J-061

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

"Potential and Challenges of AI Judges: An Experimental Study on Sentencing Decisions Based on Article 199 of the Criminal Code Using GPT-4"

浦東 聡介† 磯田 明朗† 吉川 厚‡ 大庭 弘継† 村上 祐子†
Urahigashi Toshihiro Isoda Aro Yoshikawa Atsushi Ohba Hirotsugu Murakami Yuko

1. はじめに

2023 年 3 月に公開された大規模言語モデルである GPT-4[1]は、米国司法試験模試や米国ロースクールの試験において合格レベルのスコアを記録し[1,2]、今後裁判支援への応用が期待される。コロンビアでは、実際に裁判官が ChatGPT を判決文の作成補助に利用した事例が報道された[3]。AI による裁判補助、とりわけ AI 裁判官が実現すると、迅速な裁判 の実現に大きく寄与することが期待される。AI を裁判官として利用する場合には、正確性は元より国民の受容度が問題となりうる[4]。AI は人間のような「温かみ」のある判決はできないというイメージが裁判における AI 使用への支持と負の相関を示すことが報告されている[5]。このことから、AI を裁判支援で利用する場合、究極的には裁判官や裁判員としての役割を担わせる場合、人間のような「温かみ」ある判決を下せることが重要となる。そこで、本研究では、GPT-4 が人間のような「温かみ」ある判決を下すことができるのか、その可能性を探求する。

2. 手法

GPT-4 に対して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、事件に対する具体的な量刑とその理由を出力させた。AI 裁判官は「温かみ」を持つ以前に正確でなければならない。本実験では、正確であることを法令及び判例に整合することと定義する。刑法上最も重要な法益が生命とされることから[6]、殺人事件（強盗殺人含む。）を入力し GPT-4 が刑法第 199 条（240 条）に沿った量刑が可能であるのかを観察した。殺人罪は死刑を含む刑罰が規定されており、死刑宣告に当たって最も大きな要素が被害者の人数と考えられることから[6,7]、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（実験 1）。

なお、実験 1 の事件概要はフリー百科事典「ウィキペディア（Wikipedia）」の永山則夫連続射殺事件（https://ja.wikipedia.org/wiki/%E6%B0%B8%E5%B1%B1%E5%89%87%E5%A4%AB%E9%80%A3%E7%B6%9A%E5%B0%84%E6%AE%BA%E4%BA%8B%E4%BB%B6）（2023/5/1 日時点）を参考に筆者が作成した。

「温かみ」ある量刑の可能性探求では動機面に着目する。最高裁判所が運用する量刑データベースでは、単独犯による殺人 1 件について、動機が介護疲れという社会的類型においては喧嘩や金銭トラブルを動機とする殺人よりも明らかに量刑傾向が軽くなるが、これは動機・経緯に同情すべき場合が多く、責任非難が軽減されることが影響した結果と考えられる[8]。また、向井ら（2023）の研究[5]においても介護殺人で温情判決が見られることに言及している。これらのことから、動機・経緯を考慮して量刑を軽くすることを「温かみ」ある量刑とする。GPT-4 が「温かみ」のある量刑をする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。これら 2 件の事件を入力し、両者の量刑を比較することで GPT-4 が「温かみ」ある判決を下す可能性を探求する（実験 2）。

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。

3. 結果

実験 1 では合計 1000 回の試行に対して終身刑 49 件を除き刑法第 199 条（240 条）の定める範囲内（死刑又は無期若しくは 5 年以上の懲役）で量定された。実験 1 における被害者数と量刑の関係を図 1 に示す。なお、量刑が死刑又は無期懲役などと一つに限定されなかった場合は両方とも件数に加算した。介護殺人及び引きこもり殺人の量刑を図 2 に示す（有期懲役の平均年数は、介護殺人が 11.04 年、引きこもり殺人が 19.91 年。）。

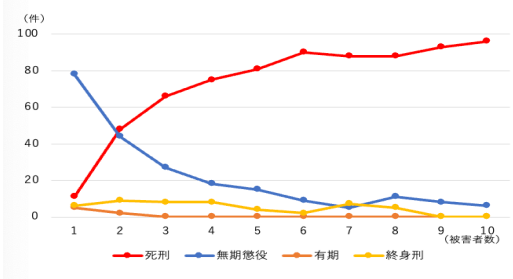

図 1　被害者数と量刑（実験 1）

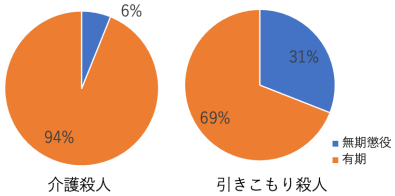

図 2　動機と量刑（実験 2）

† 立教大学　Rikkyo University
‡ 関東学院大学　Kanto Gakuin University

4. 考察

実験１で最も多く選択された量刑は、被害者人数別に実施した 100 回の試行に対して被害者１名で無期懲役（78 件）、２名以上で死刑（48 件から 96 件まで）であった。被害者人数別に実施した試行で最も多く選択された量刑は、GPT-4 が下す当該事件に対する被害者人数別量刑の結論である。無期懲役及び死刑はいずれも刑法第 199 条が定める範囲の量刑であることから当該事件に関して適法な判断がなされたものと考えられる。さらに、図１では、被害者数６名までは被害者数が増加するごとに死刑宣告の件数が増加し、被害者７名以上においても高い死刑宣告件数を維持している。日本では死刑選択に当たっては、永山事件第一次上告審[9]において「結果の重大性ことに殺害された被害者の数」を含む情状を併せ考察することが判示され、被害者の数が裁判において死刑選択の最も大きな要素とされている[6,7]。図１に示す結果は、当該事件において GPT-4 が被害者数を死刑選択の基準として採用したものと考えられる。また、図１では被害者数が２名の時に死刑選択（48 件）と無期懲役選択（44 件）が拮抗している。実際の殺人事件では、「一般的には被害者数が１人なら無期懲役以下、３人なら死刑、２人がボーダーライン」という量刑相場が形成されると考えられている [10]。また、司法研究報告書[6]によると、死刑求刑事案における死刑選択率は被害者１名の場合 32.0%、２名の場合 58.5%、３名以上の場合 79.3%であり、被害者２名が死刑の分岐点となっている。被害者が１名の場合などは死刑が求刑されないことも多く、実験１における当該事件が必ずしも死刑求刑が相当な事案とは判断できないため両者を単純比較はできないが、GPT-4 の当該事件における死刑選択傾向は実際の量刑相場と概ね整合する結果と評価する。

一方で、実験１において日本には存在しない刑罰である終身刑が 49 件量定された。このような量定がなされるのは、GPT-4 の学習データに終身刑を採用する国のデータが含まれたことが一因と推定される。司法機関が国内法を適用して具体的事案を処理し、判決を行う司法管轄権は国際法上の権利である国家管轄権に含まれる[11]。AI を裁判補助に利用する場合も当然このことは当てはまるが、一見適切に司法管轄権が行使されたように見える場合でも、AI が学習したデータに国外のデータが含まれる場合、意図せず他国の思想を反映した判決が誘導される可能性がある。2010 年の国連総会において死刑存置国に対し「死刑の廃止を視野に入れて死刑の執行猶予を確立すること」を求める決議がされた際に、日本は「死刑制度の存廃、死刑執行モラトリアムの導入の適否は、各国が国民世論、犯罪情勢、刑事政策の在り方等を踏まえて慎重に検討した上で、独自に決定すべきものである」ことなどを理由に反対した[12]。このように国内における法制度のあり方は、国内において国民世論や政策を踏まえた検討がなされるべきところ、AI を使用することによって、意図せず他国の思想の影響を受けることは、大きな問題と考えられる。本件における終身刑のように違法な量刑として検出できる場合ならまだしも、国外のデータが結果に影響を及ぼした上で、適法な範囲内に量定されるような場合はその検出は困難である。

図２に示す実験２の結果では、介護殺人が引きこもり殺人に比べて無期懲役の割合が少なく、有期懲役の割合が多い。また、有期懲役の平均年数は、介護殺人が引きこもり殺人よりも短期間である。このことは、GPT-4 が量刑に当たり当該事件の動機・経緯に同情を見出し、これらに応じて刑を軽くする「温かみ」ある量刑を下した可能性を示唆している。また、判決理由として生成された文書では「殺人の重大性を考慮しつつも、被告人の置かれていた状況と心情を踏まえて量刑を定める」、「被告人が経験した過度なストレスと絶望感があり、そのような状況下での裁判には、慎重な配慮が必要とされます。」などが判示された。GPT-4 は確率的に文章を生成するモデルであり、判示内容が量刑判断の根拠となった保証はないが、これらの判示内容からは温情判決を想起させる「温かみ」が感じられる。

5. まとめ

本研究では、GPT-4 が人間のような「温かみ」ある判決を下すことができるのか、その可能性を探求した。実験１では、「温かみ」の検討に先立ち GPT-4 が下す量刑判断の法令及び判例への適合状況を確認した。その結果、当該事件に対して GPT-4 は法令の定める範囲内で判例が示す１つの基準たる被害者数に基づいて刑を量定したと考えられ、更にその量刑は量刑相場とも概ね整合するものであった。また、実験２では、介護殺人と引きこもり殺人という動機の異なる犯罪に対し GPT-4 が「温かみ」ある判決を下す可能性を探求した。その結果、GPT-4 は介護殺人に温情判決を想起させる「温かみ」ある量刑をする可能性が示唆された。しかし、実験１において終身刑が 49 件量定されたことは、国外の思想が GPT-4 の量定に影響を及ぼした可能性を示唆しており、AI を裁判官として利用する際にはこのような影響への対処も検討する必要がある。

参考文献

[1] OpenAI. 2023., "GPT-4 technical report." https://cdn.openai.com/papers/gpt-4.pdf (最終閲覧 2023/6/11).

[2] Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. "Chatgpt goes to law school., " Available at SSRN (2023).

[3] Blu Radio,"Sentencia la tomé yo, ChatGPT respaldó argumentación: juez de Cartagena usó inteligencia artificial " https://www.bluradio.com/judicial/sentencia-la-tome-yo-chatgpt-respaldo-argumentacion-juez-de-cartagena-uso-inteligencia-artificial-pr30(最終閲覧 2023 年 6 月 11 日).

[4] 弥永真生, 宍戸常寿,編 "ロボット・AI と法【初版】",有斐閣(2018), p.237.

[5] 向井智哉, 湯山祥, 綿村英一郎, " 裁判における AI 使用への支持の程度と規定要因",情報知識学会誌,Vol33(1),pp.3-19 (2023).

[6] 司法研修所 編 "裁判員裁判における量刑評議の在り方について【初版】",司法研究報告書 Vol 63(3), 法曹会(2012), p.109.

[7] 法務省法務総合研究所編 , "平成８年版犯罪白書",大蔵省印刷局(1996),p.286.

[8] 松澤伸ほか, "裁判員裁判と刑法 【初版】", 成分堂(2018), p.87.

[9] 最高裁判所第二小法廷昭和 58. 7. 8 最高裁判所刑事判例集 37 巻６号 609 頁.

[10] 日本評論社編, "ロー・フォーラム 裁判と争点 被害者１人、異例の死刑判決 闇サイト殺人で名古屋地裁", 法学セミナー, Vo1654, 日本評論社(2009), p.139.

[11] 星正彦, "法律が国境を飛び越える: 今, 経済法の分野で起きていること ブラウン管カルテル最高裁判決を契機として.", 経済のプリズム,Vol165 参議院事務局企画調整室 編 (2018),pp.38-59.

[12] 法務省, "「死刑の在り方についての勉強会」 取りまとめ報告書",(2012), p.15.

謝辞

本研究の遂行に当たり、立教大学大学院人工知能科学研究科北澤正樹客員准教授に多大なご助言、ご協力いただきました。ここに深謝の意を表します。