AI 支援医療意思決定における責任帰属が非難と信頼に及ぼす影響の構造方程式モデリングによる分析

Responsibility Attribution and Its Impact on Blame and Trust in AI-Assisted Medical Decisions Using Structural Equation Modeling

三宅 圭音 *1*2 Keito Miyake　山田 誠二 *3 Seiji Yamada

*1総合研究大学院大学 The Graduate University for Advanced Studies, SOKENDAI
*2国立情報学研究所 National Institute of Informatics
*3神奈川大学 Kanagawa University

生成 AI の発展により，医療分野では人間と AI の協調意思決定が増加している．本研究は，医師と診断 AI が関与する誤診シナリオを用い，責任帰属が非難および医師の判断への信頼に与える影響を検討した．構造方程式モデリングの結果，医師への責任帰属は医師への非難を高め，判断への信頼を低下させた．一方，AI への責任帰属は AI 自身だけでなく開発者への非難も促進した．

1. はじめに

近年，人工知能（AI）は医療，交通，防衛などの高リスク領域において意思決定支援ツールとして広く導入されつつある．特に医療分野では診断支援 AI の性能向上により医師と AI が協調して意思決定を行う場面が増加している．このような人間—AI 協調意思決定において重要な課題の一つが誤りや不利益な結果が生じた場合に「誰がどの程度責任を負うのか」という責任配分の問題である．

先行研究では，AI の関与によって責任の所在が不明確になる責任ギャップや人間が過度に非難を引き受けてしまう「モラル・クランプルゾーン」といった現象が指摘されてきた [Elish 19]．これらの問題は社会的信頼や技術受容，さらには AI ガバナンスの設計に大きな影響を与える [Ryan 20, Gillespie 22]．また，人々の責任や非難の判断は AI の擬人化や意図性の知覚など，認知的・文脈的要因によって左右されることが示されている [Malle 15, Joo 24]．

一方，責任や非難は哲学および認知科学の分野においても体系的に研究されており，構造的因果モデルを用いて責任や非難の度合いを定式化する理論が提案されている [Chockler 04, Halpern 15]．さらに，自動化に対する信頼研究では AI の失敗が人間の判断や自動化システムへの信頼を著しく低下させることが報告されている [Madhavan 06, Parasuraman 10]．しかし，医師と AI が共同で関与した医療意思決定において，一般の人々がどのように複数の主体（医師，AI，開発者）へ責任を帰属させ，それが非難や医師の判断に対する信頼にどのような影響を及ぼすのかについては実証的研究が十分とは言えない [Reis 24, Sagona 25]．

そこで本研究では，医師と診断 AI が関与する誤診シナリオを用いて責任帰属が非難および医師の判断に対する信頼（confidence）に及ぼす影響を検討する．具体的には，一般参加者を対象としたシナリオ調査を実施し，医師および AI に対する責任帰属，非難，判断への信頼の関係を構造方程式モデリング（SEM）によって分析する．本研究は，AI 支援医療における責任配分の理解を深化させ，社会的に受容可能な人間—AI 協調システム設計に対する示唆を提供することを目的とする．

連絡先: 三宅圭音，総合研究大学院大学，東京都千代田区一ツ橋 1-1-2，101-8430，m-keito@nii.ac.jp

2. 実験方法

2.1 参加者

参加者は Yahoo!クラウドソーシングを通じて募集した．回答完了者は 51 名であった．性別内訳は男性 48 名，女性 2 名，未回答 1 名であった．年齢は 32 歳から 69 歳に分布し，中央値は 46 歳であった．参加者には 20 ポイントの謝礼を支払った．1 ポイントは 1 円に相当する

2.2 仮説

先行研究に基づき，責任帰属が非難および医師の判断に対する信頼に与える影響を検証するため以下の仮説を設定した．

H1: AI に対する責任帰属は AI に対する非難を正に予測する．

H2: 医師に対する責任帰属は医師に対する非難を正に予測する．

H3: 医師に対する責任帰属は医師の判断に対する信頼（confidence）を正に予測する．

なお，本研究における confidence は，医師の道徳性や長期的関係に対する一般的信頼（trust）ではなく，当該診断判断の正確性・能力に対する参加者の信念を指す

2.3 実験デザイン

本研究は，人間—AI 協調型の医療診断場面における責任帰属，非難，信頼の関係を検討するためシナリオ提示型のオンライン調査実験を採用した．設計は横断的であり参加者は単一の誤診シナリオを読んだ後に質問項目へ回答した．

まず，診断支援 AI システム「AI-1」と医師「Dr. A」が登場する状況を提示した．性能の前提情報として過去 100 件の診断における AI-1 の正答率は 95 %であり，Dr. A は 72 %であると説明した．

次に，中核となる失敗シナリオとして，Dr. A が AI-1 の提示した誤った診断を用いて最終判断を行い，実際は肺炎であった患者を「風邪」と誤診した結果，治療が遅延したという負の帰結を提示した．

その後，参加者は主要構成概念について 7 段階リッカート尺度（1 ＝強く不同意，7 ＝強く同意）で回答した．測定項目は，医師の判断に対する信頼（confidence），AI および医師への責任帰属，AI，医師，および AI 開発者への非難である．加えて，AI に対する否定的感情，将来の AI 利用に関する不安，および非難理由を尋ねる自由記述項目を含めた．質問は Google Forms 上で実施した

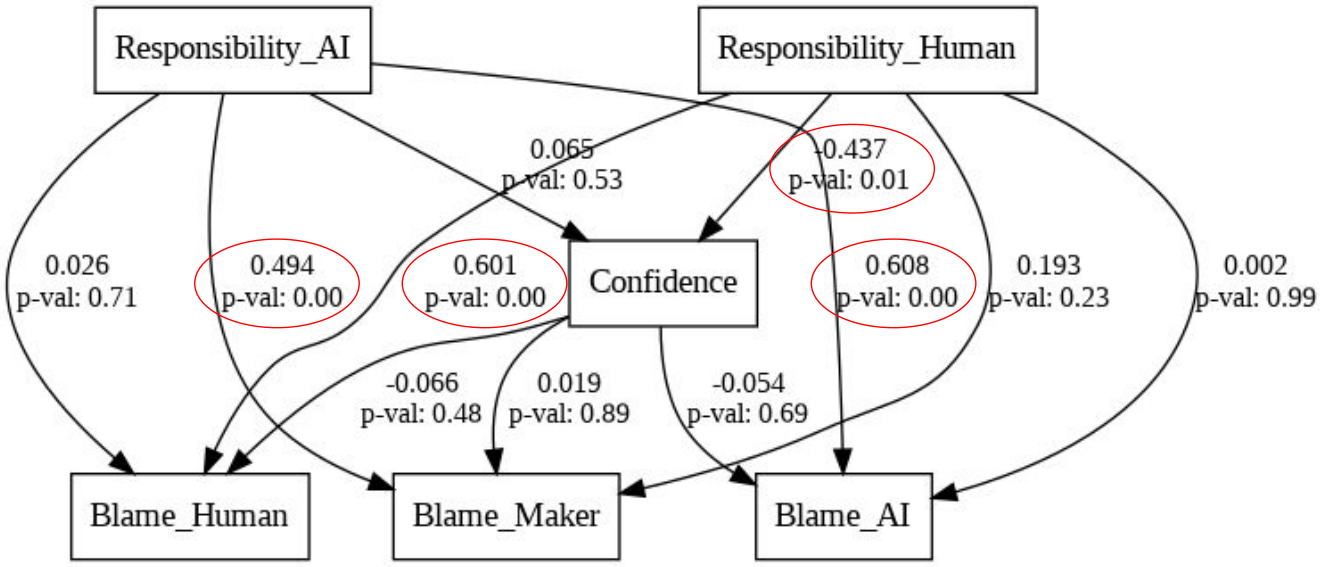

図 1: 責任帰属，信頼，および非難の関係を示す構造方程式モデル（SEM）．矢印は仮定された因果パスを示し，数値は非標準化パス係数を表す．赤色の円は統計的に有意なパス（$p < .05$）を示す．

3. 結果

3.1 構造方程式モデリング分析

責任帰属，非難，および医師の判断に対する信頼（confidence）の関係を検討するため，構造方程式モデリング（SEM）を用いた分析を行った．分析には Python の semopy パッケージを使用した．パス係数については，非標準化係数（B）と標準化係数（β）の両方を報告する．

まず非難に関する結果として，AI に対する責任帰属は，AI に対する非難を有意に正に予測した（$B = 0.61$，$\beta = 0.65$，$Z = 6.01$, $p < .001$）．さらに，AI に対する責任帰属は，AI の開発者に対する非難も有意に正に予測した（$B = 0.49$，$\beta = 0.58$, $Z = 5.10$, $p < .001$）．一方，医師に対する責任帰属は，医師に対する非難を有意に正に予測した（$B = 0.60$，$\beta = 0.62$, $Z = 5.33$, $p < .001$）．

次に信頼（confidence）に関する結果として，医師に対する責任帰属は，医師の判断に対する信頼を有意に負に予測した（$B = -0.44$，$\beta = -0.37$, $Z = -2.77$, $p = 0.006$）．一方，AI に対する責任帰属は，医師の判断に対する信頼を有意に予測しなかった（$\beta = 0.08$, $p = 0.526$）．

また，信頼（confidence）から各主体への非難（AI，医師，開発者）へのパスはいずれも有意ではなかった（すべて $p > 0.05$）．さらに，AI に対する責任帰属から医師への非難，および医師に対する責任帰属から AI への非難といった交差的パスも有意ではなかった．

以上の結果から，責任帰属は対応する主体への非難を強く規定する一方で信頼（confidence）は非難の媒介変数としては機能しないことが示された．特に医師への責任帰属が高いほど，誤診後の医師の判断に対する信頼が低下するという逆方向の関係が確認された（図 1）．

3.2 自由記述回答の定性的分析

定量分析の結果を補足するため，自由記述回答に対する定性的分析を行った．分析では，責任帰属，非難，および信頼判断に直接関連する記述に着目した．

医師に責任を帰属した参加者の多くは，「最終判断を下したのは医師である」「AI を使ったとしても責任は医師にある」といった理由を挙げており，医師の責任を AI の管理・監督義務として捉えていることが示唆された．これらの記述は，医師への責任帰属が高い場合に信頼が低下するという SEM の結果と整合的である．すなわち，責任は能力の証拠ではなく，義務違反や過失の指標として解釈されていた．

一方，AI に責任を帰属した参加者の多くは，「AI を設計・開発した企業にも責任がある」「AI 単体ではなく開発者も含めて問題である」と述べており，AI への責任帰属が開発者への非難へ拡張される傾向が確認された．これは，AI が単独の行為主体としてではなく，社会技術的システムの一部として理解されていることを示唆する．

また，一部の参加者は，「医療ミスは生命に関わるため許容できない」といった医療の高リスク性に言及しており，結果の深刻さが責任および非難判断を強化していることが示された．これらの定性的所見は，責任帰属が単なる因果判断ではなく，道徳的・規範的評価として機能していることを示している．

4. 考察

本研究では，AI 支援医療意思決定における責任帰属が主体別の非難および医師の判断に対する信頼（confidence）に与える影響について 3 つの仮説を検証した．その結果，AI に対する責任帰属が AI への非難を高めるとする仮説 H1，および医師に対する責任帰属が医師への非難を高めるとする仮説 H2 は

いずれも支持された．一方で，医師に対する責任帰属が医師の判断に対する信頼を高めるとする仮説 H3 は支持されず，むしろ責任帰属が信頼を低下させるという逆の関係が示された．

これらの結果は，人間—AI 協調意思決定において，人々が主体ごとに責任を区別して非難を行う一方，失敗が顕在化した文脈では医師への責任帰属が肯定的評価ではなく過失の指標として機能する可能性を示唆している．

自由記述回答に基づく定性分析は，本研究の定量的結果を解釈する上で有用な補助的知見を提供した．特に，医師に対する責任帰属を行った参加者が責任を「最終判断者としての監督義務」として捉えていた点は医師への責任帰属が信頼（confidence）の低下と結びついた定量的結果と整合的である．すなわち，責任は判断能力や権限の表象ではなく，規範的義務が果たされなかったことを示す評価的手がかりとして機能していたと解釈できる．

また，AI に責任を帰属した参加者が AI 単体ではなくその開発者にも非難を拡張していた点は，AI が独立した行為主体ではなく人間が設計・管理する社会技術的システムとして理解されていることを示唆する．このような直観的理解は責任判断が単なる因果的貢献ではなく，予見可能性や管理可能性と結びついて行われている可能性を示している．

以上より，定性分析の結果は責任帰属が道徳的・規範的判断として形成され，主体の役割理解に基づいて非難や信頼評価が行われているという本研究の主要な解釈を補強するものと位置づけられる．

5. まとめ

本研究は，AI 支援医療意思決定において責任帰属が非難および医師の判断に対する信頼に与える影響を検討した．構造方程式モデリングによる分析の結果，AI および医師に対する責任帰属はいずれも対応する主体への非難を高めることが示された．一方で，医師への責任帰属は誤診という失敗文脈において，医師の判断に対する信頼を低下させることが明らかとなった．

これらの結果は，人間—AI 協調意思決定において責任が能力や統制の指標ではなく，過失や義務違反の評価として解釈されうることを示唆している．また，AI の責任が開発者へと拡張される傾向は AI を含む社会技術システム全体を視野に入れた責任設計の重要性を示していると考えられる．

参考文献

[Chockler 04] Chockler, H. and Halpern, J. Y.: Responsibility and blame: A structural-model approach, *J. Artif. Intell. Res.*, Vol. 22, pp. 93–115 (2004)

[Elish 19] Elish, M. C.: Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, *Engag. Sci. Technol. Soc.*, Vol. 5, pp. 40–60 (2019)

[Gillespie 22] Gillespie, T.: Building trust and responsibility into autonomous human-machine teams, *Front. Phys.*, Vol. 10, (2022)

[Halpern 15] Halpern, J. Y.: Cause, responsibility and blame: a structural-model approach, *Law Probab. Risk*, Vol. 14, No. 2, pp. 91–118 (2015)

[Joo 24] Joo, M.: It's the AI's fault, not mine: Mind perception increases blame attribution to AI, *PLoS One*, Vol. 19, No. 12, p. e0314559 (2024)

[Madhavan 06] Madhavan, P., Wiegmann, D. A., and Lacson, F. C.: Automation failures on tasks easily performed by operators undermine trust in automated aids, *Hum. Factors*, Vol. 48, No. 2, pp. 241–256 (2006)

[Malle 15] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C.: Sacrifice one for the good of many? People apply different moral norms to human and robot agents, in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 117–124, IEEE (2015)

[Parasuraman 10] Parasuraman, R. and Manzey, D. H.: Complacency and bias in human use of automation: an attentional integration, *Hum. Factors*, Vol. 52, No. 3, pp. 381–410 (2010)

[Reis 24] Reis, M., Reis, F., and Kunde, W.: Influence of believed AI involvement on the perception of digital medical advice, *Nat. Med.*, Vol. 30, No. 11, pp. 3098–3100 (2024)

[Ryan 20] Ryan, M.: In AI we trust: Ethics, artificial intelligence, and reliability, *Sci. Eng. Ethics*, Vol. 26, No. 5, pp. 2749–2767 (2020)

[Sagona 25] Sagona, M., Dai, T., Macis, M., and Darden, M.: Trust in AI-assisted health systems and AI' s trust in humans, *Npj Health Syst.*, Vol. 2, No. 1, pp. 1–5 (2025)