UNKNOWN LAB / RESEARCH REPORT 01

医療・ケアにおける AIへの過度な依存

自動化バイアスと「確認可能な協働」に関する対象限定レビュー

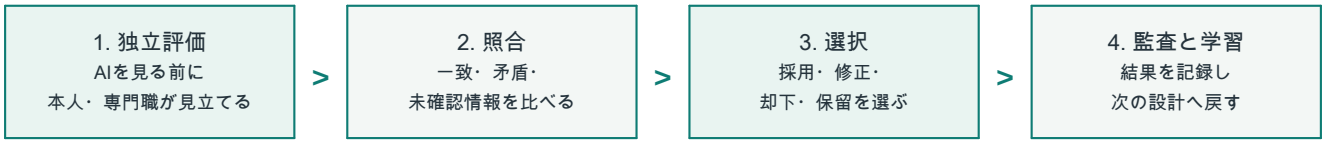
森元大樹 / Unknown Lab / ORCID 0009-0009-6444-750X
研究報告書（対象限定レビュー） Version 1.0 2026-07-12
DOI: 10.5281/zenodo.21327355

要旨と主要な結論

医療・ケアに人工知能（AI）を導入する際、「AIは補助にとどまり、人間が最終判断を行う」という原則が広く用いられる。しかし、人間が形式上の最終判断者であることは、AIの誤りを発見し、適切に修正できることを意味しない。本研究報告書は、医療・ケアにおける自動化バイアスの発生要因、臨床判断への影響、対策の実証状況を整理し、作業療法とETIC 2.0への含意を検討することを目的とした。PubMed、PMC、出版社ページおよび国際機関の公式資料を用い、自動化バイアス、確認の複雑さ、説明可能AI、信頼較正、認知的強制、AI出力の保留に関する主要なレビューと実験研究を対象限定で探索した。既存研究からは、正しいAIが臨床家の判断を改善しうる一方、誤ったAIは経験者を含む臨床家の判断を悪化させること、説明を付加しても誤誘導を十分に防げないこと、確認作業の認知負荷が高いほど人間による監督が機能しにくいことが示された。独立した初期判断、重要情報の並列表示、具体的な反証確認、低信頼出力の保留、利用後監査などは有望であるが、患者・生活アウトカムに対する効果は確立していない。作業療法・リハビリテーションを直接対象とした研究は乏しく、診断支援研究からの一般化には注意を要する。AIへの過信は個人の注意力だけでなく、人、AI、画面、業務環境、組織ルールの関係から生じる設計課題である。医療・ケアAIの評価では、AI単体の精度ではなく、誤答時を含む人間-AIチーム全体の成績と、採用・修正・却下・保留の過程を検証する必要がある。

01	正しいAIは判断を改善しうる一方、誤ったAIは経験者を含む専門職を誤答側へ動かす。
02	説明を付け、人間を最終判断者に置くだけでは、誤誘導を十分に防げない。
03	独立評価、情報の照合、出力の保留、利用後監査を組み合わせた設計が必要である。
04	作業療法への直接証拠は乏しく、ETIC 2.0への接続は設計上の推論として扱う。

図1 確認可能な協働の基本手順



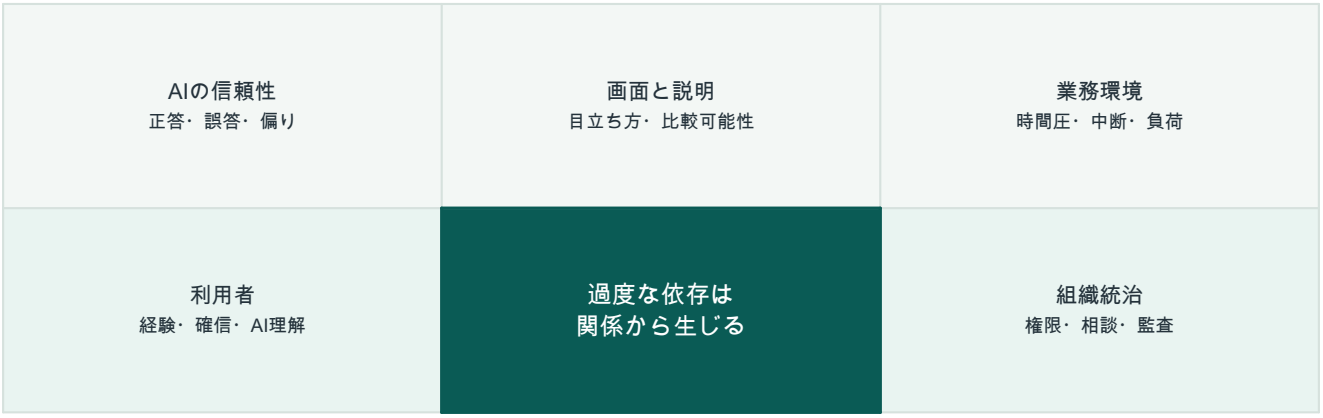
AIに従うか拒むかを一度で決めるのではなく、独立した見立て、照合、選択、事後監査を一つの循環として設計する。

Abstract

The principle that artificial intelligence (AI) should remain assistive while humans retain final decision-making authority is widely used in health care. Formal human authority, however, does not ensure that AI errors will be detected and corrected. This focused review examines the determinants and clinical effects of automation bias, the evidence for mitigation strategies, and the implications for occupational therapy and ETIC 2.0. Major reviews, experimental studies, and official guidance concerning automation bias, verification complexity, explainable AI, trust calibration, cognitive forcing, and AI output suppression were identified through PubMed, PMC, publisher websites, and official institutional sources. The reviewed evidence indicates that reliable AI can improve clinical performance, whereas erroneous or systematically biased AI can degrade clinicians' decisions, including those of experts. Explanations alone do not consistently prevent harmful reliance, and verification may impose cognitive demands equal to or greater than the original task. Independent initial assessment, side-by-side presentation of critical evidence, specific contradiction prompts, abstention or suppression of high-risk outputs, and post-deployment audit are promising, but their effects on patient- and participation-relevant outcomes remain uncertain. Direct evidence in occupational therapy and rehabilitation is sparse. Automation bias should therefore be treated not solely as an individual failure of attention but as a systems design problem involving people, AI, interfaces, workflow, and governance. Evaluation should focus on the performance of the human-AI team under both correct and erroneous AI conditions and document how recommendations are accepted, modified, rejected, or deferred.

Keywords: automation bias, artificial intelligence, clinical decision support, human-AI collaboration, explainable AI, occupational therapy, care ecosystem

図2 自動化バイアスをケア生態系の問題として捉える



過度な依存は一人の注意不足ではなく、人、AI、画面、業務環境、組織ルールの組み合わせで強まったり弱まったりする。

1. 背景

生成AIと機械学習を用いた意思決定支援は、医療記録の要約、画像診断、鑑別診断、治療選択、患者相談などへ広がっている。AIは大量の情報を短時間で処理し、見落としを補い、判断の一貫性を高める可能性がある。一方、AIが提示した答えは、利用者がその後の情報を見る枠組みそのものを変える。AIが誤っている場合、人間は誤りを訂正する安全装置になるとは限らず、正しかった初期判断をAIの誤答に合わせて変更することもある。

ParasuramanとRileyは、自動化への過度な依存をmisuseと位置づけ、監視の失敗や判断バイアスにつながると整理した[1]。Skitkaらは、自動化された監視・助言に依存することで、誤った助言を採用するコミッションエラーと、自動化が示さなかった問題を見落とすオミッションエラーが生じることを実験的に示した[2]。医療分野では、この傾向は一般に自動化バイアスと呼ばれる。

自動化バイアスと区別すべき概念として、automation-induced complacencyとアルゴリズム回避がある。前者は、自動化が監視しているという安心から人間の警戒や継続的確認が弱まる状態を指す。後者は、AIの失敗を見た後などに、AIが人間より良い成績を示していても利用を避ける傾向である[15]。したがって、安全なAI利用の目標は信頼を最大化することでも、AIを常に疑うことでもない。AIが当たりやすい条件と外しやすい条件に応じて依存度を変えられる、適切に較正された信頼が必要となる。

Kheraらは、AIが補助的な位置づけであっても、自動化バイアスを通じて臨床判断へ害を与えうるため、AI単体ではなく臨床家と組み合わせた結果を評価する必要があると指摘している[10]。

2. 目的

本報告書の目的は、次の4点を整理することである。

- 医療・ケアにおける自動化バイアスは、どのような条件で生じるか。
- 正しいAIと誤ったAIは、専門職の判断にどのような影響を与えるか。
- 説明表示、人間による最終判断、訓練などは、過度な依存を防げるか。
- 既存研究は、作業療法とETIC 2.0に何を示し、何をまだ示していないか。

3. 方法

3.1 調査方法

2026年7月11日までに公開された文献を対象に、PubMed、PubMed Central（PMC）、Crossref、出版社ページ、世界保健機関（WHO）の公式資料を探索した。主な検索概念は、automation bias healthcare、clinical decision support overreliance、AI advice reliance clinician、verification complexity、explainable AI clinician trust、trust calibration diagnostic AI、cognitive forcing AI、human oversight AI healthcareであった。

系統的レビュー、比較実験、無作為化研究、自動化研究の基礎論文を優先した。医療以外の研究は、概念の定義または対策候補の説明に必要な場合に限って採用した。全文を確認できない文献は、要旨の範囲を超える主張には用いなかった。著者、誌名、年、巻号、ページまたは記事番号、DOI、PMIDを可能な範囲で照合した。

最終的に本文で使用した主要資料は15件であり、基礎・概念研究3件、系統的レビュー3件、臨床家を対象とした実験・準実験6件、非臨床の介入実験1件、臨床論説1件、国際機関のガイダンス1件で構成された。

3.2 本調査の位置づけ

本調査は、あらかじめ定めた中心質問に関連する主要研究を整理した対象限定レビューである。PRISMAに準拠した系統的レビューではなく、全データベースを用いた網羅的検索、重複した独立査読、研究の統計的統合、正式なGRADE評価は行っていない。本文中の「比較的確立」「限定的」「未検証」は、一般読者に根拠の確かさを伝えるための編集上の区分であり、正式なエビデンス評価ではない。

4. 結果

4.1 自動化バイアスは、個人の性格だけでは説明できない

Goddardらの系統的レビューは、13,821件から74研究を採用し、自動化バイアスを左右する要因として、認知スタイル、課題経験、システムへの信頼と確信、作業負荷、課題の複雑さ、時間制約などを整理した[3]。これは、自動化バイアスが「AIを信じやすい人」の問題に限定されないことを示す。忙しい環境で、もっともらしい推奨が先に示され、その正誤を確かめるために多くの情報処理が必要な場合、AIの提案は判断の近道として使われやすい。

LyellとCoieraは、890件から40研究を採用し、そのうち医療研究は6件であったと報告した[4]。オミッションエラーを調べた研究の81%、コミッションエラーを調べた研究の91%が何らかの自動化バイアスを報告したが、非自動化条件と統計的に比較した研究は9件に限られていた。研究間の課題、指標、対象者は大きく異なり、効果量も十分に報告されていなかった。それでも、単一課題であっても正誤確認が複雑な場合には自動化バイアスが生じうること、複数課題では認知負荷が累積することが示唆された。

4.2 正しいAIは助けになるが、誤ったAIは人を誤答側へ動かす

AI支援の効果は一方向ではない。皮膚がん画像の研究では、質の高いAI支援は医師またはAI単独より診断成績を改善し、経験の浅い臨床家ほど利益が大きかった。一方、誤ったAIは専門家を含む幅広い臨床家を誤らせた[5]。

日本の医師22人が16の症例シナリオを判断した無作為化研究では、AIによる鑑別診断リストの有無で全体の診断正確性に有意差はなかった（57.4%対56.3%、p=0.91）[7]。ただし、正解がAIリストに含まれる場合には診断正確性が高まり、リスト使用時の判断の15.9%がオミッションエラー、14.8%がコミッションエラーに分類された。AIの平均的な有用性と、個々の症例で新たに生じる誤りは同時に存在しうる。

入院患者の急性呼吸不全を扱ったJabbourらの症例実験には、医師、ナースプラクティショナー、フィジシャンアシスタント計457人が参加した[9]。標準AIは基準時より診断正確性を2.9ポイント改善し、説明を付けた場合は4.4ポイント改善した。これに対し、系統的に偏ったAIは正確性を11.3ポイント低下させ、説明を付けても9.1ポイント低下した。偏ったAIに説明を付けたことによる改善は、偏ったAIのみの場合と比べて統計的に有意ではなかった。

前十字靱帯損傷のMRI画像を用いたWangらの研究では、40人の臨床家が200件をAI支援の有無で判断した[11]。AI支援により平均正確性は87.2%から96.4%へ改善したが、AI支援後に残った誤りの45.5%は自動化バイアスと関連していた。これらの結果は、「AIは役立つ」ことと「誤ったAIが人を誤らせる」ことが矛盾しないことを示している。

4.3 専門性は重要だが、専門家も免疫ではない

独力で課題を解く能力が高い利用者は、AIの失敗を見つけやすい傾向がある。しかし、皮膚がん画像研究や前十字靱帯損傷の研究では、誤ったAIの影響は経験者にも認められた[5,11]。したがって、自動化バイアスを初心者だけの問題とみなすことはできない。専門教育と経験は重要な防御要因であるが、システム設計と業務環境の代わりにはならない。

4.4 説明可能AIは、それだけでは安全策にならない

AIの結論に説明を付ければ、人間が誤りを見抜きやすくなるという考えは直感的である。しかし、臨床実験の結果は一貫していない。抗うつ薬選択の研究では、220人の臨床家に症例シナリオと機械学習による推奨を提示した[8]。AI

推奨は、臨床家が独立して判断した基準条件より全体の選択正確性を改善しなかった。誤った推奨に、理解しやすい限定的な特徴説明が付いた条件では、基準条件より正確性が低下した。

説明可能AIと臨床家の信頼に関するRosenbackeらの系統的レビューでは、778件から10研究が採用された[12]。説明によって信頼が高まった研究、変化しなかった研究、説明の複雑さや一貫性によって信頼が上下した研究が混在していた。採用研究のバイアスリスクは、いずれも中等度または中等度から高程度であった。したがって、説明可能AIの目的を「AIを信じてもらうこと」に置くのは適切ではない。必要なのは、AIの結論、根拠となる観察事実、欠けている情報、適用外条件、反証を利用者が照合できることである。

4.5 「人間が最後に確認する」は必要条件だが、十分条件ではない

AIを補助に限定し、最終責任を人間に残すことは重要である。しかし、確認の複雑さが高い場合、人間による監督は形式化しやすい。LyellとCoieraのレビューでは、責任を強調する、性能情報を示す、手作業の訓練を追加する、自動化の失敗例を教えるなどの介入は、効果がないか限定的であった[4]。確認のための認知負荷を下げない対策は、忙しい臨床環境では機能しにくいと考えられる。

日本の医師20人を対象にした信頼較正の準実験では、最終診断がAIの鑑別リストに含まれるかを確認させても、診断正確性は有意に改善しなかった（41.5%対46.0%、p=0.27）[13]。医師がAIを適切に信頼または拒否できた割合は61.5%で、AIの正確性を実際より約30%高く見積もった。単に「AIが正しいか確認してください」と求めることは、元の診断課題と同程度に広く難しい問いになりうる。

5. 過度な依存を減らす設計

現時点で、単独で自動化バイアスを防ぐことが確立した対策はない。複数の対策を組み合わせ、人間-AIチームとして検証する必要がある。

対策	根拠の現状	主な限界
AIを見る前に本人・専門職の初期判断を記録する	認知的強制の実験で過信低減の可能性[6]	非臨床課題が中心で、利用負担が増える
AIの提案と確認に必要な情報を横に並べる	確認負荷を下げる設計としてレビューが支持[4]	実臨床アウトカムの比較研究が少ない
「AIは正しいか」ではなく、具体的な矛盾を尋ねる	信頼較正研究からの発展的提案[13]	小規模研究であり効果は未確立
確信度、対象外条件、欠損情報を表示する	一部研究で信頼較正を補助[3,12]	表示だけで適切な依存になるとは限らない
危険または低信頼の出力を保留する	AI出力抑制のシミュレーションで有望[11]	前向きに実装して再検証されていない
誤答例を含む訓練を行う	理論的には妥当[3,4]	訓練単独の効果は概して小さい
採用、修正、却下、保留と結果を監査する	WHOの統治原則と整合[14]	自動化バイアス減少を直接示す比較研究が乏しい

Buçincaらは199人を対象とした非臨床実験で、AIの説明を利用者に検討させる認知的強制が、単純な説明表示より過度な依存を減らすことを示した[6]。一方、過度な依存を最も減らした設計ほど、利用者からの主観的評価は低かった。安全性を高める小さな手間は、使いやすさや利用継続と衝突する可能性がある。

Wangらは、AIが人間を誤らせる確率が高い出力を提示しない抑制戦略により、自動化バイアスを41.7%減らせると推定した[11]。ただし、この値は既存実験データに基づくシミュレーションであり、抑制機能を実装した前向き試験の結果ではない。それでも、AIが常に答えを返すのではなく、「判断を保留する」「追加情報を求める」「専門職相談へ戻す」ことを正常な動作として設計する重要性を示している。

6. 作業療法とETIC 2.0への含意

6.1 診断研究を、そのまま生活支援へ一般化しない

今回確認した臨床研究の多くは、診断、画像、薬剤選択、症例シナリオを扱っていた。作業療法士、本人、家族を含み、長期的な作業参加や生活の質を主要アウトカムとした自動化バイアス研究はほとんど確認できなかった。そのため、以下のETIC 2.0への接続は、既存研究から導かれる設計上の推論であり、ETIC 2.0自体の有効性を示す直接証拠ではない。

6.2 「人間責任」を実行可能な条件として捉える

作業療法場面でAIが「転倒リスクが高いので外出を控える」「この自主訓練を続ける」と提案したとき、最終判断者を作業療法士または本人と定めるだけでは不十分である。AIとは独立した見立てを形成できる情報と時間、AIの提案を拒否・修正・保留できる権限、他職種へ相談できる経路が必要になる。

ETIC 2.0が掲げる人間責任の保持は、責任者の名前を記録するだけでなく、責任を実行できる条件の設計として具体化する必要がある。

6.3 「説明可能性」を「照合可能性」へ広げる

生活支援では、AIの推論過程を技術的に説明するだけでなく、その提案が本人の目標、生活史、住環境、家族負担、地域資源と整合しているかを確認する必要がある。たとえば、転倒確率だけでなく、「本人はどこへ行きたいのか」「外出しないことによる不利益は何か」「住宅改修や同行支援で危険を下げられないか」を同じ画面または同じ話し合いに置く。これはAIの説明を読む作業ではなく、異なる情報を照合する作業である。

6.4 ケア生態系の問題として監査する

AIへの過度な依存は、本人または専門職一人の問題ではない。画面上でAIの結論が最も目立つ、確認時間がない、異議を述べる担当者がいない、AIを使わない選択肢がない、誤りを記録する仕組みがないといった条件によって生じる。ETIC 2.0では、人、AI、デバイス、データ、生活環境、ケア調整、組織統治の関係をケア生態系として捉える。この視点からは、AIの採用率だけでなく、修正率、却下率、保留率、誤誘導、利用者集団ごとの差を継続的に確認することが重要となる。

7. 考察

本調査から得られる中心的な示唆は、AIへの依存を「信じるか、信じないか」の二択で扱わないことである。信頼できる条件ではAIを活用し、信頼性が低い条件では立ち止まる能力が必要である。その能力は、利用者の注意力だけで生まれるものではない。AIを見る前に独立した判断を持つ、AIの根拠と反証を比較できる、AIが答えを保留できる、利用後の結果が監査されるという作業手順に支えられる。

また、AI単体の性能評価だけでは、臨床的安全性を十分に説明できない。高精度なAIであっても、誤答時に人間を強く誤らせるなら、人間-AIチームの利益は小さくなる可能性がある。逆に、不完全なAIでも、人間が誤りを見つけやすく、適切に拒否できる設計であれば有用性を保てる可能性がある。AI導入前の評価には、正答条件だけでなく、系統的な偏り、もっともらしい誤答、情報欠損、対象外集団を含む試験が必要である。

説明可能性についても、信頼を高めたかではなく、適切な採用と拒否を可能にしたかで評価すべきである。分かりやすい説明は、誤った提案を説得的に見せることがある。主観的な安心感や満足度だけを安全性の代替指標にしてはならない。

8. 限界

本調査には複数の限界がある。第一に、対象限定レビューであり、網羅的な系統的レビューではない。第二に、既存研究の多くが症例シナリオ、画像、短時間の実験であり、実際の患者アウトカムや長期的な生活への影響を直接測定していない。第三に、医師と診断課題への偏りがあり、作業療法、リハビリテーション、ケア職、本人、家族への一般化には不確実性がある。第四に、従来型の臨床意思決定支援と、対話的で出力が変化する生成AIでは、利用方法とリスクが異なる。自動化研究の人間工学的知見は参考になるが、生成AI固有のリスクを直接証明するものではない。第五に、対策研究は小規模または非臨床研究が多く、複数対策を組み合わせた実装効果は確立していない。

9. 結論

医療・ケアにおけるAIへの過度な依存は、利用者個人の注意不足だけでは説明できない。認知負荷、時間圧、AIの信頼性、説明の見せ方、正誤確認の難しさ、業務環境、組織ルールが組み合わさって生じる。正しいAIは専門職の判断を改善しうる一方、誤ったAIは経験者を含む専門職を誤答側へ動かす。説明を付け、人間に最終責任を残すだけでは、十分な安全策にならない。

今後の医療・ケアAIには、AIを見る前の独立評価、重要情報の並列表示、具体的な反証確認、低信頼出力の保留、採用・修正・却下・保留の監査を組み合わせた「確認可能な協働」が必要である。作業療法とETIC 2.0にとっては、本人の意味ある作業と参加を中心に、人、AI、データ、環境、制度の関係を設計し、その結果を生活アウトカムで検証することが次の研究課題となる。

研究倫理・利益相反・AI利用

- 本報告書は公開済み文献を対象とし、個人を対象とする研究は実施していない。
- 開示すべき利益相反はない。
- 文献探索、情報整理、書誌確認、文章作成および編集に生成AIを使用した。著者が内容を確認し、最終的な責任を負う。
- 本報告書は未査読であり、個別の診断・治療・リハビリテーションを指示するものではない。

参考文献

1 Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Human Factors. 1997;39(2):230-253. doi:10.1518/001872097778543886 (https://doi.org/10.1518/001872097778543886)

2 Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? Int J Hum Comput Stud. 1999;51(5):991-1006. doi:10.1006/ijhc.1999.0252 (https://doi.org/10.1006/ijhc.1999.0252)

3 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089 (https://doi.org/10.1136/amiajnl-2011-000089)

4 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423-431. doi:10.1093/jamia/ocw105 (https://doi.org/10.1093/jamia/ocw105)

5 Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-1234. doi:10.1038/s41591-020-0942-0 (https://doi.org/10.1038/s41591-020-0942-0)

6 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):Article 188, 1-21. doi:10.1145/3449287 (https://doi.org/10.1145/3449287)

7 Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. Int J Environ Res Public Health. 2021;18(4):2086. doi:10.3390/ijerph18042086 (https://doi.org/10.3390/ijerph18042086)

8 Jacobs M, Pradier MF, McCoy TH Jr, et al. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry. 2021;11(1):108. doi:10.1038/s41398-021-01224-x (https://doi.org/10.1038/s41398-021-01224-x)

9 Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295 (https://doi.org/10.1001/jama.2023.22295)

10 Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA. 2023;330(23):2255-2257. doi:10.1001/jama.2023.22557 (https://doi.org/10.1001/jama.2023.22557)

11 Wang DY, Ding J, Sun AL, et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. J Am Med Inform Assoc. 2023;30(10):1684-1692. doi:10.1093/jamia/ocad118 (https://doi.org/10.1093/jamia/ocad118)

12 Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI. 2024;3:e53207. doi:10.2196/53207 (https://doi.org/10.2196/53207)

13 Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. JMIR Form Res. 2024;8:e58666. doi:10.2196/58666 (https://doi.org/10.2196/58666)

14 World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. https://www.who.int/publications/i/item/9789240029200 (https://www.who.int/publications/i/item/9789240029200)

15 Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen. 2015;144(1):114-126. doi:10.1037/xge0000033 (https://doi.org/10.1037/xge0000033)