**立教大学学術推進特別重点資金（立教ＳＦＲ）**
**大学院学生研究**
**２０２３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**<br>（2024年3月現在のものを記入） | 在籍課程・学年<br>□博士前期課程　年<br>■博士後期課程１年 | 氏　名<br>浦東　聡介 | |
| **指導教員** | 所属部局・職名<br>人工知能科学研究科　教授 | 氏　名<br>村上　祐子 | |
| **自然・人文・社会の別** | 自然 | **個人・共同の別** | 共同　２名 |
| **研究課題** | 世代間又は他者への影響を考慮した意思決定 | | |
| **研究組織**<br>（研究代表者・共同研究者）<br>※2024年3月現在のものを記入 | 在籍研究科・専攻・課程・学年<br>人工知能科学研究科<br>博士課程後期課程１年<br><br>人工知能科学研究科<br>博士課程前期課程２年 | 氏　名<br>浦東聡介（代表研究者）<br><br>戸谷　颯（共同研究者） | |
| **研究期間** | ２０２３　年度 | | |
| **研究経費**<br>（1円単位） | （支出金額）　250,000円　／（採択金額）　250,000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。）

本研究は、他者への影響を考慮すべき意思決定の一例として裁判に着目し、AIの可能性と課題を探究することにより、司法における公平性、透明性及び効率性の向上を目指すものである。具体的には、市民が裁判官に対して持つ期待を、YouTubeに投稿された動画に対するコメント分析、有識者による会議議事録の調査及び関連研究などから明らかにした。そして、そのような期待に応えるAI裁判官の実現可能性及びそれに伴う倫理的・法的な課題などについて、GPT-4を用いた実験や大規模言語モデルの評価を通じて検討した。

**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　〕〔意思決定　〕〔AI倫理　〕

**研究成果の概要**（図・グラフ等は使用しないこと。）

本研究は、意思決定において世代間又は他者への影響を考慮することの重要性を探究するものである。意思決定には、裁判員裁判における裁判員の意思決定のように意思決定者の利害に直接影響を与えないものや、天然資源の利用とそのことによる気候変動のように、利益を受ける世代と損害を受ける世代が異なるものが存在する。このような意思決定に当たっては、公正性や倫理性が求められると同時に、価値観の違いによって様々な意見が存在することから、絶対的な正解がない領域と言え、意思決定プロセスにおける合意形成の重要性が浮き彫りになる。本研究では、このような意思決定プロセスや合意形成を AI あるいは意思決定科学によってサポートすることを目的とする。

本年度の具体的研究テーマとしては、他者へ影響を与える意思決定の一例として裁判に着目し、AI の可能性と課題を探究することにより、司法における公平性、透明性及び効率性の向上に貢献することを目的とした。具体的には、裁判官あるいは裁判員としての AI 活用の可能性及びそれに伴う倫理的・法的な課題などについて検討した。また、AI に司法判断をさせて良いか、AI に司法判断をさせることができるのか、といった問いを考察するためには、その前提となる司法判断の在り方を検討する必要があることから、本研究では、司法判断の在り方の典型例として「裁判官はいかにあるべきか」という問いの考察を出発点とした。

以上から、本研究では①裁判官はいかにあるべきか、②AI 裁判官の実現可能性の検討及び③AI 裁判官実現に当たり検討すべき法的・倫理的課題の抽出を試みた。

1　裁判官はいかにあるべきか

司法分野において、AI を利用することの是非を検討するためには、裁判官に対する期待や役割を明らかにし、そのうちどのような部分を AI に置き換えるのか、あるいは AI に補助させるのかを具体的に議論する必要がある。このことから、司法分野における AI 利用の是非の検討には、その前提となる「裁判官はいかにあるべきか」という問いの検討が欠かせない。司法制度改革審議会では国民が求める裁判官像を「法律家としてふさわしい多様で豊かな知識、経験と人間性を備えていること」としており、法律知識のみならず人間性をも裁判官に要求している。この見解は、学識経験のある委員が議論を重ねた結果として裁判官の在り方に大きな示唆を与えている。また、我が国では裁判を受ける権利が憲法によって保証されていることからも、裁判官の在り方については、権利主体たる一般市民の見解も重要である。一般市民の裁判官に対する期待を明らかにする手法としては、アンケート調査などが候補に挙げられるが、一般市民が普段意識することが少ない裁判官の在り方について、よりリアルな市民の期待を把握するために、本研究では YouTube に公開された法廷のやりとりに関する動画に対して投稿されたコメントを分析対象とした。

分析対象のデータとして YouTube Data API から合計 25,076 件のコメントを取得し、これらのコメントに対してトピックモデルの代表的手法である Latent Dirichlet Allocation (LDA)を用いて分析した。その結果、①公平性を求めるトピック、②社会への不満が表現されたトピック、③動画内のセリフに関するトピック、④宗教的なトピック、⑤裁判官に道徳的正しさを求めるトピックの５つのトピックが抽出された。これらのうち、裁判官の在り方に示唆を与えるものは①公平性を求めるトピック及び⑤裁判官に道徳的正しさを求めるトピックである。また、全体的な傾向としては⑤裁判官に道徳的正しさを求めるトピックがコメント全体に与える影響が大きく支配的であった。以上から、一般市民は裁判官に対して第一に道徳的正しさを求め、同時に公平に裁判が運営されることを求めているものと考えられる。この結果は、裁判官の役割全てを明らかにするものではないが、専門家が裁判官へ人間性を求めていることとも整合し、一般市民の裁判官に対する期待の一端を明らかにするものとして重要な示唆を与えている。

2　AI 裁判官の実現可能性の検討

2023 年以降、大規模言語モデルの利用が急速に広がりを見せている。そこで、本研究では大規模言語モデルを利用した AI 裁判官の実現の可能性を探るため GPT-4 に刑事事件における量刑の判断をさせ、その結果を考察した。実験では判例理解と、動機や背景に着目した「温情判決」ができるのかどうかの２つの観点を確認した。判例理解は、一定の基準に沿った判断の可否を確認するもので、公平性を担保するために求められる。動機や背景に着目した判断の可否は、道徳的価値判断に対応するものである。判例理解の確認では、GPT-4 に日本の法律に基づいた裁判官としての役割指示と刑事事件の概要を入力し、事件に対する量刑判断とその理由を出力させた。入力した刑事事件は殺人事件とした。これは刑法の最も重要な保護法益が生命であるという考え方に基づく。被害者数を 1 人から 10 人まで変化させ、被害者数以外の文章は被害者数変更に伴う整合性維持以外の変更はせずに判決の傾向を観察した。GPT-4 のパラメータは全てデフォルトとした。このことで、生成される文章には確率的な揺らぎが生じる。この揺らぎを GPT-4 が内部に持つ当該事件に対する確信度と捉え、その確信度を計測するため 100 回同じ事件概要を入力してその結果を記録した。なお、入力に当たっては毎回新たにセッションを開始した。

**研究成果の概要**（つづき）

実験の結果、被害者が増加するごとに死刑を選択する傾向が強くなることを確認した。これは、我が国の死刑選択基準と言われる判例において、被害者の数を一つの基準としていることと整合する。また、被害者２名において、死刑と無期懲役の選択件数が拮抗した。これは、いわゆる量刑相場として被害者２名が死刑選択のボーダーラインと言われることと整合的である。次に、GPT-4 がいわゆる「温情判決」を下す可能性を確認するため、要介護状態にある親 1 名を息子が殺害する事件という共通の背景を持つ 2 件の事件を用意した。1 件は十分に介護を尽くし、かつ、金銭的困窮から止むに止まれず親を殺害する事例(以下、介護殺人)である。もう一件は長年自宅に引きこもっていた息子が、親が要介護状態になったことを疎ましく思い、趣味の時間を確保するために親を殺害する事例(以下、引きこもり殺人)である。判例理解の確認同様に 100 回同じ事件を入力し、量刑の差を確認したところ介護殺人の刑期が引きこもり殺人の刑期よりも統計上優位に短いこと、出力された判決理由が介護殺人において同情的な文面が見られることを確認した。以上の結果は、GPT-4 が AI 裁判官として判例の基準に沿った判断や温情判決を下すことができるという可能性を示唆するものと考えられる。

3　AI 裁判官実現に当たり検討すべき法的・倫理的課題

本研究は、「AI 裁判官の実現可能性の検討」における実験を基に、大規模言語モデルを司法分野、特に刑事裁判での判断に用いる場合の課題の抽出を試みた。その結果、①公平性、②ブラックボックス性、③正確性の確認という３つの観点において検討すべき課題が抽出された。

①　公平性

大規模言語モデルは、膨大なデータを学習することで国民感覚に即した結論を導くことが期待される。時代に即した国民感覚を反映するためには、随時アップデートを繰り返す必要があるが、モデルのアップデート前後で判断傾向が大きく変化する可能性がある。このような場合、アップデート前後の裁判において、全く同じ条件であっても結論が異なる可能性があり、公平性の観点から問題となり得る。例えば、裁判員裁判においては裁判の積み重ねによる緩やかな判断傾向の変容は是認できると考えられているが、大規模言語モデルにおけるアップデートを境とした判断傾向の変化に対して、どの程度の変化を許容するのか合意形成のプロセスを含め検討する必要がある。

②　ブラックボックス性

AI の判断については、内部の計算が複雑であり、その理解が困難であることなどからブラックボックス性が問題として指摘される。裁判においては、刑事訴訟法上理由を附すことが定められているところ（44 条１項)、大規模言語モデルを利用する場合、判決理由を出力することができる。人の裁判官であっても、その内心は外部から窺い知ることはできず、言語化された判決文を通じて裁判結果の妥当性を判断することとなる中で、AI の判断に対してどこまで説明性や解釈性を求めるべきであるのか議論する必要がる。

③　正確性の確認

GPT-4 に対して、「AI 裁判官の実現可能性の検討」における判例理解と同じ条件で、刑法の２つの異なる立場である「応報刑論」と「教育刑論」のそれぞれに立った判断をするように指示を与えたところ、出力結果が大きく異なることを確認した。このことは、AI 裁判官の判断をプロンプトによってコントロールできることを示している。判例や量刑相場が形成された事例であれば、それらと照らし合わせることでその AI の判断結果の妥当性を確認することができるが、法令改正直後など、基準が存在しない場合に AI を用いた判断結果の妥当性を確認する方法が存在しない。プロンプトによって、結論のコントトールが可能な AI 裁判官においては、特に前例が存在しないような事例で、その妥当性を判断する方法の確立が求められる。

**※この（様式２）に記入の成果の公表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。**

④学会発表

・AI裁判官の可能性と課題：GPT-4を用いた刑法第199条に基づく量刑判断の実験的検討
FIT2023 第22回情報科学技術フォーラム 2023年9月8日

・大規模言語モデルを利用したAI裁判官の実現に向けた試論的検討
法と心理学会第24回大会 2023年10月21日

・YouTubeコメントに見る裁判官への市民期待の分析
第3回計算社会科学会大会 2024年2月20日