立教大学学術推進特別重点資金（立教ＳＦＲ）
大学院学生研究
２０２５年度研究成果報告書

研究科名	立教大学大学院　人工知能科学研究科　人工知能科学専攻		
研究代表者（2026年3月現在のものを記入）	在籍課程・学年 □博士前期課程　年 ■博士後期課程３年	氏　名 浦東　聡介	
指導教員	所属部局・職名 人工知能科学研究科　教授	氏　名 村上　祐子	
自然・人文・社会の別	自然 ・ ~~人文~~ ・ ~~社会~~	**個人・共同の別**	個人 ・ 共同　　名
研究課題	刑事司法分野へのAI導入の可能性：厳罰化志向・犯罪不安の拡散構造と社会的受容性		
研究組織（共同研究者がいる場合のみ、2026年3月現在のものを記入）	在籍研究科・専攻・課程・学年	氏　名	
研究期間	２０２５　　年度		
研究経費（1円単位）	（支出金額）　100,000円　／（採択金額）　100,000円		

研究の概要（200～300字で記入、図・グラフ等は使用しないこと。）

当該研究の研究目的を含むこと。

本研究は、AI裁判官を中心に、裁判支援AIを含む刑事司法分野におけるAI活用の社会的受容条件を明らかにするため、二つの側面から実証的検討を行った。第一に、哲学カフェにおける市民の対話を分析し、AIが司法で受容されるかどうかは、単なる賛否ではなく、事案の性質、公正性、共感・対話可能性、責任の所在に応じて変動することを明らかにした。第二に、犯罪不安や厳罰化志向に関するSNS上の言説を分析し、感情的な意見ほど可視化・拡散されやすい可能性を示した。これにより、司法におけるAIは人間を単純に代替するものではなく、社会的納得を支える仕組みと併せて検討すべきことを示した。

キーワード（研究内容をよく表しているものを３項目以内で記入。）

〔AI裁判官　　〕〔AI倫理　　〕〔法とAI　　〕

研究成果の概要（図・グラフ等は使用しないこと。）

以下の視点を含めて記載のこと。
・当該研究は何をどこまで明らかにできたのか（できなかったのか）。
・何をもって研究成果（経過）を達成できた（できなかった）と考えられるのか。
自身が設定した研究目的・目標に照らして、その根拠がわかるよう記載のこと。

本年度は、研究計画において掲げた「AI 裁判官の実現に向けた課題の検討」及び「司法領域における世論形成メカニズムの解明」を中心に進め、刑事司法分野への AI 導入可能性を、社会的受容と SNS 上の言説拡散という二つの側面から実証的に検討した。具体的には、哲学カフェによる熟議を通じて AI 裁判官に関する一般市民の受容条件と制度的・倫理的課題を明らかにするとともに、YouTube コメント分析により、犯罪不安や厳罰化志向が SNS 上でどのように可視化・拡散されるかを解明することを目標とした。

1　哲学カフェの議論分析による AI 裁判官の社会的受容条件の検討

本年度の研究では、AI 裁判官に対する受容が固定的・一様なものではなく、事案の性質、当事者の置かれた状況、判断の帰結の重大性といった文脈に応じて動的に変化することを実証的に明らかにした。とりわけ、哲学カフェにおける熟議の分析を通じて、AI 裁判官に対する評価は、単純に「人間より正確か」「効率的か」という技術的観点のみで決まるのではなく、公正性、共感・対話可能性、責任の所在という複数の規範的要素の相対的な重みづけによって構成されることが示された。すなわち、AI 受容は抽象的な賛否の問題ではなく、どのような事件で、何が裁判官に期待され、いかなる結果が生じうるのかという条件の下で再編成される「文脈依存的な受容」であることが確認された。

具体的には、抽象的・一般的な水準では AI 裁判官への一定の支持が観察され、参加者の 56.7%が AI 裁判官を支持した。しかし、この支持は具体的事例に移行すると大きく揺らぎ、特に裁判官に共感や人間的理解が期待される場面では、支持率が 11.7%にまで低下した。この大幅な低下は、AI に対する評価が単なる技術信頼ではなく、「人として話を聞いてもらえるか」「苦境や事情に心を寄せられるか」といった期待によって大きく左右されることを端的に示している。とりわけ、生活困窮や、やむを得ない事情を背景にもつ事案では、参加者は法的判断の正確さだけでなく、当事者の境遇を汲み取り、その語りに応答する姿勢を裁判官に求める傾向を示した。ここでは AI の一貫性や偏りの少なさよりも、「人間として向き合うこと」が正統性の重要な条件として前面に出ていた。

他方で、AI の受容が一律に低いわけでもない点も、本研究の重要な成果である。冤罪事案のように、事実認定の厳密さや誤判の是正が重視される文脈では、AI は一貫性やバイアス低減可能性の観点から相対的に肯定的に評価された。このことは、AI 裁判官の受容が全面的代替への賛否としてではなく、公正な判断を支える機能への期待として現れていることを示唆している。これは、AI 導入の是非が二者択一的に語られるべきではなく、どの判断機能を AI に委ね、どの判断機能を人間が保持すべきかという役割分担の問題として理解される必要があることを示している。

さらに、重大犯罪や死刑のように、判断結果が不可逆的で取り返しのつかない帰結を伴う事案では、責任の所在が受容の中核的論点として浮上した。こうした場面では、たとえ AI が高い精度や整合性を示したとしても、「最終的に誰が責任を負うのか」が曖昧である限り、AI 裁判官への留保が生じることが確認された。参加者は、生命や身体、人格に重大な影響を及ぼす判断については、単なる計算や推論能力ではなく、法的・道義的責任を引き受ける主体の存在を不可欠とみなしていた。換言すれば、不可逆的な判断領域では、受容の条件は性能の高さではなく、最終判断を人間が担い、説明責任と責任帰属を明確にする制度的枠組みによって支えられるのである。この知見は、AI 導入を考える際には透明性や説明可能性だけでなく、監督、介入、監査、救済を含む責任インフラの設計が不可欠であることを示唆している。

以上の成果は、研究計画で掲げた「AI 裁判官あるいは AI による裁判官補助の実現に向けて解決すべき課題や社会的合意を要する事項を定義する」という目的に対応するものである。すなわち本研究は、AI 導入の可否を単純な賛成・反対として把握するのではなく、受容を規定する評価軸が文脈ごとに組み替わる構造を示した点に意義がある。具体的には、公正性が優位となる場面、共感や対話可能性が期待される場面、責任の所在が最重要となる場面とでは、制度に求められる要件そのものが異なる。したがって、司法分野における AI 導入は、単に精度や処理速度を高めるだけでは足りず、判断過程の説明可能性、当事者との対話可能性、最終責任の明確化、人間と AI の役割分担を制度的に組み込むことではじめて社会的正統性を獲得しうる。本研究は、このような「適応的受容」の枠組みを通じて、司法における AI 実装を支える制度設計上の視点を具体化し、司法分野特有の課題を抽出するとともに、「AI の特性を踏まえた調整の指針」を構想するための基礎を提示したと評価できる。

研究成果の概要（つづき）

２　YouTube コメント分析による司法領域の世論形成メカニズムの検討

本年度のもう一つの柱は、司法領域における世論形成メカニズムを解明するための SNS 分析である。具体的には、犯罪関連 YouTube 動画に付されたコメントを対象として、犯罪不安や厳罰化志向に結びつく言説ほど「いいね」を獲得しやすく、その結果として可視化・拡散されやすい傾向を実証的に示した。本研究は、SNS 上の刑事司法言説を単なる多様な意見の集積ではなく、感情的傾向をもつ発話をプラットフォーム上の評価機能を通じて相対的に支持されやすい構造として捉えた点に特徴がある。とりわけ YouTube の「いいね」は、個人的共感の表示にとどまらず、他者の発言に対する社会的承認のシグナルとして機能し、どの言説が目立ち、さらに支持されるかを方向づける。本研究は、この評価メカニズムを通じて、犯罪不安や厳罰志向を帯びた発言が公共圏のなかで可視化されやすい構造を明らかにした。

方法論的には、第一に犯罪関連 YouTube 動画に投稿されたコメントとその「いいね」数を収集し、第二に各コメントがどの程度「犯罪不安」あるいは「厳罰志向」を表しているかを既存の心理尺度項目とテキストとのコサイン類似度によって数量化し、第三に得られたスコアと「いいね」数の関係をハードルモデルによって推定した。これにより、従来の質問紙調査では把握しにくかった SNS 上の大規模言説空間における感情傾向と承認構造を数量的に分析する枠組みを提示できた。特に、少なくとも一件の「いいね」を得る確率と、その後どれだけ多くの「いいね」を獲得するかを分けて扱うことで、発話の可視化と拡散の過程を区別して捉えることが可能となった。

その結果、犯罪不安を強く表すコメントも、厳罰志向を強く表すコメントも、いずれも「いいね」を通じて支持されやすく、可視性を高めやすいことが示された。この知見は、SNS 上で流通する刑事司法言説が、中立的・均等な意見市場ではなく、感情的で脅威志向的な発話が相対的に可視化されやすい場であることを示唆している。とりわけ、犯罪不安を喚起する言説は、「危険を見落とすよりも過剰に警戒するほうがコストが低い」とする Error Management Theory の観点から理解できる。すなわち、不確実な状況では人々は脅威を見逃さない方向に反応しやすく、その心理的傾向が SNS 上の評価機能と結びつくことで、脅威を強調する言説が一層可視化される構造が生じていると考えられる。同様に、厳罰化を支持する言説も、個々人の処罰感情の表出にとどまらず、社会的承認を受けやすい発話として位置づけられることが示された。これは、刑事政策に関する「世論」や「参加」が、熟議の結果としてのみ形成されるのではなく、承認されやすい発話の偏りを通じても形成されうることを意味する。

以上より、本研究は、研究計画で掲げた「犯罪への不安や厳罰志向に代表される意見が、他の意見よりも社会的に拡散されやすい傾向にあることを明らかにする」という目標に対し、約 67 万件規模の YouTube コメント分析を通じて数量的根拠を与えた点で、当初の目的を相当程度達成したといえる。とりわけ、SNS 上の世論形成が感情的・直感的反応によって増幅される可能性を検証対象としていた研究目的に照らせば、犯罪不安や厳罰志向を帯びた言説ほど可視化・承認されやすいことを示した点が、本年度の研究成果の中心である。

以上のとおり、本年度は、哲学カフェの議論分析を通じて AI 裁判官に対する社会的受容の条件と制度的論点を明らかにするとともに、YouTube コメント分析を通じて犯罪不安や厳罰志向を帯びた言説が SNS 上で可視化・拡散されやすい傾向に数量的根拠を与えた点で、研究目的の中核部分を相当程度達成したと評価できる。他方で、AI 導入時の制度設計については、哲学カフェの分析を通じて公正性・共感・責任の所在といった受容条件や制度的論点を抽出することはできたものの、それらを裁判所・開発者・監督機関のあいだの責任配分、説明・不服申立ての手続、試行導入の対象領域と段階設定といった具体的な実装経路にまで落とし込むには至っていない。また、YouTube コメント分析によって犯罪不安や厳罰志向を帯びた言説が「いいね」を通じて可視化・拡散されやすい傾向には数量的根拠を与えたが、その背後にある心理的・因果的メカニズムについては、YouTube という単一プラットフォーム・直近一年のデータに依拠しているため、他媒体・他時期・他文化圏にも通用する形で厳密に特定する段階には至っておらず、今後は参加者層の拡張や縦断的分析、実務家を交えた検証が必要である。

以上から、本年度は、刑事司法分野における AI 活用の社会的受容条件及び世論形成過程の基礎的構造を明らかにし、今後の制度設計と拡散メカニズムの精緻化に向けた基盤を得た段階に到達したと位置づけられる。

※この（様式２）に記入の成果の公開を見合わせる必要がある場合は、その理由及び差控え期間等を記入した調書（Ａ４縦型横書き１枚・自由様式）を添付すること。

研究発表（研究によって得られた研究成果を発表した①～④について、該当するものを記入してください。該当するものが多い場合は主要なものを抜粋してください。なお、成果発表を確認できる資料を合わせて研究成果報告書提出フォームより提出してください（紙媒体等、研究成果報告書提出フォームから提出できない場合は、別途リサーチ・イニシアティブセンターへ提出してください）。

①雑誌論文（著者名、論文標題、雑誌名、巻号、発行年、ページ）
②図書（著者名、出版社、書名、発行年、総ページ数）
③シンポジウム・公開講演会等の開催（会名、開催日、開催場所）
④その他（学会発表、研究報告書の印刷等）

※修士論文・博士論文は含みません。

①雑誌論文（著者名、論文標題、雑誌名、巻号、発行年、ページ）

Mukai, Tomoya, and Toshihiro Urahigashi. ″Are Punitive Comments on YouTube More Likely to Receive Likes? An Analysis Using Cosine Similarity and Hurdle Model.″ *International Criminal Justice Review* (Volume 36, Issue 2,2026,182-199)

Urahigashi Toshihiro, Atsushi Yoshikawa, Hirotsugu Ohba, and Yuko Murakami. “Contextual Justice: Citizens’ Dialogues on AI Judges in Japan.” AI and Ethics, accepted for publication. (Volume 6, article number 259 (2026))