

論文とETIC 11

AIを使った後、 人の判断には何が残るのか

判断の偏りと内視鏡の検出成績から、支援・学習・技能保持を分ける

UNKNOWN LAB / 論文とETIC 11

AIを使った後、 人の判断には何が残るのか

支援・学習・技能保持を分けて考える



Daiki Morimoto / Unknown Lab

ORCID: 0009-0009-6444-750X

Version 1.0 / 公開日 2026-09-30

DOI: 10.5281/zenodo.23049020

未査読 / 専門職としての著者確認済み

独立専門レビュー未実施 / AI支援執筆

CC BY 4.0

30秒で読む

AIと一緒に仕事をしたときの成績が高いことと、使った人の力が育ったことは同じではない。反対に、AIを外した直後に成績が下がっても、直ちに以前の技能を失ったとは言えない。支援の上乗せがなくなっただけかもしれないからである。

GlickmanとSharotの心理実験では、偏ったAIとの反復的なやり取りによって、その試行のAI回答を見る前の人の判断にも偏りが生じた。正確なアルゴリズムと接した場合には、判断の正確さが改善した。[1] 一方、Pedersenらが内視鏡医13人による5,013件の大腸内視鏡検査を追った研究では、経験の浅い医師のポリープ検出割合はAI使用中に高まったが、使用終了後の向上や低下は統計的に明確ではなかった。[2]

二つの結果は矛盾とは限らない。前者は短時間の判断過程、後者は診療の異なる時期の検出成績を測っている。本稿は「AIが人を賢くするか、衰えさせるか」という二択を避け、人の側に残る変化を、時点・課題・経験・比較相手に分けて検討する。

よい成果物ができた。その次の仕事はどうか

次の場面は論点を説明する仮想例であり、原論文に記録された事例ではない。

ある職場で、AIが作った候補を見ながら報告書を整える。短時間で必要な項目がそろい、上司の評価もよい。数週間後、別の案件をAIなしで考える機会が来る。以前より要点を見抜けるようになった人もいれば、何から考えればよいのか迷う人もいる。さらに、AIで繰り返し見た説明の型を、別の状況でも無意識に当てはめているかもしれない。

ここには少なくとも三つの違いがある。道具がその場の仕事を助けたこと、道具とのやり取りから人が学んだこと、学んだ内容が次の状況にも適切であることだ。完成した報告書だけを見れば、これらは区別しにくい。医療・教育・事務・設計のいずれでも想像できる問題だが、想像しやすさは発生

頻度の証拠ではない。具体的な研究から、言える範囲を確かめる必要がある。

前回の論評は、AIを利用している最中の得点を、比較相手と重大な失敗に分けて読んだ。今回は時間を一歩先へ進める。人とAIの組合せの性能だけではなく、その経験を経た人が次にどのように判断するかが中心の問いである。AIに頼らないことを目的にするのではない。使い続ける場合にも、誤りを見抜き、必要なら別の方法へ切り替える能力がどう変わるかは重要である。

「AIの後」を三つの時点に分ける

「AIを使った後」という言葉は広い。ある問題についてAIの回答を読んだ直後も、やり取りを繰り返した後の新しい問題も、利用をやめて数か月後も含み得る。しかし、この三つでは測っているものが違う。

測定の時点	主に確かめること	それだけでは分らないこと
同じ問題でAI回答を見た後	助言の採用、修正、共同作業の成績	人が自分で理解・判断できるようになったか
反復利用中、新しい問題のAI回答を見る前	それまでのやり取りが初期判断へ及ぼす影響	AI利用を完全にやめた後も変化が続くか
利用終了後、AIなしで行う課題	支援がない状態での成績、保持や回復の可能性	比較条件がなければ、変化がAIによるものか

表1。二つの主論文の設計を比較するためのUnknown Labによる整理である。新たな実験結果ではない。

さらに、学習とは必ずしも望ましい変化ではない。誤った手掛かりを学ぶこともある。技能低下という語を使うなら、単に支援付きより成績が低いのではなく、適切な対照や利用前の状態に比べて、何の能力が低下したのかを示す必要がある。初心者が本来得られたはずの学習機会を失うことと、熟練者が既に持っていた技能を失うことも、同じ現象として扱わない。

判断の偏りは、AIの回答を見なくても移るのか

対象と研究の組立て

GlickmanとSharotの研究は、Nature Human Behaviourに掲載された一連のオンライン実験である。論文が報告する参加者総数は補足実験を含む1,401人で、募集にはProlificを用いた。Reporting Summaryのデータ収集期間は2021年4月から2024年3月である。これは研究全体の収集期間であり、同じ人を3年間追跡したという意味ではない。[1]

標本は便宜的な募集によるもので、医療職や一般人口を代表する集団ではない。各実験の人数は予備研究に基づく検出力計算から設定され、通常または矯正視力を持つ参加者が対象だった。注意確認への誤答が多い14人を人間同士の学習条件から除外したことなどが報告されている。総数だけを大きく表示して、すべての結果が1,401人で検証されたかのように扱うべきではない。

実験	対象・手続き	AIの影響を捉える位置
表情の平均判断	基礎データ50人、人から学ぶ条件50人、やり取りの4条件各50人	新しい顔の配列への最初の回答
動く点の方向判断	主実験120人。正確・一定方向に偏る・ばらつくアルゴリズムを比較	その試行のAI出力を提示する前の推定
同じ方向判断の反復	偏った条件50人、正確な条件50人	5ブロックを通じた判断変化
職業に関する印象判断	100人をAI画像50人と対照画像50人へ無作為割付	画像への接触前後の人物選択

表2。主実験の構成。[1] 補足実験は別に報告されている。行間で効果量や参加者数を合算しない。

第一の課題では、12枚の顔を短時間だけ見せ、全体として悲しそうか、嬉しそうかを判断させた。顔画像は二つの表情の間を段階的に変形したもので、各配列の基準は画像の構成から決められる。実際の人の気持ちを正しく推測したかを測る課題ではない。この区別は、結果をケアでの共感能力に直結させないために重要である。

研究者は、人の回答を学習した画像分類用のニューラルネットワークを用意した。人の基礎データには「悲しそう」と答えるわずかな偏りがあり、学習したモデルではその偏りが大きくなった。その後、別の参加者がこのモデルとやり取りした。参加者はまず自分で答え、AIの回答を見て、答えを変えるか選んだ。

注目するのは、AIに合わせて同じ回答を選び直した回数だけではない。以前の試行でAIの判断を見た経験によって、新しい試行の最初の答えが変わったかである。研究者は人間同士のやり取りも用意し、さらに情報源の実体と「AIだ」「人だ」という説明を入れ替えた条件も比較した。助言の内容と、助言者に対する期待を分けようとした設計である。

正確さと偏りを分ける仕組み

第二の課題では、動く点のうち右向きに動く割合を推定した。正確な回答を返すもの、一方向に高く見積もるもの、方向は一定しないがばらつくものという三つのアルゴリズムを比較した。この部分のアルゴリズムは、挙動を統制するため研究者が規則で実装したものであり、現在の大規模言語モデルの実測性能ではない。

偏ったアルゴリズムとばらつくアルゴリズムは、誤差の大きさをおおむねそろえた。したがって、単に誤りが多いことと、いつも同じ向きに誤ることを区別して調べられる。参加者は自分の推定を入力してから、共同回答に自分とAIをどの程度反映させるか決め、その後でAIの具体的な推定を見た。条件の順序も参加者間で割り当てた。[1]

この仕組みにより、その試行のAI回答を丸写しするだけでは説明できない変化を測れる。ただし、人が最初に回答した時点でも「この後でAIと共同回答を作る」という実験状況は続いている。完全に独立した別の日の仕事

とは違う。AI回答提示前の判断を、人間だけで行う長期追跡試験と呼ぶことはできない。

小さな偏りが移る一方、正確な助言は学習を助けた

比較	主な結果	読み方
表情課題、AI条件の基準時点→やり取り中	「悲しそう」の回答49.9%→56.3%	AI回答提示前の回答。割合の差は6.4ポイント
同条件の最初→最後のやり取りブロック	50.72%→61.44%	実験内での推移。年単位の変化ではない
表情課題、人間同士	50.6%→51.45%、 $p=0.48$	この条件では明確な増加を検出しなかった
動く点、偏ったアルゴリズム	平均的な偏り0.62→2.66ポイント	差の95%信頼区間0.76～3.35、 $p=0.002$
動く点、正確なアルゴリズム	平均絶対誤差15.03→13.48ポイント	改善1.55ポイント、95%信頼区間0.69～2.42
職業印象、AI画像への接触	白人男性を選ぶ割合32.36%→38.20%	画像を見た直後の判断。実際の採用行動ではない

表3。原著Resultsと各図に基づく主要数値。[1] ポイントは割合の差または0～100の回答尺度上の差を表す。異なる課題の値を一つの効果として比較しない。

表情課題では、AIとのやり取りが進むにつれ、AIの回答を見る前の人の判断も偏った。一方、人間同士の条件では同程度の変化は確認されなかった。情報の実体と情報源のラベルを交差させた比較では、出力の特徴だけでなく「AIからの回答だ」と受け止めることも、時間とともに生じる変化に関係した。人が不合理だから影響される、という単純な説明では足りない。正確だと期待する情報源から学ぶ行動でも、その情報が偏っていれば判断は偏り得る。

動く点の課題では、正確なアルゴリズムとのやり取りによって、人自身の推定誤差が小さくなった。反復した追加実験でも同じ方向の結果だった。したがって、この研究を「AIに接すると人の判断が悪くなる」という根拠として使うのは不適切である。どのような信号から学んだかによって方向が異なる。

参加者は、偏ったアルゴリズムより正確なアルゴリズムの影響を強く受けたと自己評価した。しかし、研究者が行動から算出した影響の比較では、そのような差を明確に検出しなかった。これを、二つの影響が厳密に同じ大きさだと証明された、とまでは読まない。異なる指標を標準化する方法や、差を検出する精度にも依存する。それでも、本人の「影響されていない」という感覚だけで、判断の変化を把握することの難しさを示す結果である。

生成画像による印象変化は、別の意味を持つ

第三の課題では、Stable Diffusionが生成した金融管理職の画像に接した後、誰がその職業らしいと思うかを尋ねた。対照群はフラクタル画像を見た。白人男性を選ぶ割合の変化はAI画像群で大きく、群間で変化量を比べた検定は $p=0.02$ だった。[1]

これは、同じ問題への助言を直接求めなくても、反復して見る表象が後の判断へ影響する可能性を示す。ただし、点の移動方向とは違い、顔だけから個人の職業を正しく当てる客観的な正解はない。職業集団の人口構成についての推定と、個人の適性や能力の評価を混同すべきでもない。現実の採用差別や診療上の差別が何件増えたかを測った結果ではない。

対照が中立的な図形であることにも限界がある。同じ偏った人物画像を人間が作成した場合、AI生成だと知らせない場合、偏りを修正した画像を見せた場合との比較ではない。画像群への接触が短期の選択を変えたことと、その変化がAIに固有であることは分ける必要がある。

実験の強さと、まだ言えないこと

この研究の強みは、結果が出た後に人の印象を尋ねるだけでなく、判断を繰り返し記録し、出力の誤り方や情報源を操作した点にある。補足資料では、表情判断を別の方法で数値化した分析、学習用データ量を人とAIでそろえた比較、条件内の時間を考慮した分析も行っている。[1] 一つの質問文や一つの集計だけに依存しないかを点検している。

補足の計算モデルでは、過去のAI出力から学習するモデルが、学習を含まない比較モデルよりデータに適合した。しかし、同じデータへの当てはまりは、それ自体で長期の認知変化を予測する外部検証にはならない。今回比較されていない説明まで否定したことにもならない。モデルの複雑さを増やす場合には、未知の課題や別の参加者で予測が改善するかが次の問題となる。

最も重要な限界は持続期間である。著者らも、偏りがどの程度長く続くかは今後の研究課題としている。同じ実験内で変化したことから、数か月後も回復しない、能力そのものが不可逆的に損なわれた、と結論づけることはできない。注意の向け方、一時的な期待、回答尺度の使い方が変わった可能性と、長く残る学習を区別する追跡が必要である。

原著には記述上の不整合もある。情報源のラベルを入れ替えた第二条件について、Resultsの一文は第一条件と同じ説明を繰り返しているが、Methodsと図1dでは、人の回答をAIの回答として提示する条件と記載されている。本稿は後者に基づいて整理した。原文を黙って訂正したり、疑義のある一文から実験全体の無効を主張したりはしない。補足実験の番号にも本文との対応の逆転が見られるため、番号だけでなく実際の手続きを照合した。

内視鏡では、支援を外した後の成績をどう測ったか

対象、期間、介入と比較条件

Pedersenらは、2021年11月から2025年6月に、ノルウェーの大学病院、地域病院、高件数の外来内視鏡施設という3施設で前向き研究を行った。3期すべてに参加した13人の内視鏡医による5,034件から、人口統計情報が不十分な14件と同じ患者への重複検査に関する7件を除き、5,013件を解析した。患者の年齢中央値は59歳、女性は2,703人だった。[2]

医師は研究開始時の経験で分けた。経験の浅い群は過去の大腸内視鏡100～1,000件の7人、経験群は1,000件を超える6人である。全員が少なくとも100件の経験を持ち、未経験者を対象にした研究ではない。また、AIを使うかどうかを患者や医師ごとに無作為に割り付けた試験でもない。

各医師は、AI使用前、AI使用中、AI使用終了後の順に診療した。支援にはGI GeniusのCAdEソフトウェアversion 2を用いた。CAdEは内視鏡画像中のポリープ候補を見つける補助であり、患者の診断や治療方針を言語で提案するAIではない。導入前に標準的な研修を行い、原則として内視鏡を引き抜いて粘膜を観察する段階で装置を作動させた。

対象は通常の外來診療である。便潜血検査陽性を理由とする国の大腸がん検診プログラムの検査は除外された。検診では発見率が高く、実施施設もそろわないためである。一方、炎症性腸疾患などは実診療を反映するため含まれた。したがって、今回の割合を検診集団や別の施設の数値とそのまま比較することはできない。

主要指標は「5mm以上のポリープを一つ以上見つけた割合」

主要評価項目はPDR-5で、少なくとも一つの5mm以上のポリープが見つかった検査の割合を指す。ポリープの総数、病変を見落とした割合、がんを防げた患者の割合ではない。病理で確認した腺腫の発見割合であるADRとも同じではない。PDR-5はノルウェーのGastronet登録で用いられ、迅速に把握できる利点がある一方、5mmという大きさの判断は観察者による。

この違いは、技能を論じる際にも重要である。候補の発見、正確な大きさの推定、病理の判定、処置の適切さは別の能力である。ある指標に変化がなくても、すべての能力が変わらなかったとは限らない。逆に、検出割合が上がっても、重要な病変の見落としや患者の予後が改善したとまでは言えない。

医師の経験	使用前	使用中	使用終了後
経験の浅い7人	216/678件、31.9%	257/651件、39.5%	173/476件、36.3%
経験群6人	304/1,075件、28.3%	291/1,088件、26.7%	284/1,045件、27.2%
全13人	520/1,753件、29.7%	548/1,739件、31.5%	457/1,521件、30.0%

表4。PDR-5の分子・分母と割合。原著Table 2と補足Table 1sに基づく。
[2] 5,013件は検査の数で、5,013人の医師による独立した経験ではない。

経験の浅い群では、AI使用中の割合が高かった。しかし、使用終了後の36.3%を使用前の31.9%と比べ、4.4ポイントの学習効果が残ったと即断してはいけない。同じ医師でも各期の検査数は異なり、患者の年齢や検査理由も変わる。全体の割合と、医師内のまとまりを考慮した推定値は、同じ計算ではない。

「技能低下なし」の結論を、表の列まで確かめる

著者らは一般化線形混合モデルを用い、同じ医師による検査が似やすいことをランダム効果で扱った。性別、年齢、検査理由、腸管の前処置などを検討し、単変量解析で $p < 0.1$ となった説明変数を多変量解析に採用した。施設を加える感度分析でも結論は大きく変わらなかったという。[2]

ただし、「調整した研究だから、表中のすべての値が多変量調整済み」とは読めない。AI使用中の経験の浅い群のOR 1.43は多変量解析の列にある。一方、使用前と終了後を比べるOR 1.03は単変量解析の列にあり、多変量解析の対応欄は空欄である。補足Table 3sでも同じ配置を確認できる。

比較 (使用前が基準)	OR [95%信頼区間]	原著の解析欄
経験の浅い群、使用中	1.43 [1.11, 1.84]	多変量混合モデル
経験群、使用中	0.92 [0.76, 1.12]	単変量混合モデル
経験の浅い群、終了後	1.03 [0.79, 1.34]	単変量混合モデル
経験群、終了後	0.93 [0.76, 1.12]	単変量混合モデル

表5。原著Tables 3–4および補足Tables 2s–3s。[2] 単変量モデルにも医師のランダム効果は含まれるため、単純な集計だけのORではない。ただし、患者の複数属性を同時調整した推定値でもない。

ORはオッズ比であり、1.43を「見つかる確率が43%増えた」と書くのは不正確である。記述的な割合の差は表4で読み、モデルに基づく関連は表5で読む。また、経験の浅い群だけが統計的に有意だったことは、それだけで二群間のAI効果の差が有意であることを意味しない。群間の差を論じるには、その差自体の検定や推定が必要になる。

使用終了後のOR 1.03の区間は0.79～1.34である。低い場合も高い場合も含み、事前に定めた許容幅内の同等性を証明した値ではない。「今回の測定では明確な向上も低下も検出しなかった」が適切な読み方である。13人という医師数、とくに経験の浅い群の終了後検査数が476件にとどまる点にも注意が必要だ。検査を多数集めたことと、医師レベルのばらつきを十分に捉えたことは同じではない。

著者らは個人ごとの割合の増減をupskilling、deskillingとして記述している。しかし、ある医師の割合が下がったことだけで、その人がAIによって技能を失ったとは確定しない。少数検査の偶然、対象患者の違い、測定誤差も考えられる。本稿では、こうした個人別の増減を、因果的な技能獲得・喪失の人数として数え直さない。

時間が長い研究ほど、学習を証明しやすいとは限らない

この臨床研究は実診療を追った点で心理実験を補う。一方、全員が使用前、使用中、終了後の固定順で進むため、AIへの接触と時間経過を完全には分離できない。通常診療の経験、研修、担当患者、勤務配置の変化も同時に起こり得る。

経験の浅い群では、最初の検査からの平均経過月数が各期で6.8、17.7、23.8か月、経験群では2.2、7.6、11.8か月だった。[2] これらは各期に行われた検査の時点の平均であり、全員が同じ月数だけAIを使い、同じ期間だけ中止したという情報ではない。23.8か月を「AIをやめて約2年後の追跡」と説明するのは誤りである。

研究者は経過時間とPDR-5の関係も調べたが、時間の係数が有意でないことによって、自然な学習や時期による交絡が消えるわけではない。一定の直線的な変化として測ることが十分か、各人の変化が異ならないかという問題が残る。多変量解析へ変数を入れるかどうかをp値で選ぶ手続きも、交絡要因をすべて除ける保証にはならない。

著者らは、別の観察研究で報告された技能低下との違いを議論している。重要なのは、AIを使う検査と使わない検査が同じ時期に混在する状態と、AI利用の期を終えた後の状態が異なる点である。ただし、「低下は一時的で、やめれば回復する」という説明は、今回の研究で直接検証された回復機序ではない。研究間の違いを説明する仮説として留めるべきである。

ここから、利用終了後も変わらなかったから人は影響を受けない、とも、短期実験で偏ったから長期的に必ず衰える、とも言えない。何日後、どの課題で、どの程度の変化を検出できる設計だったかが必要である。二つの論文を同じ効果量に換算して平均することはしない。

時間、費用、患者の経験は十分に測られたか

内視鏡研究の補足表では、経験の浅い群の観察時間中央値は使用前10分、使用中12分、終了後9分だった。経験群は9分、10分、9.5分である。ただし、生検や処置を行った検査はこの集計から除外され、全体の2,830件が対象外となっている。[2] これを診療1件当たりの総時間や、純粋なAI追加時間と解釈することはできない。

患者が回答する結果指標の回収は2,341/5,013件、46.7%だった。疼痛は主要モデルに含まれず、観察時間も上記の対象制限のため含まれていない。検出割合が変わったかという結果と、患者が安心できたか、負担が減ったかという結果は区別する必要がある。回答しなかった患者の経験が同じとも限らない。

装置費用、追加研修の工数、勤務配置、保守費用を合わせた費用対効果は、この研究からは確定できない。心理実験の参加報酬は報告されているが、実際の職場で学習を維持する費用とは別である。どちらも体系的な面接による混合研究法の研究ではない。内視鏡の音や表示への反応に関する記述を、代表性のある質的テーマや発生率として扱わない。

利益相反も読む必要がある。内視鏡研究では筆頭著者へのESGEとMedtronicによる研究助成などが開示されている。これだけで結果を否定する根拠にはならないが、製品一種類の研究から配置戦略を勧める際には、別の製品や独立した研究での確認が重要である。GlickmanとSharotは利益相反なしと申告し、資金提供者は研究設計や解析などに関与しなかったと記載している。[1,2]

著者の結論とUnknown Labの批判を分ける

論点	著者の結論・解釈	Unknown Labの評価
偏りの反復学習	偏ったAIとのやり取りが人の判断を偏らせ得る	実験内の操作が根拠。持続性や臨床被害へは拡張しない
正確な出力からの学習	正確なアルゴリズムは人の判断を改善し得る	AI利用そのものではなく、学ぶ信号と条件を問う
内視鏡使用中の改善	経験の浅い群でPDR-5が高まった	固定順の臨床研究。無作為化による因果効果とは分ける
内視鏡使用終了後	向上も低下も観察されなかった	非有意は技能保持の証明ではない。区間・検査数・交絡を残す

表6。原著の主張と本稿の解釈を区別したもの。[1,2] 原著者がETIC 2.0を検証したわけではない。

本稿の第一の批判は、AIを外した状態を一種類として扱うことへの批判である。回答直前にAIが表示されていないことと、過去の経験から独立していることは違う。逆に、過去に使ったことがあるというだけで、現在の判断をAIの影響と呼ぶこともできない。

第二の批判は、人の技能を単一の成績へ押し込めることへの批判である。候補を見つける力は上がっても、誤った候補を退ける力は育っていないかもしれない。平均的な正答率が変わらなくても、まれだが重大な見落としが増える可能性は別に点検する必要がある。ただし、これらは今回の論文で示された追加の害ではなく、測定の不足から生じる次の問いである。

第三の批判は、測定項目を増やせば解決するという発想への批判である。毎回の判断を記録し、定期試験を課せば、利用者と職員の負担が増える。臨床に有効な支援を研究のためだけに外すことにも倫理的な問題がある。安全に再現できる模擬課題、既存の研修、業務記録などを使い、何が分かれば判断を変えるのかを先に決めた最小限の比較が必要である。

この研究からETIC 2.0が学べること

考え方と重なる点

ETIC 2.0は、技術だけではなく、人、環境、制度、支援の関係からケアを考える。[3] 今回の研究で、人の判断が固定した入力ではなく、AIとのやり取りや経験に伴って変わり得ると示されたことは、この関係的な視点と重なる。ただし、重なることは理論の直接検証ではない。既存の学習理論や助言利用の研究でも、同じ現象を説明できる可能性がある。

正確なAIから人が学ぶ場合も、偏りを受け取る場合もあるという両方向性は重要である。人とAIを組み合わせたという形式だけを価値とせず、誰がどの能力を獲得し、誰に負担が残るかを問う必要がある。支援を受ける本人、家族、専門職を同じ「利用者」にまとめれば、それぞれにとっての改善や不利益が見えにくくなる。

答えるべき問い

第一に、どの人の、何の変化を説明するのか。医師の検出技能、患者が相談内容を整理する力、家族が異変に気づく力は同じ尺度では測れない。今回の主論文から、相談前にAIを使った患者の判断がどのように変わるかを確定することもできない。対象を変えるときは、新しい根拠が必要になる。

第二に、変化をどの時間幅で扱う必要があるか。一回の助言の採用だけで十分な問いもあれば、数日後の別課題への転移が問題になる場合もある。時間を含む複雑な説明を用意する前に、単純な利用前後の測定と適切な比較群で十分かを確かめたい。

第三に、支援を外す能力をどのように評価するか。AIを使わず全業務を再現できることを常に求めるのではない。停止時に必要な支援を選び、誤りが疑われるときに相談し、危険な提案を実行しないことが重要な場合もある。業務に必要な安全な代替能力と、昔と同じ方法を維持することを区別しなければならない。

具体的な見直し案

以下はUnknown Labによる未検証の提案であり、原論文の結論でも、臨床運用の推奨基準でもない。まず「利用中の成果」と「人に残る変化」を別々に記録する。後者には、同じ問題の回答修正だけでなく、新しい問題での初期判断と、必要な場合に限り支援終了後の測定を含める。

点検したいこと	最小限の比較案	提案が弱まる、または追加不要となる条件
その場の助けと学習の違い	練習時間をそろえ、支援付き成績と新しい無支援課題を分ける	支援付き成績だけで後の必要能力を十分予測できる
一方向の誤りの学習	誤差量をそろえ、一定方向の誤りとばらつく誤りを比較する	方向による追加の違いが十分な精度で否定される
持続と一時的な適応	即時と遅延時点を事前に決め、利用履歴を記録する	即時効果が消え、重要な後続判断への影響が認められない
経験による違い	経験区分だけでなく利用前技能と接触量を測る	経験別の追加説明が未知データの予測を改善しない
点検の負担	誤りの減少と追加時間・人員を併せて比較する	既存の簡単な研修・監査と同等で、負担だけ増える

表7。将来の比較を考えるための提案。実施済みの研究計画や検証結果ではない。人を対象に実施する場合には、別途、倫理・安全・同意と実施可能性を確認する。

比較相手は「何もしない群」だけでは足りない。適切な例題、正解の解説、人による助言、同じ時間だけの独立した練習など、既存の方法で同じ改善が得られる可能性を残す。AIによる説明が長い条件と短い条件を比較するのなら、読む時間や情報量も違ってしまう。道具の名称以外に何が変わったかを確認する必要がある。

また、平均点と重大な失敗は別に記録したい。新しい課題での小さな誤差が減っても、危険な提案を退けられなくなれば、単純な合計点では評価しにくい。しかし、重大さの定義を結果を見た後で変更すれば、都合のよい

結論を作れる。どの誤りが重要かは専門職や当事者と事前に定め、AIがその場で付けた点数だけに委ねないことが必要である。

何が分かれば、この見直し案を小さくできるか

今回の文献は、人の判断を時間の中で評価する理由を与える。しかし、どの現場にも新しい大規模監視を導入する理由にはならない。もし、適切な対照を置いた研究で、利用後の差が小さく、事前に合意した重要度の範囲内に収まり、重大な失敗にも違いがなければ、追加の追跡を簡略化できる。差が有意でないことだけではなく、十分に狭い不確実性の範囲で判断したい。

反対に、AIがよく間違える方向へ人の初期判断が移り、支援を外した遅延課題でも残るなら、利用中の高い得点だけで安全を説明する考え方は弱まる。その場合も、原因を単一の理論へ帰属させず、学習、注意配分、信頼、練習機会の減少という既存の説明を比較する必要がある。

ETIC 2.0の視点を使った説明が、これらの既存の方法に比べて予測や支援判断を改善しないなら、新しい概念を追加しないことも正当な結論である。理論を守るためにすべての結果を取り込むのではなく、どの観察なら追加説明が不要になるかを残しておくことが、批判的な論評の役割である。

次に読む論文

Bastaniらの「Generative AI without guardrails can harm learning: Evidence from high school mathematics」を推薦する。[4] 高校数学の教育場面で、AIを使える練習と、使えない試験を区別している。単純なチャット型支援と、教師が用意した情報やヒントを組み込む支援を比較した点が、今回の問いを次へ進める。

次に確認したいのは、支援を取り除くこと自体ではなく、練習中の設計によって何が変わるかである。正解を直接受け取ることと、自分で考えるた

めの手掛かりを受け取ることを比べる際、教材作成に必要な人の労力も含めて読む。学校の短期試験の結果を、そのまま医療職の長期的な技能保持に移すことはしない。出版社は2025年8月に著者所属の訂正を記録しており、次の読解では訂正版と補足資料を正本にする。

結論

AIの利用によって成果物が良くなったことと、人の判断が育ったことは同じではない。今回の心理実験は、AIの出力に含まれる偏りだけでなく、正確さも人の後続判断へ移り得ることを示した。内視鏡研究は、利用中の検出成績と利用終了後の成績を分ける必要を示す一方、長期の技能保持や低下を確定するには限界があった。

問うべきなのは「AIを使う人は衰えるのか」という一般的な断定ではない。何を手掛かりに学び、どの課題へ持ち越し、どの程度続き、本人と周囲にどのような結果をもたらすのかである。支援中、回答提示前、利用終了後という時間の違いを明示するだけでも、同じデータから導きすぎる結論を減らせる。

本稿は公開研究に基づく批判的論評であり、新たな臨床試験、独立した再解析、ETIC 2.0の有効性の証明ではない。人に残る変化を測るという提案自体も、既存の簡単な方法を超える価値があるか、今後の比較に開かれている。

参考文献

1. Glickman M, Sharot T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*. 2025;9:345–359.
<https://doi.org/10.1038/s41562-024-02077-2> (原文全文、Supplementary Information、Reporting Summaryを参照)
2. Pedersen TA, Mori Y, Botteri E, et al. Learning and deskillling effects of artificial intelligence in colonoscopy among endoscopists with different levels of experience: a pragmatic, prospective trial. *Endoscopy*. 2026;58:1003–1014.
<https://doi.org/10.1055/a-2858-7084> (原文全文、補足Tables 1s–3sを参照)
3. Morimoto D. エコテクノロジカル・ インテグラティブ・ ケア(ETIC)2.0, Version 2.1改訂版: AI時代の作業療法とケア生態系設計. Unknown Lab; 2026.
<https://doi.org/10.5281/zenodo.21848456> (公開版の理論本文)
4. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*. 2025;122:e2422633122.
<https://doi.org/10.1073/pnas.2422633122> (次に読む論文。著者所属の訂正あり)