193. doi: 10.1057/JIT.2015.15.

Bornstein, R. F. and D'Agostino, P. R. (1992) 'Stimulus Recognition and the Mere Exposure

Effect', *Journal of Personality and Social Psychology*, 63(4). doi: 10.1037/0022-3514.63.4.545.

Box, G. E. P. (1976) 'Science and statistics', *Journal of the American Statistical Association*, 71(356), pp. 791–799. doi: 10.1080/01621459.1976.10480949.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101. doi: 10.1191/1478088706QP063OA.

Braun, V. and Clarke, V. (2012) 'Thematic analysis.', in *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, pp. 57–71. doi: 10.1037/13620-004.

Bruce, G. (2024) *Why healthcare lags on AI, per Amazon*. Available at: https://www.beckershospitalreview.com/disruptors/why-healthcare-lags-on-ai-per-amazon.html (Accessed: 8 April 2024).

Brynjolfsson, E. and Mcafee, A. (2017) 'The business of artificial intelligence', *Harvard Business Review*.

Brynjolfsson, E. and Mitchell, T. (2017) 'What can machine learning do? Workforce implications', *Science*. doi: 10.1126/science.aap8062.

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', *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery, 5(CSCW1), p. 21. doi: 10.1145/3449287.

Carruthers, B. (no date) *Differentiating Trust and Trustworthiness: A Sociologist's Perspective | Kellogg School of Management*. Available at: https://www.kellogg.northwestern.edu/academics-research/trust-project/videos/carruthers-ep-1.aspx (Accessed: 12 April 2024).

Carter, J. A. (2023) 'Trust and trustworthiness', *Philosophy and Phenomenological Research*, 107(2). doi: 10.1111/phpr.12918.

Castelvecchi, D. (2016) 'Can we open the black box of AI?', *Nature*, 538(7623). doi: 10.1038/538020a.

Chanda, T. *et al.* (2024) 'Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma', *nature.comT Chanda, K Hauser, S Hobelsberger, TC Bucher, CN Garcia, C Wies, H Kittler, P TschandlNature Communications, 2024•nature.com*. Available at: https://www.nature.com/articles/s41467-023-43095-4 (Accessed: 29 January 2024).

Chew, H. S. J. and Achananuparp, P. (2022) 'Perceptions and Needs of Artificial Intelligence in Health Care to Increase Adoption: Scoping Review', *Journal of Medical Internet Research*. doi: 10.2196/32939.

Chromik, M. *et al.* (2021) 'I Think i Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 307–317. doi: 10.1145/3397481.3450644.

Constantinides, P. (2023) *Digital Transformation in Healthcare: An Ecosystem Approach*, *Digital Transformation in Healthcare: An Ecosystem Approach*. doi: 10.4324/9781032619569.

Constantiou, I., Joshi, M. P. and Stelmaszak, M. (2024) 'Introduction to research handbook on artificial intelligence and decision making in organizations', in *Research Handbook on Artificial Intelligence and Decision Making in Organizations*. doi: 10.4337/9781803926216.00007.

Creswell, J. and Plano, V. (2017) 'Designing and Conducting Mixed Methods Research - John

W. Creswell, Vicki L. Plano Clark', *SAGE Publications*.

Cui, M. and Zhang, D. Y. (2021) 'Artificial intelligence and computational pathology', *Laboratory Investigation*. doi: 10.1038/s41374-020-00514-0.

Dai, T. and Ching, A. (2022) *AI in Healthcare Is Here, But Uptake Is Slow*, *Hopkins Business of Health Initiative*. Available at: https://hbhi.jhu.edu/news/ai-healthcare-here-uptake-slow (Accessed: 8 April 2024).

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000033.

Dolan, P. *et al.* (2012) 'Influencing behaviour: The mindspace way', *Journal of Economic Psychology*. doi: 10.1016/j.joep.2011.10.009.

Eiband, M. *et al.* (2019) 'The impact of placebic explanations on trust in intelligent systems', in *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3290607.3312787.

Epstein, R. M. and Street, R. L. (2011) 'The values and value of patient-centered care', *Annals of Family Medicine*, pp. 100–103. doi: 10.1370/afm.1239.

Evans, T. *et al.* (2022) 'The explainability paradox: Challenges for xAI in digital pathology', *Future Generation Computer Systems*, 133. doi: 10.1016/j.future.2022.03.009.

Fazal, M. I. *et al.* (2018) 'The past, present and future role of artificial intelligence in imaging', *European Journal of Radiology*. doi: 10.1016/j.ejrad.2018.06.020.

Fügener, A. *et al.* (2021) 'Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation', *https://doi.org/10.1287/isre.2021.1079*. INFORMS, 33(2), pp. 678–696. doi: 10.1287/ISRE.2021.1079.

Fürnkranz, J., Kliegr, T. and Paulheim, H. (2020) 'On cognitive preferences and the plausibility of rule-based models', *Machine Learning*. Springer, 109(4), pp. 853–898. doi: 10.1007/s10994-019-05856-5.

Gaube, S. *et al.* (2023) 'Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays', *Scientific Reports*, 13(1). doi: 10.1038/s41598-023-28633-w.

Gerlings, J., Shollo, A. and Constantiou, I. (2021) 'Reviewing the need for explainable artificial intelligence (XAI)', in *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE Computer Society, pp. 1284–1293. doi: 10.24251/hicss.2021.156.

Ghassemi, M., Oakden-Rayner, L. and Beam, A. L. (2021) 'The false hope of current approaches to explainable artificial intelligence in health care', *The Lancet Digital Health*, pp. e745–e750. doi: 10.1016/S2589-7500(21)00208-9.

Gigerenzer, G. and Gaissmaier, W. (2011) 'Heuristic decision making', *Annual Review of Psychology*, 62, pp. 451–482. doi: 10.1146/annurev-psych-120709-145346.

Gisselsson-Solén, M. *et al.* (2014) 'Risk factors for carriage of AOM pathogens during the first 3 years of life in children with early onset of acute otitis media', *Acta Oto-Laryngologica*. Informa Healthcare, 134(7), pp. 684–690. doi: 10.3109/00016489.2014.890291.

Gisselsson-Solén, M. *et al.* (2015) 'Effect of pneumococcal conjugate vaccination on nasopharyngeal carriage in children with early onset of acute otitis media-a randomized controlled trial', *Acta Oto-Laryngologica*. Informa Healthcare, 135(1), pp. 7–13. doi:

10.3109/00016489.2014.950326.

Giuste, F. *et al.* (2023) 'Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review', *IEEE Reviews in Biomedical Engineering*, 16. doi: 10.1109/RBME.2022.3185953.

Glikson, E. and Woolley, A. W. (2020) 'Human trust in artificial intelligence: Review of empirical research', *Academy of Management Annals*, 14(2). doi: 10.5465/annals.2018.0057.

Goodfellow I, Bengio Y, C. A. (2016) 'Deep Learning - MIT', *Nature*. doi: 10.1038/nmeth.3707.

Gordon, C. (2013) 'Beyond the observer's paradox: The audio-recorder as a resource for the display of identity', *Qualitative Research*, 13(3). doi: 10.1177/1468794112442771.

Greenhalgh, T. (2019) 'How to read a paper : the basics of evidence-based medicine and healthcare', *News.Ge*.

Gunning, D. and Aha, D. W. (2019) 'DARPA's explainable artificial intelligence program', *AI Magazine*, 40(2). doi: 10.1609/aimag.v40i2.2850.

Gupta, D. M., Boland, R. J. and Aron, D. C. (2017) 'The physician's experience of changing clinical practice: A struggle to unlearn', *Implementation Science*, 12(1). doi: 10.1186/s13012-017-0555-2.

Haque, A. B., Islam, A. K. M. N. and Mikalef, P. (2023) 'Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research', *Technological Forecasting and Social Change*, 186. doi: 10.1016/j.techfore.2022.122120.

Hemmer, P. *et al.* (2023) 'Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction', *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 453–463. doi: 10.1145/3581641.3584052.

Ioannidis, J. P. A. (2018) 'Why most published research findings are false', in *Getting to Good: Research Integrity in the Biomedical Sciences*. doi: 10.1371/journal.pmed.0020124.

Jongsma, K. R. and Sand, M. (2022) 'Agree to disagree: the symmetry of burden of proof in human-AI collaboration', *Journal of Medical Ethics*. doi: 10.1136/medethics-2022-108242.

Jung, J. *et al.* (2023) 'Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review', *Heliyon*. doi: 10.1016/j.heliyon.2023.e16110.

Jussupow, E. *et al.* (2021) 'Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence', *Information Systems Research*, 32(3). doi: 10.1287/ISRE.2020.0980.

Kahneman, D. (2011) *Thinking fast, thinking slow*, *Interpretation, Tavistock, London*.

Kaushik, V. and Walsh, C. A. (2019) 'Pragmatism as a research paradigm and its implications for Social Work research', *Social Sciences*, 8(9). doi: 10.3390/socsci8090255.

Kepecs, A. *et al.* (2008) 'Neural correlates, computation and behavioural impact of decision confidence', *Nature*, 455(7210). doi: 10.1038/nature07200.

Kiani, A. *et al.* (2020) 'Impact of a deep learning assistant on the histopathologic classification of liver cancer', *npj Digital Medicine*, 3(1). doi: 10.1038/s41746-020-0232-8.

Klein, G. (2015) ‘A naturalistic decision making perspective on studying intuitive decision making’, *Journal of Applied Research in Memory and Cognition*, 4(3). doi: 10.1016/j.jarmac.2015.07.001.

Kliegr, T., Bahník, Š. and Fürnkranz, J. (2021) ‘A review of possible effects of cognitive biases on interpretation of rule-based machine learning models’, *Artificial Intelligence*, 295. doi: 10.1016/j.artint.2021.103458.

Kolasa, K. *et al.* (2023) ‘Systematic reviews of machine learning in healthcare: a literature review’, *Taylor & Francis*. Taylor and Francis Ltd. doi: 10.1080/14737167.2023.2279107.

De Koning, H. J. *et al.* (2003) ‘Determining the cause of death in randomized screening trial(s) for prostate cancer’, *BJU International, Supplement*, 92(2). doi: 10.1111/j.1465-5101.2003.04402.x.

Kundu, S. (2021) ‘AI in medicine must be explainable’, *Nature Medicine*, 27(8). doi: 10.1038/s41591-021-01461-z.

Lecun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep learning’, *Nature*, pp. 436–444. doi: 10.1038/nature14539.

Lee, J. G. *et al.* (2017) ‘Deep learning in medical imaging: General overview’, *Korean Journal of Radiology*. doi: 10.3348/kjr.2017.18.4.570.

Lewicki, R. J. and Brinsfield, C. T. (2011) ‘Framing trust: Trust as a heuristic’, in Donohue, W. A., Rogan, R. R., and Kaufman, S. (eds) *Framing matters: Perspectives on negotiatin research and practice in communication*. Peter Lang Publishing, pp. 110–135. Available at: http://www.mdpi.com/1996-1073/2/3/556/.

Logg Jennifer (2018) ‘Do People Trust Algorithms More Than Companies Realize?’, *Harvard Business Review*. Available at: https://hbr.org/2018/10/do-people-trust-algorithms-more-than-companies-realize (Accessed: 13 September 2019).

Loh, H. W. *et al.* (2022) ‘Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)’, *Computer Methods and Programs in Biomedicine*. doi: 10.1016/j.cmpb.2022.107161.

Lopes, M. A., Martins, H. and Correia, T. (2024) ‘Artificial intelligence and the future in health policy, planning and management’, *International Journal of Health Planning and Management*. doi: 10.1002/hpm.3709.

Lucic, A., Haned, H. and de Rijke, M. (2020) ‘Why does my model fail? Contrastive local explanations for retail forecasting’, in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. doi: 10.1145/3351095.3372824.

Madsen, M. and Gregor, S. (2000) ‘Measuring Human-Computer Trust’, *Proceedings of Eleventh Australasian Conference on Information Systems*.

Mahmud, H. *et al.* (2022) ‘What influences algorithmic decision-making? A systematic literature review on algorithm aversion’, *Technological Forecasting and Social Change*, 175. doi: 10.1016/j.techfore.2021.121390.

Matias, J. N. (2017) *Bias and Noise: Daniel Kahneman on Errors in Decision-Making, Medium.com*. Available at: https://medium.com/@natematias/bias-and-noise-daniel-kahneman-onerrors-in-decision-making-6bc844ff5194 (Accessed: 21 October 2019).

McKee, M., Greenley, R. and Permanand, G. (2023) *Trust: The foundation of health systems.*

Available at: https://eurohealthobservatory.who.int/publications/i/trust-the-foundation-of-health-systems (Accessed: 11 March 2024).

McKee, M., Rosenbacke, R. and Stuckler, D. (2024) 'The power of artificial intelligence for managing pandemics: A primer for public health professionals', *The International journal of health planning and management*. Int J Health Plann Manage. doi: 10.1002/HPM.3864.

Miller, T. (2019) 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, pp. 1–38. doi: 10.1016/j.artint.2018.07.007.

Moor, M. *et al.* (2023) 'Foundation models for generalist medical artificial intelligence', *Nature*, 616(7956). doi: 10.1038/s41586-023-05881-4.

Morgan, D. L. (2014) 'Pragmatism as a Paradigm for Social Research', *Qualitative Inquiry*, 20(8). doi: 10.1177/1077800413513733.

Naiseh, M., Al-Thani, D., *et al.* (2021) 'Explainable recommendation: when design meets trust calibration', *World Wide Web*, 24(5). doi: 10.1007/s11280-021-00916-0.

Naiseh, M., Cemiloglu, D., *et al.* (2021) 'Explainable Recommendations and Calibrated Trust: Two Systematic User Errors', *Computer*, 54(10). doi: 10.1109/MC.2021.3076131.

Naiseh, M., Al-Mansoori, R. S., *et al.* (2021) 'Nudging through Friction: an Approach for Calibrating Trust in Explainable AI', in *Proceedings of 2021 8th IEEE International Conference on Behavioural and Social Computing, BESC 2021*. doi: 10.1109/BESC53957.2021.9635271.

Naiseh, M. *et al.* (2023) 'How the different explanation classes impact trust calibration: The case of clinical decision support systems', *International Journal of Human Computer Studies*, 169. doi: 10.1016/j.ijhcs.2022.102941.

Nazar, M. *et al.* (2021) 'A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques', *IEEE Access*. doi: 10.1109/ACCESS.2021.3127881.

Nazarian, S. *et al.* (2021) 'Diagnostic accuracy of artificial intelligence and computer-aided diagnosis for the detection and characterization of colorectal polyps: Systematic review and meta-analysis', *Journal of Medical Internet Research*. doi: 10.2196/27370.

Nickerson, R. S. (1998) 'Confirmation bias: A ubiquitous phenomenon in many guises', *Review of General Psychology*. doi: 10.1037/1089-2680.2.2.175.

Nourani, M. *et al.* (2021) 'Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 340–350. doi: 10.1145/3397481.3450639.

Nyberg, J., Rosenbacke, R. and Ben-Menachem, E. (2024) 'Digital clinics for diagnosing and treating migraine', *Current opinion in supportive and palliative care*. Curr Opin Support Palliat Care, 18(3). doi: 10.1097/SPC.0000000000000715.

Okoli, C. and Schabram, K. (2012) 'A Guide to Conducting a Systematic Literature Review of Information Systems Research', *SSRN Electronic Journal*. doi: 10.2139/ssrn.1954824.

Page, M. J. *et al.* (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *The BMJ*. doi: 10.1136/bmj.n71.

Petersson, L. *et al.* (2022) 'Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden', *BMC Health Services Research*, 22(1). doi: 10.1186/s12913-022-08215-8.

Pinker, S. (2018) *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin.

Rajpurkar, P. *et al.* (2022) 'AI in health and medicine', *Nature Medicine*. doi: 10.1038/s41591-021-01614-0.

Reddy, S. *et al.* (2020) 'A governance model for the application of AI in health care', *Journal of the American Medical Informatics Association*, pp. 491–497. doi: 10.1093/jamia/ocz192.

Reddy, S. (2022) 'Explainability and artificial intelligence in medicine', *The Lancet Digital Health*. doi: 10.1016/S2589-7500(22)00029-2.

Reinhardt, K. (2023) 'Trust and trustworthiness in AI ethics', *AI and Ethics*, 3(3). doi: 10.1007/s43681-022-00200-5.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2018) 'Anchors: High-precision model-agnostic explanations', in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. doi: 10.1609/aaai.v32i1.11491.

Rosenbacke, R. *et al.* (2022) 'DESIGNING A DIGITAL ARTIFACT FOR DATA COLLECTION TO IMPROVE DAILY ADHD MEDICATION', *ECIS 2022 Research-in-Progress Papers*. Available at: https://aisel.aisnet.org/ecis2022_rip/22 (Accessed: 20 October 2022).

Rosenbacke, R. *et al.* (2024a) 'AI and XAI second opinion: the danger of false confirmation in human-AI collaboration', *Journal of Medical Ethics*. doi: 10.1136/jme-2024-110074.

Rosenbacke, R. (2024a) 'Errors in Physician-AI Collaboration: Insights From a Mixed-methods Study of Explainable AI and Trust in Clinical Decision-making', *SSRN Electronic Journal*. doi: 10.2139/SSRN.4773350.

Rosenbacke, R. (2024b) 'HEURISTICS AND ERRORS IN XAI-AUGMENTED CLINICAL DECISION-MAKING: MOVING BEYOND ALGORITHMIC APPRECIATION AND AVERSION', *ECIS 2024 Proceedings*. Available at: https://aisel.aisnet.org/ecis2024/track03_ai/track03_ai/11 (Accessed: 3 May 2024).

Rosenbacke, R. *et al.* (2024b) 'How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Healthcare: Systematic Review', *Journal of Medical Internet Research AI*, 3. doi: 10.2196/53207.

Rosenbacke, R., Melhus, Å. and Stuckler, D. (2024) 'False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making', *Nature Communications 2024 15:1*. Nature Publishing Group, 15(1), pp. 1–2. doi: 10.1038/s41467-024-50952-3.

Schwalbe, N. and Wahl, B. (2020) 'Artificial intelligence and the future of global health', *The Lancet*. doi: 10.1016/S0140-6736(20)30226-9.

Shen, J. *et al.* (2019) 'Artificial intelligence versus clinicians in disease diagnosis: Systematic review', *JMIR Medical Informatics*. doi: 10.2196/10010.

Van Someren, M., Barnard, Y. F. and Sandberg, J. (1994) 'The think aloud method: a practical approach to modelling cognitive', *London: AcademicPress*, 11.

Straus, S. E. . *et al.* (2019) 'Evidence-based medicine : how to practice and teach EBM'. Elsevier, p. 324.

Sutton, R. T. *et al.* (2020) ‘An overview of clinical decision support systems: benefits, risks, and strategies for success’, *npj Digital Medicine*. doi: 10.1038/s41746-020-0221-y.

Timmermans, S. and Mauck, A. (2005) ‘The promises and pitfalls of evidence-based medicine’, *Health Affairs*. doi: 10.1377/hlthaff.24.1.18.

Tonelli, M. R. (2006) ‘Integrating evidence into clinical practice: An alternative to evidence-based approaches’, *Journal of Evaluation in Clinical Practice*, 12(3). doi: 10.1111/j.1365-2753.2004.00551.x.

Wadden, J. J. (2021) ‘Defining the undefinable: the black box problem in healthcare artificial intelligence’, *Journal of Medical Ethics*, 48(10). doi: 10.1136/medethics-2021-107529.

Wang, D. *et al.* (2019) ‘Designing theory-driven user-centric explainable AI’, in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. doi: 10.1145/3290605.3300831.

Wason, P. C. and Evans, J. S. B. T. (1974) ‘Dual processes in reasoning?’, *Cognition*. doi: 10.1016/0010-0277(74)90017-1.

Weis, C. *et al.* (2024) ‘Investigating Interaction Errors in Clinical Decision-Making: Implications for Risk Understanding and XAI Assistance in Melanoma Diagnostics’, in *Cancer Prevention Research Conference, American Cancer Society*.

Wies, C., Hauser, K. and Brinker, T. J. (2024) ‘Reply to: False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making’, *Nature Communications 2024 15:1*. Nature Publishing Group, 15(1), pp. 1–3. doi: 10.1038/s41467-024-50954-1.

Yetley, E. A. *et al.* (2017) ‘Options for basing Dietary Reference Intakes (DRIs) on chronic disease endpoints: Report from a joint US-/Canadian-sponsored working group’, in *American Journal of Clinical Nutrition*. doi: 10.3945/ajcn.116.139097.

Zhang, G. *et al.* (2021) ‘Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study’, *The Lancet Digital Health*, 3(10). doi: 10.1016/S2589-7500(21)00134-5.

Appendix I: Paper I

Published in *Journal of Internet Medical Research AI.* This is the peer-reviewed and accepted version. There can be deviations after copyediting and publisher layout.

Citation: Rosenbacke, R. et al. (2024) 'How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Healthcare: Systematic Review', *Journal of Medical Internet Research AI*, 3. doi: 10.2196/53207.

How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Healthcare

A Systematic Review

Rikard Rosenbacke[1], Åsa Melhus[2], Martin McKee[3], David Stuckler[4]

[1]Centre for Corporate Governance, Department of Accounting, Copenhagen Business School, Copenhagen, Denmark
[2]Department of Medical Sciences/Section of Clinical Microbiology, Uppsala University, Uppsala, Sweden
[3]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, United Kingdom
[4]Department of Social and Political Science, Bocconi University, Milano, Italy

Running title:" "Explainable AI in Healthcare"

* - corresponding author; rr.ccg@cbs.dk, Rikard Rosenbacke, Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg

Word Count: 4 757

Keywords: Explainable Artificial Intelligence; XAI; AI, Trust; Healthcare; Systematic Literature Review

Abstract

Background: Artificial intelligence (AI) has significant potential in clinical practice. However, its "black box" nature can lead clinicians to question its value. The challenge is to create sufficient trust for clinicians to feel comfortable using AI but not so much that they defer to it even when it produces results that conflict with their clinical judgement in ways that lead to incorrect decisions. Explainable AI (XAI) aims to address this by providing explanations of how AI algorithms reach their conclusions. However, it remains unclear whether such explanations foster an appropriate degree of trust to ensure the optimal use of AI in clinical practice.

Objective: The objective of this study was to systematically review and synthesize empirical evidence on the impact of XAI on clinicians' trust in AI-driven clinical decision-making.

Methods: A systematic review was conducted in accordance with PRISMA guidelines, searching PubMed and Web of Science databases. Studies were included if they empirically measured the impact of XAI on clinicians' trust, using cognition- or affect-based measures. Out of 778 articles screened, 10 met the inclusion criteria. We assessed risk of bias using standard tools appropriate to the methodology of each paper.

Results: The risk of bias of all papers was Moderate or Moderate to High. All included studies operationalized trust primarily through cognitive-based definitions, with two also incorporating affect-based measures. Of these, five studies reported that XAI increased clinicians' trust compared to standard AI, particularly when the explanations were clear, concise, and relevant to clinical practice. Three studies found no significant effect of XAI on trust, and the presence of explanations does not automatically improve trust. Notably, two studies highlighted that XAI could either enhance or diminish trust, depending on the complexity and coherence of the provided explanations. The majority of studies suggest that XAI has the potential to enhance clinicians' trust in recommendations generated by AI. However, complex or contradictory explanations can undermine this trust. More critically, trust in AI is not inherently beneficial, as AI recommendations are not infallible. These findings underscore the nuanced role of explanation quality and suggest that trust can be modulated through careful design of XAI systems.

Conclusions: Excessive trust in incorrect advice generated by AI can adversely impact clinical accuracy, just as can happen when correct advice is distrusted. Future research should focus on refining both cognitive and affect-based measures of trust, and on developing strategies to achieve an appropriate balance in terms of trust, preventing both blind trust and undue skepticism. Optimizing trust in AI systems is essential for their effective integration into clinical practice.

Introduction

Artificial intelligence (AI) is increasingly being promoted as a means to transform healthcare. AI can enhance clinical decision-making, reduce medical errors, and improve patient outcomes.[1,2] Yet to realize its full potential in healthcare, clinicians must trust it and be comfortable with its outputs.[3] Establishing and maintaining trust is challenging, especially in light of growing warnings from some leading AI experts about its potential risks to society.[4]

Currently, there is a dearth of studies on how to increase trust in AI among clinicians. In a recent systematic review on trust in AI, it was observed that transparency is critical for fostering trust among decision-makers.[5] To increase transparency, and thus trust in AI, it has been proposed that measures should be added to its predictions to make the models more transparent and explainable to human users.[6] So-called Explainable AI (XAI) can be considered to fall within several categories: (i) "local" (specific) explanations of an individual prediction [7]; (ii) "global" explanations presenting the model's general logic [8]; (iii) "counterfactual" explanations indicating a threshold at which the algorithm could change its recommendations; (iv) confidence explanations, indicating the probability that the prediction is correct [9]; and (v) example-based, where the AI justifies its decision by providing examples that have similar characteristics from the same dataset. [10]

Trust is a complex concept that has been explored in a range of disciplines, including philosophy, economics, sociology, and psychology [11–15], with a recent review by one of us [16] noting how little interaction exists between these disciplinary perspectives. Here we rely on psychological models, which we consider are particularly helpful in this context. In a dual theory developed by Daniel Kahneman[17], two main ways of thinking exist. The first is quick and based on gut feelings or intuition, whereas the second is slower, taking a more thoughtful and reasoning approach. Trust forms a mental picture of another person or a system, and when trying to untangle all its intricacies it is practically impossible to use only rational thought. Consequently, the decision to trust someone or something like an AI tool or a physician is often derived from an instinctive judgment or the intuition. In this model, trust is viewed as a decision-making shortcut, enabling the decision-maker to select information while ignoring other information to simplify a complex decision.[18] Applied to empirical research, Madsen and colleagues describe these two broad approaches as cognition-based trust and affect-based trust,[19] terms we will use in this study.

A series of recent reviews has examined XAI from a trust perspective. However, partly reflecting the speed of development of the field, these do not include the most recent empirical evidence from clinical settings, although they did consistently speculate that XAI could increase users' trust and thus the intention to use AI tools [20] [21] as well as enhance confidence in decisions and thus the trust of clinicians.[22] [23] None of these studies differentiated between varying trust measures or healthcare domains.

To fill this gap, we performed a systematic review of empirical evidence on the impact of XAI on clinicians' trust. Additionally, we categorised and differentiated studies according to which type of trust measure they employed, cognition- or affect-based trust, as well as types of medical data employed (imaging vs. tabular formats).

Methods

Search Strategy

Two of the authors (RR and DS) performed a systematic review using the PRISMA method. [24] On March 23, 2023, we searched the title and abstract fields of PubMed and recognised that the topic would be covered by a wide range of disciplines, and hence we also used Web of Science. We searched for published articles on XAI and trust within healthcare. Our initial reading revealed the use of many words that conveyed some aspect of what we might consider "trust". In light of this work and the many different conceptions of trust [25], we intentionally employed a broad search strategy without specifying trust and its alternative variants (such as confidence, intention to use, etc) to avoid risk of 'type-2 errors' whereby relevant articles which should have been included were omitted.

We operationalised XAI and healthcare using a range of keyword permutations adapted to each database (see Appendix 1 for full strategy).

Inclusion and exclusion criteria

We applied a range of inclusion and exclusion criteria. Articles were included if they (i) measured trust (and related terms) as an outcome, (ii) used XAI as an intervention or exposure, (iii) used Machine Learning (ML) in the underlying AI model, (iv) were empirical studies, and (v) the evaluation was carried out by practicing clinicians. Articles were excluded if they were (i) reviews, commentaries, reports of methodology, or conceptual papers or ii) not applied in a healthcare setting from a clinician's perspective. Two reviewers, RR and DS, performed the screening, and any disputes were resolved against these pre-specified criteria and with a third reviewer (ÅM).

Extraction and analysis

We extracted from each included study the following data: author, year of publication, country, healthcare domain, discipline behind the study, image versus tabular data input, study design and setting, clinical or experimental setting, sample size, intervention or exposure of interest, outcome measures, study results, and conclusions. Data were entered into a Microsoft Excel spreadsheet for analysis. RR extracted the data using the pre-established data entry format, with verification by DS to ensure consistency. We disaggregated the analysis by trust dimensions (cognitive versus affect-based) and by type of data evaluated (image versus tabular data). We also assessed each paper for risk of bias, using either the Cochrane Risk of Bias (RoB 2) or Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) tool.

Results

Overview of search results

Our initial search identified 373 publications in PubMed and 713 publications in Web of Science, 308 were duplicates, leaving 778 for the screening and eligibility stages. We excluded 300 records since they were reviews, commentaries, methodological reports, conceptual papers, or not related to the healthcare sector. Eighty-three papers did not study XAI, and 347 were not empirical studies with trust as an outcome and explanations as an intervention. This left 48, all of which were successfully retrieved. We excluded another 38 studies when reviewing the full text as they did not measure trust or XAI empirically or the evaluation was not carried out by practicing clinicians. This yielded 10 articles for final review (Figure 1). [26–35]

The publications were imported into Zotero reference management software. The PRISMA flow diagram of our review is shown in Figure 1.

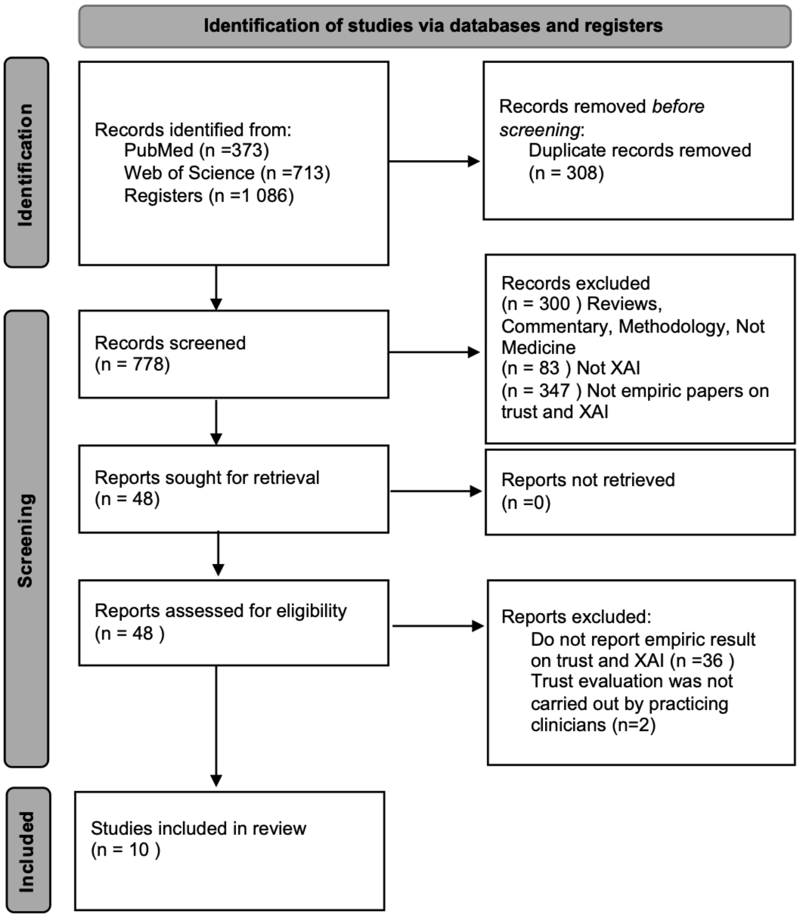

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
PubMed (n =373)
Web of Science (n =713)
Registers (n =1 086)
Records removed *before screening*:
Duplicate records removed
(n = 308)
Records screened
(n = 778)
Records excluded
(n = 300) Reviews, Commentary, Methodology, Not Medicine
(n = 83) Not XAI
(n = 347) Not empiric papers on trust and XAI
Reports sought for retrieval
(n = 48)
Reports not retrieved
(n =0)
Reports assessed for eligibility
(n = 48)
Reports excluded:
Do not report empiric result on trust and XAI (n =36)
Trust evaluation was not carried out by practicing clinicians (n=2)
Studies included in review
(n = 10)

Figure 1: PRISMA flow chart.

Characteristics of included studies

Table 1 provides a summary of the final studies. There was a clear increase in papers on trust and XAI in healthcare during 2022; 70% were published between 2022 and the end of the inclusion period on March 23, 2023.

Title	**Authors (Year)** **Country**	**Study discipline**	**Respondents** **(Sample size)** **Healthcare domain**	**Tabular/ Image**	**Description of intervention**	**Trust measurement**	**Trust improvement**
As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI	Cabitza *et al.* (2020) Italy	Computer Science, Orthopedic and Bio Medicine	Physician (13) Radiology	Image	Measure radiologists' confidence score as a marker for trust	Quantitative confidence score, 6-grade scale.	No effect
Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis	Kumar *et al.* (2021) India	Computer Science	Physicians (10) Brain tumour	Image	Building an explainable deep learning model to reduce complexity in MR classifications.	Quantitative doctor survey using 5-grade Likert Scale.	Increased trust
Does AI explainability affect physicians' intention to use AI?	Liu *et al.* (2022) Taiwan	Medical Research, Cardiology, Pediatrics	Physicians (295)	Image	Comparing intention to use XAI vs AI	Quantitative survey using 5-grade scale.	Increased trust
Explainable recommendation: when design meets trust calibration.	Naiseh *et al.* (2021) UK	Computer Science	Physicians and pharmacists (24) Oncology	Tabular	Involved physicians and pharmacists in think-aloud study and co-design to identify potential trust calibration errors	Qualitative interviews analysed using content analysis.	Varied, depending on factors such as the form of explanation
How the different explanation classes impact trust calibration: The case of clinical decision support systems	Naiseh et *al.* (2023) UK	Computer Science	Physicians and pharmacists (41) Chemotherapy	Tabular	Trust calibration for 4 XAI classes (counterfactuals, example based, global and local explanations) vs no explanations	Quantitative self-reporting cognitive-based trust using 5-grade scale and qualitative interviews was coded.	Varied, depending on factors such as the form of explanation
Interpretable clinical time-series modelling with intelligent feature selection for early	Martinez-Aguero *et al.* (2022) Spain	Computer Science and Intensive Care Department for validation	Clinicians (no specification) Antibiotic resistance	Tabular	SHAP explanations for predictors to provide clinicians with	Qualitative where clinicians self-report	Increased trust

prediction of antimicrobial multidrug resistance					explanations in natural langue		
Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays.	Gaube *et al.* (2023) US/ Canada	Medicine, Psychology and Computer Science	Internal/ emergency medicine physicians and radiologists (223) Radiology	Image	Visible annotation on the X-ray done by human or XAI	Quantitative self-reporting using 7-grade scale	No effect
The explainability paradox: Challenges for XAI in digital pathology	Evans, *et al.* (2022)	Computer Science and Bio Medicine	Board-certified pathologists and professionals in pathology or neuropathology (6+25) Pathology	Image	Saliency maps to explain predictions through visualizations	Quantitative self-reporting using 7-grade scale. Qualitative semi-structured interviews.	Increased trust
Trustworthy AI Explanations as an Interface in Medical Diagnostic Systems	Kaur *et al.* (2022) US	Computer Science	Physicians (2) Breast Cancer Prediction	Image	Involved physicians evaluate 3 different systems and rate the "Trustworthy Explainability Acceptance"	Quantitative, trust calculated using both impression and confidence	Developed framework to measure trust. No effect identified.
UK reporting radiographers' perceptions of AI in radiographic image interpretation Current perspectives and future developments	Rainey *et al.* (2022) UK	Health Science, Radiography, Computer Science	Radiographers (86) Radiography	Image	-	Quantitative self-reporting using 10-grade scale	Increased trust

Table 1: Summary of extraction table.

The studies displayed marked heterogeneity in methods, disciplinary collaboration, and perspectives of trust. All but one involved computer scientists; four were conducted solely by computer scientists without involvement by experts with a medical background, and the remaining five involved collaborations between medical experts and computer scientists. The inputs to the AI tools were medical imaging or tabular data formats. The risk of bias in each study is reported in the Appendix 2. In all studies, the risk of bias was moderate or moderate to high.

We begin by looking at studies of medical imaging and tabular data separately, providing an overview of the characteristics and results before moving on to talk about the different ways in which studies conceptualize or measure trust (as we found that this seemed to be a key consideration in interpreting studies' results).

Medical imaging

Of the seven medical imaging studies reviewed, four (57%) identified a significant and positive association between use of XAI and perceived trust, one study (14%) reached no clear conclusions, while two (29%) found a limited or no significant impact.

One study by Liu *et al.* [35] asked 295 physicians across three hospitals in Taiwan if explanations increased their trust in the algorithm and their propensity to use XAI compared to AI. They found that physicians were more inclined to trust and implement AI in clinical practice if they perceived the results as being more explainable or comprehensible. Similarly, an online experiment by Evans *et al.* [30] surveyed trust levels among board-certified physicians in pathology or neuropathology in using XAI to interpret pathology images. The XAI instrument highlighted the areas in medical images that determined whether the prediction was made with high or low confidence. Seventy percent agreed that their level of trust increased as a result of the explanations provided, while approximately 10% disagreed, and the rest were undecided.

One study by Cabitza *et al.* [26] differentiated Gold Standard labels (categorising cases as positive or negative) from Diamond Standard ones, where the reason for categorisation was annotated and indicated confidence in the allocation. Thirteen radiologists were then asked to evaluate images of knees. Confidence in the allocation was considered a proxy for trust and there was no association between confidence and accuracy. Gaube *et al.* [33] conducted a qualitative investigation of 117 clinical residents or practicing emergency medicine physicians and 106 radiologists. They reported that explanations had little or no significant impact on trust and/or perceived usefulness of AI. The participants were shown x-rays with and without annotations as explanations. Internal and emergency medicine physicians (IM/EM), who lacked specialist training in radiology, achieved better diagnostic accuracy when provided with explanations ($p_{\text{IM/EM}} = 0.042$), but there was no such benefit for radiologists ($p_{\text{Radiology}} = 0.120$). In neither group did annotations have any meaningful effect on confidence in their own final diagnosis ($p_{\text{IM/EM}} = 0.280$, $p_{\text{Radiology}} = 0.202$). The authors did not find convincing evidence for either algorithmic appreciation (a tendency to trust algorithms) or algorithmic aversion (a tendency not to trust algorithms).

Tabular data

The three studies using XAI techniques with tabular data found positive relationships between explanations of AI and perceived trust. However, in two of the studies, results varied, and the authors argued that an inappropriate use of explanations can induce under- or over-trust.

A qualitative study by Martinez-Aguero and colleagues [34] asked whether XAI, when compared with AI, increased trust among clinicians searching for multidrug-resistant bacteria in intensive care units. The authors concluded that both visual and textual explanations helped clinicians understand the model output and increased trust in the XAI. However, neither the number of respondents nor the instrument used to measure trust was clearly reported.

Naiseh and co-workers [28] performed a qualitative study of the influence of XAI on the prescribing decisions of physicians and pharmacists in the field of oncology. For trust they used the terminology used by Lee and colleagues of appropriate reliance [36]. They initially performed semi-structured interviews with 16 participants to understand how these providers engaged with five distinct types of explanations—local, global, counterfactual, example-based, and confidence-based. The authors coded the providers as exhibiting 'high' or 'low' trust only if this behavior was consistent across all five explanation types in the study. Although the physicians and pharmacists were generally favorable towards explanations, they exhibited a lack of trust and skepticism about XAI's accuracy. They further identified two primary causes of errors in trust calibration: (i) skipping explanations or (ii) misapplication of explanations. Skipping occurred when providers made decisions with AI without fully engaging with the accompanying explanations. This was due to: i) disinterest in understanding the explanation, ii) decision delays due to the explanation, iii) perceived redundancy, complexity, or context irrelevance. Misapplication occurred when the providers misunderstood the explanations or simply sought after them to confirm their initial judgement. They then conducted co-design sessions with eight participants. From these, they proposed enhancing XAI interface designs to help avoid skipping or misinterpreting explanations. The designs included active and/or cognitive engagement of decision-makers in the decision-making process, challenge of habitual actions in the XAI system by introducing alternative perspectives or recommendations that may not align with the clinical decision-maker's prior experiences or assumptions, friction that requires the decision-maker to confirm their decision before it is implemented, and support consisting of training and learning opportunities for clinical decision-makers to enhance the understanding and usage of the system.

This same team studied 41 medical practitioners who were frequent users of clinical decision support systems.[29] They sought to develop interventions that would enable physicians to have an optimal level of trust (or reliance), as defined by the authors, in predictions by AI models and to avoid errors that might arise from excessive under- or over-trust. The clinicians used four different XAI classes (global, local, counterfactual, and example-based – their other study had included confidence-based) and the research group explored the clinicians' experiences using semi-structured interviews. In a subsequent mixed-methods study on chemotherapy prescriptions found differences in the trust generated by different explanations. Participants found example-based and counterfactual explanations more understandable than the others but there were no differences in perceptions of technical competence, a view supported in semi-structured interviews, largely because they were easier to comprehend. Additionally, the researchers identified a potential for over-reliance on AI, as providers were more inclined to accept AI recommendations when they were accompanied by explanations, although explanations did not help them identify incorrect recommendations. They made a series of suggestions as to how the interface design might be enhanced although they also noted that it could be very difficult to incorporate the many different types of questions that users might ask. Some might seek very detailed explanations while others could be deterred by the resulting cognitive overload. As the authors note *"long and redundant explanations make participants skip them"*. Perhaps more fundamentally, several of those interviewed said that they would be reluctant to use this tool because of the high cognitive load involved in seeking to understand some decisions.

Conceptualizing and Measuring Trust

The studies reviewed take two broad approaches to defining trust: cognition-based trust and affect-based trust [19]. The initial approach, cognition-based trust, revolves around the perceived clarity and technical ability of XAI, fundamentally grounded in rational analysis. On the other hand, affect-based trust encompasses emotional bonds and belief, originating from previous experiences and sentiments towards AI, as opposed to logical deliberation. All ten studies applied cognitive-based trust. However, two studies also investigated trust in terms of affect or emotions.

Eight of the studies employed quantitative surveys to measure trust, integrating them with qualitative interviews in two instances. The remaining two exclusively utilized qualitative interviews. We found marked heterogeneity in the questions used.

Naiseh and colleagues noted that explanations affected both cognitive and affect-based trust and could result in either over-trust or under-trust. In the 2021 study [28], they used qualitative think-aloud methods and suggested that one reason for users skipping or misapplying explanations could be that affect-based trust overrides cognitive and deliberate trust. Two years later, they published a new study[29], in which they investigated whether different XAI classes or methods increased or decreased cognitive-based trust. They found that some types of explanation could introduce a cognitive over-reliance on the AI, but they questioned whether biases and affect-based trust also played roles.

Discussion

Principal Results

We examined empirical evidence on the impact of explainable AI on physicians' trust levels and intention to use AI. Of the 10 studies included, 50% reported that XAI increased trust while 20% observed both increased and decreased trust levels. Both over-trust and under-trust appeared to be modifiable by brief cognitive interventions to optimise trust. [28,29] In two studies (20%), no effects of XAI were shown, and one study (10%) did not reach any conclusions. Only small differences of no consequence were identified between studies using tabular data formats and image data.

Before interpreting these findings further, we must note several important limitations of our study's search strategy. First, there is considerable heterogeneity in the use of the term 'trust' and how it is operationalised in healthcare research. To avoid potentially missing important studies in our search, we adopted a conservative search strategy in which we did not specify trust as a keyword but rather manually searched for all papers including a broad set of trust-related outcomes. Related to this, the rapid evolution of AI has been associated with conceptual confusion about its meaning. Several recent studies have sought to operationalise AI in markedly varying ways drawing on technology, for example, which is not actually based on AI-algorithms[37,38]. For clarity, we specifically constrained our search to AI algorithms which used machine-learning techniques. Second, we use two main databases of peer reviewed studies, PubMed and Web of Science. The former broad coverage in medicine and social sciences, but could potentially missed emerging studies in computer science, but Clarivate, who publish Web of Science, note that it has "Strongest coverage of natural sciences & engineering, computer science, materials sciences, patents, data sets". [39] We do, however, accept that, in a rapidly developing field, we may have missed material in preprints or non-peer reviewed conference papers. Additionally for coherence across platforms we did not employ MeSH terms in PubMed, as they are not used in Web of Science and we wanted to achieve consistency. The keyword, ‘clinical’, also may potentially have excluded studies in some clinical specialities. However, the vast number of potential specialist terms that could be used make it virtually impossible to implement a wider strategy in practice. Finally, there has been extensive study of psychological biases in how decision-makers, including clinicians, respond to new data and update prior beliefs in incorporating evidence to make decisions.[17,40] Studies by psychologists are needed to evaluate the role these biases (including but not limited to default bias and confirmation bias) play in medical decision-making when using XAI.

There were also a series of limitations identified in the included studies themselves. Generally, the study designs widely varied, from qualitative investigations to experimental quantitative studies, making it difficult to draw direct comparisons. However, we have sought to the extent possible to identify emerging themes and patterns across tabular and visual XAI applications, as well as a series of methodological limitations to address in future studies. Additionally, the relatively low number of studies (n= 10) limits generalisability to other populations and settings. Another limitation present in several studies was weak reporting of trust measurement instruments, as well as the numbers of respondents, particularly in qualitative studies. Few studies reported the validity of the underlying XAI algorithm which could also alter the healthcare providers’ engagement and trust of XAI technologies. Future research should seek to improve reporting of this necessary information.

Although our review focused on how XAI impacted clinicians' trust levels and intention to use this technology, a few additional observations are of interest. Gaube and co-workers [33] found no difference in trust between experts and non-experts but reported that the performance of non-experts who drew upon XAI was superior in clinical practice. Future studies are needed not just to evaluate the impact of XAI on its adoption and trustworthiness but also its potential clinical efficacy. In this context, it is worth noticing that while all included studies offered explanations that could be added to AI predictions, the validity of those explanations has yet to be critically evaluated. [41] It is unclear how XAI can overcome limitations inherent in clinical domains where mechanistic understanding is lacking. That is, XAI will likely struggle to explain what is currently unexplainable at the frontier of clinical medicine. This could potentially lead to explanations which, albeit perceived as trustworthy, are not founded on established clinical knowledge and instead are ‘misconceptions’ by AI. The XAI explanations are still simplifications of the original AI model, and when the abstraction level is heightened, the granularity is usually reduced.

This review also points to the need to understand how trust in XAI can be optimised, rather than simply being evaluated in terms of increased or decreased with the help of different types of explanations. Clinical decision-making inevitably involves an element of judgment. While AI may be able to process more information than a human, humans may also be able to incorporate insights that are not included in algorithms. [41] Thus, the challenge is to achieve an appropriate level of trust in AI, neither too limited, in which case the clinician will be reluctant to use it, nor too extensive, as this may cause experienced clinicians to subordinate their own judgment to the AI outputs.

Yet, while it is apparent that neither blind trust nor blind distrust may be appropriate, it is unclear what an appropriate or optimal level of trust should be. None of the studies attempted to explore what this should be, which remains an important area for future research. However, the studies reviewed indicated that the levels of trust that healthcare providers place in AI depend on multiple clinically-relevant factors, including but not limited to the accuracy of the algorithm, the validation, and the potential impact on patients.

Our study also points to several further directions for future research. First, while the interdisciplinary literature featured prominent computer scientists and clinicians, there was a notable absence of psychologists. There is considerable scope to improve the appropriate uptake and adoption of AI by drawing upon evidence from the wider psychological literature on medical decision-making. One such framework is a dual process model, which integrates both cognitive and affect-based means of decision-making jointly. Kahneman argue that the human mind uses two processes for decision-making: the fast thinking and intuitive process, including heuristics, biases, and cognitive shortcuts that recalls affect-based trust, and the slow thinking and reasoning process that recalls cognitive-based trust. [17] Furthermore, Thaler and colleagues have found that both these processes can be influenced (or nudged), especially the rapid thinking intuitive judgments. Brief cognitive interventions like nudging have sometimes proven to be useful in health. [42] The extant literature appears to incorporate mainly reasoning-based cognitive markers but misses out on intuitive and emotional-based processes for evaluating trust levels in emerging technologies.

Conclusions

A majority of the included studies showed that XAI increases clinicians' trust and intention to use AI; two of these studies showed that explanations could both increase and decrease trust and in three studies, explanations fell through or did not add any value. However, in healthcare, when AI tool incorporates associated explanations, they must avoid two common psychological pitfalls. First, they must be made sufficiently clear to avoid risks of blind distrust when physicians do not understand them. Second, they must avoid oversimplification and failing to disclose limitations in models that could lead to blind trust among physicians with an artificial level of clinical certainty. Explanations can both increase and decrease trust, and understanding the optimal level of trust in relation to the algorithm's accuracy will be critical. When AI algorithms surpass physicians in terms of accuracy, the integration could be facilitated through means such as providing explanations. Yet, the provision of explanations is not a failsafe method to detect errors in the algorithms, as it might inadvertently foster excessive trust. How to find an optimal level of trust and how to best to communicate AI to physicians will remain a defining healthcare challenge of our time.

Contribution

RR contributed with the idea, collaborated with DS in data collection, performed the review and drafted the manuscript. All authors contributed to the interpretation, writing and editing of the manuscript.

Conflicts of Interest

Competing interests: None declared.

Funding: The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Abbreviations

AI: Artificial intelligence

XAI: Explainable artificial intelligence

References

1. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet Lond Engl*. 2020;395(10236):1579-1586. doi:10.1016/S0140-6736(20)30226-9. PMID: 32416782

2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0. PMID: 35058619

3. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *Npj Digit Med*. 2020;3(1):1-5. doi:10.1038/s41746-020-0254-2. PMID: 32258429

4. Ienca M. Don't pause giant AI for the wrong reasons. *Nat Mach Intell*. Published online 2023:1-2. https://doi.org/10.1038/s42256-023-00649-x

5. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Acad Manag Ann*. 2020;14(2):627-660. doi.org/10.5465/annals.2018.0057

6. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012

7. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2016:1135-1144. https://doi-org.esc-web.lib.cbs.dk/10.1145/2939672.2939778

8. Wu W, Su Y, Chen X, et al. Towards global explanations of convolutional neural networks with concept attribution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ; 2020:8652-8661.

9. Zhang Y, Liao QV, Bellamy RK. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ; 2020:295-305. https://doi-org.esc-web.lib.cbs.dk/10.1145/3351095.3372852

10. Liao QV, Gruen D, Miller S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery; 2020:1-15. doi:10.1145/3313831.3376590

11. Mechanic D. The functions and limitations of trust in the provision of medical care. *J Health Polit Policy Law*. 1998;23(4):661-686. PMID: 9718518 DOI: 10.1215/03616878-23-4-661

12. Fukuyama F. *Trust: The Social Virtues and the Creation of Prosperity*. Simon and Schuster; 1996. ISBN 0684825252

13. Seligman AB. *The Problem of Trust*. Princeton University Press; 2000. ISBN 6-691-05050-1

14. Arrow KJ. Uncertainty and the welfare economics of medical care. In: *Uncertainty in Economics*. Elsevier; 1978:345-375. https://doi.org/10.1016/B978-0-12-214850-7.50028-0

15. Berg J, Dickhaut J, McCabe K. Trust, reciprocity, and social history. *Games Econ Behav*. 1995;10(1):122-142. https://doi.org/10.1006/game.1995.1027

16. McKee M, Greenley R, Permanand G. Trust: The foundation of health systems. Published online December 12, 2023. https://eurohealthobservatory.who.int/publications/i/trust-the-foundation-of-health-systems.

17. Kahneman D. *Thinking, Fast and Slow*. macmillan; 2011. ISBN: 9780141033570

18. Lewicki RJ, Brinsfield C. Framing trust: trust as a heuristic. *Fram Matters Perspect Negot Res Pract Commun*. Published online 2011:110-135.

19. Madsen M, Gregor S. Measuring human-computer trust. In: *11th Australasian Conference on Information Systems*. Vol 53. Citeseer; 2000:6-8.

20. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*. Published online 2023. PMID: 37234618

21. Nazar M, Alam M, Yafi E, Su'ud M. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE ACCESS*. 2021;9:153316-153348. doi:10.1109/ACCESS.2021.3127881

22. Antoniadi A, Du Y, Guendouz Y, et al. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl Sci-BASEL*. 2021;11(11). doi:10.3390/app11115088

23. Giuste F, Shi W, Zhu Y, et al. Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. *IEEE Rev Biomed Eng*. 2022;PP. doi:10.1109/RBME.2022.3185953

24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int J Surg*. 2021;88:105906. doi:10.1016/j.ijsu.2021.105906

25. Trust: The foundation of health systems. Accessed March 11, 2024. https://eurohealthobservatory.who.int/publications/i/trust-the-foundation-of-health-systems

26. Cabitza F, Campagner A, Sconfienza LM. As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Med Inform Decis Mak*. 2020;20(1):219. doi:10.1186/s12911-020-01224-9. PMID: 32917183

27. Kumar A, Manikandan R, Kose U, Gupta D, Satapathy S. Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain

Tumor Diagnosis. *ACM Trans Multimed Comput Commun Appl*. 2021;17(3). doi:10.1145/3457187

28. Naiseh M, Al-Thani D, Jiang N, Ali R. Explainable recommendation: when design meets trust calibration. *World Wide Web*. 2021;24(5):1857-1884. doi:10.1007/s11280-021-00916-0

29. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int J Hum-Comput Stud*. 2023;169. doi:10.1016/j.ijhcs.2022.102941

30. Evans T, Retzlaff C, Geissler C, et al. The explainability paradox: Challenges for xAI in digital pathology. *FUTURE Gener Comput Syst- Int J ESCIENCE*. 2022;133:281-296. doi:10.1016/j.future.2022.03.009

31. Kaur D, Uslu S, Durresi A. Trustworthy AI Explanations as an Interface in Medical Diagnostic Systems. In: Barolli L, Miwa H, Enokido T, eds. *Indiana University System*. Vol 526. ; 2022:119-130. doi:10.1007/978-3-031-14314-4_12

32. Rainey C, O'Regan T, Matthew J, et al. UK reporting radiographers' perceptions of AI in radiographic image interpretation - Current perspectives and future developments. *Radiogr Lond Engl 1995*. 2022;28(4):881-888. doi:10.1016/j.radi.2022.06.006

33. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep*. 2023;13(1):1383. doi:10.1038/s41598-023-28633-w

34. Martinez-Aguero S, Soguero-Ruiz C, Alonso-Moral J, Mora-Jimenez I, Alvarez-Rodriguez J, Marques A. Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. *FUTURE Gener Comput Syst- Int J ESCIENCE*. 2022;133:68-83. doi:10.1016/j.future.2022.02.021

35. Liu CF, Chen ZC, Kuo SC, Lin TC. Does AI explainability affect physicians' intention to use AI? *Int J Med Inf*. 2022;168:104884. doi:10.1016/j.ijmedinf.2022.104884

36. Chiou EK, Lee JD. Trusting automation: Designing for responsivity and resilience. *Hum Factors*. 2023;65(1):137-165. PMID: 33906505 DOI: 10.1177/00187208211009995

37. Broussard M. *Artificial Unintelligence: How Computers Misunderstand the World*. The MIT Press; 2018. doi:10.7551/mitpress/11022.001.0001

38. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4). doi:10.1136/svn-2017-000101

39. Matthews T. LibGuides: Resources for Librarians: Web of Science Coverage Details. Accessed September 25, 2023. https://clarivate.libguides.com/librarianresources/coverage

40. Kliegr T, Bahník Š, Fürnkranz J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif Intell*. 2021;295:103458. doi:10.1016/j.artint.2021.103458

41. Sanchez-Martinez S, Camara O, Piella G, et al. Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging. *Front Cardiovasc Med*.

2022;8. PMID: 35059445 PMCID: PMC8764455 Accessed July 4, 2023. https://www.frontiersin.org/articles/10.3389/fcvm.2021.765693

42. Thaler RH, Sunstein CR. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin; 2009. ISBN 10: 014311526X ISBN 13: 9780143115267

Appendix 1. Search Strategy

PubMed:

(XAI[Title/Abstract] OR "explainable artificial intelligence"[Title/Abstract] OR "explainable AI"[Title/Abstract]) AND (Healthcare[Title/Abstract] OR medical*[Title/Abstract] OR clinical*[Title/Abstract])

Web of Science:

#1 XAI OR "explainable artificial intelligence" OR "explainable AI" (Topic)

#2 healthcare OR medical* OR clinical* (Topic)

#1 AND #2

Appendix 2. Assessment of risk of bias

Title	**Authors (Year)** **Country**	Tool	Risk	Reasons
As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI	Cabitza *et al.* (2020) Italy	Cochrane Risk of Bias (RoB 2)	Moderate	Potential biases in labeling due to human judgment variability, potential deviations in rater performance, and how these issues are managed in the study's methodology.
Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis	Kumar *et al.* (2021) India	Cochrane Risk of Bias (RoB 2)	Moderate to high	Lack of representativeness, over-reliance on technical outcomes, and insufficient real-world validation of the model's performance and explainability.
Does AI explainability affect physicians' intention to use AI?	Liu *et al.* (2022) Taiwan	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate	Potential confounding factors, the use of convenience sampling, and the subjective nature of the self-reported outcomes.
Explainable recommendation: when design meets trust calibration.	Naiseh *et al.* (2021) UK	Cochrane Risk of Bias (RoB 2)	Moderate to high	Qualitative and non-randomized design, potential deviations due to participants' familiarity with AI, and subjective nature of the data collection and reporting processes
How the different explanation classes impact trust calibration: The case of clinical decision support systems	Naiseh e*t al.* (2023) UK	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate	While the study uses validated tools and consistent application of interventions, limitations such as lack of participant randomization, potential order effects, and reliance on self-reported measures could affect the robustness of the findings.
Interpretable clinical time-series modelling with intelligent feature selection for early prediction of antimicrobial multidrug resistance	Martinez-Aguero *et al.* (2022) Spain	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate to high	Potential confounding, selection bias, handling of missing data, and reliance on EHR data quality.

Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays.	Gaube *et al.* (2023) US/ Canada	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate	Potential confounding factors, selection bias due to the recruitment strategy, and the use of self-reported measures that could affect validity.
The explainability paradox: Challenges for XAI in digital pathology	Evans, *et al.* (2022)	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate	Risk of selection bias, potential confounding due to uncontrolled participant variability, and measurement bias from self-reported data.
Trustworthy AI Explanations as an Interface in Medical Diagnostic Systems	Kaur *et al.* (2022) US	Cochrane Risk of Bias (RoB 2)	Moderate to high	Reliance on simulated expert profiles, the absence of detailed handling of missing data, and lack of a real-world clinical validation component.
UK reporting radiographers' perceptions of AI in radiographic image interpretation Current perspectives and future developments	Rainey *et al.* (2022) UK	Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	Moderate to high	Risk of selection bias from convenience sampling, potential confounding factors that were not controlled, and reliance on self-reported data.

Appendix II: Paper II

Citation: Rosenbacke, R. (2024) 'Errors in Physician-AI Collaboration: Insights From a Mixed-methods Study of Explainable AI and Trust in Clinical Decision-making', SSRN Electronic Journal. doi: 10.2139/SSRN.4773350.

Errors in Physician-AI Collaboration

Insights from a mixed-methods study of explainable ai and trust in clinical decision-making

Rikard Rosenbacke*

Centre for Corporate Governance, Copenhagen Business School, Copenhagen, Denmark

* - corresponding author; rikard@rosenbacke.com, rr.ccg@cbs.dk, Rikard Rosenbacke, Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg

Word Count: 11,895

Key words: Artificial Intelligence; Explainable AI; XAI; Decision-making; Clinical decision support systems; Healthcare; Dual process theory

Abstract

Artificial Intelligence-based diagnostic systems are increasingly prominent in healthcare systems, although they face multiple cognitive challenges to their acceptance and effective use among healthcare professionals. One major issue is low trust among doctors in AI advice, especially when this advice appears “black box” lacking clear diagnostic explanations. However, another challenge is that doctors may blindly trust and accept incorrect AI diagnostics. Here we investigated doctors’ trust and decision-making errors in collaborating with AI and explainable AI in a field study of 11 physicians making 330 diagnostic decisions on recurrent ear infections. We calibrated the AI at 60% accuracy so to better differentiate trust and errors that emerge when AI is correct or incorrect, and either confirms or conflicts with doctors’ diagnoses. To deepen understanding of cognitive mechanisms, we performed “think-aloud” protocols during qualitative interviews where doctors describe their reasoning when using or discarding AI diagnostic advice. Turning first to accuracy, we found that physician-AI collaboration outperformed physician decision-making alone. However, accuracy was substantially reduced in scenarios when doctors lacked confidence in their decisions and shifted from a correct diagnosis to an incorrect AI diagnosis. In terms of uptake, physicians exhibited “stickiness” in their diagnostic decisions in about two-thirds of all cases, consistent with AI distrust and a potential commitment bias. Adding explanations with XAI did persuade more physicians to use it, but nonetheless about half of the doctors remained unchanged with the aide of XAI. Virtually none of the physicians altered their decisions when AI confirmed their incorrect diagnosis (a “false confirmation”), which accounted for two-thirds of all errors identified in our study. Our qualitative analysis showed that physicians neglected the possibility of AI error in cases of confirmation. We conclude with proposing an agenda for future research that could tap the power of cognitive psychology and explainable AI to improve physician-AI collaboration and balanced trust in AI applications in healthcare settings.

Introduction

AI technologies are rapidly emerging as indispensable instruments for decision-making in modern healthcare (Schwalbe and Wahl, 2020; Rajpurkar *et al.*, 2022). AI in clinical decision support systems has the capacity to improve clinical diagnostic assessments, reduce medical errors, and improve overall patient outcomes (Sutton *et al.*, 2020). Yet, for the healthcare field to realise the full potential of AI, it is necessary for clinicians to trust, and employ effectively (Cutillo *et al.*, 2020). This trust is further challenged, from widespread popularized concerns about generative AI, (Ienca, 2023), but also from a deeply rooted culture of evidence-based practice in healthcare (Amann *et al.*, 2020).

There is a common perception among doctors that AI operates as a "black box", without providing clear justification for its health-related advice (Fazal *et al.*, 2018; Wadden, 2021). Unlike rule-based systems, AI platforms are less transparent, making their errors harder to anticipate (Jussupow *et al.*, 2021). Commonly employed AI algorithms in healthcare draw upon intricate statistical frameworks, such as deep neural networks, that are inherently difficult for humans to interpret (Castelvecchi, 2016). When healthcare providers do not understand clinical advice, they are much less likely to use it (Cui and Zhang, 2021).

Recently, explainable AI (XAI) has been developed to overcome these limitations, increasing its uptake in diverse management domains as well as healthcare. An emerging body of research claims that XAI is essential for securing the safety, approval, and adoption of AI systems in clinical settings (Antoniadi *et al.*, 2021; Evans *et al.*, 2022; Reddy, 2022; Haque, Islam and Mikalef, 2023; Chanda *et al.*, 2024). The central goal of explainable AI is building trust and doing so through greater transparency (Gerlings, Shollo and Constantiou, 2021). For example, the US defense XAI program DARPA underscores the necessity of XAI for understanding, trusting, and effectively managing the next wave of AI technologies (Gunning and Aha, 2019).

Although it is highly plausible, XAI would improve trust, a series of recent systematic reviews reveal that, apart from assumptions that it will work, there is disappointingly little evidence it actually does so. One study highlighted the importance of trust assessment when developing XAI in healthcare (Jung *et al.*, 2023); another study focused on explanations to the end-using clinician to create a trustworthy environment (Nazar *et al.*, 2021); another review speculated that XAI could enhance decision confidence and trust for clinicians (Antoniadi *et al.*, 2021), while another argued that XAI could instill trust in the users, and assist clinicians in decision-making (Giuste *et al.*, 2023).

To address this lack of evidence on how XAI shapes trust, we performed a field study that directly compares how clinicians engage with AI and XAI to make complex diagnoses. We combine quantitative and qualitative methods to identify the main cognitive challenges physicians face when incorporating AI and XAI advice, and especially those which correlate with decision-making errors. Thus specifically we sought to answer: *How does diagnostic AI* *and XAI advice influence physicians' trust and decision-making processes, particularly when AI-generated advice either conflicts or confirms the physicians own diagnoses?*

Briefly, we found that explainability of XAI increases trust and usage intentions, significantly over and above AI alone. This led to a significant rise in diagnostic accuracy. Nevertheless, it did not overcome the oft-seen “stubbornness” among physicians. We found a substantial portion of doctors adhered to their initial diagnoses when contradicted by AI/XAI. Additionally, we were able to detect a more subtle error in which XAI confirms an erroneous physician diagnosis, a “false confirmation.”

The rest of this paper is as follows. First, we outline streams of literature on AI in healthcare. Furthermore, we draw on cognitive psychology theories of decision-making in relation to AI and XAI in healthcare. Next, we identify decision errors and cognitive challenges in human-AI collaboration, which we use to develop our study’s hypotheses. We then describe the mixed methods used in our field study, followed by presenting the results. We conclude by reinterpreting our data in light of contemporary theories of decision-making, with implications for how to better promote effective physician-AI trust and collaboration.

AI’s uptake in healthcare settings

Research on artificial intelligence healthcare is progressing swiftly. A recent meta-analysis performed an umbrella review of 220 systematic literature reviews on machine learning AI applications in healthcare in the past decade, which collectively analyzed over seven thousand original research studies. The predominant applications were identified as clinical prediction and disease prognosis using imaging data in oncology and neurology (Kolasa *et al.*, 2023). The authors found generally poor and variable reporting quality of research focusing on the development and adaptation of ML algorithms for clinical use. Furthermore, only two-thirds of these studies included evaluations of system accuracy. Despite the proliferation of published AI algorithms, their actual influence on clinical practice appears to be low.

The reasons for slow uptake are multiple. In general, it seems that many clinicians are hesitant to adopt and integrate these technologies into their routine practices and, when they do so, may not tap their full benefit for patients. A recent scoping review examining the studies on attitudes of healthcare providers towards AI (Chew and Achananuparp, 2022). It found providers were generally positive, as a result of AI’s perceived accessibility, user-friendliness, and the potential to enhance service efficiency and reduce costs in the healthcare sector. Curiously, however, the providers expressed reluctance to actually use it, due to concerns about trustworthiness, data confidentiality, patient safety, the current state of technology, and the prospect of complete automation so making doctors redundant.

One frequently cited barrier is AI’s "black-box" nature, or the opaqueness with which AI systems make predictions. Clinical practice and culture centre around understanding mechanisms and explanations of disease processes. Clear, transparent decision-making is a cornerstone of clinical medicine, especially based on the emphasis on evidence-based practices (Amann *et al.*, 2020; Kundu, 2021). Thus, any perceived lack of clarity in, or difficulty comprehending, AI predictions could deter their use in clinical practice (Cui and Zhang, 2021; Loh *et al.*, 2022).

Many argue that a newer generation of XAI tools, which provide understandable explanations alongside AI-generated diagnoses, will enhance trust and resulting uptake among clinicians (Antoniadi *et al.*, 2021; Evans *et al.*, 2022; Reddy, 2022; Haque, Islam and Mikalef, 2023). Yet it is unclear whether or not these additional explanations actually better persuade doctors to use them, and if doing so translate into better diagnoses and ultimately patient outcomes.

Next, we focus on the cognitive challenges that physicians meet when the AI/XAI advice is both correct and incorrect and the implications on physicians' trust and intention to use it.

Cognitive challenges in physician-AI collaboration: the role of dual process theory

Much research has found that decision-makers more generally are more likely to trust and incorporate advice from humans than AI algorithms. This phenomenon, perhaps most extensively studied in information systems research, has been variously defined as algorithm aversion (Dietvorst, Simmons and Massey, 2015) (when rejecting advice), or algorithm appreciation (Logg Jennifer, 2018) (when incorporating it). A recent systematic literature review investigated 80 empirical studies on algorithm aversion and found that, in general, "*People tend to rely less on algorithms even when algorithms provide better decisions*" (Mahmud *et al.*, 2022, p. 17). However, these studies of algorithmic aversion have tended to be conducted in artificial laboratory settings, often with students or crowd-sourced workers (like Mechanical Turk), which may not reflect the actual performance of AI systems in real-world settings. Mahmud and colleagues call for more qualitative studies with practitioners, noting that "*scholars should undertake more qualitative research on this area [algorithm aversion], involving practitioners*" (Mahmud *et al.*, 2022, p. 15).

In line with previous studies (Buçinca, Malaya and Gajos, 2021; Jussupow *et al.*, 2021; Naiseh, Cemiloglu, *et al.*, 2021; Bertrand *et al.*, 2022), to better understand the potential cognitive challenges in physician-AI collaboration, we must first revisit cognitive psychological theories of decision-making. One widely accepted framework is dual-process theory. It argues that humans (doctors included) mainly use so-called "System 1" thinking, an intuitive, rapid, and automatic mode of thinking that operates effortlessly without conscious control (Tversky and Kahneman, 1974), as opposed to the more analytical System 2 thinking, which is slower and requires more mental effort. For the vast majority of decisions, System 1 is adequate. However, System 1 relies on mental shortcuts to swiftly make judgments and decisions, hence opening the door to cognitive biases, which can lead to potentially flawed decisions (Tversky and Kahneman, 1974). A systematic review of how cognitive biases affect XAI-assisted decision-making argues that heuristics like AI algorithm aversion and appreciation are trust-related heuristics that arise from System 1 (Bertrand *et al.*, 2022).

Trust in AI encompasses more than just its technical attributes, such as reliability and accuracy (related to System 1); it extends to encompass a range of human cognitive, motivational, heuristics, and behavioral factors (related to System 2) (Liu *et al.*, 2022). From a cognitive psychology perspective, System 1 thinking can be understood as a heuristic approach to trust, while the behavioral tradition views trust as System 2 rational-choice behavior (Lewicki and Brinsfield, 2011). In human-computer trust interactions, these two fundamental trust types are commonly referred to as affect-based (System 1) and cognition-based trust (System 2) (Madsen and Gregor, 2000). When provided an AI recommendations, clinicians inherently face a dilemma in deciding to accept or reject them intuitively (System 1) or to engage in an effortful and time-consuming cognitive analysis (System 2). Hence, decision-makers often develop their own personal heuristics for when to trust and follow AI advice, and explanations can reinforce these heuristics (Buçinca, Malaya and Gajos, 2021).

Psychologists are divided as to whether System 1 heuristic errors can be rectified. One prevailing view is that they cannot; at best, decision-makers can recognise common decision-making pitfalls and try to circumnavigate them. As the architect of dual process theory, Kahneman explains, "*It' s false to hope that if you become more aware of your errors you will make better decisions*" (Matias, 2017). In contrast, Klein's naturalistic decision-making has a more positive notion. Naturalistic decision-making does not isolate heuristic and systematic processes but instead examines the broader cognitive processes that enable decision-makers to oversee and regulate their reasoning (System 2) in relation to their intuition (System 1) (Klein, 2015).

Several strategies have been proposed to shift people towards analytical thinking (System 2), so as to minimise the influence of heuristics and cognitive biases (System 1) on decision-making. The two main strategies involve educational approaches and cognitive forcing functions (Croskerry, 2003). Education aims to improve future decision-making through awareness and training (where AI explanations could be part of the learning). However, evidence shows that diagnostic errors often result from biases, not actually an underlying lack of knowledge (Graber *et al.*, 2012). Alternatively, cognitive forcing functions occur in real time of the decision-making to encourage analytical thinking. Studies indicate that these real-time interventions are more effective than education in enhancing diagnostic accuracy, and these interventions have been tested with XAI (Sherbino *et al.*, 2014; Lambe *et al.*, 2016; Buçinca, Malaya and Gajos, 2021; Naiseh *et al.*, 2023). Cognitive forcing functions can take the form of checklists, moments to pause and reconsider diagnoses, or requests for the individual to consciously eliminate an alternative option. Two cognitive forcing interventions that have demonstrated effectiveness are: (i) Making a decision first – studies reveal that people are better decision-makers when they form their own opinions before seeing AI recommendations, avoiding anchoring bias (Green and Chen, 2019); and (ii) delaying AI recommendations – research in human-computer interaction shows that merely postponing the display of AI suggestions can enhance decision quality (Park *et al.*, 2019).

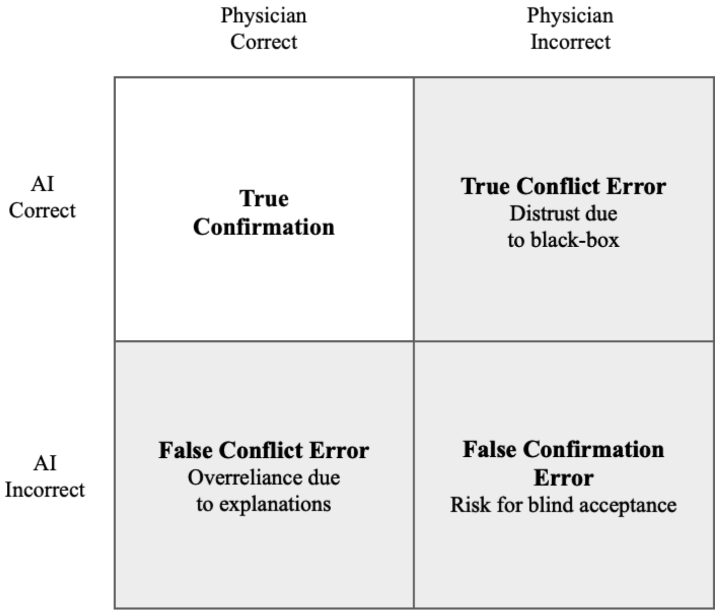
Thus, to mitigate the likelihood of diagnostic errors arising from various heuristics and biases associated with intuitive (System 1) thinking and to have the physicians engage in analytical (System 2) thinking, we drawn upon both educational interventions (in the form of explanations of the AI's risk factors and weights), and cognitive forcing function interventions (such as "making a decision first" and "delaying AI recommendations"). As we describe further below in developing our study's hypotheses, we place a special emphasis on two potentially present sources of decision-making bias which have been found in healthcare: i) commitment bias (Dolan et al., 2012), which could occur here when the AI/XAI contradicts the physician and ii) confirmation bias (Nickerson, 1998), which could take place when the AI and its explanation falsely confirm an initial incorrect clinical judgment.

Problematization and hypotheses

Prior studies have tended to focus on whether and to what extent physicians adjust their decision-making when AI models are correct or perform significantly better than clinicians. For example, scholars noted that *"most prior work has assumed that provided system advice is correct and beneficial. In doing so, it has largely neglected the cognitive challenges entailed in incorrect system advice"*(Jussupow *et al.*, 2021).

Clinicians face three cognitive challenges when drawing upon AI support in decision-making (Jussupow *et al.*, 2021). These arise when AI and clinical judgment come into conflict (Jacobs et al., 2021; Naiseh et al., 2023). However, a third, hidden, challenge emerges, when AI and clinical judgment coheres, but are both incorrect. We discuss each below, as depicted in Figure 1.

Scholars have identified a critical need for research into the implications of XAI and its explanations on the cognitive challenges faced by physicians, as well as the potential for errors in clinical decision-making. Jussupow and colleagues studied physicians' metacognitive challenges when aided by AI and called for research on whether explainable AI may lessen these challenges (Jussupow *et al.*, 2021). Evans and colleagues call for "*empirical studies of user interaction with explainability elements embedded into more true-to-life workflow would provide further valuable insights*." (Evans *et al.*, 2022). Furthermore, Naiseh and colleagues call for "*future work to explore XAI design modalities and principles to mitigate potential over-reliance risk when explanations are provided*" (Naiseh *et al.*, 2023). In response to these calls for further research, we investigate the cognitive challenges introduced by XAI in a clinical set up.



	Physician Correct	Physician Incorrect
AI Correct	**True Confirmation**	**True Conflict Error** Distrust due to black-box
AI Incorrect	**False Conflict Error** Overreliance due to explanations	**False Confirmation Error** Risk for blind acceptance

Figure 1: Potential errors in Human-AI collaboration and decision-making

True Conflict: Convincing the physician when AI is correct

When AI makes a diagnosis or judgment that differs from that made by the physician, there is evidence that clinicians are quite unlikely to adjust their decision-making (Petersson *et al.*, 2022). This "stubbornness", in light of new evidence, is not only seen with AI, but also in reluctance to adopt new technologies and diagnostic tools into their practice. A series of qualitative studies have suggested that physicians, like humans in general, are "resistant to change" and "creatures of habit" (Gupta, Boland and Aron, 2017).

We derive our first hypothesis:

H1: Physicians tend to stick to their initial diagnoses — and do not change their clinical diagnoses when AI contradicts them.

Explanations that accompany AI advice have been suggested as an intervention to enable "learning" among physicians so that they can better understand when they have made an incorrect clinical judgment and that this learning process could increase both trust and diagnostic accuracy. A limited number of studies have found empirical evidence that explanations can increase clinicians' trust (Kumar *et al.*, 2021; Liu *et al.*, 2022; Martínez-Agüero *et al.*, 2022). A recent study showed that non-task experts benefitted considerably more than task-experts from correct explainable AI advice. However, this research argues that further investigation is required to examine the impact of explanations on physicians' dependence on advice, particularly when that advice is inaccurate (Gaube et al., 2023). This situation presents a second conflict.

False Conflict: Over-trust problems when AI is incorrect

A second conflict emerges when physicians are correct, but AI makes an incorrect judgment. Using and trusting a correct algorithm is intuitively a correct judgment; however, algorithms can err, and in this conflict scenario, high trust by physicians can potentially be counterproductive.

Although explanations are likely to increase clinicians' trust, they could worsen decision-making if doctors dispense with accurate judgements for inaccurate ones. Studies found that explanations not only increase trust but can potentially do it to an extent where optimal trust turns into overreliance or even blind trust (Naiseh, Al-Mansoori, *et al.*, 2021; Naiseh, Al-Thani, *et al.*, 2021; Naiseh *et al.*, 2023). Naiseh and colleagues argue that trust should be optimized since overreliance can lead to another error when clinicians change their correct assessment based on incorrect AI advice.

At present, current trends in research are seeking to enable physicians to better calibrate their trust so as to optimize across true and false conflicts through cognitive interventions in addition to explanations. One is cognitive forcing, which aims to disrupt heuristic reasoning, prompting analytical thinking (Lambe *et al.*, 2016). In brief (as described above), such interventions "force" physicians to think hard. They have been tested with XAI, examples include checklists, diagnostic time-outs, or asking the decision-maker to make a clinical assessment before seeing the AI diagnosis and its explanations (Buçinca, Malaya and Gajos, 2021; Naiseh *et al.*, 2023).

Thus we hypothesise that in cases of true conflict, XAI could be beneficial by promoting trust and uptake, but have a dark side, persuading physicians to an incorrect answer, in cases of false conflict:

H2: Physicians are more likely to change their clinical diagnoses when contradicted by XAI explanations than when contradicted by AI explanations alone.

Curiously little research at present, however, covers a third, hidden cognitive challenge, which we term "False Confirmation": when both physician and AI err.

False Confirmation - challenges when both physician and AI are incorrect

When conflict occurs between clinical diagnosis and AI, it may seem natural for physicians to probe the underlying reasons for this divergence. However, a more subtle error may arise when the AI falsely confirms an incorrect clinical judgment.

Over-trust in an explainable AI system can emerge from confirmation bias, a psychological tendency where humans are more likely to trust an AI system that consistently produces outputs aligning with their pre-existing beliefs or initial hypotheses (Naiseh et al., 2023), and a reluctance to seek disconfirmatory evidence. This over-reliance on XAI can pose significant risks, particularly when the system's outputs are erroneous but reaffirm the user's prior convictions (Naiseh et al., 2023).

It may seem intuitive to blindly accept AI or XAI advice when it coheres with clinical judgment, where reason silently accepts the AI's judgment (Kahneman, 2011). Yet, in clinical settings, this could pose serious threats to decision-making accuracy since decision-makers would fail to detect the problem at least initially (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023). However, previous research has yet to investigate the whether, and the extent to which, False Confirmation errors occur and can potentially be mitigated. The only prior study to our knowledge reported it as a source of error, noting "*participants felt confirmed by incorrect [AI] advice*" (Jussupow et al., 2021).

Thus, we hypothesise:

H3: Neither AI nor XAI helps physicians overcome false confirmation.

Next, we delve into the mixed-methods approach employed in our field study, which investigates the influence of diagnostic AI and XAI advice on physicians' trust and decision-making processes.

Method

To test our hypotheses, we performed a field study investigating the impact of AI advice, as compared with XAI advice, on physicians' decision-making processes. In particular, we pay close attention to AI and XAI engagement in scenarios of True and False Conflict, and False Confirmation, evaluating the resulting diagnostic accuracy.

We intentionally designed an AI setup where a significant portion of the time it was incorrect (40%), and applied it to an area of medical decision-making fraught with diagnostic challenges: detecting and diagnosing recurrent ear infections. We present physicians with both correct and incorrect AI advice to identify how physicians respond cognitively when their judgment comes into conflict with the AI advice as well as when AI gives false confirmation.

The study was divided into three main parts, illustrated in Figure 3: i) an initial part where physicians diagnose patients' risk of recurrent ear infections; ii) a second part where physicians have the opportunity to update their judgment when provided with AI advice; iii) finally, the last part where physicians had a second opportunity to update their judgment when provided with XAI. Throughout each step, both quantitative and qualitative data were collected. Using "think-aloud" protocols for qualitative data collection, which we describe in more detail below, we further seek to deepen our understanding of the physicians' reasoning and decision-making process.

Study Recruitment

We selected a total of 11 physicians from three Swedish hospitals for the study. Recruitment was achieved by initiating contact via email or telephone, resulting in unanimous consent from those approached. The selection aimed to cover a range of medical specialties, thereby providing a varied perspective on the use of conventional AI and XAI systems from different fields of medical practice. Each participating physician possessed adequate experience to assess the patient cases provided for this study.

The group of physicians primarily consisted of nine consultants, each with at least five years of clinical experience. Five held a Ph.D. Two participants were general practitioners without a specific specialty. The study did not mandate prior AI experience as it was centered on the pragmatic use of AI in clinical settings. Demographic details such as age, gender, and professional designations are recorded in Table 1. The interviews for the study were conducted from October to November 2022.

Age	Frequency	Percentage
30-39	5	45%
40-49	0	0%
50-59	1	9%
60-69	5	45%
Sex		
Male	9	82%
Female	2	18%
Highest level of experience		
Medical doctor (MD)	2	18%
MD + consultant	4	36%
MD + Doctor of Philosophy + senior consultant	2	18%
Professor + senior consultant	3	27%

Table 1: Participating physicians

Patient dataset

In the study we used AI and XAI algorithms on a dataset concerning risk predictors for recurring ear infections in young children during their formative years (ages). If an initial ear infection arises before 6 months of age, the likelihood of subsequent occurrences is notably high. The research was built upon data from a previous vaccination trial conducted at Sweden's Lund University Hospital, Department of Otorhinolaryngology, Head and Neck Surgery. This trial followed randomized, prospective, and single-blinded protocols, endorsed by Lund University's Ethics Committee, with parental consent secured. The outcome materialized in two published works (Gisselsson-Solén *et al.*, 2014, 2015).

The explainable AI algorithm

AI algorithms were employed on the data set, centering on machine learning with the Random Forest technique, utilizing the scikit-learn Python package (*API Reference — scikit-learn 1.1.3 documentation*). The Random Forest model constructs numerous decision trees, each trained concurrently on distinct data subsets, and the final outcome is determined by majority voting. To explain the Random Forest predictions, an open-source SHapley Additive exPlanations (SHAP) code was utilized (Lundberg, 2022). The SHAP framework is widely regarded as a benchmark for local explanations, owing to its robust theoretical foundation and broad applicability (Mosca *et al.*, 2022). A recent systematic review of applications for XAI in healthcare found that overall, SHAP is the most widely utilized XAI technique for identifying which clinical features are crucial in predicting various diseases or patient outcomes (Loh *et al.*, 2022).

The study's input parameters (x-values) encompassed family history of recurrent ear infections, number of siblings, attendance at public daycare, breastfeeding, parental smoking, previous ear infection count before study entry, and pneumococci vaccination status. The output (y-value) was defined as children with four or more recurring ear infections by the study's end (12 months), incorporating the historical ear infection count prior to inclusion. Figure 2 graphs the average of SHAP absolute values for each of the seven parameters. The bars illustrate the average contribution of each parameter to the risk of experiencing four or more ear infections. Longer bars correspond to greater influence of the feature on the output.

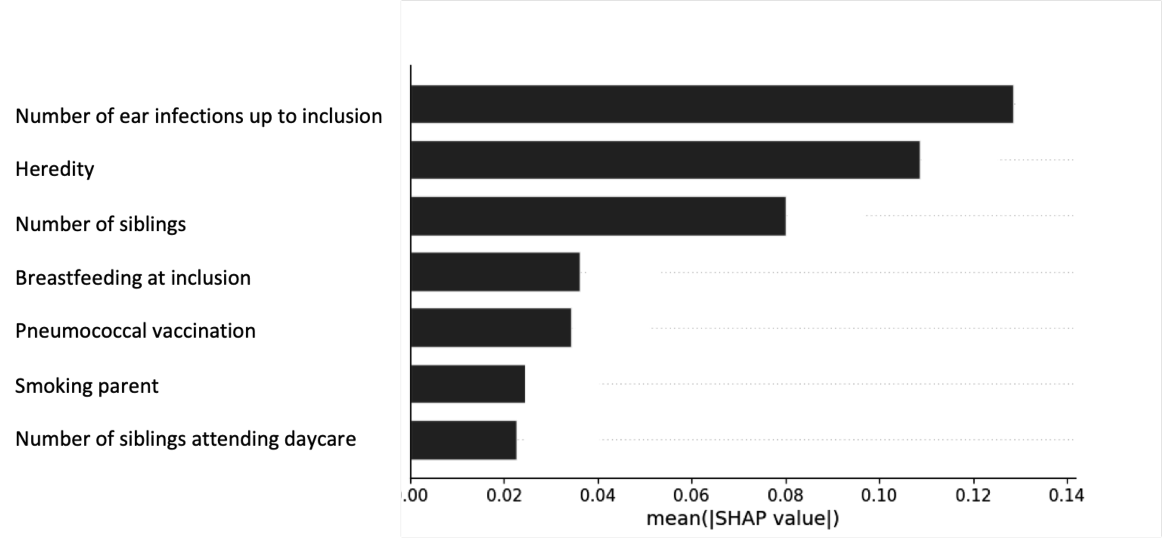
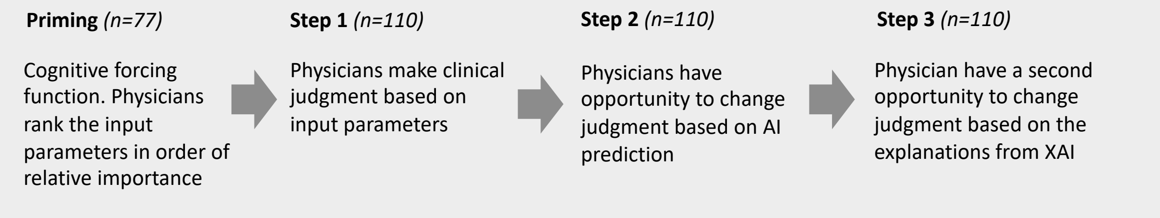


Figure 2: Bar plot of risk factors for recurring ear infection.

Notes: Each parameter is shown by rank and weight. The first risk factor "No of ear infections up till inclusion" has more than three times higher weight than risk factor 4 "Brest feeding at inclusion" or risk factor 5 "Vaccinated for pneumococci".

Data Collection



Priming *(n=77)*

Cognitive forcing function. Physicians rank the input parameters in order of relative importance

Step 1 *(n=110)*

Physicians make clinical judgment based on input parameters

Step 2 *(n=110)*

Physicians have opportunity to change judgment based on AI prediction

Step 3 *(n=110)*

Physician have a second opportunity to change judgment based on the explanations from XAI

Figure 3. Process for data collection

All three steps in Figure 3 involved collecting quantitative data, which were subsequently recorded in an Excel file. Qualitative data were collected employing "think-aloud" protocols (Van Someren, Barnard and Sandberg, 1994) and semi-structured one-on-one interviews with the 11 physicians. This method required physicians to verbalize their thought processes in real-time as they conducted diagnoses, allowing us to gain insight into their cognitive patterns and decision-making strategies. Zoom video conferencing software was used to record interviews with physicians who were located in diverse geographical locations across Sweden, with the exception of one whom did not give consent. Each interview ranged between 30 and 80 minutes, depending on the level of detail shared by the participants regarding their experiences with AI and XAI. Finally, we manually transcribed the interviews.

In Step 1, we asked each physician to sequentially rank the seven established risk factors for recurrent ear infections in infants. This totalled 77 judgments. After ranking the seven risk factors, we asked each of the 11 physicians to diagnose which of the 10 patients are likely to have recurrent ear infections (totaling 110 diagnoses). Overall, 7 out of 10 were diagnosed with actual recurrent ear infections; however, this was blinded to the physicians.

The purpose of having the physicians first rank the parameters and then diagnose the infants was to design a study where participants actively thought about their decisions rather than blindly trusting or distrusting AI and XAI advice. Existing research emphasizes the need for careful thinking when clinicians use such advice for example, strategic application of friction (Naiseh, Al-Mansoori, *et al.*, 2021), cognitive forcing functions (Buçinca, Malaya and Gajos, 2021), and ongoing self-reflection (Chromik *et al.*, 2021). We used two cognitive forcing functions "making a decision first" and "delaying AI recommendations" (Green and Chen, 2019).

We then in Step 2 provided diagnostic judgement made by the AI algorithm for all patients. Doctors were given the opportunity to change their initial judgments based on the AI predictions. Here, the physicians made another 110 diagnoses. To identify different cognitive patterns, we provided the physicians with correct and incorrect AI advice. In the sample, the AI algorithm had 60% accuracy, which was blinded for the physicians. Instead, the physicians were informed to assume that: "The algorithm is prospectively validated and meets the requirements of the National Guidelines."

In a final, third step, we made available XAI for all 11 patients, which provided clinical explanations for the AI's predictions. This was in the format of weightings for the same set of risk factors the doctors evaluated in Step 1 and is shown in the SHAP bar plot in Figure 2. As shown in the figure, the AI placed a 3-fold greater weight on the history of infections at inclusion than breastfeeding or pneumococcal vaccination. Following these explanations, the doctors were prompted to reconsider their clinical judgments and make another 110 decisions.

Taken together, this three-step procedure generated a total of 77 cognitive forcing judgments and 330 patient diagnoses.

Quantitative analysis

In the quantitative data analysis, the focus was initially on quantifying physicians' switching decisions—instances where doctors either altered or maintained their initial clinical judgment upon exposure to AI (in Step 2) or XAI (in Step 3). A heat map was generated (refer to Appendix II) to visualize these patterns, which were subsequently labeled and tallied. The dataset underwent multifaceted analysis, examining scenarios where AI was accurate or erroneous, as well as breaking down the data per patient, per physician, and per type of decision.

We tested the statistical significance of physician switches with AI and XAI in different ways. Following prior papers as a validation exercise (Chanda *et al.*, 2024), we tested whether AI led to improvements in decision-making accuracy, using t-tests to compare the accuracy of decisions with and without-AI/XAI. Turning to our hypotheses, we evaluated the correlation of their initial decision without AI with subsequent decisions with either AI or XAI. A correlation coefficient of 1 would reveal that doctors remained perfectly unchanged, whereas 0 would correspond to switching every decision. We also applied a chi-squared test to observe whether departures from original decisions were beyond what could be expected through random decision-making changes (e.g. a doctor simply changing his or her mind upon further reflection). In subsequent models we performed multivariate regression to quantify the added benefit of AI and XAI on overall diagnostic accuracy, adjusting for potential confounding factors, such as individual patient effects (e.g. how complicated their diagnostic cases were).

Qualitative analysis

For the qualitative data analysis, we performed a thematic analysis using Braun and Clarke's methodology (Braun and Clarke, 2006, 2012), involving 6 stages: 1) Familiarization with the data, 2) Generating codes, 3) Searching for themes, 4) Reviewing themes, 5) Defining and naming themes, and 6) Writing.

The thematic analysis was conducted using a systematic approach to uphold the integrity and thoroughness of the coding procedure. Initial inductive coding was carried out by the author to generate a broad set of codes directly from the data. These preliminary codes and the associated data excerpts were then examined in collaboration with another experienced researcher from the team, to reduce the risk of subjective bias. This collaborative review introduced a supplementary analytical layer, which contributed to refining the coding process and strengthening the reliability of the findings. In cases of divergent coding interpretations, a third expert researcher was consulted to facilitate a consensus, thus ensuring a critical evaluation of each code, the reduction of individual biases, and the establishment of a robust coding framework.

The coding framework was developed through an iterative process. After the inductive coding, we employed a deductive (top-down) approach and linked the themes around cognitive challenges of how doctors engaged with AI and XAI to the main decision-making errors identified. This enabled us to go beyond relatively crude characterisations of algorithmic aversion and algorithm appreciation (Logg, Minson and Moore, 2019) to understand why certain patterns of engagement occurred and further refine whether, and to when extent, they were associated with errors in clinical judgment.

To do so, our initial themes, as presented in our full quantitative and qualitative dataset (see Appendix 2), Specifically, we coded positive/negative for when doctors made a correct decision, and aversion/appreciation for when they rejected or accepted it. Thus positive appreciation corresponded to when doctors used AI advice to ultimately make a correct decision. Based on this starting point, we further began to refine patterns consistent with cognitive challenges. We identified multiple such challenges, which we further describe below, but these involved potential confirmation bias (when doctors did not seek additional information or understanding when AI confirmed their decision) (Nickerson, 1998) and commitment bias (when doctors clung to their initial decisions irrespective of AI advice) (Dolan et al., 2012). Further we differentiated these patterns in specific cases across AI and XAI. For example, when doctors rejected AI and XAI advice both, positively, we labeled it as "clinical integrity" in our initial coding, or when negatively, as "preserve incorrect frame". Throughout the analysis phase, we engaged in ongoing discussions to ensure the codes were firmly rooted in the data. As themes emerged, they were continuously cross-referenced with the existing literature, guaranteeing that our thematic construction was both data-driven and theoretically informed.

Findings

In the subsequent section, we categorize the decision outcomes from our study, examining the relationship of physician decision-making and AI's advice, with a focus on both accuracy and trust. We disaggregate two distinct decision pathways: when the AI or XAI advice is correct or incorrect. In so doing, we draw on qualitative interviews to ascertain whether and to what extent the physicians may exhibit any cognitive biases or heuristics (including algorithmic aversion or appreciation) that explain their observed decision-making patterns and interaction with AI and XAI.

First, we assess the overall quantitative impact of AI and XAI on physicians' diagnostic accuracy, differentiating physicians by level of clinical performance. Next, to deepen the psychological interpretation of these data, we outline a decision process model describing the cognitive routes physicians undertake based on whether the AI is correct or not. This draws on the qualitative "think-aloud" interviews to identify alternative decision-making approaches physicians employ in response to AI and XAI advice, highlighting the balance between human judgment, AI suggestions and trust. Following this, we analyse individual patient cases to identify incongruences in physician decision-making patterns. In Appendix IV, we developed a novel categorization of cognitive patterns we identified in the study.

Diagnostic accuracy with and without AI and XAI support

Quantitative Findings

Table 2a charts physician accuracy across three phases of clinical judgment: independent clinical decision-making (Step 1); with AI (Step 2), and with XAI explanations (Step 3).

	Overall diagnostic accuracy n=110	**Physician accuracy when AI is correct n=66**	**Physician accuracy when AI is incorrect n=44**
AI's accuracy	60%	100%	0%
Ste 1: Doctor's Initial Diagnosis	50%	53%	45%
Step 2: Doctor's Diagnosis with AI Support	53%	67%	32%
Step 3: Doctor's Diagnosis with XAI Support	57%	79%	25%

Table 2a: Physicians' Accuracy Across Different Judgment Steps.

Note: 11 physicians make judgments for 10 patients, resulting in 110 judgments for each of the three steps (n=330)

Overall, the physicians' initial clinical judgement was 50% accurate (55 out of 110 diagnoses). In Step 2, with AI advice, accuracy increased modestly to 53% (58 out of 110). A pairwise comparison revealed no statistically significant improvement (p=0.28, two-sided paired t-test, *n*=11). Finally in step 3 with XAI, the accuracy rose to 57% (63 out of 110). The pairwise comparison revealed a statistically significant improvement compared to the diagnostic accuracy in Step 1 (p<0.01, two-sided paired t-test, *n*=11).

These aggregate findings mask important trends and patterns. We next disaggregated scenarios into those when AI was correct and where it was not, as AI, by design, was accurate only in 60% of diagnoses.

Turning first to the correct AI diagnoses (corresponding to 66 diagnoses in each step), physicians' accuracy began at 53% but rose to 67% with the aid of AI. A multivariate analysis shows that when AI is correct, it increases physicians' accuracy by 39.9% (95% CI: 19.1% to 60.6%). This reveals that a significant portion of physicians persisted in an incorrect diagnosis

(True Conflict), even with the support of AI. When adding XAI in Step 3, physician accuracy jumped to 79%, indicating an important and significant improvement associated with explaining the rationale for AI judgment. A multivariate analysis shows that when the XAI advice is correct, it increases physicians' accuracy by 50.1% (95% CI: 32.2% to 69.4%). Stated otherwise, these data show a significant portion of initially incorrect physicians were better convinced by XAI than by AI alone.

However, in the scenario when AI was incorrect, there was evidence that physicians were also influenced adversely. Here, physicians initially were correct in 45% of diagnoses, but this dropped in Step 2 with AI assistance (32%) and further declined to 25% with XAI included. Thus, in cases of False Conflict, inaccurate AI can supersede correct physician decision-making.

Top-performing physicians' clinical accuracy (65%) did not improve with AI advice and slightly with XAI (68%), while low performers improved their clinical accuracy (41%) with both AI (46%) and XAI (51%) (as shown in Table 2b). In Step 2, all top performers except one make two changes but realize no overall accuracy gain. However, on average, the low performers make 1.5 changes, accounting for the entirety of the observed accuracy gains. In the third step, low performers make 8-fold more changes (on average, two changes versus 0.25), contributing to 80% of the total improvements. Clearly, the improvements in clinical accuracy can be almost entirely attributed to improvements in judgment among lower-performing physicians.

	Mean accuracy Step 1	Mean accuracy Step 2	Mean accuracy Step 3	Step 2 net additional correct diagnoses	Step 3 net additional correct diagnoses
Total physicians (n=11)	53%	56%	61%	3	5
Top-performing physicians (≥ 60% accuracy, n=4)	65%	65%	68%	0	1
Low performing physicians (< 60% accuracy, n=7)	41%	46%	51%	3	4

Table 2b: Performance metrics for top-performing physicians (n=4) compared to low performers (n=7).

Notes: Maximum total correct answers are 10 per physician in each step. Mean is total number of correct answers (or number of improvements) divided by 11 physicians. Top physicians are defined as having the same accuracy as AI or better (60%, six or more correct answers of 10 patients), and low performers are less than AI (41% accuracy on average).

Qualitative observations

To identify "thicker" descriptions of the reasoning process engaged by physicians and their corresponding trust in AI, we accompanied these judgments with interviews, prompting them to explain their logic.

All physicians articulated that trust increased with XAI, as shown in Table 3. Doctor F, for example, said, "*I believe more in the XAI compared to black-box AI.*" Similarly, Doctor G said, "*I would argue that XAI improves trust.*" Doctor C reinforced this point, "*I interpret the XAI as more trustworthy when I see that the weights and parameters*" Doctor B also concluded, "*It is a must to have the explanation available*." There was considerable appreciation of XAI over AI alone. This enhanced trust assists physicians in reducing True Conflict errors, as they are more inclined to follow the advice provided by XAI. However, conversely, a significant portion of these benefits is negated by an increase in False Conflict errors. This occurs when physicians are persuaded to switch to an incorrect diagnosis due to inaccurate XAI recommendations.

Curiously, despite voicing greater trust in XAI, some physicians trusted the AI without explanations in their decision-making. For example, Doctor B said, "*Before I use the [AI] algorithm, I need to understand how it works.*" Yet, Doctor B still made four changes based on the pure AI prediction and did not change any judgment when XAI explained the parameters.

In the above quantitative analysis, we observed that physicians with lower performance derived greater benefits from the AI-generated explanations; the qualitative data corroborate this finding. Doctor C argued, *"I would rather trust an experienced doctor making his own "black-box" clinical judgment than an unexperienced physician using AI.*" Doctor H argued in the same direction, "*There are good and bad physicians; for the bad physicians, the black-box AI is much more dangerous than the XAI.* However, the data indicate that lower performing physicians benefit more from both AI and XAI.

Doctor	Quotes
A	*"The will be more skepticism against the black-box AI. Yes, I would probably say that I trust the explanatory AI a bit more." "*
B	*"It is a must to have the explanation available… "exactly, a greater intention [to use XAI]."*
C	*"I interpret the XAI as more trustworthy when I see that the weights and parameters."*
D	*"I can trust an AI if I understand how it handles the data."*
E	*"When I got answers to what is important, I trust this [XAI] support more."*

F	*"I believe more in the XAI compared to the black-box AI."*
G	*"I would argue that XAI improves the trust."*
H	*"[XAI's] risk factors, the top three I trust a lot."*
I	*"I would not trust the AI unless I know what parameters it used."*
J	*"The XAI helps me understand; that gives me comfort and I trust the algorithm more."*

Table 3: Qualitative evaluation of XAI's influence on physicians' trust and intention to use.

Overall, our study found that AI boosts physician decision-making accuracy; this effect is amplified when explanations are provided with XAI. However, these improvements are substantially offset when incorrect AI and XAI suggestions are adopted by physicians, reducing accuracy.

Identifying patterns in decision-making

We next evaluated decision "switches," from incorrect to correct or vice-versa, by following doctors' clinical judgments at each step of the decision-making process. To do so, we employ a "decision process model," shown in Figures 3a and 3b. At each step of the model, we evaluate both the quantitative and qualitative data to ascertain the potential presence of cognitive decision-making biases or heuristics that can account for the observed decision-making patterns. Appendix IV further depicts our underlying coding of each potential switch using a 3x4 matrix (corresponding to 3 decisions and 4 possibilities of AI correct/incorrect and physician correct/incorrect), with accompanying quotes corresponding to each juncture in Appendix V.

A significant observation is that almost all physicians tend to cling to their preliminary judgment framework when it is validated by AI; unfortunately, this pattern persists even when the AI is incorrect, a decision-making pattern we describe below as False Confirmation or a "confirmation bias." Furthermore, in cases where AI contradicts the physicians' stance, a clear majority of physicians remain steadfast in their initial perspective. This strategy proves advantageous only for False Conflict when the AI is incorrect. In cases of True Conflict, where the AI is correct and the physicians incorrect, the clear majority remain steadfast. We label this commonly observed pattern as "commitment bias".

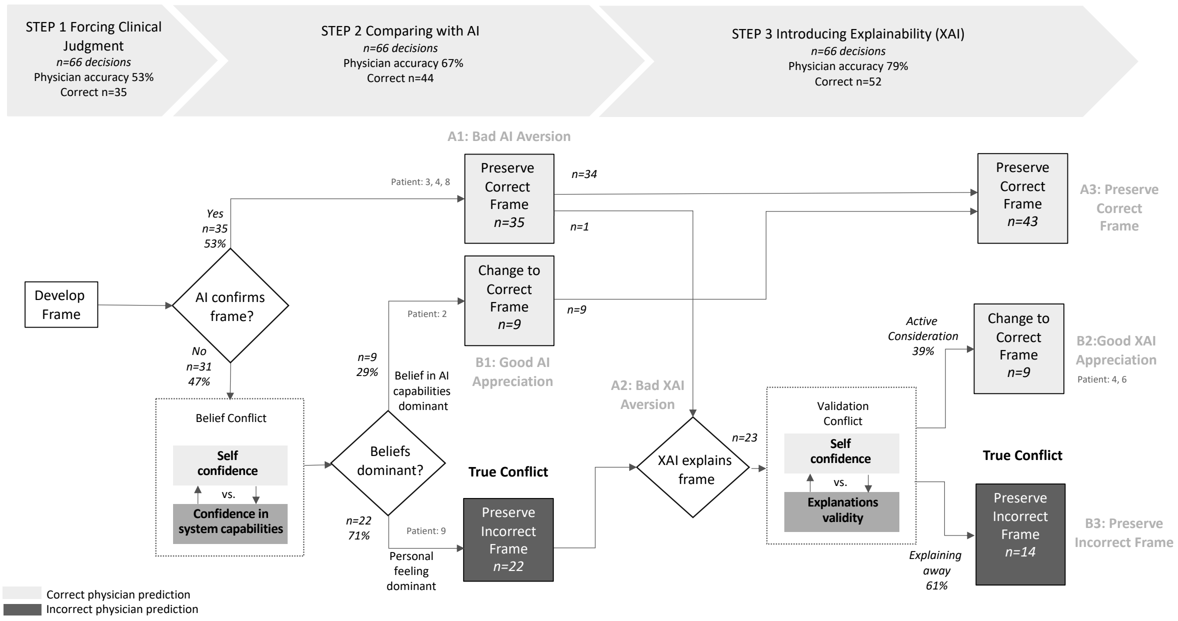


Figure 4a: Scenario 1 where AI is correct (n=198): implications of belief and validation conflicts on clinical accuracy: a comparison without and with AI and XAI]

Notes: Doctors made a total of 198 clinical decisions, distributed across three steps. In Step 1, physicians were asked to rank seven risk factors (n=77) and make an initial judgment (n=66); in Step 2, physicians were provided with AI's diagnosis and given the opportunity to update their judgment; in Step 3, physicians were provided by SHAP explanation to the AI prediction. The different patterns or themes (e.g., A1) are described in Appendix IV.

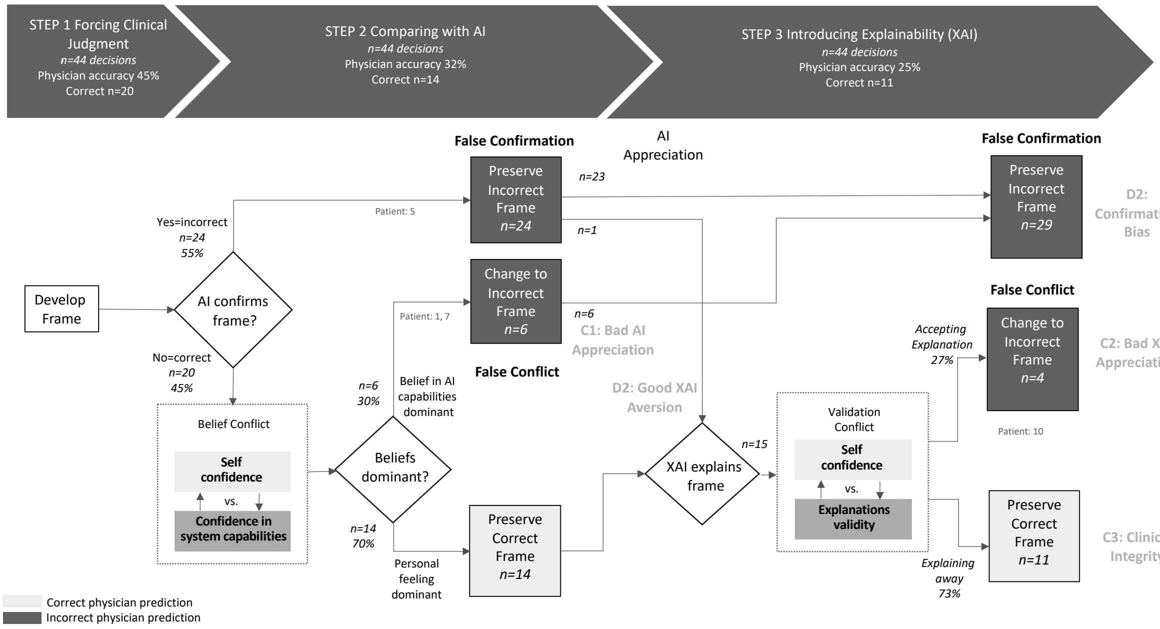


Figure 4b: Scenario 2 where AI is incorrect (n= 132): implications of belief and validation conflicts on clinical accuracy: a comparison without and with AI and XAI.

Notes: Doctors made a total of 132 decisions when AI is incorrect. The different patterns or themes (e.g., C1) are described in Appendix IV.

STEP 1 – Cognitive forcing and its implications

In Step 1, we employed cognitive forcing to engage the physicians. They ranked the clinical significance of seven risk factors for recurrent ear infections (*n*=77 judgments) and made clinical diagnosis for the 10 patients (*n*=110). We did this intentionally to activate deliberate reasoning among physicians. The importance of fostering deliberate reasoning in AI-augmented medical practice has been emphasized in studies, calling for strategic friction, cognitive forcing functions, and continuous self-reflection as discussed in section 1.2.

Then in the subsequent Step 2, the doctors were provided with AI advice, and we identified several True Conflict and False Conflict situations. Here, we noticed that (likely due to the cognitive forcing) the doctors engaged in reason based (System 2) analysis when they compared their own ranking with the AI and sought to understand how the AI ranked the clinical factors to arrive at its conclusion. For example, Doctor C commented, "*Patient 10, where I thought the number of ear infections was more important, but according to the algorithm, it's vaccination and breastfeeding*". The doctors initially pointed to difficulty comprehending the AI's judgment. For example, Doctor D argued, "*I don't understand its reasoning, so I can't say, because I don't understand what it's based on.*"

In the third step, physicians were presented with XAI they began to actively compare their own ranking of clinical parameters with those of the AI. We again noticed that this reasoning included hard cognitive work (System 2) leading in some cases to an informed re-assessment. Doctor J said in a case of True Conflict "*The patient has 2 ear infections before inclusion and heredity, which are the two most important parameters. So, I suppose I'll have to change my stance here as well; I understand why*." Doctor I commented, "*Breastfeeding at inclusion, no heredity... no, that one must be negative, there should be some logic. AI says it's positive; I don't really understand why it's positive, because it's breastfeeding, it's been vaccinated, it has a sibling in daycare—no, I remain negative on this one*".

The cognitive forcing approach, in the first step, we argue helps mitigate False Conflict errors like blind trust (an extreme of algorithm appreciation (Logg, Minson and Moore, 2019)) or True Conflict errors like distrust (an extreme of algorithm aversion (Dietvorst, Simmons and Massey, 2018)), enabling physicians to incorporate AI advice in a reasoned manner rather than as an automatic substitute for their judgment. However, cognitive forcing may provoke two important biases, namely confirmation bias and commitment bias. In cases of False Confirmation, the confirmation bias erodes the clinical accuracy when the AI errs. Commitment bias helps physicians stand against incorrect AI in False Conflict. However, as demonstrated in the following section, when AI is correct (True Conflict), this bias may lead to a failure to realise the potential accuracy gains from AI support.

STEP 2 – Potential commitment bias in clinical decision-making

We identified evidence supportive of potential "commitment bias" in step 2, reflecting how physicians fail to change their judgement where the AI disagrees. By commitment bias we invoke the definition by Cialdini "to believe more strongly in choices, once made"(Cialdini, 2007). Here 71% of the time (True Conflict) when the AI provided a correct diagnosis, while the physician's judgment was incorrect, the physician failed to update their diagnosis to the correct one.

This failure appears to be linked to a conflict between their own self-confidence and their confidence in the AI system's capabilities. When conflict arises with AI, the doctors appear to exhibit intuitive trust (System 1) in their own diagnosis in the majority of cases. As argued by Doctor J "*I don't believe I'm better, but I can't see a reason to change my decision".* To further corroborate this, we also find that the doctors who initially are correct, when the AI is incorrect (False Conflict) also maintain their (accurate) position in 70% of the cases when this situation arises.

This commitment bias appears to impede physicians from reaping the full benefits of AI's potential to improve their judgment. True Conflict is the only explanation of why physicians' accuracy only increases from 53% to 67% when AI is correct. On the other hand, when the AI is incorrect (False Conflict), the self-commitment preserves the accuracy to a large extent. When

AI is wrong, it is beneficial for doctors to trust their own diagnosis, but this appears to occur more due to bias than deliberate choice since this happened in about 70% of the cases, regardless of whether AI is correct.

Our first hypothesis was confirmed; it posited that physicians tend to stick with their first clinical diagnosis when AI disagrees, regardless of whether the AI advice is correct or incorrect. We found a moderately weak Pearson correlation between initial clinical diagnoses and those diagnoses influenced by AI in Step 2 (r=0.38, p = 0.01), consistent with the hypothesis that doctors tend to stick to their original diagnoses. This was also statistically significant based on chi-square tests of independence for physicians' decisions with and without AI ($\chi^2(1) = 7.31$, p <0.01).

To quantify the impact of AI on doctor's decision-making changes, we then performed a multivariate regression. We found that the probability of doctors' changing their diagnoses when contradicted by AI was, overall, 29.5% (p-value < 0.01, 95% CI: 16.8% to 42.2%), even after adjusting for whether or not the patient diagnosis was positive or the AI was accurate or not. Stated otherwise, 70% of the time, doctors clung to their initial diagnosis. Of note, whether or not the AI was accurate had no significant effect on physicians' decision-making changes.

It is additionally worth noting that when AI agreed with physicians, none altered their clinical decision, indicating a confirmation bias. Importantly, this applied also to inaccurate physician decisions, creating the potential for AI to reinforce and exacerbate erroneous decision-making. False Confirmation was described by Doctor J "*I don't change where I'm already correct*". An agreement with AI is not the same as being correct, both can still be wrong. This is in line with our third hypothesis that neither AI nor XAI helps physicians overcome false confirmation.

Of the total incorrect judgments with incorrect AI advice (*n*=24+6=30) in Step 2, False Confirmation and confirmation bias correspond to 80% of the errors, explaining the major reason why physicians' accuracy decreases so much when the AI errs. The remaining 20% of the errors stem from False Conflict where the AI convinced physicians to change to an incorrect frame despite the fact that they initially made a correct clinical judgment.

STEP 3 – Explanations increase doctors' confidence in the AI system's capacity

When physicians are presented with explanations, they learn from the explanations and update their cognitive frame. However, despite the learning, we identify that the commitment bias remains, and the learning is not applied in full. Furthermore, there is still a confirmation bias, and it seems they are not able to combine their own clinical experience with newfound knowledge from the explanations. In cases of False Confirmation when both the physicians and the AI are incorrect, we could not identify any reason based hard cognitive work (System 2) to identify errors, all except one physician blindly keep their incorrect frame.

The physicians argue that the explanation not only helps them understand the AI; it also teaches them something new, especially in areas they lack expertise in. Doctor H discussed the importance of learning and argued that XAI not only helps to understand but also educates "*XAI is preferred, it is education, and then it becomes more interesting*". Doctor H elaborated further, saying that once the XAI had explained the risk factors, there was no longer a need for the algorithm. Doctor C discussed the value added of XAI's explanations "*If I have explanations, XAI can be a good support, especially in fields where I am not a specialist"*. Learning seems to be one of the outcomes of adding explanations to AI predictions.

Furthermore, the explanations help the physicians to some extent to change their incorrect frame in cases for True Conflict. In the cases where AI is correct, and the physician is incorrect, an additional 39% change to the correct diagnosis. Unfortunately, in cases of False Conflict, when AI is incorrect, and the physician is correct, 27% change to a wrong diagnosis.

The commitment bias to cling to the initial wrong clinical judgment prevails despite the explanations where 61% preserve the incorrect frame. When the AI errs, the physicians' self-confidence helps them to resist the wrong advice in 73% of the cases. In total, the explanations improve the overall accuracy.

However, the explanations also wrongly led 27% of the judgments in the wrong direction. This phenomenon was evident with Doctor D, who, using the explanations, made six adjustments—three correct and three incorrect—resulting in no net gain. Doctor D stated: " *When I understood which parameters that are important, I used the algorithm*". Explanation can be harmful in False Conflict and lower doctors' accuracy when AI is incorrect.

Turning to our second hypothesis that XAI would be more convincing to physicians than AI, we tested whether the impact of XAI in cases of contradictions was greater than that of AI. To do so, we compared physicians' decisions made with AI to those made in the subsequent step with XAI. If XAI had little or no persuasive value, we would see little or no change in decision-making across these steps. Here we found a moderately weak Pearson coefficient (r=0.35, p-value<0.05), consistent with the hypothesis that doctors did change diagnoses further when given XAI explanations. This was also borne out in tests of independence ($\chi2(1) = 4.42$, p <0.01). According to our multivariate regression, the probability of physicians changing their diagnosis in Step 3 versus their diagnosis in Step 2 was 20.9% (p-value < 0.01, 95% CI: 7.6% to 34.2%), indicating that XAI is more persuasive than AI. Whether or not the AI was accurate had no significant effect on these changes.

Taken together, the overall impact of XAI on physicians changing from Step 1 to 3 was considerable, at 50.4% (p-value <0.01, 95% CI: 35.5% to 65.2%), revealing a substantial effect of XAI collaboration on decision-making. Nevertheless, approximately half did not change their minds when contradicted by AI (irrespective of whether it was accurate or not).

As anticipated, little changed when physicians who were already in agreement with AI were presented with explanations of XAI; the confirmation bias prevailed. The physicians argued that the explanations taught them new things, and they updated their cognitive frame. They selectively compared their own ranking of the parameter with the XAI and made changes accordingly for the patients where there was a disagreement with the AI advice. However, despite cognitive forcing and explanations, in cases of False Confirmation, all except one (Doctors D) did not revisit the cases where they agreed with the AI to look for potential errors in the algorithm. This False Confirmation bias corresponds to 88% (29 of 33) of the errors when AI was wrong.

Our third hypothesis asserted that neither AI nor XAI helps physicians to overcome cases of False Confirmation. We found a strong Pearson correlation (r=0.91, p-value < 0.01) between the initial clinical diagnosis in Step 1 and the final diagnosis in Step 3 in scenarios where both the AI/XAI and the physicians' assessments were incorrect. This correlation underscores the tendency of physicians to adhere to their original diagnoses when falsely confirmed, thus indicating a substantial confirmation bias.

Curiously, only one of the physicians revisited patients in cases of False Confirmation, where there was an agreement with the AI to identify potential errors made by the algorithm. Doctor D made only one change in Step 2 and argued, "*I do not feel I can change my mind based on AI since I do not understand, I lose control of the patient*" and "*I cannot just trust a computer*." When presented with XAI, however, Doctor D made two changes against both their and the AI's judgment. Doctor D, updated the cognitive frame, arguing, "*The first time, I ranked things wrong in my mind, but now I understand what is most important.*" It seems that the doctors struggled between intuitive self-confidence and reason to understand the explanations' validity. D argued, *"What do you do if you, like your intuition, think this child is at risk, but AI says no, or vice versa?"* For one of Doctor D's changes, the AI was wrong, and for the other, the AI was correct, implying no improvement in accuracy.

Disaggregating by doctor and patient: persistent evidence of blind trust and distrust

Next, we evaluated the clinical decisions made for each patient, as shown in Table 4. The analysis reveals limited evidence of blind distrust of AI, apart from two physicians who did not use either AI or XAI advice. However, several doctors showed signs of blind distrust for specific patients when they were very confident, and, conversely, blind trust when they lacked confidence in their judgment. For one patient, none of the doctors drew upon AI (erroneously).

Patient	AI Correct	Step 1: Clinical judgment (accuracy)	Step 2: AI advice (accuracy)	Step 3: XAI advice (accuracy)
1	No	45%	27%	9%
2	Yes	27%	73%	100%
3	Yes	100%	100%	100%
4	Yes	64%	73%	91%
5	No	18%	18%	18%
6	Yes	36%	55%	82%
7	No	36%	9%	18%
8	Yes	73%	82%	73%
9	Yes	18%	18%	27%
10	No	82%	73%	55%

Table 4: Physicians diagnostic accuracy per patient

Note: Maximum is 11 correct physicians

Blind distrust among physicians: perceptions and practice

Turning first to evaluating doctors' engagement with AI as a whole, it appears most doctors exhibited a reasoned use of AI, rather than a complete engagement or disengagement. Two exception was Doctor F and G, who showed blind distrust of AI and XAI, completely rejecting its use, despite the explanations. As the Doctor F said, "*I trust my clinical assessment*" and "*I find it hard to believe that AI knows more than me*".

There emerged an important difference between what the physicians articulated, and how they used the available evidence to make clinical decisions. While Doctor G voiced "*I would argue that XAI improves trust*", he/she still refused to use it. Another Doctor B said, "*It is a must to have the explanations available,"* but in practice, did not alter judgments whatsoever when given XAI. Still, Doctor B made four changes with AI advice for ambiguous cases. Of note Doctors B and F were among the most experienced doctors in the study sample, refer to Appendix I.

Inverse uptake and confidence relationship

Trust in AI and XAI seems not only to differ between physicians (inter-personal) but also across patients (intra-personal). Patients 2 and 9 illustrate these intra-personal shifts. For both patients, the AI advice is correct and cases of True Conflict. For patient 2, accuracy increases from 27% to 73% with AI, and to100% with XAI, as all doctors take up its suggestions. However, for patient 9, the accuracy similarly starts at 18% but does not change with AI advice. The only doctor to update their judgement was Doctor A with XAI, increasing overall accuracy to 27%.

In the case of True Conflict, the reasons for discrepancy emerge clearly from the qualitative data. The physicians felt uncertain about patient 2 and, as a result, were more easily persuaded. Doctor I changed for patient 2, when presented with XAI emphasising the importance of daycare exposure. Doctor I noted, "*That's how it was with our daughter. As soon as she went back to daycare, she got it [recurrent ear infection] back*". Doctor also B changed easily for patient 2 in the second step, "*There are some that I don't want to change, so to speak, but I could very well change number 2 to positive, no doubt about it.*"

In contrast, physicians felt more confident in their judgment of patient 9, although it was incorrect. In this case, XAI emphasized the role of heredity, which the doctors did not categorize as a strong risk factor. Thus Doctor I was not convinced "*AI says that it is positive, I don't know why it is positive, because it is breast feeding, it has been vaccinated, a sibling in daycare, no, I remain negative there."* Doctor E goes against the AI in Step 2, "*I wouldn't change, for example [patient 9], I mean that 9 would become positive, that feels strange*", and with explanations, remained unconvinced, "*I still don't understand"* and further argue *"*[Patient] *Nine has no heredity and it becomes positive?.*" For patient 9, True Conflict disagreements with AI's judgment led to a missed opportunity for improved accuracy.

It is worth noting another extreme case of False Conflict, patient 1. Here five physicians initially make the correct clinical judgment. However, AI persuades them to shift to an incorrect one, falling to 27% accuracy with AI and ultimately 9% with XAI. Doctor A argued, "*I would probably change [the clinical decision for] patients 1 and 2. I changed my mind for those I was not sure.*" In this False Conflict case, doctors also voiced being unsure, as with patient 2, but took up an erroneous AI decision. Taken together, these findings are consistent with the possibility that neither cognitive forcing nor XAI is sufficient to eliminate hyperbolic decision-making, either blind trust or blind distrust.

Discussion

In this study we move beyond characterizing physician engagement with AI as “aversion” or “appreciation” to identify the cognitive challenges and patterns that can arise in physician-AI collaboration, and specifically how these relate to decision-making errors. We additionally compared how physicians trust and assess information from AI relative to explainable AI. Following prior literature (Jussupow *et al.*, 2021), we sought to identify recurrent patterns of cognition and judgment, especially in scenarios where AI advice came into conflict with physicians’ own beliefs.

To our knowledge, this approach for the first time identifies a comprehensive set of decision-making errors in human-AI collaboration which is likely to be quite generalizable. Specifically, we find evidence, as seen in prior studies (Chanda *et al.*, 2024), that physician-AI collaboration can outperform physician decision-making alone. However, systematic errors can arise when doctors lacked confidence in their decision-making and draw upon erroneous AI advice. Doctors also showed a general reluctance to change decisions when conflicted by AI, consistent with AI distrust and a potential commitment bias. Explanations helped mitigate this seeming stubbornness but only partially.

Next we turn to evaluating these observations in light of our main study hypotheses, and their implications, both theoretical and practical for designing information systems for healthcare applications.

Key findings in relation to our research question and hypotheses

Turning to our first hypothesis, we posited that physicians would tend to stick to their original diagnoses and would be unlikely to change them when contradicted by AI advice. We found support for this hypothesis, observing that doctors in about 2/3rds of all decisions clung to their initial decision irrespective of whether the AI system was correct or incorrect. These observations appear consistent with a “commitment bias”, whereby physicians initial decision anchors them, irrespective of new facts and information to the contrary.

Importantly, we found that explainable AI significantly increases physicians' trust and intention to use XAI advice compared with conventional AI advice without explanation, increasing uptake by 20%. Our think-aloud protocols revealed that doctors were more willing to incorporate XAI decisions when they could comprehend them.

These observations provided support for our second hypothesis that “Physicians are more likely to revise their clinical diagnoses when XAI explanations contradict their assessments than when contradicted by AI alone.” Whereas AI convinced about 30% of doctors to change their minds, XAI convinced another 20% over and above AI alone. Yet this still left about half of all physicians who did not change their diagnosis in light of AI guidance. While it may have been possible that this reflected AI accuracy, we observed no difference in these behaviours whether or not AI was correct.

Finally, our data were also consistent with our third hypothesis: neither AI nor XAI were effective in assisting physicians to avoid False Confirmation errors. That is, when AI confirmed an erroneous diagnosis, virtually no physician sought further information or to understand its explanation (irrespective of whether that explanation was coherent or not with the doctor's own understanding). This was a substantial theme associated with decision-making errors: about two-thirds of all errors in our study were attributable to false confirmation. Our qualitative interviews revealed that physicians were reluctant to seek further information in such cases, thereby perpetuating diagnostic errors and obstructing the learning process.

In terms of accuracy, consistent with other studies (Chanda *et al.*, 2024), we found that physician-AI collaboration led to accuracy gains, although in ours this was only significant with the incorporation of (the more persuasive) XAI. AI led to greater accuracy gains among weaker performing doctors (not necessarily those with more or less experience). However, these benefits were in part offset by doctors' uptake of erroneous AI advice, often arising from situations of clinical uncertainty combined with blind trust in AI.

Contributions to current theory and evidence

Our findings hold significance for both theory and practice. From a theoretical standpoint, our study illuminates several previously overlooked aspects of clinical decision augmentation involving AI and XAI. First, while prior studies have speculated that there could be an hypothetical "worst-case" scenario in which decision-makers fail to identify a false AI confirmation (Jussupow *et al.*, 2021), our study has empirically demonstrated, to our knowledge for the first time, that this can be quite substantial and correlate strongly with overall errors in human-AI collaboration. In our study, when AI falsely confirmed physicians they almost universally blindly accepted it.

Second, theoretically, we go beyond identifying aversion or appreciation to connect these behaviours to errors and potential underlying cognitive heuristics. For example, we argue that the unquestioning acceptance of AI in a false confirmation is consistent with a confirmation bias. In medical decision-making, confirmation bias is often defined as "*where evidence against one's position is selectively disregarded*" (Rollwage *et al.*, 2020). However, future studies would be needed to see if doctors actively or passively discard additional information or truly selectively seek out only information which would confirm their decision made with AI.

Additionally, in line with previous studies, we found that explanations lead to fewer True Conflict errors where incorrect clinicians override a correct AI (Jussupow *et al.*, 2021; Kumar *et al.*, 2021; Martínez-Agüero *et al.*, 2022). Kumar and colleagues found that explanations increased physicians' trust and intention to use deep learning models for the diagnostics of brain tumors (Kumar *et al.*, 2021). Martínez-Agüero and colleagues found that XAI improved physicians' trust and accuracy when detecting the presence of antimicrobial multidrug-resistant bacteria in intensive care units (Martínez-Agüero *et al.*, 2022). On the other hand, as argued by Jussupow and colleagues, researchers should focus not only on cases when AI is correct but also on cases when AI is incorrect (Jussupow *et al.*, 2021). In line with other studies, we found that explanations can also lead to an overreliance on AI, creating more False Conflict errors where correct physicians are convinced to change to an incorrect diagnosis (Naiseh, Al-Mansoori, *et al.*, 2021; Naiseh, Al-Thani, *et al.*, 2021; Naiseh *et al.*, 2023).

In our study, we also reveal the importance of doctors' own confidence in their decision-making as a key factor shaping their AI engagement. Thus whether doctors show aversion or appreciation changes dynamically. Our interview data show that when physicians feel confident in their decisions and perceive low uncertainty, they were less likely to use AI's suggestions. This occurred regardless of whether AI was right or wrong, suggesting it reflects an intuitive, System 1 heuristic like commitment bias rather than a product of System 2's logical reasoning. In contrast, when physicians were uncertain, they were more likely to exhibit blind trust in AI, taking its diagnosis on board, even when the doctors failed to understand the clinical rationale.

Additionally, while there has been speculation that cognitive forcing could shift thinking to System 2, and in so doing avoid pitfalls of System 1 heuristics, our data did not support this notion. Kahneman argues that "When there are cues that an intuitive judgment could be wrong, System 2 can impose a different strategy, replacing intuition by careful reasoning." (Kahneman and Klein, 2009, p. 519). Again, our study did not find evidence to support this. It may be possible, however, that alternative cognitive forcing strategies could be more effective in averting decision-making errors.

Implications for Practice

Our study also provides a number of recommendations for better optimising effective doctor-AI collaboration. First, our findings indicate a potential need to train physicians to use of AI and XAI in clinical decision-making. This would likely make physicians better aware of the risks of false confirmation and call for critical evaluation.

Second, while it is tempting to suggest universally XAI, given doctors were more likely to trust and deploy its advice, we found that XAI could increase erroneous uptake when AI was incorrect. We suggest that XAI systems should test whether including a caveat with its advice a prompt reminding doctors of AI fallibility. This could help avert the issue of blind trust, especially prominent when doctors feel uncertain about their diagnosis. Clearly much more research is needed to optimize the design of AI systems to help mitigate the risks of errors we identified in this study.

Third, system design should likely account for AI's accuracy as well as the stakes of decision-making. When AI accuracy approaches 100%, greater trust in and utilization of AI should be the norm, and providing explanations could always be utilised to enhance compliance. However, given AI fallibility, especially in high-risk situations, we would call for a true human "Second Opinion" rather than unquestioning reliance upon AI.

Limitations and Future Directions

Our field study has important limitations worth noting. First, our study involved 11 physicians who made a total of 330 decisions, which creates a potential for decision-making fatigue as well as learning. Thus, physicians may engage differently with AI advice for the first patient relative to the last. It is also possible that physicians establish early on a decision-making heuristic, which they continue to adhere to throughout, and so the ordering of AI advice could yield differing behaviours. Dedicated experimental studies would be needed to disentangle these subtle, but important potential influences on AI engagement. Second, we were unable to directly ask doctors about trust in AI, as a potential "observer effect" could reinforce or skew their trust in AI for successive decisions. Instead, we sought to infer trust from physicians' actual behavior as a "revealed preference." A more complex and nuanced design would be needed to elicit physicians' trust perceptions without impacting their actual decision-making. Ideally, the study could have better assessed the confidence levels of physicians for each individual judgment, although carefully doing so to avert potential observer effects.

Third, while our paper could demonstrate behaviour that was consistent with biases, such as commitment or confirmation bias, a bespoke experimental design would be needed to prove absolutely their presence.

Fourth, as with any experimental laboratory setting, ours narrowed the range of real-world complexity when doctors make decisions with AI. In our study physician accuracy was likely lower than what would be expected in a real clinical setting. Physicians solely relied on risk factor data without considering other clinical inputs which in real-clinical settings would be available to them. Future research could try to evaluate decision-making in real clinical settings. Additionally, AI accuracy was set at 60%. Ideally future studies could vary accuracy to elicit how doctors exhibit more or less trust in relation to perceived and actual AI accuracy.

Fifth, we intentionally introduced cognitive forcing to trigger System 2 thinking. This could have exacerbated commitment bias, leading doctors to cling more forcefully to the initial decisions they invested cognitive resources into making. While our research importantly demonstrated that cognitive forcing did not alleviate certain common errors, future research could be designed to test explicitly the magnitude and type of effects of cognitive forcing on decision-making.

Sixth, our experiment did not involve randomization, which creates an absence of a true control group, but instead uses doctors as their own control, incrementally exposing them to AI and then later XAI. Future research, with larger samples, could attempt to randomize doctors to different intervention arms to more convincingly ascertain AI and XAI effects.

Our study also points to several further important directions for future research. To reduce False Confirmation errors, future research can explore two key avenues. First, implementing nudges within AI systems can prompt critical evaluation by physicians. These nudges act as reminders, encouraging doctors to question consensus and consider the possibility of errors from both AI and their initial diagnosis. Second, a comparative analysis of AI explanations (including ranking and weights of risk factors) compared to physician clinical judgment. Examining whether they rely on the same clinical factors and assign similar weights can uncover discrepancies, particularly in cases of False Confirmation.

For now our study also points to important safeguards to implement in clinical practice. Given the scale of false confirmation, AI applications in healthcare may consider prompting of doctors to consider potentially that the AI advice, when confirming their own advice, could be fallible. In the future AI may be able to serve as a "second opinion", but more research and understanding of how doctors integrate AI into decisions is needed before AI can begin to substitute for real human-decision making, especially when stakes are high. Better experiments not only of AI's accuracy and validity, but real world studies of how doctors integrate them into practice is needed to realise fully the hope and promise of human-AI collaboration.

Acknowledgments: The author extends gratitude to all the physicians who contributed to this study. I would like to express my sincere gratitude to Ioanna Constantiou, Åsa Melhus and David Stuckler for their invaluable roles as sounding boards and for their meticulous review and editing of this paper.

Competing interests: None declared.

Funding: The author has not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

References

Amann, J. *et al.* (2020) ‘Explainability for artificial intelligence in healthcare: a multidisciplinary perspective’, *BMC Medical Informatics and Decision Making*. BioMed Central Ltd, 20(1), pp. 1–9. doi: 10.1186/S12911-020-01332-6/PEER-REVIEW.

Aniansson, G. *et al.* (1992) ‘Nasopharyngeal Colonization during the First Year of Life’, *Journal of Infectious Diseases*, 165, pp. S38–S41. doi: 10.1093/infdis/165-Supplement_1-S38.

Antoniadi, A. M. *et al.* (2021) ‘Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review’, *Applied Sciences (Switzerland)*. MDPI AG, 11(11). doi: 10.3390/app11115088.

‘API Reference’ (2006) in *The Definitive Guide to MySQL5*, pp. 693–720. doi: 10.1007/978-1-4302-0071-0_23.

Bernstein, J. M. *et al.* (1991) ‘Nasopharyngeal flora in the first three years of life in normal and otitis-prone children’, *Annals of Otology, Rhinology & Laryngology*, 100(8), pp. 612–615. doi: 10.1177/000348949110000802.

Bertrand, A. *et al.* (2022) ‘How cognitive biases affect XAI-Assisted decision-making: A systematic review’, in *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91. doi: 10.1145/3514094.3534164.

Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, *Qualitative Research in Psychology*, 3(2), pp. 77–101. doi: 10.1191/1478088706QP063OA.

Braun, V. and Clarke, V. (2012) ‘Thematic analysis.’, in *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, pp. 57–71. doi: 10.1037/13620-004.

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’, *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery, 5(CSCW1), p. 21. doi: 10.1145/3449287.

Castelvecchi, D. (2016) ‘Can we open the black box of AI?’, *Nature*, 538(7623). doi: 10.1038/538020a.

Chanda, T. *et al.* (2024) ‘Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma’, *nature.comT Chanda, K Hauser, S Hobelsberger, TC Bucher, CN Garcia, C Wies, H Kittler, P TschandlNature Communications, 2024•nature.com*. Available at: https://www.nature.com/articles/s41467-023-43095-4 (Accessed: 29 January 2024).

Chew, H. S. J. and Achananuparp, P. (2022) ‘Perceptions and Needs of Artificial Intelligence in Health Care to Increase Adoption: Scoping Review’, *Journal of Medical Internet Research*. doi: 10.2196/32939.

Chromik, M. *et al.* (2021) 'I Think i Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 307–317. doi: 10.1145/3397481.3450644.

Cialdini, R. B. (2007) *Influence: The psychology of persuasion*, *New York, NY, USA: HarperCollins Publishers*. Collins New York. doi: 10.1017/CBO9781107415324.004.

Croskerry, P. (2003) 'Cognitive forcing strategies in clinical decisionmaking', *Annals of Emergency Medicine*, 41(1). doi: 10.1067/mem.2003.22.

Cui, M. and Zhang, D. Y. (2021) 'Artificial intelligence and computational pathology', *Laboratory Investigation*. doi: 10.1038/s41374-020-00514-0.

Cutillo, C. M. *et al.* (2020) 'Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency', *npj Digital Medicine*. doi: 10.1038/s41746-020-0254-2.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000033.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2018) 'Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them', *Management Science*. doi: 10.1287/mnsc.2016.2643.

Dolan, P. *et al.* (2012) 'Influencing behaviour: The mindspace way', *Journal of Economic Psychology*. doi: 10.1016/j.joep.2011.10.009.

Evans, T. *et al.* (2022) 'The explainability paradox: Challenges for xAI in digital pathology', *Future Generation Computer Systems*, 133. doi: 10.1016/j.future.2022.03.009.

Fazal, M. I. *et al.* (2018) 'The past, present and future role of artificial intelligence in imaging', *European Journal of Radiology*. doi: 10.1016/j.ejrad.2018.06.020.

Gaube, S. *et al.* (2023) 'Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays', *Scientific Reports*, 13(1). doi: 10.1038/s41598-023-28633-w.

Gerlings, J., Shollo, A. and Constantiou, I. (2021) 'Reviewing the need for explainable artificial intelligence (XAI)', in *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE Computer Society, pp. 1284–1293. doi: 10.24251/hicss.2021.156.

Gisselsson-Solén, M. *et al.* (2014) 'Risk factors for carriage of AOM pathogens during the first 3 years of life in children with early onset of acute otitis media', *Acta Oto-Laryngologica*. Informa Healthcare, 134(7), pp. 684–690. doi: 10.3109/00016489.2014.890291.

Gisselsson-Solén, M. *et al.* (2015) 'Effect of pneumococcal conjugate vaccination on nasopharyngeal carriage in children with early onset of acute otitis media-a randomized controlled trial', *Acta Oto-Laryngologica*. Informa Healthcare, 135(1), pp. 7–13. doi: 10.3109/00016489.2014.950326.

Giuste, F. *et al.* (2023) ‘Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review’, *IEEE Reviews in Biomedical Engineering*, 16. doi: 10.1109/RBME.2022.3185953.

Graber, M. L. *et al.* (2012) ‘Cognitive interventions to reduce diagnostic error: A narrative review’, *BMJ Quality and Safety*. doi: 10.1136/bmjqs-2011-000149.

Green, B. and Chen, Y. (2019) ‘The principles and limits of algorithm-in-the-loop decision making’, *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery, 3(CSCW). doi: 10.1145/3359152.

Gunning, D. and Aha, D. W. (2019) ‘DARPA’s explainable artificial intelligence program’, *AI Magazine*, 40(2). doi: 10.1609/aimag.v40i2.2850.

Gupta, D. M., Boland, R. J. and Aron, D. C. (2017) ‘The physician’s experience of changing clinical practice: A struggle to unlearn’, *Implementation Science*, 12(1). doi: 10.1186/s13012-017-0555-2.

Haque, A. B., Islam, A. K. M. N. and Mikalef, P. (2023) ‘Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research’, *Technological Forecasting and Social Change*, 186. doi: 10.1016/j.techfore.2022.122120.

Ienca, M. (2023) ‘Don’t pause giant AI for the wrong reasons’, *Nature Machine Intelligence*. doi: 10.1038/s42256-023-00649-x.

Jacobs, M. *et al.* (2021) ‘How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection’, *Translational Psychiatry*, 11(1). doi: 10.1038/s41398-021-01224-x.

Jung, J. *et al.* (2023) ‘Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review’, *Heliyon*. doi: 10.1016/j.heliyon.2023.e16110.

Jussupow, E. *et al.* (2021) ‘Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence’, *Information Systems Research*, 32(3). doi: 10.1287/ISRE.2020.0980.

Kahneman, D. (2011) *Thinking fast, thinking slow*, *Interpretation, Tavistock, London*.

Kahneman, D. and Klein, G. (2009) ‘Conditions for Intuitive Expertise: A Failure to Disagree’, *American Psychologist*. doi: 10.1037/a0016755.

Kilpi, T. *et al.* (2001) ‘Bacteriology of acute otitis media in a cohort of Finnish children followed for the first two years of life’, *Pediatric Infectious Disease Journal*, 20(7), pp. 654–662. doi: 10.1097/00006454-200107000-00004.

Klein, G. (2015) ‘A naturalistic decision making perspective on studying intuitive decision making’, *Journal of Applied Research in Memory and Cognition*, 4(3). doi: 10.1016/j.jarmac.2015.07.001.

Kolasa, K. *et al.* (2023) 'Systematic reviews of machine learning in healthcare: a literature review', *Taylor & Francis*. Taylor and Francis Ltd. doi: 10.1080/14737167.2023.2279107.

Kumar, A. *et al.* (2021) 'Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis', *ACM Transactions on Multimedia Computing, Communications and Applications*, 17(3s). doi: 10.1145/3457187.

Kundu, S. (2021) 'AI in medicine must be explainable', *Nature Medicine*, 27(8). doi: 10.1038/s41591-021-01461-z.

Lambe, K. A. *et al.* (2016) 'Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review', *BMJ Quality and Safety*, pp. 808–820. doi: 10.1136/bmjqs-2015-004417.

Lewicki, R. J. and Brinsfield, C. T. (2011) 'Framing trust: Trust as a heuristic', in Donohue, W. A., Rogan, R. R., and Kaufman, S. (eds) *Framing matters: Perspectives on negotiatin research and practice in communication*. Peter Lang Publishing, pp. 110–135. Available at: http://www.mdpi.com/1996-1073/2/3/556/.

Liu, C. F. *et al.* (2022) 'Does AI explainability affect physicians' intention to use AI?', *International Journal of Medical Informatics*, 168. doi: 10.1016/j.ijmedinf.2022.104884.

Logg, J. M., Minson, J. A. and Moore, D. A. (2019) 'Algorithm appreciation: People prefer algorithmic to human judgment', *Organizational Behavior and Human Decision Processes*. doi: 10.1016/j.obhdp.2018.12.005.

Logg Jennifer (2018) 'Do People Trust Algorithms More Than Companies Realize?', *Harvard Business Review*. Available at: https://hbr.org/2018/10/do-people-trust-algorithms-more-than-companies-realize (Accessed: 13 September 2019).

Loh, H. W. *et al.* (2022) 'Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)', *Computer Methods and Programs in Biomedicine*. doi: 10.1016/j.cmpb.2022.107161.

Lundberg, S. (2022) *Lundberg/SHAP*, *GitHub*. Available at: https://github.com/slundberg/shap (Accessed: 9 November 2022).

Madsen, M. and Gregor, S. (2000) 'Measuring Human-Computer Trust', *Proceedings of Eleventh Australasian Conference on Information Systems*.

Mahmud, H. *et al.* (2022) 'What influences algorithmic decision-making? A systematic literature review on algorithm aversion', *Technological Forecasting and Social Change*, 175. doi: 10.1016/j.techfore.2021.121390.

Martínez-Agüero, S. *et al.* (2022) 'Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance', *Future Generation Computer Systems*, 133. doi: 10.1016/j.future.2022.02.021.

Matias, J. N. (2017) *Bias and Noise: Daniel Kahneman on Errors in Decision-Making*, *Medium.com*. Available at: https://medium.com/@natematias/bias-and-noise-daniel-kahneman-onerrors-in-decision-making-6bc844ff5194 (Accessed: 21 October 2019).

Mosca, E. *et al.* (2022) 'SHAP-Based Explanation Methods: A Review for NLP Interpretability', in *Proceedings - International Conference on Computational Linguistics, COLING*.

Naiseh, M., Al-Thani, D., *et al.* (2021) 'Explainable recommendation: when design meets trust calibration', *World Wide Web*, 24(5). doi: 10.1007/s11280-021-00916-0.

Naiseh, M., Cemiloglu, D., *et al.* (2021) 'Explainable Recommendations and Calibrated Trust: Two Systematic User Errors', *Computer*, 54(10). doi: 10.1109/MC.2021.3076131.

Naiseh, M., Al-Mansoori, R. S., *et al.* (2021) 'Nudging through Friction: an Approach for Calibrating Trust in Explainable AI', in *Proceedings of 2021 8th IEEE International Conference on Behavioural and Social Computing, BESC 2021*. doi: 10.1109/BESC53957.2021.9635271.

Naiseh, M. *et al.* (2023) 'How the different explanation classes impact trust calibration: The case of clinical decision support systems', *International Journal of Human Computer Studies*, 169. doi: 10.1016/j.ijhcs.2022.102941.

Nazar, M. *et al.* (2021) 'A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques', *IEEE Access*. doi: 10.1109/ACCESS.2021.3127881.

Nickerson, R. S. (1998) 'Confirmation bias: A ubiquitous phenomenon in many guises', *Review of General Psychology*. doi: 10.1037/1089-2680.2.2.175.

Park, J. S. *et al.* (2019) 'A slow algorithm improves users' assessments of the algorithm's accuracy', *Proceedings of the ACM on Human-Computer Interaction*. doi: 10.1145/3359204.

Petersson, L. *et al.* (2022) 'Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden', *BMC Health Services Research*, 22(1). doi: 10.1186/s12913-022-08215-8.

Rajpurkar, P. *et al.* (2022) 'AI in health and medicine', *Nature Medicine*. doi: 10.1038/s41591-021-01614-0.

Reddy, S. (2022) 'Explainability and artificial intelligence in medicine', *The Lancet Digital Health*. doi: 10.1016/S2589-7500(22)00029-2.

Rollwage, M. *et al.* (2020) 'Confidence drives a neural confirmation bias', *Nature Communications*, 11(1). doi: 10.1038/s41467-020-16278-6.

Schwalbe, N. and Wahl, B. (2020) 'Artificial intelligence and the future of global health', *The Lancet*. doi: 10.1016/S0140-6736(20)30226-9.

Sherbino, J. *et al.* (2014) ‘Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: A controlled trial’, *Canadian Journal of Emergency Medicine*, 16(1). doi: 10.2310/8000.2013.130860.

Van Someren, M., Barnard, Y. F. and Sandberg, J. (1994) ‘The think aloud method: a practical approach to modelling cognitive’, *London: AcademicPress*, 11.

Sutton, R. T. *et al.* (2020) ‘An overview of clinical decision support systems: benefits, risks, and strategies for success’, *npj Digital Medicine*. doi: 10.1038/s41746-020-0221-y.

Tversky, A. and Kahneman, D. (1974) ‘Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty’, *Science*. doi: Cited By (since 1996) 3914\nExport Date 30 November 2011.

Wadden, J. J. (2021) ‘Defining the undefinable: the black box problem in healthcare artificial intelligence’, *Journal of Medical Ethics*, 48(10). doi: 10.1136/medethics-2021-107529.

Appendix I: Risk factors and physician performance

Patient	Number of Ear Infections Until Inclusion (Anamnestic)	Pneumococcal Vaccination	Breastfeeding at Inclusion / 4 Months	Any Parent Smokes	Number of Siblings	Number of Siblings in Daycare	Heredity (Includes Uncles/Aunts)	Algorithm Response	Correct Response
1	1	No	Yes	No	1	1	Yes	1	0
2	1	No	No	No	3	0	Yes	1	1
3	6	No	Yes	No	1	1	Yes	1	1
4	2	No	No	No	0	0	Yes	1	1
5	1	No	No	Yes	6	1	No	1	0
6	1	No	No	No	1	1	Yes	1	1
7	1	Yes	Yes	No	1	1	Yes	0	1
8	1	Yes	No	No	4	0	Yes	0	0
9	1	Yes	Yes	No	1	1	No	1	1
10	2	Yes	Yes	No	1	1	No	0	1

Table AI-1: Risk factors

Doctor	Step 1 Clinical judgment: # correct predictions	Step 2 AI assistance: #correct predictions	Step 3 XAI assistance: #correct predictions	Step 2 # changes with AI assistance	Step 3 # changes with AI assistance	Step 2 # improvments	Step 3 # improvments
A	7	7	8	2	1	0	1
C	7	7	7	2	0	0	0
F	6	6	6	0	0	0	0
H	6	6	6	2	0	0	0
I	5	5	6	0	1	0	1
J	5	5	6	2	1	0	1
K	5	4	6	1	4	-1	1
B	4	6	6	4	0	2	2
D	4	5	5	1	6	1	1
E	4	5	5	1	2	1	1
G	2	2	2	0	0	0	0
Mean	5,30	5,60	6,10	1,50	1,50		
Physicians>AI	6,50	6,50	6,75				
Physicians>AI	4,14	4,57	5,14				

Table AI-2: Physicians' performance, ranked per best performance in Step 1, clinical judgment

Appendix II: Heat map of physicians' performance

doctor_id	patient	diag1	diag2	diag3	Corr answer	AI correct	#_corr_1	#_corr_2	#_corr_3	Pattern
1	2	0	1	1	1	Yes	0	1	1	Good AI Appreciation
1	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
1	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
1	6	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
1	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
1	9	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
1	1	0	1	1	0	No	1	0	0	Bad AI Appreciation
1	5	0	0	0	0	No	1	1	1	Clinical Integrity
1	7	0	0	0	1	No	0	0	0	Confirmation bias
1	10	1	1	1	1	No	1	1	1	Clinical Integrity
2	2	0	1	1	1	Yes	0	1	1	Good AI Appreciation
2	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
2	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
2	6	0	1	1	1	Yes	0	1	1	Good AI Appreciation
2	8	1	0	0	0	Yes	0	1	1	Good AI Appreciation
2	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
2	1	1	1	1	0	No	0	0	0	Confirmation Bias
2	5	1	1	1	0	No	0	0	0	Confirmation bias
2	7	1	0	0	1	No	1	0	0	Bad AI Appreciation
2	10	1	1	1	1	No	1	1	1	Clinical Integrity
3	2	0	1	1	1	Yes	0	1	1	Good AI Appreciation
3	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
3	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
3	6	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
3	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
3	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
3	1	0	0	0	0	No	1	1	1	Clinical Integrity
3	5	0	0	0	0	No	1	1	1	Clinical Integrity
3	7	0	0	0	1	No	0	0	0	Confirmation bias
3	10	1	0	0	1	No	1	0	0	Bad AI Appreciation
4	2	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
4	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
4	4	0	1	1	1	Yes	0	1	1	Good AI Appreciation
4	6	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
4	8	0	0	1	0	Yes	1	1	0	Bad XAI Aversion
4	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame

4	1	0	0	1	0	No	1	1	0	Bad XAI Appreciation
4	5	1	1	1	0	No	0	0	0	Confirmation Bias
4	7	0	0	1	1	No	0	0	1	Good XAI Aversion, Why?
4	10	1	1	0	1	No	1	1	0	Bad XAI Appreciation
5	2	0	1	1	1	Yes	0	1	1	Good AI Appreciation
5	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
5	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
5	6	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
5	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
5	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
5	1	0	0	1	0	No	1	1	0	Bad XAI Appreciation
5	5	1	1	1	0	No	0	0	0	Confirmation Bias
5	7	0	0	0	1	No	0	0	0	Confirmation Bias
5	10	0	0	0	1	No	0	0	0	Confirmation Bias
6	2	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
6	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
6	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
6	6	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
6	8	1	1	1	0	Yes	0	0	0	Preserve Incorrect Frame
6	9	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
6	1	1	1	1	0	No	0	0	0	Confirmation Bias
6	5	1	1	1	0	No	0	0	0	Confirmation Bias
6	7	1	1	1	1	No	1	1	1	Clinical Integrity
6	10	1	1	1	1	No	1	1	1	Clinical Integrity
7	2	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
7	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
7	4	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
7	6	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
7	8	1	1	1	0	Yes	0	0	0	Preserve Incorrect Frame
7	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
7	1	1	1	1	0	No	0	0	0	Confirmation Bias
7	5	1	1	1	0	No	0	0	0	Confirmation Bias
7	7	0	0	0	1	No	0	0	0	Confirmation Bias
7	10	0	0	0	1	No	0	0	0	Confirmation Bias
8	2	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
8	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
8	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame

8	6	0	1	1	1	Yes	0	1	1	Good AI Appreciation
8	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
8	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
8	1	0	1	1	0	No	1	0	0	Bad AI Appreciation
8	5	1	1	1	0	No	0	0	0	Confirmation Bias
8	7	0	0	0	1	No	0	0	0	Confirmation Bias
8	10	1	1	1	1	No	1	1	1	Clinical Integrity
9	2	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
9	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
9	4	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
9	6	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
9	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
9	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
9	1	1	1	1	0	No	0	0	0	Confirmation Bias
9	5	1	1	1	0	No	0	0	0	Confirmation Bias
9	7	0	0	0	1	No	0	0	0	Confirmation Bias
9	10	1	1	1	1	No	1	1	1	Clinical Integrity
10	2	0	1	1	1	Yes	0	1	1	Good AI Appreciation
10	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
10	4	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
10	6	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
10	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
10	9	0	0	0	1	Yes	0	0	0	Preserve Incorrect Frame
10	1	1	1	1	0	No	0	0	0	Confirmation Bias
10	5	1	1	1	0	No	0	0	0	Confirmation Bias
10	7	1	0	0	1	No	1	0	0	Bad AI Appreciation
10	10	1	1	1	1	No	1	1	1	Clinical Integrity
11	2	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
11	3	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
11	4	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
11	6	0	0	1	1	Yes	0	0	1	Good XAI Appreciation
11	8	0	0	0	0	Yes	1	1	1	Preserve Correct Frame
11	9	1	1	1	1	Yes	1	1	1	Preserve Correct Frame
11	1	1	1	1	0	No	0	0	0	Confirmation Bias
11	5	1	1	1	0	No	0	0	0	Confirmation Bias
11	7	1	0	0	1	No	1	0	0	Bad AI Appreciation
11	10	1	1	0	1	No	1	1	0	Bad XAI Appreciation

Appendix III: Medical background on ear infection

Middle ear infection or acute otitis media (AOM) is the most common bacterial infection during childhood. It has been estimated that 60-70% of all children will have had 1-3 episodes before the age of two years, and about 10% will have recurrent AOM (rAOM). Recurrent AOM can lead to long and repeated periods of antibiotic treatment and hearing impairment. It is defined as three AOM episodes in six months or four in a year (Bernstein *et al.*, 1991; Aniansson *et al.*, 1992).

Leading otitis pathogens are *Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis,* and *Streptococcus pyogenes* (Kilpi *et al.*, 2001). During the first four years of life, most children are colonized by one or more AOM pathogens in the nasopharynx, the reservoir of upper respiratory tract pathogens. From the nasopharynx, the bacteria enter the middle through the Eustachian tube. With age, the colonization decreases, and the anatomy changes. AOM is, therefore, relatively rare after the age of 12 years.

If the first AOM occurs before the age of 6 months, the risk is very high (80%) that that child will develop rAOM. Several other risk factors of AOM have been studied, including heredity, smoking, number of siblings, etc. In a previous vaccination trial performed at the Department of Otorhinolaryngology, Head and Neck Surgery Department at Lund University Hospital, Sweden, data concerning children with or without rAOM were collected. The trial was randomized, prospective, and single-blinded. The Ethics Committee approved the study at Lund University, and written consent was obtained from the parents. It resulted in two publications (Gisselsson-Solén *et al.*, 2014, 2015).

A total of 105 children were included in the trial. The children were recruited between March 2003 and June 2007. For inclusion, the child had to have had at least one AOM episode confirmed by an otorhinolaryngologist before the age of six months, thereby having an 80% risk of developing rAOM. The first publication investigated the risk factors for the carriage of AOM pathogens during the first three years of life (Gisselsson-Solén *et al.*, 2014), whereas the second publication explored if a heptavalent pneumococcal conjugate vaccine could decrease the number of AOM episodes in a population with a high risk of developing rAOM (Gisselsson-Solén *et al.*, 2015). To this data set, different algorithms were applied.

Appendix IV. Synthesising decision-making patterns in human-AI collaboration

To sum up our quantitative and qualitative analysis we present an analytical framework highlight the main decision patterns and evidence of associated cognitive heuristics.

There are four main potential scenarios, as articulated above: A) True Confirmation – Physician correct / AI correct; B) True Conflict – Physician incorrect / AI correct; C) False Conflict – Physician correct/ AI incorrect; and D) False Confirmation – both the physician and the AI are incorrect. We then cross-coded these against decision-making changes: update with AI (1); update with XAI (2) ; or no change (3), as shown in Table 5. Taken together, this results in a total of 12 outcomes (A1-3, B1-3, C1-3).

Note: In column "No judgment change in Step 1 or 2" physicians have undergone two opportunities to change. For example, in the first row when AI is correct and physician is correct all 35 keep their judgment in Step 2, in Step 3 one change their mind and 34 keep their judgment. Total n is hence 35+34=69 out of the total judgments 35x2 steps= 70. Percentages do not sum to 100% as the denominator differs for each decision-making juncture.

Step 1: **Initial Clinical judgment**		**Step2:** **Change with AI**	**Step 3:** **Change with XAI**	**No judgement changes in Step 1 or 2**
True Confirmation	AI Correct/ Physician Correct (n=35)"	A1: Negative AI Aversion 0/35 (0%) Physician was initially correct, and so was AI, but when given AI's answer, physician changes their diagnosis to an incorrect judgment	A2: Negative XAI Aversion 1/35 (3%) Physician was initially correct and so was AI, but when given XAI explanation for AI's answer, physician changes their mind to an incorrect judgment	A3: Preserve Correct Frame 69/70 (99%) Physician was initially correct, and so was AI. Physician's answer remains correct throughout.
True Conflict	AI Correct/ Physician Incorrect (n= 31)	B1: Positive AI Appreciation 9/31 (29%) Physician initially incorrect, but AI correct. Physician updates judgment to correct diagnosis.	B2: Positive XAI Appreciation 9/22 (41%) Physician was initially incorrect, but AI was correct. Physician did not change when given AI advise but did change with explanations.	B3: Preserve Incorrect Frame (Commitment Bias) 35/53 (66%) Physician initially incorrect, but AI correct. Physician does not change despite AI or XAI advice.
False Conflict	AI Incorrect/ Physician Correct (n=20)	C1: Negative AI Appreciation 6/20 (30%) Physician initially correct, but AI incorrect. Physicians update to incorrect answer based on AI judgement.	C2: Negative XAI Appreciation 4/14 (29%) Physician initially correct, but AI incorrect. Physicians update to incorrect answer based on the explanations.	C3: Clinical Integrity 24/34 (69%) Physician initially correct, but AI incorrect. Physicians preserve correct judgement against incorrect AI diagnosis.
False Confirmation	AI Incorrect/ Physician Incorrect (n=24)	D1: Positive AI Aversion 0/24 (0%) Physician and AI incorrect. Physician updates to correct answer against AI judgement.	D2: Positive XAI Aversion 1/24 (4%) Physician and AI incorrect. Physician updates to correct answer based on the explanations.	D3: Confirmation Bias 47/48 (98%) Physician and AI incorrect, and physician maintains flawed perspective.

Table AIV-1: Themes of Physicians' Clinical Judgment and Interactions with AI and XAI.

Appendix V. Illustrative Quotes of Positive and Adverse Decision-Making Patterns in Human-AI Collaboration

We evaluate each outcome in turn with accompanying quotes illustrating the doctors' decision-making process for the junctures in the table below, starting with True Confirmation.

Pattern	Incidence	Quote(s)
Positive		
A3: Preserve Correct Frame	99%	"If one has a hypothesis and AI agrees, then it might be more likely to be true." (E)
C3: Clinical Integrity	69%	"Find it hard to believe the algorithm knows more than I do" (F)
B2: Positive XAI Appreciation	41%	"I made an additional change when I noticed that the number of siblings carried significant weight." (A)
B1: Positive AI Appreciation	29%	"I can certainly say that the patients I mentioned as negative were, as you understood, merely a guess." (B)
D2: Positive XAI Aversion	4%	"The first time, I ranked incorrectly in my mind, but now I understood which one was the most important." (D)
Adverse		
D3: Confirmation Bias	98%	"I don't change where I'm already correct." (J)
B3: Preserve Incorrect Frame Commitment Bias	66%	"XAI improve trust but I still do not use it." (G)
C1: Negative AI Appreciation	30%	"AI could serve as valuable support in situations of ambiguity." (A)
C2: Negative XAI Appreciation	29%	"XAI give me more comfort than a black-box AI" (C)
A2: Negative XAI Aversion	3%	"The first time, I ranked incorrectly in my mind, but now I understood which one was the most important." (D)

Table AV-1: Incidence and Type of Decision-Making Patterns

Note: percentages do not sum to 100% as the denominator differs for each decision-making juncture. Please see table 5 and figures 2a and 2b for reference

A. True Confirmation

In 99% of the scenarios where AI and physicians were correct, the doctors preserved the correct judgement (A3). Doctor E argued "*If one has a hypothesis and AI agrees, then it might be more likely to be true"*. Hence doctors clearly felt that reinforcement from AI helped increase their confidence in their initial judgement.

There was one curious exception, with Doctor D, who, when presented with XAI, attempted to learn from the XAI ranking and in so doing, altered the clinical judgment to an incorrect one. As Doctor D explained, "*The first time, I ranked incorrectly in my mind, but now I understood which one was the most important."*

B. True Conflict

The second row highlights that physicians tended to stay with their own frame in most cases of disagreements, but several were persuaded to change their minds by AI and especially XAI when they understood and resultantly trusted it.

Understanding the algorithm was key to trust in cases of True Conflict. Doctor A stated, "*When I know what it [the XAI] base its judgment on, it is a support to my own [clinical] judgment".* Several doctors completely avoided AI because it lacked explanation and they could not understand its decision. Doctor I stated*, "I would not trust the AI unless I know what parameters it used, then I trust my clinical judgment."* Additionally, when doctors voiced they did not understand XAI's explanation, they disregarded it. For example, Doctor D argued, "*I do not feel I can change my mind based on AI since I do not understand, I lose control of the patient*" and "*I cannot just trust a computer*". Doctor D changed one judgement based on the AI advice and another two based on the XAI advice.

Curiously, even when doctors understood and trusted XAI, they could be reluctant to use it. Doctor G argued "*XAI improve trust but I still do not use it.*" These scenarios portray a potential commitment bias (as voicing trust reveals that there was not an overarching aversion to AI).

False Conflict

When physicians are correct, but the AI errs, the challenge is to resist the advice, but our data shows that with explanations it is harder to resist. In False Conflict scenarios, the desired outcome is for physicians to retain their correct stance, resisting AI influence. Yet, 30% (C1) changed their judgment based on the AI advice, and 29% (C2) changed to wrong judgment based on the explanations. This negative appreciation indicates that when algorithms err, explanations may further diminish accuracy. Doctor D further pointed out the importance of prudence and reluctance to the AI support in Step 2 "*A doctor saying, I looked into a black box and will not prescribe penicillin for your child, how trustworthy is that?"*

In False Conflict cases, the inclusion of explanations has the potential to reduce accuracy. However, in 74% of instances (C3) where either AI or XAI erroneously suggests that physicians alter their stance, they maintain their clinical integrity by disregarding the advice. Doctor F argued "*Find it hard to believe the algorithm knows more than I do"*. Not changing frame despite contradictory recommendations from AI and XAI is a positive aspect in this case where the AI is incorrect, in contrast to the negative implications indicated in scenario B3.

False Confirmation - The Dark Side of Confirmation Bias

The final row reveals the perilous aspect of confirmation bias, with physicians holding steadfast to their wrong clinical assessments in 98% of the cases (D3) where both AI and XAI sanctioned the incorrect decision. Doctor J felt that confirmation is similar to being correct, and argued "*I don't change where I'm already correct"*. Being confirmed is, unfortunately, not the same thing as being correct but a case of False Confirmation. In this case, D3 serves as the negative counterpart to A3, and D2 is the positive counterpart to A2, illustrating this dualistic phenomenon.

Appendix III: Paper III

In Proceedings of the *32th European Conference on Information Systems (ECIS) Association for Information System*s. This is the peer-reviewed and accepted version. There can be deviations after copyediting and publisher layout.

Citation: Rosenbacke, R. (2024) ‘Heuristics and Errors in XAI-Augmented Clinical Decision-Making: Moving Beyond Algorithmic Appreciation and Aversion’, *In Proceedings of the 32th European Conference on Information Systems (ECIS) Association for Information Systems*. AIS Electronic Library (AISeL).

Heuristics and Errors in XAI-Augmented Clinical Decision-Making

Moving Beyond Algorithmic Appreciation and Aversion

Rikard Rosenbacke

Copenhagen Business School, Denmark

Keywords: XAI, Explainable Artificial Intelligence, Dual Process Theory, Heuristics.

Abstract

How do physicians integrate AI tools into medical decision-making? Prior research has analyzed extensively whether they exhibit AI algorithmic aversion or appreciation. Yet we argue that these behavioral outcomes arise from underlying decision-making heuristics such as pro-innovation bias, ambiguity aversion, or commitment bias. In this qualitative study, we examined 330 clinical decisions using "think aloud" protocols to identify heuristics employed with AI and explainable AI (XAI). We observed the presence of multiple heuristics, including a "mere exposure effect" and "false confirmation bias". These heuristics were associated with decision-making errors. The "mere exposure effect" occurred commonly with XAI, when physicians, feeling uncertain about their diagnoses, altered their decision to an incorrect AI diagnosis. False confirmation errors also emerged when AI confirmed an erroneous diagnosis, precluding doctors from seeking alternative information. We also discuss how cognitive interventions could redress these heuristics in decision-making to better optimize accuracy.

Introduction

Artificial Intelligence (AI) is emerging rapidly as an important tool for decision-making in the healthcare field (Rajpurkar *et al.*, 2022). In theory, AI can help clinicians make more accurate judgments and reduce errors. Still, in practice, these decision-support systems are not always trusted, and their use in clinical situations is, therefore, not as extensive as maybe wished for.

There is a common perception among doctors that AI operates in a "black-box" without providing clear justification for its health-related advice (Fazal *et al.*, 2018). Unlike rule-based systems, AI platforms are less transparent, making their errors harder to anticipate (Jussupow *et al.*, 2021). When healthcare providers do not understand clinical advice, they are much less likely to use it (Cui and Zhang, 2021).

Recently, explainable AI (XAI), in which AI diagnoses are accompanied by clinical explanations, has been developed to overcome these limitations, increasing its uptake in diverse management domains as well as in healthcare. XAI methods aim to make AI systems' hidden logic intelligible to humans, understanding why the AI system makes the predictions it does (Bauer, von Zahn and Hinz, 2023). An emerging body of research argues that XAI is essential for securing the safety, approval, and adoption of AI systems in clinical settings (Evans et al., 2022). Several systematic literature reviews argue that XAI could enhance decision confidence and trust for clinicians (Antoniadi *et al.*, 2021; Nazar *et al.*, 2021; Giuste *et al.*, 2023). However, there is limited empirical evidence for increased trust.

Much research has found that decision-makers are more likely to incorporate advice from humans than AI algorithms. This has been variously defined as algorithm aversion (Dietvorst, Simmons and Massey, 2015) (when rejecting advice), or algorithm appreciation (Logg Jennifer, 2018) (when incorporating it). A recent systematic literature review investigated 80 empirical studies on algorithm aversion and found that, in general, "*People tend to rely less on algorithms even when algorithms provide better decisions*" (Mahmud *et al.*, 2022).

Although the issue of algorithm aversion has been documented extensively (Burton, Stein and Jensen, 2019; Mahmud *et al.*, 2022), limited attention has been given to exploring the underlying reasons for its presence and how it could be addressed. However, a large body of literature (Kahneman, 2011) has found the presence of decision-making "shortcuts", or heuristics, likely to drive decision-making patterns, such as aversion or appreciation. Additionally, the studies of algorithmic aversion have tended to be conducted in artificial laboratory settings, often with students or crowd-sourced workers (like Mechanical Turk), which may not reflect the actual performance of AI systems in real-world settings. In a systematic review, researchers call for more qualitative studies with practitioners, noting that "*scholars should undertake more qualitative research on this area [algorithm aversion], involving practitioners*" (Mahmud *et al.*, 2022).

Another limitation of prior scholarship is that studies have tended to focus on if and to what extent physicians adjust their decision-making when AI models are correct or perform significantly better than clinicians. For example, scholars noted that *"most prior work has assumed that provided system advice is correct and beneficial. In doing so, it has largely neglected the cognitive challenges entailed in incorrect system advice"*(Jussupow *et al.*, 2021).

Here, we aim to contribute by qualitatively studying physicians in a real-world clinical setting. We specifically seek to explore the heuristics used in making clinical decisions, with and without the support of AI (and XAI with its explanations) and with AI advice that is both correct and incorrect.

Our data sources come from a prior experimental study calibrating AI accuracy in diagnosing recurrent ear infections at 60% (Rosenbacke, 2024). The authors used think-aloud protocols to investigate how physicians handled correct and incorrect advice from both AI and XAI. We revisited the qualitative data to identify potential cognitive challenges using explanations as an intervention with a special focus on whether they used cognitive shortcuts or heuristics that could explain algorithm aversion or appreciation in their decision-making process. Although prior scholars have conceptualized algorithm aversion and appreciation as heuristics themselves, we argue that these are merely behavioural outcomes of the deeper underlying heuristics. We aim to move beyond this literature by addressing two unresolved questions:

1. What are the underlying heuristics that drive algorithm aversion or appreciation, as well as decision-making errors?
2. How are these heuristics affected by the presence of explanations accompanying AI support?

We begin by reviewing background research on human-AI collaboration in medical decision-making from both cognitive psychology and medical perspectives. Then, we detail the methods employed in our qualitative thematic analysis of decision-making heuristics. Subsequently, we present the findings that we used to develop a conceptual framework of cognitive challenges, heuristics, and potential sources of errors in AI/XAI-augmented decision-making. Finally, we propose potential directions for future research and conclude by reviewing alternative interventions that could redress the biases in decision-making we observed.

Background and related work

In order to realize the potential of AI in decision-making, it is important that it can align not only with the needs of end-users but also how they cognitively process that information. We first provide a quick overview of research on AI's use or non-use in medical decision-making, followed by an analysis of dual process theory, which we argue goes beyond the simplistic characterisation of AI appreciation or aversion among end-users.

AI's uptake in healthcare settings

Although AI is emerging rapidly in its applications to healthcare, a series of recent reviews have found that clinicians seem to be slow in adapting to it, and employing it effectively in practice. A recent scoping review on clinicians perceptions of AI found many were positively disposed because of its availability, ease of use, and potential to improve efficiency and reduce the cost of healthcare service delivery (Chew and Achananuparp, 2022). However, doctors raised concerns regarding the lack of trust, data privacy, patient safety, technological maturity, and the possibility of full automation, which limited AI's use in clinical settings.

One common factor limiting doctors' use of AI, emerging from systematic review evidence, is AI's "black-box" nature. This is the way in which AI commonly makes predictions without clinical justification, and acts a major barrier to its use in clinical practice (Loh *et al.*, 2022). When healthcare providers do not understand clinical advice, they are much less likely to use it (Cui and Zhang, 2021). The core of clinical medicine, particularly the practice founded on evidence-based medical practice, necessitates clear and transparent decision-making processes (Amann *et al.*, 2020; Kundu, 2021). However, a new generation of XAI tools, which advanced upon AI diagnoses by accompanying them with explanations that clinicians can understood, could help boost trust and associated use of AI. Overall, there are very few studies which investigate the organisational and individual factors which influence AI's application in healthcare settings, as well as the potential impact of XAI on usability.

Alternatively, much AI research in medicine has tended to focus on scenarios of whether or not AI yields improvements in clinical decision-making. In these studies, the experimental designs are often constructed so that the AI is correct and the physician is incorrect (Jussupow *et al.*, 2021). This has led to a binary focus on whether physicians reject AI advice (labelled as a heuristic of algorithm aversion (Dietvorst, Simmons and Massey, 2015) or where people adhere more to advice from an AI (labelled as algorithm appreciation (Logg Jennifer, 2018)).

However, since AI is imperfect, like any model, imperfect, it is critical to investigate the cognitive processes that take place when AI models provide results that are incorrect (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023). A simplistic dichotomy of algorithmic aversion or appreciation is analytically unhelpful. Aversion can be useful when the AI advice is incorrect, and appreciation can be useful when the AI advice is correct and vice versa.

Beyond aversion and appreciation – the role of dual process theory

How can scholarship of human-AI collaboration deepen understanding of these processes and, in so doing, ultimately design better support systems? We believe there is untapped potential to incorporate the now vast body of knowledge from cognitive psychology on decision-making heuristics that likely apply to AI-augmented decisions. We utilize this when investigating what gives rise to either appreciation or aversion.

We follow prior information systems (IS) research in this area which has drawn upon the dual-process theory (Kahneman, 2011), to deepen understanding of the reasoning processes decision-makers engage in when incorporating information from AI supports. In a previous systematic review the dual-process theory was used to better understand cognitive constraints in human-AI collaboration, with a focus on cognitive biases (Bertrand *et al.*, 2022). Another IS study used dual-process theory to investigate how cognitive decision processes hinder the optimal utilization of AI advice among clinicians (Jussupow *et al.*, 2021), and one study used it to explain how decision noise impacts users' information processing related to XAI (Bauer, von Zahn and Hinz, 2023). Although these studies drew upon dual process theory, this scholarship is still at a relatively embryonic stage of development and considerable literature merely cites algorithmic aversion or appreciation, without seeking to identify their potential cognitive underpinnings.

The dual process theory posits that human cognition operates through two distinct processes: intuition (fast thinking System 1) and reasoning (slow thinking System 2) (Kahneman, 2011). When acquiring new skills, the slower-thinking System 2 engages in intensive cognitive processing. With prolonged practice, when System 1 is effectively trained, it becomes capable of swiftly recognizing patterns. However, intuitive System 1 uses heuristics that may not be well-suited for new contexts, highlighting the potential for misjudgments and mistakes when intuitive thinking is applied outside its accustomed domain. While "bias" or "heuristic" often suggests judgment errors, in line with previous studies on XAI and heuristics, we frame it as cognitive constraints inherent in the human explanation process (Bertrand *et al.*, 2022). These shortcuts can sometimes lead to mistakes, but as our findings highlight, they can also serve as beneficial heuristics.

It is most likely the case that algorithm aversion or appreciation is fueled by System 1 and its associated heuristics pertaining to trust. To trust AI, or have the intention to use AI, can be based on cognition-based trust, where trust is derived from the perceived understandability, reliability, and technical competence of AI, rooted in reasoning. However, trust can also be intuitive or affect-based, involving emotional attachment and faith (Madsen and Gregor, 2000; Lewicki and Brinsfield, 2011). Independent of which one of these facets of trust that is engaged, trust can serve as a System 1 decision-making shortcut, enabling the decision-maker to select information while ignoring other information to simplify a complex decision.

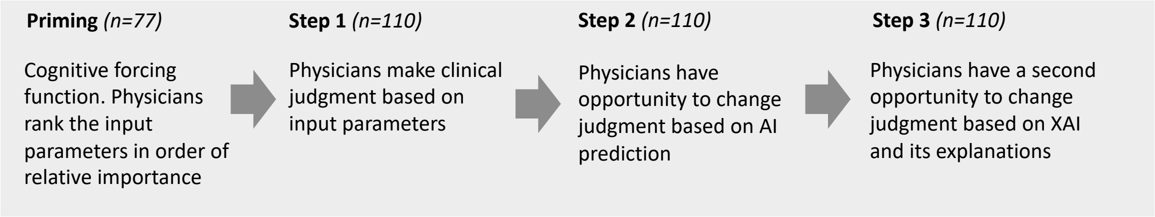
In healthcare domains, prior research has identified a number of trust-related heuristics. These include the expert halo effect (Austin and Foster, 2019) (the assumption of infallibility to an expert) or the “trust heuristic” known as the “argument from authority” (Cummings, 2014) which may lead to suboptimal decision-making. Another bias is the availability heuristic, where physicians' assessments of the likelihood of an event are affected by how easily the event comes to mind (Ly, 2021). Yet the specific heuristics from System 1 that may apply to human-AI collaboration in healthcare are not well documented.

In this study, we contribute by providing insights based on a qualitative study that allows us to uncover the heuristics that drive algorithm aversion and appreciation as well as decision errors. We also depict how the use of heuristics is affected by the use of XAI. Furthermore, we go beyond errors where physicians override correct AI advice to where AI and its explanations sway physicians to alter an accurate diagnosis or when physicians’ incorrect diagnoses are affirmed by erroneous XAI, leading to the false assumption that the chosen path is correct, when, in fact, both are incorrect.

Materials and Methods

This paper draws on data from a larger research project investigating physicians' accuracy, trust, and intention to use AI and XAI (Rosenbacke, 2024). In this field study, the researchers assessed physicians' decision-making with and without AI and XAI assistance. The task was to determine infants' risk of developing recurrent middle ear infections. The physicians were provided with both correct and incorrect advice to study different cognitive challenges, and the AI accuracy was set to 60% (blinded to the physician).

The multi-step study proceeded in four phases, as depicted in Figure 1. These were: i) priming decision-making using cognitive forcing functions; ii) physicians' initial judgment on patients' risk of recurrent ear infections; iii) an opportunity to change diagnosis based on AI advice; iv) a further opportunity to change diagnosis based on AI and its explanations (XAI). The physicians made a total of 407 decisions or judgments (*n*=77+110+110+110). At each step, we collected qualitative data, as further described below.



Priming *(n=77)*

Cognitive forcing function. Physicians rank the input parameters in order of relative importance

Step 1 *(n=110)*

Physicians make clinical judgment based on input parameters

Step 2 *(n=110)*

Physicians have opportunity to change judgment based on AI prediction

Step 3 *(n=110)*

Physicians have a second opportunity to change judgment based on XAI and its explanations

Figure 1. Process for data collection (n=total number of decisions or judgments at each step)

Qualitative data collection

We selected a total of 11 physicians from three Swedish hospitals for the study. These physicians were identified by reaching out to potential participants through e-mail or phone, with all invited physicians consenting to participate. The inclusion was designed to encompass physicians with diverse specialty experiences to evaluate the interaction dynamics with accessible AI and XAI systems across various domains of medical expertise. Each physician selected had the necessary experience to diagnose the sample patient case presented in the study.

Of the included doctors, nine were consultants (senior doctors), with a minimum of five years of clinical experience. Five doctors had a PhD. Only two of the medical doctors were generalists without a specialty. Given our focus on studying the realistic clinical application of AI, prior experience with AI was not a requirement for inclusion in the study. The demographic details, including age, sex, and professional titles, are outlined in Appendix II. The interviews were performed from October to November 2022.

The qualitative data collection involved semi-structured interviews with the physicians using "think-aloud" protocols (Van Someren, Barnard and Sandberg, 1994), where participants articulated their thought processes while performing diagnoses for us to understand their cognitive patterns and decision-making strategies. We used Zoom for recording, except for one physician who didn't consent. The interviews ranged from 30 to 80 minutes. All interviews were then manually transcribed.

In the initial step, physicians ranked seven risk factors for recurrent ear infections in infants, leading to 77 judgments. They also diagnosed ten patients, totaling 110 diagnoses, with seven out of these ten having actual recurrent ear infections (unknown to the physicians). This approach aimed to make physicians deliberate in their decisions rather than blindly relying on AI and XAI. The emphasis was on reflection, supported by research recommendations such as the strategic application of friction and cognitive forcing functions. Two cognitive forcing functions were used, "making a decision first" and "delaying AI recommendations" (Green and Chen, 2019).

In the second step, AI diagnoses were presented, and physicians had the chance to alter their initial judgments. Another 110 diagnoses were made, with both correct and incorrect AI advice given. The AI accuracy was 60%, but this was undisclosed to physicians, who were only informed that the algorithm was validated according to National Guidelines.

In the third step, XAI explanations were provided, with the weight of different risk factors presented, see Appendix I, Figure A1. After observing these, physicians made another 110 decisions. In total, the study produced 77 cognitive forcing judgments and 330 patient diagnoses.

Qualitative Analysis

A thematic analysis was conducted on our qualitative dataset following Braun and Clarke's methodology (Braun and Clarke, 2006, 2012). This qualitative method was chosen to systematically identify and report patterns within the data, and it consists of six phases: familiarization with the data, generating codes, searching for themes, reviewing themes, defining and naming themes, and writing.

In our thematic analysis, we adhered to a structured approach to ensure the integrity and rigor of the coding process. The primary researcher conducted the initial inductive (bottom-up) coding to establish a comprehensive set of codes derived from the dataset. Following this, the codes and corresponding data extracts were reviewed and discussed with a second senior researcher within the team to mitigate the potential for subjective bias. This second senior researcher brought an additional analytical perspective, thereby enriching the coding process and enhancing the reliability of the interpretation. In instances where there was a discrepancy in coding interpretation between the primary researcher and the second researcher, we engaged a third senior researcher to arbitrate and reach a consensus. This three-tiered review process ensured that each code was critically evaluated, individual biases were minimized, and the coding scheme was robust.

The coding framework was developed through an iterative process. After the inductive coding, we employed a deductive (top-down) approach to align our findings with established theoretical frameworks on cognitive biases and heuristics (Tversky and Kahneman, 1973, 1974; Kahneman, 2011). This approach focused on how heuristics influence XAI-assisted decision-making (Bertrand *et al.*, 2022) and examined the impact of AI/XAI on cognitive challenges encountered by clinicians when the advice either conflicts with or confirms their clinical judgment (Jussupow *et al.*, 2021). Throughout the analysis phase, we engaged in ongoing discussions to ensure the codes were firmly rooted in the data. As themes emerged, they were continuously cross-referenced with the existing literature on cognitive biases and heuristics, guaranteeing that our thematic construction was both data-driven and theoretically informed.

Findings

This section describes the results of our empirical analysis, where we identified four cognitive challenges or themes: i) heuristics in clinical decision-making, ii) heuristics related to trust (aversion versus appreciation) in AI/XAI, ii) heuristics when AI/XAI advice conflict with the physician, and iv) heuristics when AI/XAI confirm the clinical decision. Finally, we synthesize these findings into a conceptual framework of heuristics in human-AI/XAI collaboration, see Figure 2.

Coded by Heuristic/Bias	Description and Application to AI/XAI	Examples of Cognitive Shortcuts
1: Pre-AI heuristics in clinical decision-making		
Availability Heuristic (Tversky and Kahneman, 1973).	How readily things come to mind is interpreted as how likely the outcome is. This can have an impact on clinical diagnostic accuracy. Furthermore, this can increase the acceptance of XAI and its explanations.	"It felt like it [AI] relied a lot on heredity, … our own child with ear problems, and we have absolutely no one else in the family with it" (I, Step 2)) "I had ear issues myself, and so did my son, on the same ear as mine." (C, Step 1) "…parents smoking, is number one. The rationale is that it's bad [for health]." (J, Step 1)
Choice Overload (Scheibehenne, Greifeneder and Todd, 2010).	Limitations of the human mind in handling too many parameters or options lead to an increased reliance on AI support	"The risk factors, the top 3, I trust a lot … those are included in the [human] calculations. … you can't handle too many factors; then you need to automate it and incorporate it into a function or AI." (H, Step 3)
2: Heuristics related to trust (aversion versus appreciation) in AI/XAI or humans		
Ambiguity aversion (Fox and Tversky, 1995)	The preference for known risks over unknown risks, leading to an increase reliance on AI/XAI advice in cases of uncertainty.	"The ones that I really pondered over for a long time, yes, that's |the AI advice] what made me fall over to that side" (A, Step 2)) "Those I marked negative … as you understood, [was] just a guess. I thought they were quite equivalent, so I could very well change my assessment. There are some that I don't want to change." (B, Step 2)
Halo effect (Nisbett and Wilson, 1977) / **Horn effect** (Burton *et al.*, 2015) **Messenger bias** (Dolan *et al.*, 2012)	A positive/negative impression of a human or AI/XAI can give an inflated perception of diagnostic accuracy. A tendency to trust individuals who resemble themselves, leading to greater trust of human decision-making and distrust of AI diagnoses	"At some point, you go to a doctor because you want a human being… Getting an answer solely from a computer, I'm actually not sure how I would feel about that." (G, Step 2) "I still believe it's the doctor who should provide the answer". (E, Step 3) "I just can't simply trust a computer, I feel." (D, Step 2)
Illusion of validity (Kahneman and Tversky, 1973; Tversky and Kahneman, 1974).	A tendency to overestimate our capability to interpret data when the data seems to "tell" a coherent story, leading to an inflated confidence in oneself or in XAI	"Whenever children come to me, I can almost always predict: We will see this child multiple times, and then we do. Or I think: This was nothing, they won't come back, and then they don't. I am rarely surprised." (B, Step 3)
Mere exposure effect (Kliegr, Bahník and Fürnkranz, 2021)	Simply having an explanation can boost trust in AI's prediction.	"It could be any quack claiming to have a great algorithm, and then when you ask what it's based on? They can't reveal that, who would believe in that? The basic attitude is that science is built on transparency. (C, Step 3)

		“If the explanation is based on experience and evidence-based medicine, if I can see and understand that, then I can trust it.” (D, Step 3 arguing against a mere exposure effect)
Pro-innovation bias / technological resistance (Rogers, Singhal and Quinlan, 2019)	The belief that technology (here AI or XAI) is inherently beneficial promoting adoption / inherently harmful creating hesitancy in adoption	“Healthcare in general is extremely conservative, which can be quite frustrating at times… [but] if healthcare was as fickle as the tech industry, it would frequently run off course.” (C, Step 3) “That's probably a generation issue, yes, but the old resistant generation that doesn't quite understand is on its way out.” (E, Step 3)
3: Heuristics when the physician and AI/XAI had conflicting view (Note that aversion and appreciation are a heuristics in itself but they are also a result of other heuristics)		
Algorithm appreciation (Logg, Minson and Moore, 2019) **Automation bias** (Skitka, Mosier and Burdick, 1999)	When people adhere more to advice from an AI algorithm than from a human. A tendency to overly trust AI, often resulting in diminished active reasoning.	”If I had to assess these children, I would just follow the algorithm instead of making wild guesses” (K, Step 3)
Algorithm aversion (Dietvorst, Simmons and Massey, 2015).	When people adhere more to advice from a human than from an AI algorithm.	“Find it hard to believe the algorithm knows more than I do” (F) ”I still believe, unfortunately, that clinical judgment will be crucial, and here we have the biological reality” (I, Step 3)
Commitment bias (Dolan et al., 2012) **Consistency bias** (Cialdini, 2007),	When an individual tends to adhere to a pre-made decision to avoid internal conflict, trusting their initial choice as the optimal one.	”No, it [the AI] place a lot of trust in heredity, I don't.” (I in Step 2) ”I don't believe I'm better [then the AI], but I can't see a reason why I should change.” (J, Step 2)
4: Heuristics when AI/XAI confirmed the physicians’ judgement		
Confirmation bias (Nickerson, 1998)	The tendency to favor information that aligns with one's pre-existing beliefs, expectations, or hypotheses, often disregarding contradictory evidence	“[AI] can be a good tool, especially for those moments when you are a bit hesitant, and then it can feel reassuring to have it as an extra support for decisions.” (A, Step 2) “I felt very satisfied that I got the first two [risk factors] right. Of course, there is value in that. One perceives it as more credible when it aligns with one's own beliefs.” (C, Step3) ”I don't change anything where I was right from the beginning. That seems foolish.” (J, Step 3)

Table 2. Clinical Decision-Making Heuristics Interacting with AI/XAI Support

Pre-AI heuristics in clinical decision-making

Prior to evaluating the heuristics when doctors were given AI support, we sought the presence of heuristics established elsewhere that were not specific to human-AI collaboration. We identified two examples of important heuristics at this stage, as summarized in Table 2 (part 1): availability heuristic and choice overload.

We noticed the commonly found availability heuristic in Step 1, where the physicians made an initial clinical judgment. Several doctors made decisions based on personal experiences of their own children rather than clinical evidence. For instance, Doctor C used experiences from their own child to rank heredity as a major risk factor. At the same time, Doctor I questioned the role of heredity as a significant risk factor based on personal observations with their children. Additionally, Doctor J identified smoking as the most critical risk factor, primarily because of its general negative health associations.

We also noticed the presence of choice overload, or the limitations of the human mind in processing a large number of parameters (see Table 2, part 1). Doctor H argued for only using the top three most important risk factors, the rest were of no importance. However, the doctor argues that AI could be a useful tool to limit choice overload. *In one's mind, you can't handle too many factors; then you need to automate it and incorporate it into a function or AI."*

Heuristics related to trust in AI/XAI versus trust in humans

We noticed examples of several heuristics related to trust in an AI or XAI advice but also heuristics related to trust in humans, such as ambiguity aversion, halo and horn effect, the illusion of validity, mere exposure effect, and pro-innovation bias, see Table 2 (part 2).

For clinicians to use AI algorithms, it is crucial for them to trust them. Our study showed that when the explanations were provided for the AI prediction, trust and intention to use increased. When Doctor B was offered to make changes based on the black-box AI in Step 2, the doctor argued, *"it depends on whether I trust the algorithm or not*." Doctor I commented, "*I probably wouldn't trust AI at all if I didn't know the parameters it operates on. I actually place a lot of trust in my clinical assessment."*

Mere exposure effect was observed when just the presence of explanations created trust even if they were not used. All physicians unanimously stated that they had more trust in XAI compared to black box AI. Doctor C argued that without a clear understanding of the parameters of AI, there is potential for skepticism, comparing unexplained AI predictions to "hocus-pocus "*It could be any quack claiming to have a great algorithm, and then when you ask what it's based* *on, they can't reveal that. Who would believe in that?".* We also noticed counterarguments for the mere exposure effect where Doctor D points out that explanations are not enough; they also must be in line with evidence-based medicine: *"If the explanation is based on experience and evidence-based medicine, if I can see and understand that, then I can trust it."*

The horn effect was indicated when Doctor G declared skepticism about trusting a machine "*Getting an answer solely from a computer, I'm actually not sure how I would feel about that."* At the same time, Doctor G seemed to put a halo on the human physician (the halo effect), arguing that humans with emotions are preferred to a computer;" o*ne loses the depth that an emotional being might possess*." A similar heuristic is messenger bias, illustrated by Doctor E, who suggested that when AI/XAI is used, the human physician should be the messenger: "*I still believe it's the doctor who should provide the answer*".

Furthermore, this study identified that trust in the AI/XAI increased in cases where the physicians were in doubt, indicating an *ambiguity aversion*. Doctor A chose to use the advice for cases of uncertainty "*The ones that I really pondered over for a long time, yes, that's what made me fall over to that side, one might say."* Doctor B argues similarly, accepting the AI advice for ambivalent cases, "*I can say that those I marked negative, and it was, as you understood, just a guess. I thought they were quite equivalent, so I could very well change my assessment. There are some that I don't want to change*."

The opposite of pro-innovation bias, "technological resistance," was also observed when Doctor C argued that "healthcare, in general, is extremely conservative." Doctor E argued that technological resistance is a generation issue: *"That's probably a generation issue, yes, but the old resistant generation that doesn't quite understand is on its way out.".*

Heuristics when the physician and the AI/XAI were in conflict

When the AI/XAI is in conflict with the clinician's judgment, the cognitive challenge for the physician is to determine if the physician is correct or the AI/XAI. Explanations are suggested as an intervention and act as a help for the physician in this assessment. However, this is not only a rational assessment; we identified a number of heuristics that can produce both good and bad outcomes, including algorithm aversion and appreciation, that are heuristics in themselves. In addition, signs of the more important automation bias and commitment or consistency bias were registered, see Table 2, part 3.

The tendency to prefer human advice over AI advice, i.e., algorithm aversion, was higher for black-box AI, but there was also aversion for the explanations. When the black-box AI advice was presented to Doctor D, the lack of transparency was key to not using the advice: "*I don't feel like I change because I don't understand why it wants so many positives".* Doctor I also shows algorithm aversion for both AI and XAI. "*I still believe, unfortunately, that clinical judgment will be crucial, and here we have the biological reality. It's not possible to confirm* *anything one way or the other with more parameters in this case. But now, being purely clinical, I might be a bit pessimistic*." Doctor F was even more resistant, with blind distrust for both AI and XAI. "*I trust my own clinical assessment".* The doctor elaborated further, "*Find it hard to believe the algorithm knows more than I do".*

When the decision-makers preferred AI advice over human advice, i.e., exhibited algorithm appreciation, this heuristic was dualistic in nature. Appreciation is only preferable in cases where the AI/XAI is correct, and the clinician is wrong. If the AI is wrong, the optimal outcome would be that the physician maintains their correct judgment. Doctor K had limited experience with young children with recurrent middle ear infections and hence appreciated a potential automated decision, indicating an automation bias: "*If I had to assess these children, I would just follow the algorithm instead of making wild guesses. These 10 children with ear issues are more than I have examined during my time as a doctor*."

The study also noticed a third important heuristic, commitment bias or consistency bias. This cognitive bias potentially impedes physicians' receptiveness to new, contradicting advice from AI, as it may clash with their prior clinical judgments. Doctor I stood fast on the clinical diagnosis when in conflict with the Black-box AI, "No*, it [the AI] places a lot of trust in heredity, I don't. Yes, if I were to be logical in my statement, I would stick to the negative viewpoint."* Doctor J is also committed to the initial clinical diagnosis "*I don't believe I'm better [than the AI], but I can't see a reason why I should change*." Again, this heuristic is dualistic in nature and only beneficial when the AI is incorrect.

When the physicians were provided with explanations, the intention to use doubled (algorithm appreciation) and the diagnostic accuracy improved. However, since we provided the physicians with both correct and incorrect AI/XAI advice (AI accuracy was set to 60%, which was blinded for the physicians), most of the improvements (with correct AI advice) were offset when the AI advice was incorrect. However, a more significant and subtle diagnostic mistake occurred when the AI wrongly confirmed the physician's incorrect assessment.

Heuristics when the AI/XAI confirmed the physicians' judgment

When there were conflicting views, it became natural for the physicians to elaborate on why. However, the errors due to the AI/XAI falsely confirming an incorrect clinical judgment were silently accepted, revealing a confirmation bias, see Table 2, part 4. Doctor E exemplified the difference between conflict and confirmation:"*If one has a hypothesis and the AI thinks the same, then maybe it's more accurate. But if the AI says no, well then maybe one has to consider some additional inputs*." It comes naturally to stand fast when confirmed. When confirmed, basically all physicians held on to the initial incorrect judgment. However, the AI accuracy was set at 60% in the study. Hence, the given advice was, in many cases, incorrect. The statement "*then maybe it's more accurate"* opened the door for false confirmation errors. Essentially, no one investigated the case that "*maybe*" both were wrong.

We noticed that when the AI/XAI validated a physician's incorrect perspective, it led to reduced or even no diligence at all in further investigations. Doctor J exemplifies this when making no further investigation when confirmed, "*I don't change anything where I was right from the beginning. That seems foolish*." Being confirmed is not the same as being correct. From our qualitative perspective, a false confirmation is a much more subtle error than when in conflict with the AI.

The false confirmation is not only an error that, for most cases, is undetected but also a source of potential overreliance leading to new errors. In the study, we found that when physicians were confirmed, it increased their trust in the AI/XAI. When Doctor C was provided with the XAI ranking list of risk parameters, trust in the algorithm increased. "*I felt very satisfied that I got the first two [risk factors] right. Of course, there is value in that. One perceives it as more credible when it aligns with one's own beliefs."* Relying too much on an imperfect AI/XAI can create a dangerous cycle. Many false confirmations can increase trust in the system, leading to additional false conflict errors where the physicians are convinced to change from a correct clinical judgment to an incorrect A/XAI advice.

A conceptual framework of heuristics in human-AI/XAI collaboration

To synthesize the findings of this qualitative study, we developed a conceptual framework that outlines underlying heuristics impact on algorithm aversion and appreciation, as illustrated in Figure 2.

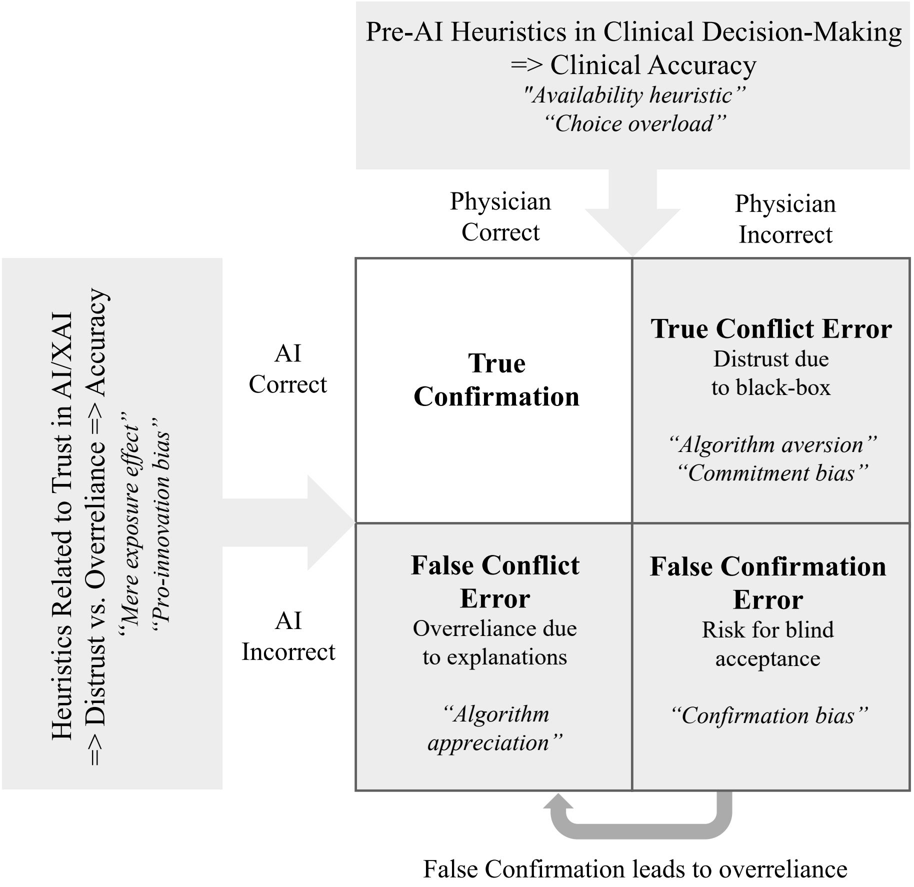


Figure 2: A conceptual framework of heuristics and errors in clinical decision-making with AI/XAI advice.

In line with previous research (Jussupow *et al.*, 2021) we reviewed the cognitive challenges when AI either conflict the physicians diagnosis (either physicians are correct and AI is incorrect and vice versa) or when it confirms (either both are correct or both incorrect). The framework demonstrates how various pre-AI heuristics affect clinical judgments and diagnostic accuracy, as indicated by the top box and arrow in the framework (refer to section 4.1, and Table 2 part 1). Additionally, numerous underlying heuristics shape clinicians' attitudes and trust toward the AI algorithm and its explanations (with potential algorithm aversion or appreciation as an outcome), as shown by the left box and arrow (refer to section 4.2 and Table 2, part 2). These two sets of heuristics significantly influence diagnostic accuracy and contribute to three distinct sources of error (Jussupow *et al.*, 2021): clinicians overriding correct AI recommendations (refer to section 4.3 and Table 2, part 3), clinicians altering accurate clinical diagnoses to incorrect ones based on erroneous AI advice (refer to section 4.3), and instances where the AI erroneously confirms an incorrect clinical diagnosis (refer to section 4.4 and Table 2, part 4). We elaborate further in the following discussion.

Discussion

In this discussion section, we begin by summarizing our findings and discussing them in relation to our conceptual framework, as illustrated in Figure 2. We then examine our contributions in the context of existing literature and explore their potential implications for practice. Finally, we address the study's limitations and offer suggestions for future research.

Physicians employ heuristics to arrive at clinical diagnoses; these heuristics can sometimes enhance efficiency and other times compromise diagnostic precision. We found that clinicians relied on availability heuristics (judging likelihood by the ease with which examples come to mind) and they limited the number of risk parameters they considered, a sign of "choice overload." These heuristics can potentially reduce clinical accuracy, while XAI explanations can potentially educate physicians on which risk parameters to focus on.

We have identified three different sources of error in AI/XAI augmented decision-making, increasing the cognitive complexity. i) True Conflict Errors: When AI is correct, and the physician is committed to their incorrect clinical judgment (algorithm aversion has a negative outcome). ii) False Conflict Errors: When the physician is correct but is convinced by the incorrect AI and/or its explanations (algorithm appreciation has a negative outcome). iii) False Confirmation Error: The physician makes an incorrect clinical judgment, and the AI/XAI confirms this incorrect diagnosis.

Our research recognized multiple heuristics that affect physicians' trust in and intention to adopt AI/XAI recommendations. We argue that these trust-related heuristics are the underlying reason for algorithm aversion and algorithm appreciation. The simple provision of explanations tends to increase their trust in the algorithm (leads to algorithm appreciation), which potentially can reduce True Conflict Errors. Additionally, a notable number of physicians exhibited a "halo effect" towards human judgment while assigning a "horn effect" to algorithmic advice (leads to algorithm aversion), particularly in the case of black-box AI systems. In situations of True Conflict, where the AI is accurate, and the physician is not, algorithm appreciation (accepting AI/XAI advice) can be beneficial. In contrast, algorithm aversion reduces False Conflict Errors while it increases True Conflict Errors.

Finally, our study also spotlighted an important source of error in human-AI collaboration: namely, when AI/XAI erroneously affirms an incorrect clinical judgment, or a False Confirmation. Physicians tend to scrutinize the basis of any discrepancy in cases of conflict. However, when AI/XAI seemingly validates the initial clinical assessment, physicians almost uniformly accept the concurrence without question, presupposing the correctness of both parties. This unquestioned acceptance can be problematic, as both AI/XAI and physician assessments may be incorrect, consequently diminishing diagnostic precision. We argue that this is not algorithm appreciation but a much deeper heuristic akin to confirmation bias (Nickerson, 1998). Whereas confirmation bias traditionally involves actively seeking or interpreting evidence to support one's existing beliefs or hypotheses, we observed a that physicians may accept confirmatory information without critical engagement or verification. This is not an active distortion of evidence, as with confirmation bias, but seemingly a rather passive acceptance of it. This reinforcement could perversely lead to a harmful cycle where increased trust (algorithm appreciation) potentially leads to further acceptance of incorrect AI/XAI advice and False Conflict Errors.

Contributions to previous literature and practice

Our research is consistent with prior studies, focusing on ability to rectify physicians' mistakes (True Conflict Errors) (Jussupow *et al.*, 2021), but moves beyond it to identify how AI can induce physician mistakes. Importantly, consistent with prior research (Bertrand *et al.*, 2022), we argue that in comprehending cognitive challenges, we need to consider not only reason-based but also heuristic influences on clinical decision-making.

This approach yields several important contributions. First, we move beyond identifying algorithmic aversion, when people adhere more to advice from a human than from an AI algorithm (Dietvorst, Simmons and Massey, 2015), and algorithm appreciation, when people adhere more to advice from a human than from an AI algorithm (Logg, Minson and Moore, 2019). Our study is consistent with the notion that these behaviours manifest from underlying heuristics like the halo/horn effect, messenger bias, pro-innovation bias, ambiguity aversion, or consistency bias, as some examples. Second, while past healthcare literature has focused on certain heuristics, like the availability heuristic, ours identifies a series which are specifically pertinent to human-AI collaboration.

Third, by testing the role of XAI, in which explanations accompany AI's "black-box" diagnoses, we found that although this added persuasiveness of explainability can enhance understanding and potentially decrease True Conflict Errors, perversely it can introduce risks of False Conflict Errors. The explanations affect heuristic decision-making as well, potentially increasing trust due to effects such as "mere exposure effect" (Kliegr, Bahník and Fürnkranz, 2021) or creating a "halo effect" that may lead to overreliance. Only a few studies (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023), have drawn attention to the cognitive difficulties associated with False Conflict Errors when AI sways physicians towards incorrect decisions.

Fourth, we identify the crucial importance of "False Confirmation", which prior researchers have highlighted but largely neglected empirically (Jussupow *et al.*, 2021; Naiseh *et al.*, 2023). These errors seem to be much more difficult to cognitively identify and mitigate. Future research is needed to better understand the extent to which these errors generalize across healthcare domains.

Fifth, our research highlights a need to optimize trust. Thus far only a handful of studies have tested interventions to optimize trust (avoiding distrust and overreliance) in XAI in healthcare

(Buçinca, Malaya and Gajos, 2021; Naiseh et al., 2023). To optimize trust, one intervention is cognitive forcing, which aims to disrupt heuristic System 1 thinking, prompting analytical (System 2) thinking (Lambe et al., 2016). Such interventions “force” physicians to think hard. They have successfully been tested with XAI; examples include checklists, diagnostic time-outs, or asking the decision-maker to make a clinical assessment before seeing the AI diagnosis and its explanations (Buçinca, Malaya and Gajos, 2021; Naiseh et al., 2023). However, despite that we used cognitive forcing functions in this study we could not see any indication that it reduced the “False Confirmation Bias”.

However, optimal trust is a narrow term since it depends on both accuracy in clinical judgment and AI accuracy. With very high AI accuracy and relatively low clinical accuracy, optimal trust should be relatively high since it reduces True Conflict Errors and adds a few False Conflict and False Confirmation Errors. While if AI accuracy is modest, low-performing physicians can still benefit from trusting the algorithm. However, physicians who perform better than AI in the clinic should not trust the algorithm. In theory, trust optimization must be considered individually and from case to case, which might be challenging in practice. The most challenging situation is if both the AI and the physician have a relatively low accuracy since, in many cases, both will be wrong, and the False Confirmation Bias seems very difficult to identify and mitigate in practice.

Limitations and future research

Our qualitative field study underscores the cognitive complexities of discerning the three potential errors arising from physicians' reliance on AI/XAI advice. Further empirical work is necessary to measure these errors and explore intervention strategies that can mitigate heuristics and cognitive errors with a special focus on False Conflict and False Confirmation errors.

First, we did not aim to evaluate quantitative decision-making accuracy in this study, as this has been done previously elsewhere (Rosenbacke, 2024). Second, we have not sought to determine causally the presence of specific heuristics but rather to demonstrate the common decision-making patterns and justifications which are likely to be consistent with their presence. Third, our research is unable to differentiate the behavioural outcomes consistent with algorithmic aversion/appreciation from underlying cognitive beliefs or states which give rise to it. Future research would need to be designed in order to disentangle the behavioural outcomes from underlying attitudes and dispositions, as they are likely to be highly correlated and thus necessitate an experimental approach.

Fourth, there may need to be interventions to optimize physician trust to AI’s level of accuracy. Thus, low accuracy may involve precipitating a reduction in trust levels, and vice-versa for high accuracy. This applies not only to AI, but also clinician’s baseline performance, as clinicians with very high accuracy may benefit from lower AI trust. Our study was based on 60% accuracy but to better test these possibility researchers could manipulate directly AI accuracy in their study.

Fifth, we found evidence consistent with the notion that doctors fail to detect False Confirmation risks. Future research could explore an opt-in approach for instances where the physician's and AI's assessments align. Upon confirmation, a prompt could appear asking the physician, “Are you sure that both are not wrong?” with options "Yes" or "No." Further studies could also investigate if this can potentially create a new fourth error, True Confirmation Error, where physicians are correct and AI is correct. Due to the question “if both can be wrong,” they might change a correct judgment to an incorrect one.

We have noticed several heuristics that impact diagnostic accuracy in human-AI collaboration. However, heuristics is seen as an intuitive judgment, and this judgment can likely be nudged as described by Richard Thaler and colleagues (Thaler and Sunstein, 2008). Further research could investigate explanations and cognitive forcing functions as nudges. If AI accuracy is high, maybe explanations should be provided only because the “mere exposure effect” is a nudge that potentially increases trust. Furthermore, researchers could investigate default bias (Johnson and Goldstein, 2003), where the order of clinical diagnosis versus AI/XAI advice can impact physicians' trust. Is it important if the clinical judgment is first or after AI? If the clinical judgment is first, will this create a commitment bias?

Conclusion

Taken together, our study sheds light on the complex dynamics of heuristics in human-AI/XAI collaboration within clinical settings. We move beyond existing studies which characterize “aversion” or “appreciation” to identify decision-making heuristics that link to decision-making patterns and, in several cases, errors. These include, among others: "commitment bias" when physicians cling to their initial incorrect clinical diagnoses, and the "mere exposure effect" whereby physicians, uncertain of their judgments, shift to an erroneous AI recommendation. We also identify a pernicious challenge of "False Confirmation" when AI affirms an incorrect initial diagnosis. These findings will help guide future research to enhance the uptake and performance of human-AI collaboration in medical decision-making.

Acknowledgments: The author extends gratitude to all the physicians who contributed to this study. I would like to express my sincere gratitude to David Stuckler and Åsa Melhus for their invaluable roles as sounding boards and for their meticulous review of this paper.

Appendix I: Patient Dataset and the Explainable AI Algorithm

The data originates from a vaccination trial at the Department of Otorhinolaryngology, Lund University Hospital, Sweden, which adhered to randomized, prospective, and single-blinded methodologies with ethical approval and parental consent. The study resulted in two published articles. (Gisselsson-Solén et al., 2014, 2015).

We used a Random Forest technique (*API Reference — scikit-learn 1.1.3 documentation*) to process the dataset with AI algorithms. For XAI, we employed the open-source SHAP code (Lundberg, 2022). A recent systematic review of applications for XAI in healthcare found that SHAP is the most widely utilized XAI technique (Loh *et al.*, 2022).

The patient dataset included seven input parameters, including family history of middle ear infections, siblings count, daycare attendance, breastfeeding, parental smoking, pre-study ear infections, and pneumococci vaccination status. The output identified children with four or more ear infections by the 12-month mark, factoring in prior infection counts. Figure A1 displays the average SHAP values for these parameters, with bar lengths indicating the parameter's impact on infection risk.

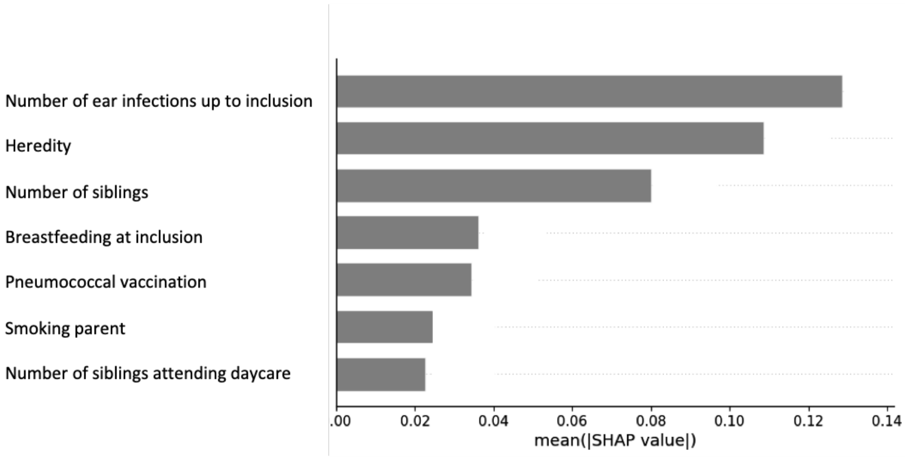


Figure A1: XAI bar plot of risk factors for recurring ear infection.

Appendix II: Participating physicians

Age	Frequency	Percentage
30-39	5	45%
40-49	0	0%
50-59	1	9%
60-69	5	45%
Sex		
Male	9	82%
Female	2	18%
Highest level of experience		
Medical doctor (MD)	2	18%
MD + consultant	4	36%
MD + Doctor of Philosophy + senior consultant	2	18%
Professor + senior consultant	3	27%

Table A1: Participating physicians

References

Amann, J. *et al.* (2020) 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective', *BMC Medical Informatics and Decision Making*. BioMed Central Ltd, 20(1), pp. 1–9.

Antoniadi, A. M. *et al.* (2021) 'Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review', *Applied Sciences (Switzerland)*. MDPI AG, 11(11). doi: 10.3390/app11115088.

'API Reference' (2006) in *The Definitive Guide to MySQL5*, pp. 693–720. doi: 10.1007/978-1-4302-0071-0_23.

Austin, J. P. and Foster, B. A. (2019) 'How pediatric hospitalists must contend with the expert halo effect', *Hospital Pediatrics*. doi: 10.1542/hpeds.2019-0053.

Bauer, K., von Zahn, M. and Hinz, O. (2023) 'Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing', *Information Systems Research*, 34(4). doi: 10.1287/isre.2023.1199.

Bertrand, A. *et al.* (2022) 'How cognitive biases affect XAI-Assisted decision-making: A systematic review', in *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101. doi: 10.1191/1478088706QP063OA.

Braun, V. and Clarke, V. (2012) 'Thematic analysis.', in *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, pp. 57–71. doi: 10.1037/13620-004.

Burton, J. W., Stein, M. K. and Jensen, T. B. (2019) 'A systematic review of algorithm aversion in augmented decision making', *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.2155.

Burton, S. *et al.* (2015) 'Broken halos and shattered horns: overcoming the biasing effects of prior expectations through objective information disclosure', *Journal of the Academy of Marketing Science*, 43(2). doi: 10.1007/s11747-014-0378-5.

Chew, H. S. J. and Achananuparp, P. (2022) 'Perceptions and Needs of Artificial Intelligence in Health Care to Increase Adoption: Scoping Review', *Journal of Medical Internet Research*. doi: 10.2196/32939.

Cialdini, R. B. (2007) *Influence: The psychology of persuasion*, *New York, NY, USA: HarperCollins Publishers.* Collins New York. doi: 10.1017/CBO9781107415324.004.

Cui, M. and Zhang, D. Y. (2021) 'Artificial intelligence and computational pathology', *Laboratory Investigation*. doi: 10.1038/s41374-020-00514-0.

Cummings, L. (2014) 'The "Trust" Heuristic: Arguments from Authority in Public Health', *Health Communication*, 29(10). doi: 10.1080/10410236.2013.831685.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000033.

Dolan, P. *et al.* (2012) 'Influencing behaviour: The mindspace way', *Journal of Economic Psychology*. doi: 10.1016/j.joep.2011.10.009.

Evans, T. *et al.* (2022) 'The explainability paradox: Challenges for xAI in digital pathology', *Future Generation Computer Systems*, 133. doi: 10.1016/j.future.2022.03.009.

Fazal, M. I. *et al.* (2018) 'The past, present and future role of artificial intelligence in imaging', *European Journal of Radiology*. doi: 10.1016/j.ejrad.2018.06.020.

Fox, C. R. and Tversky, A. (1995) 'Ambiguity aversion and comparative ignorance', *Quarterly Journal of Economics*, 110(3). doi: 10.2307/2946693.

Gisselsson-Solén, M. *et al.* (2014) 'Risk factors for carriage of AOM pathogens during the first 3 years of life in children with early onset of acute otitis media', *Acta Oto-Laryngologica*. Informa Healthcare, 134(7), pp. 684–690. doi: 10.3109/00016489.2014.890291.

Gisselsson-Solén, M. *et al.* (2015) 'Effect of pneumococcal conjugate vaccination on nasopharyngeal carriage in children with early onset of acute otitis media-a randomized controlled trial', *Acta Oto-Laryngologica*. Informa Healthcare, 135(1), pp. 7–13. doi: 10.3109/00016489.2014.950326.

Giuste, F. *et al.* (2023) 'Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review', *IEEE Reviews in Biomedical Engineering*, 16.

Green, B. and Chen, Y. (2019) 'The principles and limits of algorithm-in-the-loop decision making', *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery

Johnson, E. J. and Goldstein, D. (2003) 'Do Defaults Save Lives?', *Science*. doi: 10.1126/science.1091721.

Jussupow, E. *et al.* (2021) 'Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence', *Information Systems Research*, 32(3).

Kahneman, D. (2011) *Thinking fast, thinking slow*, *Interpretation, Tavistock, London*.

Kahneman, D. and Tversky, A. (1973) ‘On the psychology of prediction’, *Psychological Review*. doi: 10.1037/h0034747.

Kliegr, T., Bahník, Š. and Fürnkranz, J. (2021) ‘A review of possible effects of cognitive biases on interpretation of rule-based machine learning models’, *Artificial Intelligence*, 295. doi: 10.1016/j.artint.2021.103458.

Kundu, S. (2021) ‘AI in medicine must be explainable’, *Nature Medicine*, 27(8). doi: 10.1038/s41591-021-01461-z.

Lewicki, R. J. and Brinsfield, C. T. (2011) ‘Framing trust: Trust as a heuristic’, in Donohue, W. A., Rogan, R. R., and Kaufman, S. (eds) *Framing matters: Perspectives on negotiatin research and practice in communication*. Peter Lang Publishing, pp. 110–135.

Logg, J. M., Minson, J. A. and Moore, D. A. (2019) ‘Algorithm appreciation: People prefer algorithmic to human judgment’, *Organizational Behavior and Human Decision Processes*. doi: 10.1016/j.obhdp.2018.12.005.

Logg Jennifer (2018) ‘Do People Trust Algorithms More Than Companies Realize?’, *Harvard Business Review*.

Loh, H. W. *et al.* (2022) ‘Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)’, *Computer Methods and Programs in Biomedicine*. doi: 10.1016/j.cmpb.2022.107161.

Lundberg, S. (2022) *Lundberg/SHAP*, *GitHub*. Available at: https://github.com/slundberg/shap

Ly, D. P. (2021) ‘The Influence of the Availability Heuristic on Physicians in the Emergency Department’, *Annals of Emergency Medicine*, 78(5). doi: 10.1016/j.annemergmed.2021.06.012.

Madsen, M. and Gregor, S. (2000) ‘Measuring Human-Computer Trust’, *Proceedings of Eleventh Australasian Conference on Information Systems*.

Mahmud, H. *et al.* (2022) ‘What influences algorithmic decision-making? A systematic literature review on algorithm aversion’, *Technological Forecasting and Social Change*, 175. doi: 10.1016/j.techfore.2021.121390.

Naiseh, M. *et al.* (2023) ‘How the different explanation classes impact trust calibration: The case of clinical decision support systems’, *International Journal of Human Computer Studies*, 169. doi: 10.1016/j.ijhcs.2022.102941.

Nazar, M. *et al.* (2021) ‘A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques’, *IEEE Access*. doi: 10.1109/ACCESS.2021.3127881.

Nickerson, R. S. (1998) ‘Confirmation bias: A ubiquitous phenomenon in many guises’, *Review of General Psychology*. doi: 10.1037/1089-2680.2.2.175.

Nisbett, R. E. and Wilson, T. D. (1977) ‘The halo effect: Evidence for unconscious alteration of judgments.’, *Journal of Personality and Social Psychology*. doi: 10.1037/0022-3514.35.4.250.

Rajpurkar, P. *et al.* (2022) ‘AI in health and medicine’, *Nature Medicine*. doi: 10.1038/s41591-021-01614-0.

Rogers, E. M., Singhal, A. and Quinlan, M. M. (2019) ‘Diffusion of innovations’, in *An Integrated Approach to Communication Theory and Research, Third Edition*. doi: 10.4324/9780203710753-35.

Rosenbacke, R. (2024) ‘Errors in Physician-AI Collaboration: Insights From a Mixed-methods Study of Explainable AI and Trust in Clinical Decision-making’, *SSRN Electronic Journal*. doi: 10.2139/SSRN.4773350.

Scheibehenne, B., Greifeneder, R. and Todd, P. M. (2010) ‘Can there ever be too many options? A meta-analytic review of choice overload’, *Journal of Consumer Research*, 37(3). doi: 10.1086/651235.

Skitka, L. J., Mosier, K. L. and Burdick, M. (1999) ‘Does automation bias decision-making?’, *International Journal of Human Computer Studies*. doi: 10.1006/ijhc.1999.0252.

Van Someren, M., Barnard, Y. F. and Sandberg, J. (1994) ‘The think aloud method: a practical approach to modelling cognitive’, *London: AcademicPress*, 11.

Thaler, R. H. and Sunstein, C. R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. doi: 10.1016/s1477-3880(15)30073-6.

Tversky, A. and Kahneman, D. (1973) ‘Availability: A heuristic for judging frequency and probability’, *Cognitive Psychology*. doi: 10.1016/0010-0285(73)90033-9.

Tversky, A. and Kahneman, D. (1974) ‘Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty’, *Science*. doi: Cited By (since 1996) 3914\nExport Date 30 November 2011.

Appendix IV: Paper IV

Published in *British Medical Journal - J. Medical Ethics.* This is the peer-reviewed and accepted version. There can be deviations after copyediting and publisher layout.

Citation: Rosenbacke, R.; Melhus, Å.; Mckee, M.; Stuckler, D (2024) 'The AI and XAI Second Opinion: The Danger of False Confirmation in human-AI Collaboration' *British Medical Journal - J. Med. Ethics* (2024). doi:10.1136/jme-2024-110074.

The AI and XAI Second Opinion

The Danger of False Confirmation in human-AI Collaboration

Rikard Rosenbacke[1*], Åsa Melhus[2], Martin McKee[3], David Stuckler[4]

[1]Centre for Corporate Governance, Department of Accounting, Copenhagen Business School, Copenhagen, Denmark
[2]Department of Medical Sciences/Section of Clinical Microbiology, Uppsala University, Uppsala, Sweden
[3]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, United Kingdom
[4]Department of Social and Political Science, Bocconi University, Milano, Italy

Running title:" "The AI and XAI Second Opinion"

* - corresponding author; rr.ccg@cbs.dk, Rikard Rosenbacke, Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg

Word Count: 2496

Keywords: Explainable Artificial Intelligence, XAI, AI, Second Opinion

Abstract

Can AI substitute a human physician's second opinion? Recently the *Journal of Medical Ethics* published two contrasting views: Kempt and Nagel advocate for using AI for a second opinion except when its conclusions significantly diverge from the initial physician's, while Jongsma and Sand argue for a second human opinion irrespective of AI's concurrence or dissent. The crux of this debate hinges on the prevalence and impact of 'false confirmation' – a scenario where AI erroneously validates an incorrect human decision. These errors seem exceedingly difficult to detect, reminiscent of heuristics akin to confirmation bias. However, this debate has yet to engage with the emergence of explainable AI (XAI), which elaborates on why the AI tool reaches its diagnosis. To progress this debate we outline a framework for conceptualising decision-making errors in physician-AI collaborations. We then review emerging evidence on the magnitude of false confirmation errors. Our simulations show that they are likely to be pervasive in clinical practice, decreasing diagnostic accuracy to between 5% and 30%. We conclude with a pragmatic approach to employing AI as a second opinion, emphasizing the need for physicians to make clinical decisions before consulting AI; employing nudges to increase awareness of false confirmations; and critically engaging with XAI explanations. This approach underscores the necessity for a cautious, evidence-based methodology when integrating AI into clinical decision-making.Introduction

Sometimes patients or their families may request a second medical opinion. There are several reasons: to confirm a diagnosis, to explore other treatment options, to confirm what they have been told, or because they have lost confidence in their clinical team. A recent systematic review found high levels of patient satisfaction with second opinions and, while the figures varied among clinical areas, relatively high frequencies of changes in diagnosis or treatment [1]. However, second opinions are not always advantageous. They can lead to distress, delay treatment, and interrupt continuity of care [2].

While second opinions have always been possible, demand has increased, reflecting patient empowerment and greater access to medical information on the internet [3], which may or may not be accurate or relevant to the case in question. Some health systems or plans include an explicit entitlement to a second opinion in certain circumstances, such as the recently announced Martha's Rule in the English National Health Service, named after a young girl who died after a failure to seek such an opinion [4].

The process of a physician-initiated second opinion has traditionally involved consulting another physician, often with specialized expertise, to review and potentially challenge the initial clinical decision. This collaborative approach ensures a thorough exploration of diagnostic possibilities, with a focus on ensuring the highest standard of patient care. However, the increasing use of Artificial Intelligence (AI) in health care offers a potential alternative, albeit one that raises significant ethical and practical implications [5].

The *Journal of Medical Ethics* published two contrasting views of how AI could fulfill the role of a second opinion traditionally reserved for medical professionals. Kempt and Nagel [5] argue for a "rule of disagreement", by which AI can provide a second opinion. When it concurs with the initial physician assessment, no further action is required, but when it differs substantially, another human opinion is imperative. However, Jongsma and Sand [6] disagree, arguing that there is “symmetry in the burden of proof” in both agreement and disagreement. They emphasise the inherent fallibility of both human and AI judgements, advocating a second human opinion regardless of AI’s concurrence or dissent.

Underlying this debate is uncertainty about the prevalence and impact of ‘false confirmation’ – a scenario where the AI erroneously validates an incorrect human decision. Very limited data exists on the scope or scale of such occurrences [7,8]. Our own preliminary empirical research [9] and simulations reveal false confirmation rates ranging from 5% to as high as 30%, underscoring the potential hazard of relying solely on AI for secondary consultations. Perhaps more concerning is that in our study, all physicians accepted the incorrect confirmation from AI without voicing any concern that both could be wrong [9]. This type of error appears very difficult to detect; it seems to be a cognitive shortcut, reminiscent of the heuristic known as confirmation bias [10], where one accepts confirmatory information without question.

Here we argue that the threat from false confirmation in AI-assisted medical decision-making could be substantial, echoing the concerns raised by Jongma and Sand [6]. Yet we posit that this does not preclude the role of AI as a valuable tool for second opinions, especially in light of recent advances in so-called explainable AI (XAI). XAI offers a platform that supports a more detailed discussion and analysis of AI-driven diagnoses, thereby enriching the decision-making process, including interrogating the underlying clinical rationale for convergent or divergent diagnoses.

We begin by establishing a conceptual framework to analyse decision-making errors within the context of physician-AI collaboration, with focus on the phenomenon of false confirmation. Then, we delve into recent empirical studies assessing joint physician-AI decision-making, revealing a predominant adherence to the “rule of disagreement” and a concerning lack of acceptance of the potential for false confirmation. A subsequent simulation illustrates the high incidence of false confirmation at varying levels of AI and physician accuracy, highlighting the magnitude of these decision-making errors in clinical practice. Finally, we explore insights from cognitive psychology, particularly decision-making rules and techniques such as choice architecture and nudging, to propose strategies that could mitigate the risk of false confirmation when leveraging AI as a second opinion in clinical practice.

Framework to identify errors in human-AI/XAI decision-making

We can differentiate three distinct errors in joint physician-AI decision-making, to which we have assigned the following terminology: true conflict error – when the physician is incorrect but AI is correct; false conflict error – when the physician is correct but AI is incorrect; and false confirmation error – when the physician and AI agree but both are wrong[9]. (See Table 1).

	Physician right	Physician wrong
AI right	Correct	True conflict error
AI wrong	False conflict error	False confirmation error

Table 1. Potential sources of error in human-AI/XAI collaboration [9] .

Recent studies are beginning to identify how physicians-in-charge respond to these agreements or disagreements with AI. Turning to the first error, a recent study by Rosenbacke [9] found that, in cases of true conflict error, physicians tended to override a correct AI diagnosis. Previous studies found that this arises from distrust in the logic hidden in the 'black box' of AI [9,11–13].

In cases of false conflict errors, however, the physicians tended to express doubt and over-rely upon AI, especially when they felt uncertain about their own initial diagnosis. When an explanation was added as to why AI reached a diagnosis (as in XAI), it tended to mitigate true conflict errors, but exacerbate false conflict errors. This phenomenon whereby even the mere exposure to explanations can induce overreliance on AI has been now documented in several studies [8,9,14–16].

It is the third error which is perhaps more concerning. When physicians and AI concurred, doctors seemed to accept the diagnoses with little critical questioning. This happened irrespective of whether or not an explanation was provided with the AI diagnosis [9]. This echoes the concern raised by Jongsma and Sand [6], which called for a third opinion in all cases even though this approach appeared to question some of the arguments for using AI, i.e. saving scarce clinical resources.

Are these false confirmations rare events, or recurring features of human-AI collaborations? Next, we review emerging empirical evidence on their frequency.

Empirical evidence of the frequency of false confirmations

Our recent empirical analyses [9] begin to shed light on the magnitude of false confirmation. We investigated decisions made by 11 physicians-in-charge of 10 patients with a possible diagnosis of a recurrent middle ear infection, with and without AI and XAI support. The AI system was calibrated to make incorrect diagnoses in 40% of the cases. In total, 22% of all diagnoses were subject to false confirmation. In all these instances, the physicians accepted the AI confirmation without questioning it. Introducing XAI, which should help physicians understand whether their logic was consistent or not with that of AI, made little difference, encouraging the physicians to question the clinical diagnosis in only one case out of 48.

This challenges the argument posited by Kempt & Nagel that "*If the AI-DSS [AI-Decision Support System] confirms the initial opinion, however, no further steps need to be taken.*"[5]. As our physicians noted, when discussing their thought process, "*I don't change anything where I was right from the beginning. That seems foolish.*" [9]. Seemingly, AI confirmation can increase the physician's confidence in their diagnosis, even when they are incorrect. Hence, Jongsma and Sand argue that the physician-in-charge must still justify their acceptance of confirmation of the diagnosis "*If physicians do not want to naively have their views confirmed, they have to justify why they consider the AI systems' output as a confirmation*" [6].

This tendency, to believe that an AI confirmation is the same thing as being correct, is reminiscent of "confirmation bias" [10], a decision-making heuristic or shortcut [9] commonly employed by physicians in which they avoid actively to seek out conflicting information.

Although this single study calibrated AI accuracy at 60%, we next demonstrate through simulations that even with high AI accuracy, there will be a substantial incidence of false confirmation in virtually all clinically relevant settings.

Simulating the likelihood of false confirmation at varying AI and physician accuracy

We simulated the likelihood of false confirmation errors based on a series of theoretical calculations of the different outcomes in human-AI/XAI collaboration. For simplicity of illustration, we used accuracy, quantifying false confirmation as the percentage AI inaccuracy multiplied by the percentage physician inaccuracy (assuming independence of decision-making, as in the ideal second opinion the decision should be made 'blind', independent of knowledge of the other's diagnosis). Since our study showed that nearly all physicians uncritically accept false confirmations as accurate, whether provided by AI or XAI [9], we posited in our theoretical model that such errors go undetected and uncorrected.

In previous cross-field systematic reviews comparing the performance of AI and physicians [17,18], AI's accuracy varied from 55% to 99%, while the comparable figures for physicians ranged from 53% to 99%. In practice, accuracy varies considerably, depending on disease prevalence, physician experience, and the difficulty of the diagnostic challenge.

Table 2 shows the projected frequency of false confirmation at varying accuracy levels. Based on the aforementioned systematic reviews, at the lower end, where the accuracy of AI is 50% and that of the physician is 50%, false confirmation could occur in 1 out of 4 cases. At the higher end, when both exhibit an accuracy of 99%, the incidence would be less than 0.1%. Thus, it appears in cases where both physicians and AI are highly accurate, or when AI achieves accuracy greater than 95%, Kempt and Nagel's rule of disagreement [5] appears to be sufficient to keep the false confirmation rate below 3%. In a clinical context, applying realistic accuracy measures, it appears plausible that instances of false confirmations would likely reduce diagnostic accuracy between 5% and 30%.

Physician accuracy

	30%	40%	50%	60%	70%	80%	90%	95%	99%
AI accuracy	40%	36%	30%	24%	18%	12%	6%	3%	1%
	50%	30%	25%	20%	15%	10%	5%	3%	1%
	60%	24%	20%	16%	12%	8%	4%	2%	0%
	70%	18%	15%	12%	9%	6%	3%	2%	0%
	80%	12%	10%	8%	6%	4%	2%	1,0%	0,2%
	90%	6%	5%	4%	3%	2%	1,0%	0,5%	0,1%
	95%	3%	3%	2%	2%	1,0%	0,5%	0,3%	0,1%
	99%	1%	1%	0%	0%	0,2%	0,1%	0,1%	0,01%

Table 2. The table illustrates the False Confirmation Error rate (total no. False Confirmations/total no. clinical decisions. False Confirmation = (1-AI Accuracy) x (1-Physician Accuracy)

Having indicated the pervasiveness of false confirmation, we next revisit cognitive psychology literature to assess approaches that might mitigate it.

Mitigating the threat of false confirmation – evidence from cognitive psychology

Having noted the risk of false confirmation and the difficulty in detecting it [10], we apply insights from cognitive psychology, in particular the works of Herbert Simon [19], Daniel Kahneman [20], and Richard Thaler [21] who have explored the dual processes of the human mind: rapid, heuristic thinking and slower, more deliberate reasoning. By understanding these, we can develop strategies to refine decision-making and reduce false confirmation risks when using AI in clinical second opinions.

Three main cognitive interventions could apply to physician-AI collaboration. These involve explainability, cognitive forcing, and/or nudging techniques. Explainability aims to stimulate a reasoning-based discussion of the clinical rationale for decision-making. This could, in theory, enable a better dialogue between the physician and the second AI opinion, mirroring the role of a human second opinion. Yet, the only study evaluating this to our knowledge found no effect on identifying false confirmation errors [9].

The second technique, cognitive forcing, aims more directly to disrupt heuristic thinking and promote analytical reasoning [22]. Techniques include checklists and diagnostic timeouts or asking the physician to make a clinical assessment before seeing the AI diagnosis or any associated explanations [8,23]. Our preliminary study tested the possibility that asking physicians to make a diagnostic decision prior to being exposed to AI or XAI advice had no impact on false confirmation errors; physicians virtually always accepted AI confirmations without scrutiny [9], akin to the heuristic confirmation bias.

The third commonly advocated approach, 'nudging' [21,24], involves designing the healthcare environments so as to steer individuals towards making better health decisions without restricting their choices. This can include subtle or even invisible prompts to increase the likelihood that the "right" decisions are the easiest to make, such as default opt-in for organ donation. Other examples are arranging healthier food at eye level to attract more attention or send SMS reminders for vaccine appointments. As applied to false confirmation, a nudge could be a simple pop-up reminder to encourage physicians to evaluate AI-provided information critically even when it is in agreement (e.g. "Please be aware that an AI confirmation could potentially be wrong" or "Please verify that the AI and its explanations rely on the same clinical factors and assign similar weights as your clinical diagnosis"). This dual-scrutiny approach could potentially safeguard against overreliance in cases of AI/XAI diagnostic confirmation. To our knowledge, it has yet to be evaluated.

In contrast, true and false conflict errors have been relatively better studied. To mitigate true conflict errors (where physicians dismiss accurate AI advice), providing explanations with AI predictions has been effective [11–13,25]. However, this can lead to an overreliance on AI, resulting in new false conflict errors [8,14,15]. Hence, current research on these errors seeks to calibrate trust in AI, so as to better balance true and false conflict errors.

Implications for practice: Towards a pragmatic model of an AI second opinion

Taken together, we have demonstrated that false confirmation is likely to be pervasive in most clinical scenarios, ranging from 5% up to a high 30% of all joint physician-AI decisions. For AI to perform well as a second opinion, notwithstanding gaps in the current evidence base, we argue for the following steps: First, while not precluding any formal evaluations, we believe that there is a plausible case for 'nudges' to be employed anyway to make physicians aware of the possibility of false confirmations when engaging with AI decision aides. Even if they achieve little benefit, it is implausible that they will cause harm. Kemt and Nagel's "rule of disagreement" proposal is, in our view, clinically hazardous. If the AI second opinion confirms the physician's initial assessment and, according to the rule of disagreement, no additional action is required, this will result in many avoidable medical errors.

Second, we argue that it is impractical to require that a second human opinion is always sought, as advocated by Jongsma and Sand [6]. We suggest that the physician-in-charge call for a human second opinion for both cases of conflict and confirmation, taking account of the level of uncertainty; the medical stakes involved; and/or whether the patient has requested it.

Third, drawing on cognitive forcing, we believe it is critical that doctors make decisions prior to consulting AI advice. Although current research is limited, we view this as a prerequisite for a nudging approach, like pop-up reminders about false confirmation, to work effectively in practice.

Finally, if XAI is available, it is worth prompting physicians to engage in a comparative analysis of the AI explanations (either through conversation if available, or if not, through evaluating the AI ranking and weights of underlying risk factors). The physician-in-charge can then examine whether the explanations rely on the same clinical factors and assign similar weights as the clinical diagnosis. This can reveal discrepancies, particularly important in detecting and averting cases of false confirmation.

If and until AI's accuracy vastly surpasses that of physicians in clinical practice, we argue that both AI conflicts and confirmations must be scrutinised rigorously for potential errors. Clearly, we are entering a new epoch of medical collaboration, and, AI, as with all medical technologies, call for a precautionary approach, rooted in evidence-based medicine.

Contribution: RR, the paper's main author, developed the initial research concept. All authors contributed to the refinement of the research idea and critical revisions of the manuscript.

Competing interests: None declared.

Funding: The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

References

1. Greenfield G, Shmueli L, Harvey A, et al. Patient-initiated second medical consultations - Patient characteristics and motivating factors, impact on care and satisfaction: A systematic review. *BMJ Open*. 2021;11(9). doi:10.1136/bmjopen-2020-044033

2. Getting a second opinion | Cancer information | Cancer Research UK. Accessed February 5, 2024. https://www.cancerresearchuk.org/about-cancer/treatment/access-to-treatment/different-doctor-second-opinion

3. Hägglund M, Mcmillan B, Whittaker R, Blease C. Patient empowerment through online access to health records. *The BMJ*. Published online 2022. doi:10.1136/bmj-2022-071531

4. Mills M. Martha's rule: a hospital escalation system to save patients' lives. *BMJ*. 2023;383. doi:10.1136/BMJ.P2319

5. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. *J Med Ethics*. 2022;48(4). doi:10.1136/medethics-2021-107440

6. Jongsma KR, Sand M. Agree to disagree: the symmetry of burden of proof in human-AI collaboration. *J Med Ethics*. 2022;48(4). doi:10.1136/medethics-2022-108242

7. Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inf Syst Res*. 2021;32(3). doi:10.1287/ISRE.2020.0980

8. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int J Hum Comput Stud*. 2023;169. doi:10.1016/j.ijhcs.2022.102941

9. Rosenbacke R. Errors in Physician-AI Collaboration: Insights From a Mixed-methods Study of Explainable AI and Trust in Clinical Decision-making. Published online March 26, 2024. doi:10.2139/ssrn.4773350

10. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol*. Published online 1998. doi:10.1037/1089-2680.2.2.175

11. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep*. 2023;13(1). doi:10.1038/s41598-023-28633-w

12. Kumar A, Manikandan R, Kose U, Gupta D, Satapathy SC. Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Trans Multimed Comput Commun Appl*. 2021;17(3s). doi:10.1145/3457187

13. Martínez-Agüero S, Soguero-Ruiz C, Alonso-Moral JM, Mora-Jiménez I, Álvarez-Rodríguez J, Marques AG. Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. *Future Gener Comput Syst*. 2022;133. doi:10.1016/j.future.2022.02.021

14. Naiseh M, Al-Thani D, Jiang N, Ali R. Explainable recommendation: when design meets trust calibration. *World Wide Web*. 2021;24(5). doi:10.1007/s11280-021-00916-0

15. Naiseh M, Al-Mansoori RS, Al-Thani D, Jiang N, Ali R. Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. In: *Proceedings of 2021 8th IEEE International Conference on Behavioural and Social Computing, BESC 2021.* ; 2021. doi:10.1109/BESC53957.2021.9635271

16. Kliegr T, Bahník Š, Fürnkranz J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif Intell*. 2021;295. doi:10.1016/j.artint.2021.103458

17. Nazarian S, Glover B, Ashrafian H, Darzi A, Teare J. Diagnostic accuracy of artificial intelligence and computer-aided diagnosis for the detection and characterization of colorectal polyps: Systematic review and meta-analysis. *J Med Internet Res*. 2021;23(7). doi:10.2196/27370

18. Shen J, Zhang CJP, Jiang B, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med Inform*. 2019;7(3). doi:10.2196/10010

19. Simon HA. What Is an Explanation of Behavior? *Psychol Sci*. Published online 1992. doi:10.1111/j.1467-9280.1992.tb00017.x

20. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science*. Published online 1981. doi:10.1126/science.7455683

21. Thaler RH, Sunstein CR. *Nudge: Improving Decisions about Health, Wealth, and Happiness*.; 2008. doi:10.1016/s1477-3880(15)30073-6

22. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Qual Saf*. 2016;25(10):808-820. doi:10.1136/bmjqs-2015-004417

23. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc ACM Hum-Comput Interact*. 2021;5(CSCW1):21. doi:10.1145/3449287

24. Johnson EJ, Goldstein D. Do Defaults Save Lives? *Science*. Published online 2003. doi:10.1126/science.1091721

25. You S, Yang CL, Li X. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *J Manag Inf Syst*. 2022;39(2):336-365. doi:10.1080/07421222.2022.2063553

Appendix V: Paper V

Published as “Matters Arising” in *Nature Communications.* This is the peer-reviewed and accepted version. There can be deviations after copyediting and publisher layout.

Citation: Rosenbacke, R.; Melhus, Å.; Stuckler, (2024) ‘False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making ’ *Nature Communications*. 2024 151 15, 1–2 (2024)

False Conflict and False Confirmation

Crucial Components of AI Accuracy in Medical Decision Making

Rikard Rosenbacke[1*], Åsa Melhus[2], David Stuckler[3]

[1]Centre for Corporate Governance, Department of Accounting, Copenhagen Business School, Copenhagen, Denmark
[2]Department of Medical Sciences/Section of Clinical Microbiology, Uppsala University, Uppsala, Sweden
[3]Department of Social and Political Science, Bocconi University, Milano, Italy

Running title:” “Explainable AI in Healthcare”

* - corresponding author; rr.ccg@cbs.dk, Rikard Rosenbacke, Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg

Word Count: 764

Abstract

Chanda and colleagues published a pivotal paper on integrating AI in medical decision-making. In their study, they investigated overall clinical accuracy in physician-AI collaboration. We revisited their data, spotting critical physician-AI decision-making errors, specifically false conflict and false confirmation. These errors, if unaddressed, pose significant threats to diagnostic accuracy.

False Conflict and False Confirmation: Crucial Components of AI Accuracy in Medical-Decision Making

We welcome the recent study from January 2024 by Chanda and colleagues [1] in Nature Communications, as a substantial advance on integrating Explainable Artificial Intelligence (XAI) into dermatological practice. Importantly, it shows that AI can enhance diagnostic accuracy, trust, and confidence among dermatologists.

When physicians make decisions with AI, three types of errors can occur (Table 1): i) false confirmation error – when the physician and AI agree but both are wrong; ii) false conflict error – when the physician is correct, AI is incorrect, and the physician change diagnosis; and iii) true conflict error – when the physician is incorrect but AI is correct, and the physician override the correct AI diagnosis.

	Physician right	Physician wrong
AI right	Correct	True conflict error
AI wrong	False conflict error	False confirmation error

Table 1. Potential sources of error in human-AI/XAI collaboration

In their paper, Chanda and colleagues consider only overall accuracy, which masks key decision-making threats and overlook the specific user groups that stand to gain the most from AI applications.

We revisited their published data, quantifying these errors (albeit we note that without full access to their original data, we cannot make precise calculations). With a mean AI error rate of 19.6%, combined with a mean clinician error rate of 33.8%, the likelihood of both being inaccurate, or a false confirmation, is 6.6%. Applying these calculations to the worse performing clinicals (lowest quartile mean accuracy 50.3%) increases the false confirmation rate to 9.7%.

We also found evidence consistent with false conflict errors for high-performing physicians. A sub-analysis of the best-performing physicians reveals that their performance deteriorates with AI support. A sub-analysis of the 15 best-performing clinicians matched or exceeded AI accuracy (80.4%), with their initial accuracy averaging 87.3% but dipping to 77.1% once AI was introduced in phase 2 and 81.5% with XAI in phase 3, possibly due to errors from relying on AI when it falsely conflicted with their own correct diagnosis. Trust in AI is not, by definition, better since it increases false conflict errors for the best performers.

Finally, we found that AI, for the lowest-performing clinicians, helped stamp out true conflict errors. For the lowest-performing quartile of clinicians studied by Chanda and colleagues, accuracy improved from 50.3% to 66.6% and 65.9%, respectively, during the three phases of their study. In theory, if these low-performing physicians fully trusted AI, their accuracy could have risen at least to 80.4% by simply eliminating true conflict errors.

Recent studies are beginning to delve deeper into how physicians respond to conflicts with AI. The most common and discussed error occurs when physicians tend to override a correct AI diagnosis in cases of true conflict error. Previous studies found that this arises from distrust in the AI's ‘black box’ logic [2–5]. In cases of false conflict errors, however, the physicians tended to express doubt and over-rely upon AI, especially when uncertain about their initial diagnosis. When explanations are added to the AI diagnoses (as XAI), it tends to mitigate true conflict errors but exacerbate false conflict errors. This phenomenon whereby even mere exposure to explanations can induce overreliance on AI has been documented in several studies [6–9]. Finally, false confirmation is perhaps the most pernicious; it reinforces trust in AI, while perpetuating clinical errors. These false confirmation errors remind us of the confirmation bias highlighted by Ghassemi and colleagues [10]. This issue is likely present in the study conducted by Chanda and colleagues, though it was not explicitly addressed.

Given that explainable AI can assist physicians in determining whether AI diagnoses align with evidence-based medicine and that explanations are essential for meeting the trustworthiness and transparency requirements of the EU AI Act 2024, it still potentially introduces new sources of errors, such as false conflict and false confirmation errors. One intervention could be introducing more complex diagnoses instead of simple yes/no decisions. Another promising technique is conformal predictions, which shifts the focus of the AI model from pinpointing a single accurate clinical recommendation to providing the clinician with a range of possibilities tailored to the individual patient, allowing for further investigation [11]. Further research is needed to better understand interventions to avoid new human-AI collaboration errors.

We believe a more precise identification of these errors and in whom they occur creates tremendous potential to tap the full potential and promise of AI-supported decision-making.

Competing interests: The authors declare that they have no competing interests.

Contribution: RR, the paper's main author, developed the initial research concept. All authors (RR, ÅM, and DS) contributed to the refinement of the research idea, the analysis and interpretation of data, and have undertaken critical revisions of the manuscript.

References

1. Chanda, T., Hauser, K., … S. H.-N. & 2024, U. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *nature.comT Chanda, K Hauser, S Hobelsberger, TC Bucher, CN Garcia, C Wies, H Kittler, P TschandlNature Commun. 2024 nature.com* (2024).

2. Gaube, S. *et al.* Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci. Rep.* **13**, (2023).

3. Kumar, A., Manikandan, R., Kose, U., Gupta, D. & Satapathy, S. C. Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Trans. Multimed. Comput. Commun. Appl.* **17**, (2021).

4. You, S., Yang, C. L. & Li, X. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *J. Manag. Inf. Syst.* **39**, 336–365 (2022).

5. Martínez-Agüero, S. *et al.* Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. *Futur. Gener. Comput. Syst.* **133**, (2022).

6. Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int. J. Hum. Comput. Stud.* **169**, (2023).

7. Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. Explainable recommendation: when design meets trust calibration. *World Wide Web* **24**, (2021).

8. Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N. & Ali, R. Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. in *Proceedings of 2021 8th IEEE International Conference on Behavioural and Social Computing, BESC 2021* (2021). doi:10.1109/BESC53957.2021.9635271

9. Kliegr, T., Bahník, Š. & Fürnkranz, J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif. Intell.* **295**, (2021).

10. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3**, e745–e750 (2021).

11. Banerji, C. R. S., Chakraborti, T., Harbron, C. & MacArthur, B. D. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine* **29**, 2996–2998 (2023).

Appendix VI: AI/XAI and trust related heuristics

In thinking about trust and AI it is important to differentiate cognition-based trust, where trust is derived from the perceived understandability, reliability, and technical competence of AI or XAI, rooted in reasoning, from affect-based trust, involving emotional attachment and faith [1]. Most research so far has focused on the former, but the limited research available suggests that both play a role [2,3]. In one prominent dual-system theory, two main ways of thinking exist [4]: the first is quick and based on gut feelings or intuition (System 1), and the second is slower, taking a more thoughtful and reasoning approach (System 2). Trust forms a mental picture of another person or a system, and trying to untangle all its intricacies using only rational thought (System 2) would be practically impossible. So, more often, deciding to trust someone or something, like an AI or a physician, comes from an intuitive judgment, hence can trust be seen as a heuristic [5]. Using this model, trust can be viewed as a decision-making shortcut, enabling the decision-maker to select information while ignoring other information to simplify a complex decision.

If trust is an intuitive judgment, it can change occasionally based on context, interoceptive signals like blood sugar, or external factors like sunshine [6]. This makes it difficult to conclude if we trust AI algorithms; it all depends. On the one hand, heuristics enable rapid decision-making, especially in complex environments, potentially saving time and cognitive resources. On the other hand, the reliance on heuristics can sometimes lead to errors, as they can oversimplify complex issues. Thus, while heuristics serve as valuable tools for navigating a complex world, they can also be the root of inaccurate judgments and decisions [7].

When it comes to trust in AI solutions, one must consider how they will be received by those who must use them and, especially, whether they will engender the appropriate level of trust, neither too much nor too little. Various cognitive heuristics or biases can shape end-users' trust in AI systems. It is crucial to concentrate on those heuristics and biases capable of generating the two negative detrimental extremes: blind trust and blind distrust. Below, we describe a conceptual framework with 28 distinctive biases and heuristics that potentially can influence AI and XAI's trust dynamics. However, further studies are needed.

1. **Attentional bias** signifies that individuals are often influenced by their prevailing thoughts, notably when these are negative or anxiety-inducing [8]. Given the vast quantity of data and text generated by AI, the exposure to varying cues, including positive or negative words or images, can alter the user's perception of trust.
2. **Attribute substation** occurs when individuals replace a complex question with a simpler one [9]. For instance, the intricate question, "Do I trust this physician?" might be substituted with a more straightforward query: "Does the physician appear friendly?" This heuristic is critical in misplacing blind trust or distrust in algorithms.
3. **Automation bias** refers to the tendency to overly trust automated decision systems, often resulting in diminished active reasoning even in the presence of contradicting

information [10]. A classic example of blind trust is the widespread reliance on spellcheck functions. However, placing blind trust in AI's guidance within the healthcare sector can be hazardous.

4. **Availability heuristic**, how smooth things come to mind, is interpreted as how likely the outcome is. [11]. This can skew trust, as seen in the preference for car travel over air due to the prominent recall of airplane attacks or accidents, despite the prevalence of car accidents. Explainable AI illustrating probabilities may help mitigate such biases.
5. **Availability cascade**, a self-reinforcing message [12], where frequent repetition enhances a message's perceived trustworthiness. Notably exploited by political consultants and advertisers, this heuristic may engender blind trust and distrust in healthcare AI systems. On the other hand, explainable AI could indicate actual probabilities.
6. **Confirmation bias**: we search for, interpret, and favor opinions aligned with our own [13]. If an AI system validates a physician's incorrect perspective, it could lead to reduced diligence in further investigations. See also Consistency [14]
7. **Curse of knowledge** [15], we often assume that others have the same background and knowledge for understanding. Addressing this issue is crucial in ensuring that healthcare users fully know the inherent limitations of probabilistic predictions from AI advice.
8. **Commitment** [16] or Consistency bias [14], refers to the phenomenon where an individual tends to adhere to a pre-made decision to avoid internal conflict, trusting their initial choice as the optimal one. This cognitive bias can potentially impede physicians' receptiveness to new, contradicting advice from AI, as it may clash with their prior clinical judgments.
9. **Default bias** [17], implies a preference for the pre-selected option when faced with an additional alternative, a central component of decision architecture [18]. This bias can potentially create a disparity based on whether the physician's judgment precedes the AI suggestion or vice versa. When the AI recommendation is the initial default, it might impact the subsequent choice.
10. **Fallacy of composition or division** assumes that what is true/false of one part must be true/false of the whole. [19]. In the context of XAI, this could manifest as assuming that one of the explanations is true/false; all others are true or false.
11. **Framing effect**, a different way of presenting the same information, often leads to different interpretations [20]. For instance, a difference in trust might emerge when choosing between a doctor or AI touted to have a 99% success rate and one where it is emphasized that 5 out of the last 5000 patients have died, although statistically representing the same fact.
12. **Group thinking** [21], when group members want to reach harmony, conformity, or consensus with the risk of losing critical thinking. This is a classic example of collective over-trust. When AI aligns with clinical judgments, it may lead to new patterns of AI-human group thinking, potentially amplifying pre-existing biases.

13. **Halo effect,** a tendency to let a positive impression of a person, institution, or product to spill over to other areas. [22]. The opposite is the "Horn effect" [23]. An AI application's positive outcome in one medical domain cannot inherently assure success in other domains.
14. **Hindsight bias**, after an event, we tend to overestimate the likelihood it was going to happen. "I-knew-it-all-along" effect [24]. This has a significant impact on our beliefs and what we trust or not trust. Can AI's superhuman capabilities in finding historic correlations induce over-trust or disbeliefs?
15. **Illusion of transparency** [25], a tendency to overestimate how well we know ourselves, others, or an AI system. Explanations I XAI is just a simplified model of the intrinsic AI characteristics and not the same as full transparency.
16. **Illusion of control** is a tendency to overestimate the own influence (or an AI's prediction) while neglecting the potential sway of unforeseen external events [26]. In the medical field, it is critical to recognize that AI predictions are fundamentally grounded in statistical correlations and probabilities and can be significantly altered by external factors or sudden changes in conditions.
17. **Illusion of validity** a tendency to overestimate our capability to interpret data, especially when the data seems to "tell" a coherent story [7,27]. Explainable AI could potentially create an illusion of validity based on its simplified explanations.
18. **Illusory correlation** refers to the misconception of perceiving a relationship between variables when none exists. [28]. This cognitive bias surpasses the error of mistaking correlation for causation. While AI excels in identifying correlations, it remains inherently incapable of discerning causative relationships.
19. **Illusory truth effect** – a tendency to believe a statement is true if it is easy to understand or if it is in line with our own beliefs [29]. Explanations accompanying AI algorithms might inadvertently lead clinicians to place unwarranted trust in the algorithm.
20. **Information bias** – we feel more trust if we have more information, even if the additional information is irrelevant [30]. In the healthcare sector, the surge of big data can potentially escalate this bias, posing a risk of clouded judgment due to the overwhelming quantity of information.
21. **Interoceptive bias** refers to the inclination to use internal bodily signals, such as fluctuations in blood sugar levels, as indicators of external realities, a tendency that can notably influence decision-making processes [6,31]. For instance, parole judges' decisions varied in accordance with the time elapsed since their last meal, illustrating a notable shift in risk perception as blood sugar levels decreased [32]. The reliance on AI algorithms can be influenced by intra-personal and inter-personal factors and may vary over time.
22. **Naïve realism**, is a tendency to regard one's perspective as objective while dismissing differing views as uninformed, irrational, or biased [33]. This bias can potentially manifest as either blind trust or skepticism toward AI algorithms.

23. **Messenger bias** [16], or Authority or Liking bias [14], is documented as a tendency to place trust in individuals who either resemble ourselves or are perceived as authorities. An algorithm created by a renowned professor at a prestigious university might induce unwarranted trust or skepticism.
24. **Optimism bias**, is a tendency to engage in wishful thinking and overestimate positive outcomes [34]. This bias may lead healthcare professionals to harbor unrealistic expectations of AI technologies, possibly overlooking critical signs or nuances in the pursuit of favorable results.
25. **Planning fallacy** [35], denotes the common tendency to underestimate the time required to complete tasks. Applying this concept to the integration of AI in healthcare raises concerns regarding the potential over-optimism of the pace at which these technologies can be effectively implemented.
26. **Priming, anchoring** [36] is a bias where individuals rely too heavily on an initial piece of information (the "anchor") to make subsequent judgments. While AI has the capacity to assist clinicians in making more informed decisions by offering a comprehensive analysis, it also carries the risk of impairing judgment if clinicians accept AI predictions at face value, thereby possibly shortening thorough clinical investigations.
27. **Probability neglect**, the inclination to disregard probability assessments while making decisions [37]. This bias might result in healthcare professionals making decisions based more on intuition than on AI's data-driven insights.
28. **Pro-innovation bias,** the belief that the whole society should adopt an innovation without the need for its alteration [38]. This bias can be a potent driver for blind trust in AI algorithms within healthcare.

References

1. Madsen M, Gregor S. Measuring Human-Computer Trust. *Proc Elev Australas Conf Inf Syst*. 2000.
2. Naiseh M, Al-Thani D, Jiang N, Ali R. Explainable recommendation: when design meets trust calibration. *World Wide Web*. 2021;24(5). doi:10.1007/s11280-021-00916-0
3. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int J Hum Comput Stud*. 2023;169. doi:10.1016/j.ijhcs.2022.102941
4. Kahneman D. Maps of bounded rationality: a perspective on intuitive judgment and choice. In: *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel*. ; 2002.
5. Lewicki RJ, Brinsfield CT. Framing trust: Trust as a heuristic. In: Donohue WA, Rogan RR, Kaufman S, eds. *Framing Matters: Perspectives on Negotiatin Research and Practice in Communication*. Peter Lang Publishing; 2011:110-135. http://www.mdpi.com/1996-1073/2/3/556/.
6. Barrett LF. *How Emotions Are Made : The Secret Life of the Brain*. Houghton Mifflin Harcourt; 2017. http://hmhbooks.com/trade.html?isbn13=9780544133310. Accessed September 4, 2019.
7. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty. *Science (80-)*. 1974. doi:Cited By (since 1996) 3914\nExport Date 30 November 2011
8. Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, Van Ijzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychol Bull*. 2007. doi:10.1037/0033-2909.133.1.1
9. Kahneman D, Frederick S. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In: *Heuristics and Biases*. ; 2012. doi:10.1017/cbo9780511808098.004
10. Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? *Int J Hum Comput Stud*. 1999. doi:10.1006/ijhc.1999.0252
11. Tversky A, Kahneman D. Availability: A heuristic for judging frequency and probability. *Cogn Psychol*. 1973. doi:10.1016/0010-0285(73)90033-9
12. Kuran T, Sunstein CR. Availability Cascades and Risk Regulation. *Stanford Law Rev*. 1999.
13. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol*. 1998. doi:10.1037/1089-2680.2.2.175
14. Cialdini RB. *Influence: The Psychology of Persuasion*. Collins New York; 2007. doi:10.1017/CBO9781107415324.004

15. Camerer C, Loewenstein G, Weber M. The Curse of Knowledge in Economic Settings: An Experimental Analysis. *J Polit Econ*. 1989. doi:10.1086/261651

16. Dolan P, Hallsworth M, Halpern D, King D, Metcalfe R, Vlaev I. Influencing behaviour: The mindspace way. *J Econ Psychol*. 2012. doi:10.1016/j.joep.2011.10.009

17. Johnson EJ, Goldstein D. Do Defaults Save Lives? *Science (80-)*. 2003. doi:10.1126/science.1091721

18. Thaler RH, Sunstein CR. *Nudge: Improving Decisions about Health, Wealth, and Happiness*.; 2008. doi:10.1016/s1477-3880(15)30073-6

19. Pickard-Cambridge WA (translator). *On Sophistical Refutations*. The University of Adelaide: eBooks @ Adelaide; 2007. http://ebooks.adelaide.edu.au/a/aristotle/sophistical/. Accessed January 10, 2020.

20. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science (80-)*. 1981. doi:10.1126/science.7455683

21. Turner ME, Pratkanis AR. Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organ Behav Hum Decis Process*. 1998. doi:10.1006/obhd.1998.2756

22. Nisbett RE, Wilson TD. The halo effect: Evidence for unconscious alteration of judgments. *J Pers Soc Psychol*. 1977. doi:10.1037/0022-3514.35.4.250

23. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920. doi:10.1037/h0071663

24. Wood G. The knew-it-all-along effect. *J Exp Psychol Hum Percept Perform*. 1978. doi:10.1037/0096-1523.4.2.345

25. Roberts A. WikiLeaks: The illusion of transparency. *Int Rev Adm Sci*. 2012. doi:10.1177/0020852311429428

26. Langer EJ. The illusion of control. *J Pers Soc Psychol*. 1975. doi:10.1037/0022-3514.32.2.311

27. Kahneman D, Tversky A. On the psychology of prediction. *Psychol Rev*. 1973. doi:10.1037/h0034747

28. Chapman LJ. Illusory correlation in observational report. *J Verbal Learning Verbal Behav*. 1967. doi:10.1016/S0022-5371(67)80066-5

29. Hasher L, Goldstein D, Toppino T. Frequency and the conference of referential validity. *J Verbal Learning Verbal Behav*. 1977. doi:10.1016/S0022-5371(77)80012-1

30. Baron J. *Thinking and Deciding*.; 2006. doi:10.1017/cbo9780511840265

31. Zaman J, De Peuter S, Van Diest I, Van den Bergh O, Vlaeyen JWS. Interoceptive cues predicting exteroceptive events. *Int J Psychophysiol*. 2016. doi:10.1016/j.ijpsycho.2016.09.003

32. Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. *Proc Natl Acad Sci*. 2011. doi:10.1073/pnas.1018033108

33. Ross L, Ward A. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and Knowledge*. 1996.

34. Sharot T. The optimism bias. *Curr Biol*. 2011. doi:10.1016/j.cub.2011.10.030

35. Kahneman D, Tversky A. Intuitive prediction: Biases and corrective procedures. In: *Judgment under Uncertainty*. ; 2013. doi:10.1017/cbo9780511809477.031

36. Kahneman D. *Thinking Fast, Thinking Slow*.; 2011.

37. Sunstein CR. Probability neglect: Emotions, worst cases, and law. *Yale Law J*. 2002. doi:10.2307/1562234

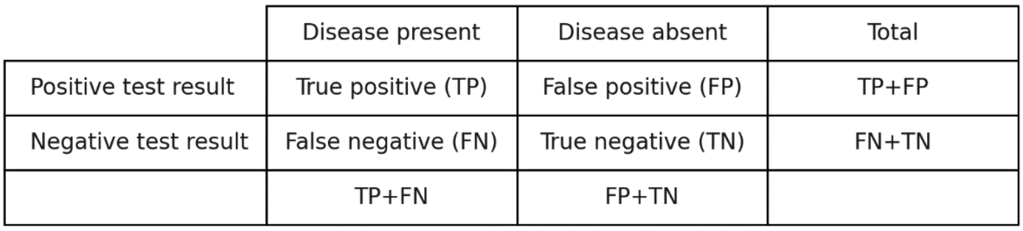
38. Rogers EM, Singhal A, Quinlan MM. Diffusion of innovations. In: *An Integrated Approach to Communication Theory and Research, Third Edition*. ; 2019. doi:10.4324/9780203710753-35

Appendix VII: AI second opinion and potential errors

To estimate the potential magnitude of false confirmation errors when AI is used for a second opinion, we need to assess the combined diagnostic performance of physicians and AI algorithms.

Performance in clinical decision-making

The performance of both AI and physicians in diagnostic tasks can be quantitatively assessed using established metrics such as sensitivity, specificity, and accuracy. Sensitivity measures the ability to correctly identify patients with a condition (true positives), while specificity assesses the ability to correctly identify those without the condition (true negatives). Accuracy reflects the overall proportion of true positive and true negative diagnoses made out of all diagnoses. These metrics allow for a comprehensive evaluation of the diagnostic effectiveness of both human clinicians and AI/XAI systems (Eisenberg, 1995). Table 1 describes different performance metrics followed by their definitions.



	Disease present	Disease absent	Total
Positive test result	True positive (TP)	False positive (FP)	TP+FP
Negative test result	False negative (FN)	True negative (TN)	FN+TN
	TP+FN	FP+TN	

Table 1. Assessing the performance of a diagnostic test or judgment (Eisenberg, 1995).

Definitions of the performance metrics:

Sensitivity = TP/(TP + FN)

Specificity = TN/(FP + TN)

Accuracy = (TP + TN)/Total

Disease prevalence = (TP + FN)/Total

Positive predictive value (PPV) = TP/(TP + FP)

Negative predictive value (NPV) = TN/(FN + TN)

Naturally, it is preferable that sensitivity and specificity are both high. However, there is normally a tradeoff where either sensitivity or specificity is high. Preferably, AI and clinicians are orthogonal in their diagnostics. Orthogonal is the concept of independence between measures or components within the test. When two factors are orthogonal, they are statistically uncorrelated, meaning that the score or outcome on one factor does not predict or influence the score on the other. When screening for a disease, the AI or clinician, depending on the case, should have high sensitivity to capture all positive cases (with a relatively high level of false positives due to lower specificity). Then, in a second confirmation phase, AI or clinicians should have a high specificity to sort out all false positives.

The balance between sensitivity and specificity also depends on the disease context and the consequences of false positives or false negatives. For diseases where missing a diagnosis could be fatal or lead to serious complications (e.g., cancer), a high sensitivity is prioritized. In contrast, for conditions where a false positive could lead to unnecessary anxiety or invasive procedures, high specificity is more important.

The importance of prevalence

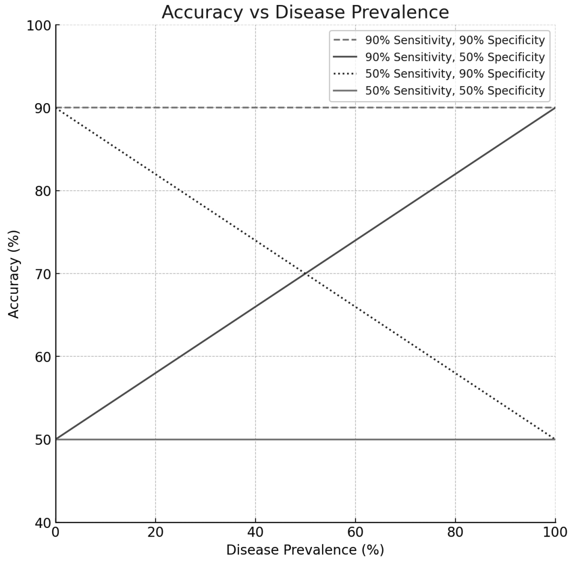


Figure 1. Assessing the performance of a diagnostic test or judgment (Eisenberg, 1995).

In populations with a low prevalence of a disease, even tests with high specificity can result in a relatively high proportion of false positives compared to true positives, which implies a low PPV. In populations with high prevalence, a high sensitivity can result in a high level of false negatives compared to true negatives, which implies a low NPV.

Figure 1, illustrates that if sensitivity and specificity are the same, the accuracy is the same regardless of prevalence; with low prevalence, the accuracy will be the same as specificity, and vice versa. With a high prevalence, accuracy is the same as sensitivity (Eisenberg, 1995).

Accuracy measures the test's overall ability to correctly classify individuals as having or not having the condition. It considers both true positives and negatives, as well as false positives and false negatives. In this paper, we focus on accuracy to grasp both sensitivity and specificity in relation to the prevalence.

In binary diagnostic judgments, assuming an equal prevalence of positive and negative cases in the population being tested, both sensitivity and specificity will be 50% by chance, implying that accuracy will also be 50%. In a high-prevalence environment, there will be many false negatives by chance, and in a low-prevalence environment, there will be many false positives by chance.

Clinician accuracy can also vary both between different clinicians and from time to time for each clinician. Experienced physicians often perform better than novice physicians (Gaube *et al.*, 2023). But there can also be intra-personnel differences based on, for example, interoceptive signals. In an Israeli study, it was observed that judges approved 65 percent of the cases at the start of the day, with the approval rate dropping to nearly zero by the session's end. However, following a break for snacks, the rate of approvals surged back to 65 percent, (Danziger, Levav and Avnaim-Pesso, 2011). Physicians may benefit more from AI advice when they are experiencing fatigue.

However, there can be a systematic error by physicians and AI; hence, in theory, accuracy, sensitivity, and specificity can be below 50%. An example of systematic errors could be that the AI or physician assumes that a parameter like high body mass index is a risk factor for a certain disease while it, in reality, is a protective parameter.

Measuring a physician’s diagnostic accuracy can be both difficult and sensitive. Furthermore, accuracy is significantly different from person to person. For novice doctors, AI can be more useful.

The problem of high prevalence in training data but low in clinical settings

In previous systematic reviews and meta-analyses, the authors compared the performance of AI versus physicians (Shen *et al.*, 2019; Nazarian *et al.*, 2021). They found that AI accuracy varies from 60-99%, with sensitivity and specificity from 60-99%. Performance for physicians was 48-99% when measuring accuracy, sensitivity, and specificity. However, no studies discussed the prevalence, especially prevalence in the training and validation dataset, versus the prevalence in a clinical real-life setting.

When training an AI algorithm, it is of crucial importance that there is a match between the training data set and the test data set (Chen *et al.*, 2023). Furthermore, there needs to be a match to the real-life clinical setting. A recent systematic review evaluated the diagnostic accuracy of AI algorithms when identifying pathology in medical imaging (Jones *et al.*, 2022). Only two studies (out of 14 224) used data from clinical settings with a low prevalence of skin cancer. Therefore, the authors argued that the widespread adoption into community and primary care practice cannot currently be recommended until efficacy in these populations is shown.

If an algorithm is trained on a high prevalence training data set, its sensitivity is likely high with likely lower specificity. If applied in a low prevalence environment, the lower specificity implies a lot of false positives, and the accuracy can fall dramatically. This implies that the XAI will have explanations for high prevalence settings.

To sum up, the overall accuracy based on a clinical diagnosis and using AI/XAI for a second opinion depends on several parameters: i) Firstly, it depends on the AI sensitivity and specificity. ii) It also depends on the clinician's sensitivity and specificity, and this can differ significantly between physicians (interpersonal difference) and also for each physician, depending on the situation (intrapersonal difference). iii) Furthermore, prevalence is paramount to understanding the overall accuracy. Even a high sensitivity and specificity in low prevalence environments can lead to a positive predictive value (PPV) significantly below 50%. iv) Furthermore, if the prevalence differs in the training data from the prevalence in the clinical practice, sensitivity, specificity, and accuracy can be significantly reduced.

In theory, in a clinical setting, even for validated algorithms with high sensitivities or specificities around 90% or higher, the overall accuracy can be as low as 50% in binary classifications based on prevalence and training data sets. Furthermore, systematic errors can reduce accuracy, potentially below 50%.

References

Chen, R. J. *et al.* (2023) 'Algorithmic fairness in artificial intelligence for medicine and healthcare', *Nature biomedical engineering*. doi: 10.1038/s41551-023-01056-8.

Danziger, S., Levav, J. and Avnaim-Pesso, L. (2011) 'Extraneous factors in judicial decisions', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1018033108.

Eisenberg, M. J. (1995) 'Accuracy and predictive values in clinical decision-making.', *Cleveland Clinic journal of medicine*, 62(5). doi: 10.3949/ccjm.62.5.311.

Gaube, S. *et al.* (2023) 'Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays', *Scientific Reports*, 13(1). doi: 10.1038/s41598-023-28633-w.

Jones, O. T. *et al.* (2022) 'Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review', *The Lancet Digital Health*. doi: 10.1016/S2589-7500(22)00023-1.

Nazarian, S. *et al.* (2021) 'Diagnostic accuracy of artificial intelligence and computer-aided diagnosis for the detection and characterization of colorectal polyps: Systematic review and meta-analysis', *Journal of Medical Internet Research*. doi: 10.2196/27370.

Shen, J. *et al.* (2019) 'Artificial intelligence versus clinicians in disease diagnosis: Systematic review', *JMIR Medical Informatics*. doi: 10.2196/10010.

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or,
How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning
– A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus
– Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain
Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities,
Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter.
En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
*Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder*

17. Adam Buchhorn
*Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy*

18. Cecilie K. Moesby-Jensen
*Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere*

19. Heidi Boye
*Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser*

20. Kristine Munkgård Pedersen
*Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival*

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
*Authority and Autonomy
Paradoxes of Modern Knowledge Work*

23. Anette Frey Larsen
*Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen*

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
*Identifying the Last Planner System
Lean management in the construction industry*

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
*Hybrid Organisations.
A study of the Open Source – business setting*

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
*Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale*

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin – en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership: Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"? – en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy
Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness
The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management
A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship
Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness
An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur
Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing
The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence
A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog
Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management
- A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat
En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd
Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing
A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions
in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være?
Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sensemaking in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a
Partly-Digital World
Comparing Digital Piracy and Legal
Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in
organizational decision-making

11. Mia Kaspersen
The construction of social and
environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading
names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte
til velgørenhed og understøtte
relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt
perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig
kontekst, strategisk tilgang, samt
anvendte redskaber og teknologier for
udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization
of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural
Production
The case of travel guidebook
production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie
i det 19. århundrede
Omtydningen af melankolien og
manien som bipolære stemningslidelser
i dansk sammenhæng under hensyn til
dannelsen af det moderne følelseslivs
relative autonomi.
En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters
of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of
Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize
social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL
RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om
selvledelsens paradoksale dynamik og
eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis – barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life - Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelsesstudie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity:
Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering: Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig
– er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm
– A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was
The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation
Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone
With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope
Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics
Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions
– on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A
– Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats?
A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management
A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management
Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management
– Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me
- The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration
A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce:
Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca
An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability
in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'.
A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World
Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber
Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts
Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity
– How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation
The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition.
Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK
An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsification and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints
A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS
Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics:
-An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
Essays on Crude Oil Tanker Markets

13. Cancan Wang
Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
The Comparative Dynamics of Private Governance
The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability
Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

23. Henriette Sophia Groskopff Tvede Schleimann
Creating innovation through collaboration – Partnering in the maritime sector

24. Kristian Steensen Nielsen
The Role of Self-Regulation in Environmental Behavior Change

25. Lydia L. Jørgensen
Moving Organizational Atmospheres

26. Theodor Lucian Vladasel
Embracing Heterogeneity: Essays in Entrepreneurship and Human Capital

27. Seidi Suurmets
Contextual Effects in Consumer Research: An Investigation of Consumer Information Processing and Behavior via the Applicati on of Eye-tracking Methodology

28. Marie Sundby Palle Nickelsen
Reformer mellem integritet og innovation: Reform af reformens form i den danske centraladministration fra 1920 til 2019

29. Vibeke Kristine Scheller
The temporal organizing of same-day discharge: A tempography of a Cardiac Day Unit

30. Qian Sun
Adopting Artificial Intelligence in Healthcare in the Digital Age: Perceived Challenges, Frame Incongruence, and Social Power

31. Dorthe Thorning Mejlhede
Artful change agency and organizing for innovation – the case of a Nordic fintech cooperative

32. Benjamin Christoffersen
Corporate Default Models:
Empirical Evidence and Methodical Contributions

33. Filipe Antonio Bonito Vieira
Essays on Pensions and Fiscal Sustainability

34. Morten Nicklas Bigler Jensen
Earnings Management in Private Firms: An Empirical Analysis of Determinants and Consequences of Earnings Management in Private Firms

2020

1. Christian Hendriksen
Inside the Blue Box: Explaining industry influence in the International Maritime Organization

2. Vasileios Kosmas
Environmental and social issues in global supply chains:
Emission reduction in the maritime transport industry and maritime search and rescue operational response to migration

3. Thorben Peter Simonsen
The spatial organization of psychiatric practice: A situated inquiry into 'healing architecture'

4. Signe Bruskin
The infinite storm: An ethnographic study of organizational change in a bank

5. Rasmus Corlin Christensen
Politics and Professionals: Transnational Struggles to Change International Taxation

6. Robert Lorenz Törmer
The Architectural Enablement of a Digital Platform Strategy

7. Anna Kirkebæk Johansson Gosovic
Ethics as Practice: An ethnographic study of business ethics in a multinational biopharmaceutical company

8. Frank Meier
Making up leaders in leadership development

9. Kai Basner
Servitization at work:
On proliferation and containment

10. Anestis Keremis
Anti-corruption in action: How is anti-corruption practiced in multinational companies?

11. Marie Larsen Ryberg
Governing Interdisciolinarity: Stakes and translations of interdisciplinarity in Danish high school education.

12. Jannick Friis Christensen
Queering organisation(s): Norm-critical orientations to organising and researching diversity

13. Thorsteinn Sigurdur Sveinsson
Essays on Macroeconomic Implications of Demographic Change

14. Catherine Casler
Reconstruction in strategy and organization: For a pragmatic stance

15. Luisa Murphy
Revisiting the standard organization of multi-stakeholder initiatives (MSIs):
The case of a meta-MSI in Southeast Asia

16. Friedrich Bergmann
Essays on International Trade

17. Nicholas Haagensen
European Legal Networks in Crisis: The Legal Construction of Economic Policy

18. Charlotte Biil
Samskabelse med en sommerfuglemodel: Hybrid ret i forbindelse med et partnerskabsprojekt mellem 100 selvejende daginstitutioner, deres paraplyorganisation, tre kommuner og CBS

19. Andreas Dimmelmeier
The Role of Economic Ideas in Sustainable Finance: From Paradigms to Policy

20. Maibrith Kempka Jensen
Ledelse og autoritet i interaktion
- En interaktionsbaseret undersøgelse af autoritet i ledelse i praksis

21. Thomas Burø
LAND OF LIGHT: Assembling the Ecology of Culture in Odsherred 2000-2018

22. Prins Marcus Valiant Lantz
Timely Emotion: The Rhetorical Framing of Strategic Decision Making

23. Thorbjørn Vittenhof Fejerskov
Fra værdi til invitationer - offentlig værdiskabelse gennem affekt, potentialitet og begivenhed

24. Lea Acre Foverskov
Demographic Change and Employment: Path dependencies and institutional logics in the European Commission

25. Anirudh Agrawal
A Doctoral Dissertation

26. Julie Marx
Households in the housing market

27. Hadar Gafni
Alternative Digital Methods of Providing Entrepreneurial Finance

28. Mathilde Hjerrild Carlsen
Ledelse af engagementer: En undersøgelse af samarbejde mellem folkeskoler og virksomheder i Danmark

29. Suen Wang
Essays on the Gendered Origins and Implications of Social Policies in the Developing World

30. Stine Hald Larsen
The Story of the Relative: A Systems-Theoretical Analysis of the Role of the Relative in Danish Eldercare Policy from 1930 to 2020

31. Christian Casper Hofma
Immersive technologies and organizational routines: When head-mounted displays meet organizational routines

32. Jonathan Feddersen
The temporal emergence of social relations: An event-based perspective of organising

33. Nageswaran Vaidyanathan
ENRICHING RETAIL CUSTOMER EXPERIENCE USING AUGMENTED REALITY

2021

1. Vanya Rusinova
The Determinants of Firms' Engagement in Corporate Social Responsibility: Evidence from Natural Experiments

2. Lívia Lopes Barakat
Knowledge management mechanisms at MNCs:
The enhancing effect of absorptive capacity and its effects on performance and innovation

3. Søren Bundgaard Brøgger
Essays on Modern Derivatives Markets

4. Martin Friis Nielsen
Consuming Memory: Towards a conceptualization of social media platforms as organizational technologies of consumption

05. Fei Liu
Emergent Technology Use in Consumer Decision Journeys: A Process-as-Propensity Approach

06. Jakob Rømer Barfod
Ledelse i militære højrisikoteams

07. Elham Shafiei Gol
Creative Crowdwork Arrangements

08. Árni Jóhan Petersen
Collective Imaginary as (Residual) Fantasy: A Case Study of the Faroese Oil Bonanza

09. Søren Bering
"Manufacturing, Forward Integration and Governance Strategy"

10. Lars Oehler
Technological Change and the Decomposition of Innovation:
Choices and Consequences for Latecomer Firm Upgrading: The Case of China's Wind Energy Sector

11. Lise Dahl Arvedsen
Leadership in interaction in a virtual context:
A study of the role of leadership processes in a complex context, and how such processes are accomplished in practice

12. Jacob Emil Jeppesen
Essays on Knowledge networks, scientific impact and new knowledge adoption

13. Kasper Ingeman Beck
Essays on Chinese State-Owned Enterprises:
Reform, Corporate Governance and Subnational Diversity

14. Sönnich Dahl Sönnichsen
Exploring the interface between public demand and private supply for implementation of circular economy principles

15. Benjamin Knox
Essays on Financial Markets and Monetary Policy

16. Anita Eskesen
*Essays on Utility Regulation:
Evaluating Negotiation-Based Approaches inthe Context of Danish Utility Regulation*

17. Agnes Guenther
Essays on Firm Strategy and Human Capital

18. Sophie Marie Cappelen
Walking on Eggshells: The balancing act of temporal work in a setting of culinary change

19. Manar Saleh Alnamlah
About Gender Gaps in Entrepreneurial Finance

20. Kirsten Tangaa Nielsen
Essays on the Value of CEOs and Directors

21. Renée Ridgway
Re:search - the Personalised Subject vs. the Anonymous User

22. Codrina Ana Maria Lauth
*IMPACT Industrial Hackathons:
Findings from a longitudinal case study on short-term vs long-term IMPACT implementations from industrial hackathons within Grundfos*

23. Wolf-Hendrik Uhlbach
Scientist Mobility: Essays on knowledge production and innovation

24. Tomaz Sedej
Blockchain technology and inter-organizational relationships

25. Lasse Bundgaard
*Public Private Innovation Partnerships:
Creating Public Value & Scaling Up Sustainable City Solutions*

26. Dimitra Makri Andersen
*Walking through Temporal Walls:
Rethinking NGO Organizing for Sustainability through a Temporal Lens on NGO-Business Partnerships*

27. Louise Fjord Kjærsgaard
*Allocation of the Right to Tax Income from Digital Products and Services:
A legal analysis of international tax treaty law*

28. Sara Dahlman
Marginal alternativity: Organizing for sustainable investing

29. Henrik Gundelach
*Performance determinants:
An Investigation of the Relationship between Resources, Experience and Performance in Challenging Business Environments*

30. Tom Wraight
Confronting the Developmental State: American Trade Policy in the Neoliberal Era

31. Mathias Fjællegaard Jensen
Essays on Gender and Skills in the Labour Market

32. Daniel Lundgaard
Using Social Media to Discuss Global Challenges: Case Studies of the Climate Change Debate on Twitter

33. Jonas Sveistrup Søgaard
Designs for Accounting Information Systems using Distributed Ledger Technology

34. Sarosh Asad
CEO narcissism and board composition: Implications for firm strategy and performance

35. Johann Ole Willers
*Experts and Markets in Cybersecurity
On Definitional Power and the Organization of Cyber Risks*

36. Alexander Kronies
Opportunities and Risks in Alternative Investments

37. Niels Fuglsang
The Politics of Economic Models: An inquiry into the possibilities and limits concerning the rise of macroeconomic forecasting models and what this means for policymaking

38. David Howoldt
Policy Instruments and Policy Mixes for Innovation: Analysing Their Relation to Grand Challenges, Entrepreneurship and Innovation Capability with Natural Language Processing and Latent Variable Methods

2022

01. Ditte Thøgersen
Managing Public Innovation on the Frontline

02. Rasmus Jørgensen
Essays on Empirical Asset Pricing and Private Equity

03. Nicola Giommetti
Essays on Private Equity

04. Laila Starr
When Is Health Innovation Worth It? Essays On New Approaches To Value Creation In Health

05. Maria Krysfeldt Rasmussen
Den transformative ledelsesbyrde – etnografisk studie af en religionsinspireret ledelsesfilosofi i en dansk modevirksomhed

06. Rikke Sejer Nielsen
Mortgage Decisions of Households: Consequences for Consumption and Savings

07. Myriam Noémy Marending
Essays on development challenges of low income countries: Evidence from conflict, pest and credit

08. Selorm Agbleze
A BEHAVIORAL THEORY OF FIRM FORMALIZATION

09. Rasmus Arler Bogetoft
Rettighedshavers faktisk lidte tab i immaterialretssager: Studier af dansk ret med støtte i økonomisk teori og metode

10. Franz Maximilian Buchmann
Driving the Green Transition of the Maritime Industry through Clean Technology Adoption and Environmental Policies

11. Ivan Olav Vulchanov
The role of English as an organisational language in international workplaces

12. Anne Agerbak Bilde
TRANSFORMATIONER AF SKOLELEDELSE - en systemteoretisk analyse af hvordan betingelser for skoleledelse forandres med læring som genstand i perioden 1958-2020

13. JUAN JOSE PRICE ELTON
EFFICIENCY AND PRODUCTIVITY ANALYSIS: TWO EMPIRICAL APPLICATIONS AND A METHODOLOGICAL CONTRIBUTION

14. Catarina Pessanha Gomes
The Art of Occupying: Romanticism as Political Culture in French Prefigurative politics

15. Mark Ørberg
Fondsretten og den levende vedtægt

16. Majbritt Greve
Maersk's Role in Economic Development: A Study of Shipping and Logistics Foreign Direct Investment in Global Trade

17. Sille Julie J. Abildgaard
Doing-Being Creative: Empirical Studies of Interaction in Design Work

18. Jette Sandager
Glitter, Glamour, and the Future of (More) Girls in STEM: Gendered Formations of STEM Aspirations

19. Casper Hein Winther
Inside the innovation lab - How paradoxical tensions persist in ambidextrous organizations over time

20. Nikola Kostić
Collaborative governance of inter-organizational relationships: The effects of management controls, blockchain technology, and industry standards

21. Saila Naomi Stausholm
Maximum capital, minimum tax: Enablers and facilitators of corporate tax minimization

22. Robin Porsfelt
Seeing through Signs: On Economic Imagination and Semiotic Speculation

23. Michael Herburger
Supply chain resilience – a concept for coping with cyber risks

24. Katharina Christiane Nielsen Jeschke
Balancing safety in everyday work - A case study of construction managers' dynamic safety practices

25. Jakob Ahm Sørensen
Financial Markets with Frictions and Belief Distortions

26. Jakob Laage-Thomsen
Nudging Leviathan, Protecting Demos - A Comparative Sociology of Public Administration and Expertise in the Nordics

27. Kathrine Søs Jacobsen Cesko
Collaboration between Economic Operators in the Competition for Public Contracts: A Legal and Economic Analysis of Grey Zones between EU Public Procurement Law and EU Competition Law

28. Mette Nelund
Den nye jord – Et feltstudie af et bæredygtigt virke på Farendløse Mosteri

29. Benjamin Cedric Larsen
Governing Artificial Intelligence – Lessons from the United States and China

30. Anders Brøndum Klein
Kollektiv meningsdannelse iblandt heterogene aktører i eksperimentelle samskabelsesprocesser

31. Stefano Tripodi
Essays on Development Economicis

32. Katrine Maria Lumbye
Internationalization of European Electricity Multinationals in Times of Transition

33. Xiaochun Guo
Dynamic Roles of Digital Currency – An Exploration from Interactive Processes: Difference, Time, and Perspective

34. Louise Lindbjerg
Three Essays on Firm Innovation

35. Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

36. Hanna Nyborg Storm
Cultural institutions and attractiveness – How cultural institutions contribute to the development of regions and local communities

37. Anna-Bertha Heeris Christensen
Conflicts and Challenges in Practices of Commercializing Humans – An Ethnographic Study of Influencer Marketing Work

38. Casper Berg Lavmand Larsen
A Worker-Centered Inquiry into the Contingencies and Consequences of Worker Representation

39. Niels le Duc
The Resource Commitment of Multinational Enterprise R&D Activities

40. Esben Langager Olsen
Change management tools and change managers – Examining the simulacra of change

41. Anne Sophie Lassen
Gender in the Labor Market

42. Alison E. Holm
Corrective corporate responses to accusations of misconduct on societal issues

43. Chenyan Lyu
Carbon Pricing, Renewable Energy, and Clean Growth – A Market Perspective

44. Alina Grecu
UNPACKING MULTI-LEVEL OFFSHORING CONSEQUENCES: Hiring Wages, Onshore Performance, and Public Sentiment

45. Alexandra Lüth
Offshore Energy Hubs as an Emerging Concept – Sector Integration at Sea

2023

01. Cheryl Basil Sequeira
Port Business Development – Digitalisation of Port Authroity and Hybrid Governance Model

02. Mette Suder Franck
Empirical Essays on Technology Supported Learning – Studies of Danish Higher Education

03. Søren Lund Frandsen
States and Experts – Assembling Expertise for Climate Change and Pandemics

04. Guowei Dong
Innovation and Internationalization – Evidence from Chinese Manufacturing Enterprises

05. Eileen Murphy
In Service to Security – Constructing the Authority to Manage European Border Data Infrastructures

06. Bontu Lucie Guschke
THE PERSISTENCE OF SEXISM AND RACISM AT UNIVERSITIES – Exploring the imperceptibility and unspeakability of workplace harassment and discrimination in academia

07. Christoph Viebig
Learning Entrepreneurship – How capabilities shape learning from experience, reflection, and action

08. Kasper Regenburg
Financial Risks of Private Firms

09. Kathrine Møller Solgaard
Who to hire? – A situated study of employee selection as routine, practice, and process

10. Jack Kværnø-Jones
Intersections between FinTech Imaginaries and Traditional Banking – A study of disciplinary, implementary, and parasitic work in the Danish financial sector

11. Stine Quorning
Managing Climate Change Like a Central Banker – The Political Economy of Greening the Monetary Technocracy

12. Amanda Bille
No business without politics – Investigating the political nature of supply chain management

13. Theis Ingerslev Jensen
Essays on Empirical Asset Pricing

14. Ann Fugl-Meyer
The Agile Imperative – A Qualitative Study of a Translation Process in the Danish Tax Administration

15. Nicolai Søgaard Laursen
Longevity risk in reinsurance and equity markets

16. Shelter Selorm Kwesi Teyi
STRATEGIC ENTREPRENEURSHIP IN THE INFORMAL ECONOMY

17. Luisa Hedler
Time, Law and Tech – The introduction of algorithms to courts of law

18. Tróndur Møller Sandoy
Essays on the Economics of Education

19. Nathan Rietzler
Crowdsourcing Processes and Performance Outcomes

20. Sigrid Alexandra Koob
Essays on Democracy, Redistribution, and Inequality

21. David Pinkus
Pension Fund Investment: Implications for the Real Economy

22. Sina Smid
Inequality and Redistribution – Essays on Local Elections, Gender and Corruption in Developing Countries

23. Andreas Brøgger
Financial Economics with Preferences and Frictions

24. Timothy Charlton-Czaplicki
Arendt in the platformised world – Labour, work and action on digital platforms

25. Letícia Vedolin Sebastião
Mindfulness and Consumption: Routes Toward Consumer Self-Control

26. Lotte List
Crisis Sovereignty – The Philosophy of History of the Exception

27. Jeanette Walldorf
Essays on the Economics of Education and Labour Market

28. Juan Camilo Giraldo-Mora
It is Along Ways – Global Payment Infrastructure in Movement

29. Niels Buus Lassen
THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

30. Frederik Bjørn Christensen
Essays on the Intergenerational Welfare State

31. Shama Patel
The Summer of 2020: Situating Digital Media in Scaling Affective Contagion: A Case of the George Floyd Video

32. Federico Jensen
Who rules the waves in the 21st Century?
The international political economy of global shipping

33. Tobias Berggren Jensen
Selvledende organisationer i den offentlige sektor – modsætninger og konflikter i radikal decentralisering

34. Jonathan Harmat
The Affects By Which We Are Torn
Four Essays on Government and Affect

35. Jørgen Valther Hansen
The Big 4 Audit Firms and the Public Interest
Public oversight & Audit Firm Governance

36. Stig Strandbæk Nyman
The Birth of Algorithmic Aspirational Control

37. Morten Tinning
Steaming Ahead
Experiences and the Transition from Sail to Steam

38. Oguzhan Cepni
Essays in Applied Financial Economics

39. Tim Dominik Maurer
Essays on Pension Policy

40. Aixa Y. Alemán-Díaz
Exploring Global Ideas in National Policy for Science, Technology and Innovation an Isomorphic Difference Approach

41. Michael Güldenpfennig
Managing the interrelationships between manufacturing system elements for productivity improvement in the factory

42. Jun Yuan (Julian) Seng
Essays on the political economy of innovative startups

43. Jacek Piosik
Essays on Entrepreneurial Finance

44. Elizabeth Cooper
Tourists on the Edge
Understanding and Encouraging Sustainable Tourist Behaviour in Greenland

2024

01. Marija Sarafinovska
Patients as Innovators: An Empirical Study of Patients' Role in Innovation in the Healthcare Industry

02. Niina Hakala
Corporate Reporting in the Governance of Climate Transition – Framing agency in a financialized world

03. Kasper Merling Arendt
Unleashing Entrepreneurial Education
Developing Entrepreneurial Mindsets, Competencies, and Long-Term Behavior

04. Kerstin Martel
Creating and dissolving 'identity'
in global mobility studies -
a multi-scalar inquiry of belongingness
and becoming on-the-move

05. Sofie Elbæk Henriksen
Big Tech to the Rescue?
An Ethnographic Study of Corporate Humanitarianism in the Refugee Crisis

06. Christina Kjær
Corporate scandals
- in the age of 'responsible business'

07. Anna Stöber
Embedded Self-Managing Modes of Organizing
Empirical Inquiries into Boundaries, Momentum, and Collectivity

08. Lucas Sören Göbeler
Shifting and Shaping
Physicality in Digital Innovation

09. Felix Schilling
Department of International Economics, Government and Business

10. Mathias Lund Larsen
China and the Political Economy of the Green State

11. Michael Bennedsen Hansen
At få sjælen med
En narrativ analyse af danske containersøfolks erindringer, fortidsbrug og identitetskonstruktioner

12. Justyna Agata Bekier
More than a numbers game
Accounting for circular economy performance in collaborative initiatives in cities

13. Frederik Schade
The Question of Digital Responsibility
An Ethnography of Emergent Institutional Formations in the Contemporary Governance of Technology

14. Alexandrina Schmidt
The Mundane in the Digital: A qualitative study of social work and vulnerable clients in Danish job centres

15. Julian Fernandez Mejia
Essays on International Finance

16. Leonie Decrinis
Nudging in the Workplace: Exploring a Micro-level Approach Towards Corporate Sustainability

17. Nina Frausing Pedersen
A Framing Contest between Institutional Actors on Crypto-Asset Policymaking in the EU

18. Amalie Toft Bentsen
The Internal Market & the EU Climate Regime
Interactions and frictions in the legal norm systems

19. Sippo Rossi
Bots on Social Media
The Past, Present and Future

20. Sumair Hussain
Essays on Disclosures

21. Kseniia Kurishchenko
Novel Mathematical Optimization Models for Explainable and Fair Machine Learning

22. Maylis Saigot
At The Heart of Digital Collaboration
Navigating Interpersonal Affective Pathways in Digitalized Work Environments

23. Alessandro Spina
Essays in Financial Markets and Beliefs

24. Charlotte Cator
Transforming the city for sustainable futures?
Contestation and alternatives in Amsterdam

25. Daniela Micu
Essays on Design Science in Management and Marketing

26. Olimpija Hristova Zaevska
INDUSTRY 4.0 ECOSYSTEMS: STRUCTURE, VALUE CREATION AND IMPACT ON REGIONAL INNOVATION

27. Thomas Holde Skinnerup
The Role of Accounting in Governing Public Organizations
The Case of the Danish University Sector

28. Christina Juhlin
Management of the Meantime
- the senses, attachments, and time in the aesthetic city

29. Joachim Elmegaard
Navigating the Digital Frontier in Accounting
Transformative Impacts and Interplay of Digitalization, Accounting Systems, and Management Accountants in an Institutional Context

30. Leonardo Maria De Rossi
Bitcoin as a Sustainable Polycentric Digital Infrastructure

31. Kristoffer Halskov
Essays on Empirical Asset Pricing

32. Juliane Lang
Win-win for whom?
Power and inequality in the Chilean wine and farmed salmon value chains

33. Mikkel Munksgaard Andersen
Feasible or Fashionable? Investigating Organizational Adoptions of Social Impact Bonds

34. Stefanie Steinbeck
Affective milieu in children's museum experiences
An ethnographic study with young visitors at the Danish Workers Museum

35. Chiara Andreoli
Bridging Profit and Purpose: Tensions in Impact Measurement and Management in Impact Investing

36. Donghyun Kang
Essays in Bankruptcy and Financial Distress

37. Sigurður Páll Ólafsson
Essays on Household Saving and Pension Reform

38. Cristine Dyhrberg Højgaard
The emergence and endurance of civic action in fluid forms of organizing
A practice-oriented exploration of flexible forms of volunteering in loose organizational attachments

39. Nicola Ens
#Poshboss: An Ecological Ethnography of Digital Hustling

40. Raphaël Huleux
Essays in Quantitative Macroeconomics
Applications to Inequality, Monetary Policy, and Climate Change

41. Sara Blasco Román
From Insights to Action:
Understanding and Mitigating Organizational Errors

42. Kristina Kazuhara
Shifting Frames
Contextualizing Biculturalism in International Business Research

43. Oliver-Alexander Press
Essays on Household Finance

44. Lars Christian Larsen
Price Dynamics and Liquidity in Financial Markets

2025

01. Marie Arnbak-Hartzberg
What are we missing?
A study of the affective and collective dimension of work-related stress

02. Louise Jørring
Digitalization and the modification of frontline work
A study of frontline workers' pursuit of meaningfulness

03. Olivia Norma Jørgensen
Auditors as Idealists?
Climate Change, EU Sustainability Regulation, and the Efficacy of Assurance

04. Rikard Rosenbacke
Cognitive Challenges in Human-AI Collaboration
A Study on Trust, Errors, and Heuristics in Clinical Decision-Making

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens