UNKNOWN LAB / 論文とETIC #8

「上書きできる」は、医療AIを統制できている証明ではない

健康領域42文献・168 passagesとHuman-AI組織設計から、誰が可能にし、行使し、評価するかを読む

Daiki Morimoto / Unknown Lab / ORCID 0009-0009-6444-750X
批判的論評 Version 1.0 2026年9月10日
DOI: 10.5281/zenodo.22667804

要旨

健康領域のmeaningful human controlを扱う42文献・168 passagesのレビューでは、臨床家は統制者として25回、患者・代表者は5回言及された一方、評価者は163 passagesで特定されなかった。著者らは統制を可能にする、実行する、評価する3過程を提案し、MHCという表示が倫理的正当性を代替する危険を指摘した。

PuranamのHuman-AI組織設計論は、分業を逐次 / 並列と専門化 / 同一仕事で、学習をfeedback依存とcommunicationで分ける。2論文を重ねると、上書き権限だけでなく、時間、情報、対象外事例、feedback、独立評価、停止後の訂正を検証する必要が見える。ただし両論文はETIC 2.0の有効性を証明していない。

医療AIの画面に、大きな「上書き」ボタンがあるとする。最終決定は人が行い、AIの提案を採用しない自由もある。書類には「human in the loop」と記載され、責任者の欄には医師の職名が入っている。

それでも、その人がAIを本当に統制できているとは限らない。

判断まで30秒しかなく、AIが何を根拠にしたか分からず、上書きすると理由説明だけを人が求められ、システム停止の権限は別部署にあるかもしれない。上書き後に何が起きたかを監査する人がいなければ、同じ問題は次の版にも残る。形だけ人を置くことと、意味のある統制を設計することは同じではない。

今回読むケア領域論文は、Hille、Hummel、Braunによる健康領域のmeaningful human control（以下MHC）レビューである。5データベースとAI倫理ガイドライン集合から42件を選び、168の文章区間を分析した。臨床家は統制者として25回言及された一方、統制を評価する主体は168区間中163区間で特定されなかった。著者らは、統制を可能にすること、実際に行うこと、機能したか評価することを分ける必要があると論じる。[1]

もう1本は、PuranamによるHuman–AI collaborative decision-making（HACD）の組織設計論である。人とAIが何を分担するかだけでなく、誰にどのfeedbackが返り、互いの経験からどう学ぶかを別の設計軸として示す。[2]

この2本を組み合わせると、「人が最後に決める」という一文では落ちる問題が見える。誰が統制を可能にするのか。誰がいつ行使するのか。誰が結果を評価するのか。そして、人とAIは何を見て次の判断を変えるのか。本稿はMHCを称賛するためではなく、この問いがまだどこまで未解決なのかを読む。

30秒で分かる研究の要点

項目	Hille et al.	Puranam
論文の種類	健康領域のMHCに関するsystematic reviewとcontent analysis	Human–AI意思決定を組織設計として捉えるPoint of View
対象	42 records、168 passages	参加者・datasetなし
期間	2021年9月27日まで検索、収載文献は2016ー2021年	2021年公表の概念論文
方法	2検索経路、3 coding cycles、9 code groups	分業2軸と学習配置2軸の類型化
主な数値	controller不明138/168、evaluator不明163/168、clinician 25、patient/representative 5	効果量なし。2つの2×2は概念上の設計空間
著者の結論	健康領域に頑健なMHC概念はなく、enable・enact・evaluateを反復して設計する必要がある	分業と学習配置を交差させれば、HACDの探索を体系化できる
直接言えること	2021年までの対象文献では、統制の役割・条件・評価者の記述が多くの場合不明確	組織設計として区別すべき論点が提示された
言えないこと	MHCが安全性や健康成果を改善する、42件が医療AI全体を代表する	どの配置が安全・有効・公平であるか
重要な限界	厳密な検索語、英語・ドイツ語、アクセス可能文献、passage頻度	概念論文で、医療・ケア固有の権利や救済を検証していない

Hille論文の頻度は、原則として168 passagesに付けたcodeの数である。たとえばclinician 25は「42文献の25件が臨床家を正式な統制責任者にした」という意味ではない。1つのpassageに複数のcodeが付き得るため、独立した人数や施設数でもない。この単位を取り違えると、精密に見える数字から誤った制度実態を作ってしまう。

English Summary

Hille, Hummel, and Braun reviewed 42 records and coded 168 passages concerning meaningful human control (MHC) over AI in health. Clinicians were coded as agents in control 25 times, patients or representatives five times, while the evaluator was unspecified in 163 passages. The authors propose three iterative strands: enabling, enacting, and evaluating control, and warn that an MHC label cannot replace ethical assessment of whether a system should be used.

Puranam conceptualizes human-AI collaborative decision-making as an organization design problem. Division of labor varies by sequential versus parallel interdependence and by specialization, while learning configurations vary by independent versus interdependent feedback and communication constraints. Combining the papers shows why an override button alone is insufficient: authority, time, information, excluded cases, feedback, independent evaluation, correction, and remedy must be tested. Neither paper validates ETIC 2.0.

Keywords: meaningful human control, health artificial intelligence, human-AI collaboration, organization design, feedback, patient rights, ETIC 2.0

本稿の位置づけ

本稿は公開済み論文、補足資料、原図をUnknown Labが批判的に読み解いた未査読の論評である。文献内容の頻度と現場の実装率、概念提案と因果効果、各著者の結論とUnknown Labの見直し案を分ける。個別の診断、治療、AI製品の安全認証や導入を指示しない。

月曜の補強記事から、何を深めるのか

直前の補強記事は、「AIに影響される人が、AIそのものを変更できないとき、何を変更可能と呼べるか」を扱った。変更対象をモデルの重みだけに狭めず、入力、出力の扱い、利用規則、説明、異議申立て、停止、調達や再設計まで層に分けた。[3]

木曜枠で一歩深める問いは、変更経路が存在するだけで十分か、である。

上書きボタンがあっても、使う時間がない。異議申立て窓口があっても、誰が再評価するか決まっていない。現場が誤りを報告しても、開発者へ届かず、次版の変更理由も戻らない。変更可能性は、操作項目の一覧だけでなく、権限、時間、情報、feedback、評価、責任の配置として検証しなければならない。

月曜記事は、本人が触れられる層を広げた。今回の2論文は、その経路を誰が支え、誰が使い、誰が評価し、組織がどう学ぶかを問う。月曜を要約し直すのではなく、変更可能性を運用可能な統制へ進める役割を持つ。

なぜ2本を組み合わせるのか

木曜の「論文とETIC」は、今回から1本の論文紹介ではなく、ケア領域の具体的な問題と、一般理論 / Human–AI研究を組み合わせる。目的は、ETICが正しいと宣伝することではない。既存研究で説明できる部分、まだ答えていない部分、反証できる形へ変えるべき主張を見つけることである。

Hille論文だけなら、健康領域のMHCが誰を統制者、可能化する主体、評価者として記述したかは分かる。しかし、分業後にどんな情報が誰へ返り、人とAIがどの経験から学ぶかは粗い。

Puranam論文だけなら、逐次 / 並列、専門化、feedback依存、communicationという設計軸は分かる。しかし、患者の権利、医療倫理、拒否、救済、責任の正当性は中心課題ではない。

2本は同じ答えを繰り返すのではなく、互いの空白を見せる。ケア領域の倫理と役割配置を、一般の組織設計で細分化する。同時に、一般理論が落とす権利と制度を、健康領域の文献から戻す。この往復が今回の読解の中核である。

Hilleらは、42件をどう選んだのか

Hilleらは2つの検索経路を使った。第1経路はGoogle Scholar、PubMed、ScienceDirect、Scopus、Web of Scienceである。健康、医療、臨床を表す語と、完全一致のmeaningful human controlを組み合わせ、2021年9月27日までを検索した。

第1検索の段階	件数
database検索で同定	583
access不能150件を除外	433
英語・ドイツ語以外9件を除外	424
重複26件を除外	398
title/abstractで健康との関連を確認	56
全文でMHCを扱う	39

第2経路は、Jobinらが集めた84件のAI倫理ガイドラインである。アクセスできない2件を除き82件を全文確認し、MHCに触れない76件を除いた。残る6件の要約・導入を健康との関連で読み、4件を採用した。第1経路39件と合わせて43件となり、経路間の重複1件を除いて最終42件となった。[1]

収載文献は2016〜2021年で、80%以上が2019年以降だった。著者らは42件からMHCを扱う168 passagesを抜き出し、Atlas.tiを用いて3回のcodingを行った。帰納的に作ったcodeを9 groupsへ整理し、各codeは少なくとも別の1人がcross-checkした。

ただし、formalなinter-rater reliability係数は報告されていない。また、42件中28件はMHCを1〜2段落だけ扱っていた。つまり、このレビューは豊富な実装研究を統合したというより、散在し、しばしば短く触れられた概念の使われ方を整理した研究である。

168 passagesは、何を示したのか

9 groupsのうち、本稿の中心となる値を再構成すると次のようになる。

問い	主なcodeと頻度	読み方
誰が統制するか	clinician 25、patient/representative 5、designer 4、society 3、不明138	臨床家への言及が中心だが、大半のpassageは主体を特定しない
誰が可能にするか	designer 19、legislative authority 6、researcher 4、不明137	統制は現場の意思だけでなく設計・法・研究条件に依存する
何のためか	ethical/legal alignment 23、responsibility attribution 22、不明124	責任と規範適合が主な目的だが、多くは目的不明
どんな条件か	tracing 17、informedness 13、tracking 12、dynamic task allocation 11、不明111	理由・役割への接続と、システム応答の追跡が別条件になる
どう行うか	exercising 29、monitoring 19、不明132	その場の介入と、経路の監視は同じではない
何が難しいか	algorithmic opacity 8、supervisor inefficiency 5、transparency 5、不明150	人を置くことで生じる処理負担自体も課題になる
誰が評価するか	executive authority 2、designer 1、legislative authority 1、researcher 1、不明163	評価者の記述が特に乏しい
いつ決めるか	forward 12、continuous 8、backward 3、不明146	事前、継続、事後の時間設計が多くは不明
MHCへの態度	肯定59、批判6、不明103	肯定的な言及はあるが、適否を評価しない記述が多い

最も目を引くのはevaluator unspecified 163である。統制を行う人を配置しても、その統制が目的を満たしたか評価する主体が見えなければ、仕組みは自己宣言になる。上書きできた回数だけでは、必要な場面で上書きできたか、見落としを減らしたか、新しい害を作らなかったかは分からない。

もう一つ重要なのは、trackingとtracingである。trackingは、システムが人の意図や理由に応答し続けることに関係する。tracingは、システムの状態や行為を、適切な道徳的理解を持つ人の設計・関与へたどれることに関係する。著者らによれば、この2条件をともに扱ったのは42件中8件だけだった。

ここでも、8/42はMHCが8件だけ成功したという意味ではない。対象文献が2条件をともに記述した数である。概念の記述密度と、現場の安全性を分けて読む必要がある。

統制を、可能化・実行・評価に分ける

Hilleらは結果を、次の3過程として整理した。

過程	中心の問い	例
統制を可能にする	何のための統制か。必要な条件を誰が作るか	設計者、立法、研究者、組織が情報・時間・権限・追跡可能性を用意する
統制を実行する	誰が、いつ、どの方法で介入・監視するか	臨床家、患者、代表者等が採用、上書き、保留、停止、照会を行う
統制を評価する	目的を満たしたと誰が、何で判断するか	独立監査、組織評価、規制、利用者feedback、害と格差の確認

3過程は一直線ではなく、反復する。評価結果が、次の設計条件と実行方法を変える。予期しない失敗が起きたとき、誰がどこへ戻し、何を変更し、再開を誰が認めるかまで含めて初めて循環になる。

この整理の強みは、統制を個人の注意力から制度へ広げる点にある。臨床家がAIを見抜けなかった、と後から責任を集中させる前に、十分な時間、情報、教育、代替経路、停止権限、監査が用意されていたかを問える。

一方、3過程の図を描くだけで実装は完成しない。Hilleら自身が、データは各要素を表面的にしか照らしていないと認める。誰を独立評価者とするか、患者の統制権をどこまで置くか、必要技能をどう確かめるか、評価結果を他の主体へどう戻すかは未解決である。

MHCというラベルは、倫理評価を代替しない

Hilleらの議論で最も重要なのは、統制の有無と、AI利用の正当性を分けた点である。

あるシステムにMHCが存在すると仮定しても、その場面でシステムを使うことが責任ある選択かは別問題である。人が十分に監視し、上書きできても、目的自体が不当、対象から特定の人が排除される、非AI経路が失われる、負担が患者や家族へ移る、といった問題は残り得る。

著者らは、MHCという表示だけで道徳的・法的正当性を得たように見せる危険をlabelling gapと呼ぶ。これはETIC 2.0にも直接関係する。チェック欄が埋まったこと、担当者名があること、上書き機能があることは、倫理的に適切、安全、有効、公平であることの代わりにならない。

さらに健康領域には、すでに医療倫理、責任、害の低減、患者の権利、専門職の手続がある。AI統制は、それらの外に新しい万能原則を置く作業ではない。既存の権利と制度が誰にどの統制を認めているかを確認し、AI導入で狭まらないようにする作業である。

Puranamは、人とAIの分業をどう分けたのか

PuranamはHACDを、複数のagentが1つの判断を作る組織として捉える。対象は、株式推薦、投資、量刑、採用screeningなど、判断が第三者によって実行されるknowledge workである。医療AIを直接研究した論文ではない。[2]

分業は2軸で表される。

	異なる仕事へ専門化	同じ仕事を行う
逐次	AIの出力を人が確認し、最終判断へ渡す等	同じ判断を順に見直し、後段が採否を決める等
並列	AIと人が異なる部分を作り、最終的に統合する	AIと人が同じ予測を独立に行い、両方を集約する等

逐次配置では、最後に位置するagentが他方の出力を採用・拒否できるため、decision rightsを誰が最後かという問題として扱える。しかし、最後に置かれたことと、意味のある判断ができることは同じではない。情報が欠け、時間がなく、拒否の不利益が大きければ、形式上の権限にとどまる。

並列に同じ判断を行う構成では、人とAIの誤りが完全には重ならないなら、集約によって精度が上がる可能性がある。だが、著者は可能な設計空間を示しているのであって、常に並列が優れると検証したのではない。人とAIが同じ偏ったdataや同じ組織目標に依存すれば、誤りは相殺されない。

分業だけでなく、学習の配置を見る

Puranamの次の論点は、人もmachine learningも経験から変化するため、協働を静止画として見てはいけないという点である。ここでいう学習はperformance向上に限らず、経験によりbeliefやbehaviorが変わることを指す。

学習配置も2軸になる。

Feedbackの関係	Communicationに制約	Input・process・output・feedbackを伝えられる
独立feedback	isolated learning	vicarious learning
相互依存feedback	coupled learning	coupled + vicarious learning

たとえば人とAIが共同で1つの報告を作り、全体が良かったか悪かったかだけを受け取る場合、両者の学習はcoupledになる。どちらの部分が結果へ寄与したか分からなければ、誤った成功や失敗から学ぶsuperstitious learningが起こり得る。

逆に、互いのinput、process、output、feedbackを伝えられるなら、一方が他方の経験から学ぶvicarious learningが可能になる。ただし、AIの内部処理が説明できない、情報量が人の処理能力を超える、専門領域が違い会話が成立しない場合、communicationは形式だけになる。

逐次配置には、さらに選別の問題がある。上流のAIが「低リスク」と判断した事例を人が見ない設計では、人が受け取る事例はAIによってcensorされる。人の学習機会は、AIの判断に依存する。後段の人が最終決定者でも、AIが見せなかった失敗からは学べない。

これは医療AIの統制に重要である。上書き率、AIとの一致率、最終判断の正答率だけでなく、誰がどの事例を見たか、どのfeedbackを受けたか、見なかった事例の転帰を誰が追うかを記録しなければならない。

2本を重ねると、何が見えるのか

Hilleの3過程とPuranamの設計軸を重ねると、MHCは次のように具体化できる。

過程	分業の問い	学習の問い
可能化	逐次か並列か。誰が何を担当し、誰が最後に決めるか	誰にどのfeedbackを返せるようにするか。communicationの障害は何か
実行	どの時点で人が確認、上書き、照会、停止できるか	その場の判断で、過去のどの経験と結果を見られるか
評価	統制者と別の評価者がいるか。対象外事例を誰が見るか	個別・全体feedbackを分け、誤りの帰属と次版への変更をどう記録するか

この表から、統制の失敗は3種類に分かれる。

1つ目は、権限の失敗である。ボタンはあるが使えない、停止できない、非AI経路へ戻せない。

2つ目は、認識の失敗である。介入すべき場面を見つける情報、時間、技能、比較対象がない。

3つ目は、学習の失敗である。個別の訂正はできても理由と結果が残らず、組織や次のmodel versionが同じ失敗を繰り返す。

MHCを実装すると言うなら、少なくともこの3種類を別に試験する必要がある。

著者の結論、観察結果、因果効果を分ける

Hilleらの結論

健康領域には、現時点で頑健なMHC概念がない。統制を可能にする、行う、評価する3過程を、関係するactorと制度を含めて反復的に検討すべきである。理論的参照点が曖昧なtheoretical gapと、MHC表示が倫理的正当性の代わりになるlabelling gapを避ける必要がある。[1]

Puranamの結論

人とAIの協働判断は、分業と学習配置を交差させた組織設計問題として検討できる。可能な組合せの多くはまだ十分に理論化・実証されておらず、研究者と実務家がdataで検討する余地が大きい。[2]

研究から直接確認できること

- Hilleの検索条件で、2021年9月までに42 recordsが採用された。
- 168 passagesでは主体、目的、条件、時間、評価者のunspecified codeが多かった。
- 臨床家、設計者、責任、規範適合、実行、監視、opacity等が明示的に言及された。
- Puranamは分業と学習配置の2つの類型を提示した。
- Puranam論文ではdatasetを生成・分析していない。

因果効果として言えないこと

- MHCを導入すれば診断、治療、安全、生活、費用、公平性が改善する。
- clinician 25やpatient 5の差が、実際の医療現場の権限差を数量化する。
- enable、enact、evaluateの3過程が、別の統治理論より優れる。
- 逐次 / 並列、専門化 / 同一仕事、4つの学習配置のどれが最善である。
- 人が最終判断者なら責任問題が解決する。
- この2論文がETIC 2.0の妥当性を証明する。

Unknown Labの批判的論評

1. 評価者の不在は強い警告だが、制度不在の直接測定ではない

evaluator unspecified 163/168は、MHC文献が評価主体をほとんど明示しなかったという重要な結果である。しかし、この値から「医療AIの163件で評価者がいなかった」とは言えない。分析単位はpassageであり、実装施設でもsystemでもない。

正確な批判は、対象文献がMHCを評価可能な要求へ変換できていなかった、である。次の研究では、誰が評価し、どのdataを見て、どの閾値で停止・再開・変更を決めるかを前向きに登録する必要がある。

2. 患者の言及が少ないことと、患者へ最終責任を渡すことは別である

controllerとしてclinician 25、patient/representative 5という差は、患者の関与を問う理由になる。ただし、患者をすべての技術判断の最終責任者にすれば解決するわけではない。

患者に必要なのは、理解できる説明、拒否、訂正、保留、別の人による再検討、非AI経路、代理判断の境界、報復なく異議を述べる手段である。専門職・組織・開発者が負うべき安全責任を、選択という名で本人へ移してはいけない。

3. 検索期限と出版年の距離を隠してはいけない

Hille論文は2026年の巻号に掲載されているが、検索期限は2021年9月27日である。原稿は2023年3月受領、同年8月受理、9月online公開という履歴を持つ。したがって「2026年時点の医療AI統制研究を網羅した」と紹介するのは不正確である。

生成AIの普及、EU AI Act、医療AIの変更管理、contestability研究など、2022年以降の変化は別に追う必要がある。本レビューの価値は現在の完全な地図ではなく、初期文献が何を明示できていなかったかを定量化した点にある。

4. Puranamの設計空間は権利の正当性を決めない

逐次配置の最後に人を置けば、decision rightsの所在は説明できる。しかし、その人が決めることが法的・倫理的に正しいか、患者が異議を申し立てられるか、責任を負える条件があるかは別である。

組織設計は、権限の配置を見えるようにする。健康領域の倫理と制度は、その配置が正当かを問う。どちらか一方で代替してはいけない。

5. 学習を「性能向上」と呼ぶと、悪い適応を見落とす

Puranamが学習をbeliefやbehaviorの変化と定義し、必ずしもperformance向上としない点は重要である。人がAIへ過度に合わせる、AIの対象外事例を見なくなる、上書きが組織的に敬遠されることも、経験による行動変化である。

導入後に利用率が安定した、AIとの一致率が上がった、という数字だけで「学習した」と肯定してはいけない。異論、訂正、見逃し、対象外事例、負担、非AI経路の利用も同じ期間で追う必要がある。

6. 費用・人員・責任・継続可能性は、図の外に残っている

Hilleはactorを示し、Puranamはagent配置を示す。しかし、独立評価者を何人置くか、夜間・休日を誰が担うか、上書き理由の確認に何分かかるか、監査と非AI経路を維持する費用はいくらかを測っていない。

MHCを追加業務として現場へ置けば、時間不足が統制を形骸化させる可能性がある。継続可能性を確かめるには、必要人員、教育、交代、保守、事故対応、離職、更新ごとの再検証を、倫理の付記でなく運用成果として測るべきである。

この研究からETIC 2.0が学べること

区分	2論文から受け取ること	ETIC 2.0で明確にすること
考え方と重なる点	統制はAI単体でなく、人、制度、情報、時間、権限の配置で成立する	ケア生態系全体を評価単位にし、本人・家族・専門職・組織・開発者の役割を分ける
考え方と重なる点	統制は可能化、実行、評価を反復する	導入前、利用時、導入後を同じ変更履歴でつなぐ
考え方と重なる点	人とAIはfeedbackにより互いの判断を変える	一致率だけでなく、誰が何を見て学んだかを記録する
答えるべき問い	evaluatorがほぼ未特定	統制が機能したと誰が独立に判定し、誰が再開を認めるのか
答えるべき問い	逐次配置は下流の学習事例を選別する	AIが見せなかった事例の転帰と見逃しを誰が追うのか
答えるべき問い	MHCは倫理的正当性のラベルになり得る	統制の存在と、安全・有効・公平・目的の妥当性をどう別判定するか
具体的な見直し案	上書き機能を単独の確認項目にしない	権限、時間、情報、操作、記録、独立評価、停止後の訂正を一組で検証する
具体的な見直し案	group-level feedbackは誤った学習を生み得る	個別判断、AI出力、最終判断、転帰、訂正を対応付け、帰属不能を明記する
具体的な見直し案	patientの役割を最終責任へ還元しない	拒否、異議、再検討、代理、非AI経路を権利として実装する

具体的な見直し案:「統制可能性記録」を追加する

記録欄	確認すること
目的	何の害・責任gap・不公正を減らす統制か
可能化	必要情報、考える時間、教育、代替経路、停止権限を誰が用意したか
実行	誰がいつ採用、上書き、照会、保留、停止できるか
対象外	AIが人へ見せない事例、利用できない人、非AI経路をどう追うか
Feedback	個別・全体の結果を誰へ返し、誤りの帰属をどう保留するか
評価	統制者と別の評価者が、何を閾値に機能・失敗を判定するか
訂正	判断、record、運用、prompt、model、契約のどこを変更し、再検証するか
救済	不利益を受けた人が訂正、説明、再判断、補償へどう到達するか

この表は、有効性が検証済みの尺度ではない。2論文を踏まえたUnknown Labの具体的な見直し案である。項目を増やすだけで現場が良くなるとは限らず、記録負担、役割衝突、対応時間、改善された失敗と新たに生じた失敗を比較する必要がある。

今後のケア研究に残る問い

以下は、今回の2論文を踏まえてUnknown Labが整理した研究上の問いである。問いへの答えや、提案の有効性が検証されたわけではない。

第1の問いは、応答した人と、評価した人を分けられているかである。同じ主体が行動し、その成功も自分で判定する構成では、統制の自己評価になる。

第2の問いは、観察できなかった事例をどう扱うかである。AIが上流で選別した結果だけを人が見るなら、人の応答dataは全体を代表しない。

第3の問いは、feedbackの宛先である。集団全体の結果だけでは、どの判断、どのagent、どの環境条件が失敗へ寄与したか決められない。

第4の問いは、異議と停止が次の設計変更へつながるかである。個別に救済されても、記録がsystem updateへ届かなければ、学習するケア生態系とは呼びにくい。

これらは、ケアの仕組みを設計・評価するとき、追加した構造が本当に説明力、誤り検出、停止、訂正を改善するかを比較する問いである。

何が確認できれば、この見直し案は弱まるのか

批判的論評も反証可能でなければならない。

- 2022年以降を含む再レビューで、統制者、可能化する主体、独立評価者、患者の権利、時間、feedbackが一貫して具体化されていれば、「健康領域のMHCは未成熟」という時点依存の評価は弱まる。
- 独立評価者や停止後の訂正経路を追加しても、誤り検出、害、格差、再発、負担が改善しなければ、今回の見直し案の実装的価値は支持されない。
- 分業とfeedback配置を区別しても、Human–AI協働の失敗、見逃し、負担移転をよりよく説明・予測できなければ、Puranamの類型をETICへ取り込む必然性は弱まる。
- 患者の異議・非AI経路を整えても利用可能性や救済到達が改善しない場合、権利の存在だけでなく、時間、支援、費用、組織応答の別要因を検討しなければならない。

「問いを増やせば安全になる」ことも未検証である。追加項目が現場の注意を分散し、記録だけを増やす可能性を含めて測る必要がある。

次に読む論文

次に読む論文として、Krakowskiらの*Human-Centered Artificial Intelligence: A Field Experiment*を挙げる。[4]

この研究は多国籍製薬企業の72人のsales experts、12 business unitsを対象に、同じAIを使いながら、人とのinteractionを変えた。work procedure、decision-making authority、training、incentivesを個人のcognitive styleに合わせる群、合わせない群、従来ITのcontrolを比較し、2012年第1四半期から2017年第2四半期まで22四半期を追った。

主分析では、合わせないAI群はcontrolより1日0.46 meeting少なく、合わせた群は0.84 meeting多かった。これは「良いAIを置く」より、権限、手順、教育、incentiveをどう組むかが結果を変え得るという、今回の概念議論を現場へ進める材料になる。

ただしrandomization単位は12 business unitsと小さく、著者ら自身がpower制約から因果解釈に注意を求める。対象は製薬salesであり、患者の健康成果や医療AIの安全を直接検証していない。次回は効果の大きさだけでなく、配置のtailoringが誰に適用され、役割衝突、利用率、売上指標、倫理的境界をどう扱ったかまで読む必要がある。

このDOI 10.1287/mnsc.2022.03849は、これまで「論文とETIC」の主論文・次に読む論文として登録していない。

結論

人が最後にいること、上書きできること、担当者名があることは、医療AIを意味のある形で統制できている証明ではない。

Hilleらが読んだ健康領域42 records、168 passagesでは、統制者、目的、条件、時間、特に評価者が多くの場合特定されていなかった。著者らは、統制を可能にする、実行する、評価する3過程を反復する必要があると論じた。同時に、MHCというラベルはAI利用の倫理的正当性を保証しないと警告した。

Puranamは、人とAIの協働を分業だけでなく学習配置として捉えた。逐次か並列か、専門化するか、feedbackが独立か相互依存か、互いの経験を伝えられるかによって、誰が何を見て学ぶかが変わる。

2本を重ねると、統制はボタンでなく経路として見える。権限、認識、学習の3種類の失敗を分け、統制者とは別の評価者、対象外事例、個別feedback、停止後の訂正と救済まで追う必要がある。

ただし、今回の見直し案も未検証である。項目を増やすことが安全、生活、費用、公平性を改善するとはまだ言えない。次に必要なのは、役割と学習配置を変えたとき、どの失敗が減り、どの負担が増え、誰が異議と救済へ到達できたかを比較する研究である。

参考文献

1 Hille EM, Hummel P, Braun M. Meaningful Human Control over AI for Health? A Review. J Med Ethics. 2026;52(e1):e50-e58. doi: 10.1136/jme-2023-109095. PMCID: PMC13151516.

2 Puranam P. Human–AI collaborative decision-making as an organization design problem. J Organ Des. 2021;10:75-80. doi: 10.1007/s41469-021-00095-2.

3 Morimoto D. AIに影響されるのに、私たちはAIを変えられないのか. Unknown Lab ETIC 2.0 設計・検証ノート01. 2026. https://unknownlab.pages.dev/articles/ai-modifiability-care/. doi: 10.5281/zenodo.22487001.

4 Krakowski S, Haftor D, Luger J, Pashkevich N, Raisch S. Human-Centered Artificial Intelligence: A Field Experiment. Management Science. 2026;72(1):57-72. Published online June 9, 2025. doi: 10.1287/mnsc.2022.03849.

本稿は公開済み研究を批判的に読み解いた未査読の論評であり、MHC、ETIC 2.0の有効性を検証した研究ではない。個別の診断、治療、AI製品の安全認証や導入を指示しない。
執筆: Unknown Lab編集部（AI支援） / 著者確認済み / 独立した専門レビューは未実施