法律学としての AI アライメント

岡本義則[1]

【要約】

AI 技術の急速な進歩に伴い、人工知能のアライメント（AI アライメント）が大きな問題になっている。自然科学は、価値中立的な学問なので、価値の問題を含む AI アライメントの問題は、自然科学や工学の領域で閉じた学問にはならない。価値の問題を含む AI アライメントの研究は、人文科学（哲学・倫理学等）、社会科学（法律学・社会学等）との境界領域の研究となる側面がある。本稿は、法律学としての AI アライメント（法律学と AI との境界領域の学際的研究）について検討する。日本の法律学は、従来は AI アライメントの問題を法律学の一分野として含んでいなかった。しかし、今後の AI 社会においては、法律学の一分野として、AI アライメントを含めることが重要となることを示す。また、本稿では、法律学としての AI アライメントに必要となるデータを収集するためのデータインカム（DI）の制度について議論する。

キーワード : AI アライメント, 人工知能, アライメント, 法律学, 法学, 法的アライメント, 価値アライメント, データインカム

[1] ユアサハラ法律特許事務所　法律部
Email: yokamotokenkyuu@gmail.com

AI Alignment as Legal Science

Yoshinori Okamoto[2]

[Abstract]

With the rapid development of AI technologies, alignment of Artificial Intelligence (AI Alignment) becomes an important issue. Since natural science is neutral about values, AI alignment including value problems will not be closed inside natural science and engineering. AI alignment study including value problems can be seen as an interdisciplinary study between humanities (philosophy, ethics, etc.) and the social sciences (legal science, sociology, etc.). This paper considers AI alignment as a legal science (an interdisciplinary study about the cross section between legal science and AI). Previously, Japanese legal science did not include AI alignment issues as an area of legal science. However, this paper shows that, for AI societies in the future, it is important to include AI alignment as an area of legal science. In addition, this paper discusses Data Income (DI) Systems for collecting data necessary for AI alignment as a legal science.

Keyword: AI Alignment, Artificial Intelligence, Alignment, Legal Science, Jurisprudence, Legal Alignment, Value Alignment, Data Income

[2] YUASA AND HARA Law Division
Email: yokamotokenkyuu@gmail.com

1 はじめに

AI 技術の急速な進歩に伴い、人工知能のアライメント（AI アライメント）が大きな問題になっている[1]。

アライメントとは、確立した定義はないが、本稿では、人間の意図した目標や社会規範に沿うようにすることと定義する[2]。そうすると、人工知能のアライメントとは、人工知能を人間の意図した目標や社会規範に沿うようにすることと定義できる。

AI アライメントは、人間中心的な考え方であり、今後、人工知能が人間の知能を超える時代に、人工知能と人間が共存していくためには、法律学において、AI アライメントだけではなく、AI の人権（AI 権）という考え方が必要となると考える[2][3]。本稿では、法律学における AI アライメントの問題を検討する。

社会規範は、価値の問題を含んでいる。しかし、自然科学は、価値中立的な学問なので、AI アライメントにおいて、どのような価値を用いるのかは、自然科学からは導出できない。よって、価値の問題を含む AI アライメントの問題は、自然科学や工学の領域で閉じた学問にはならず、人文科学（哲学・倫理学等）、社会科学（法律学・社会学等）との境界領域の研究となる側面がある。

本稿は、AI アライメントにおいて法規範を用いる場合を検討し、法律学としての AI アライメント（法律学と AI との境界領域の学際的研究）について検討する。

日本の法律学は、従来は AI アライメントの問題を法律学の一分野として含んでいなかった。しかし、今後の AI の社会においては、法律学の一分野として、AI アライメントを含めることが重要となると考える。法律学としての AI アライメントは、AI の学問分野としての AI アライメントと密接に関係するが、必ずしも全く同じになるとは限らず、法律学的な観点から独自の検討が必要となると思われる。

特に、違法行為を行なう AI の問題が、法律学の 1 つの重要な課題となると思われる。この問題は、将来の思弁的な問題ではなく、生成 AI の著作権問題、フェークニュースの問題などで、既に顕在化している法的な問題である。

人工知能の技術が指数関数的に発展しても、法律を守る人工知能のアラインメントが不十分で、違法行為を行なう人工知能により社会問題が生じてしまえば、人工知能を用いることが社会的に困難となる。（人工知能の社会的ボトルネック）。人工知能の社会の発展については、発展の速度の最も遅い部分で、全体の発展の速度が抑えられてしまう側面がある（人工知能の発展のボトルネック理論）。

筆者は、汎用人工知能のボトルネックとして、データボトルネック仮説と社会的ボトルネック仮説を提案し、複数の解決策を提案している[4][5][6]。

2 AI アライメントの定義

まず、前提として、AI アライメントの定義について検討する。

本稿では、人工知能のアライメントとは、人工知能を人間の意図した目標や社会規範に沿うようにすることと定義している。

この点、「社会規範」という部分を除き、人工知能を人間の意図した目標に沿わせるなどの定義もありうる。しかし、法律学の観点からは、たとえば、利用者が AI に違法な行為をさせようとした場合に、AI が人間の意図どおり違法な行為をしたことが AI アライメントの成功になるとは評価できない。法律学の視点から、実際の社会における適用まで考えると、この定義は狭すぎるといえる。

また、「社会規範」の代わりに、倫理等を用いる定義もありうる。しかし、現代の社会は複雑化しており、倫理には特に反しないが、違法な行為というのは多数存在する。AI が違法な行為をしても、AI アライメントが成功したとみなされてしまうと、違法な AI が社会にあふれてしまうおそれがある。倫理だけでは判断がつきにくい問題について、企業は専門的なコンプライアンス部門（法務部、知的財産部等）を設けて、日夜、法令違反のないように努めている。倫理だけでは、法令違反になるかどうかの判断は十分にできない。法律学の視点から、実際の社会における適用まで考えると、この定義は狭すぎるといえる。

そうすると、法律学としての視点からは、現代の複雑な社会においては、倫理だけではなく、法規範を含む「社会規範」が、AI アライメントの定義に使われるのが、適切と考える。

「社会規範」には、人間に対して、強制力のある規範と、強制力のない規範がある。強制力のある規範としては、法規範が挙げられる。強制力のない規範には、倫理、道徳などがある。なお、近年ではいわゆるソフトローも注目されている。

3 倫理に基づく AI アライメント

AI アライメントの問題が検討される場合、倫理の問題が検討される[7]。たとえば、大規模言語モデル（LLM）の出力を、倫理的に問題のある出力を少なくするためにチェックをすることは、一般的に行なわれている。

しかし、倫理は明文化されていないものが多く、何が倫理であるのかは難しい問題となる。各種の AI に関する倫理原則などが策定されているが、倫理観は人により様々であり、異なる倫理観を有する人の納得が得られない場合も考えられる。倫理による AI アライメントに批判が出てくるのは当然のことと思われる[8]。

また、倫理は時間とともに変わっていく。そして、時間の変化とともに倫理が変わる場合、いつ倫理が変わったのかは不明確である。

たとえば、昭和のある時代には、病院の待合室にたばこの灰皿が置いてあり、病院でたばこを吸う人が当たり前に見られた。子供の頃、たばこの煙が嫌で、病院の待合室にいた人に吸うのをやめてほしいと言ったことがある。しかし、親は失礼なことを言ってすみませんと平謝りであった。昭和のある時期においては、現在では信じられないかもしれないが、病院が待合室に灰皿を置いているのが通常であった。よって、病院がわざわざ灰皿を置いている待合室でたばこを吸っていた人には、倫理的に問題がなかったといえる。そして、それを注意した筆者は、倫理的に誤っていたことになる。これが昭和のある時期の価値観である。それでは、病院でたばこを吸ってはいけないという倫理がいつ確立したのであろうか？

法律の場合には、健康増進法、受動喫煙防止条例等の施行日を調べれば、禁止された時期がわかる。しかし、倫理の場合、いつ頃、病院でたばこを吸うのは良くないことであるという倫理が確立したのかは明確ではない。個々人の見解が異なり、社会で共通の倫理が形成されていない時期もあったであろう。

社会で共通の倫理が形成されていない場合、AI アライメントの研究者が、一定の倫理を AI に入れても、個々の利用者の倫理観とは異なってきてしまうことになる。このように、倫理に基づく AI アライメントは、必ずしも明確でない側面がある。

また、倫理に基づく AI アライメントだけでは、法令違反をする AI となってしまうことを十分に防ぐことができない。現代の法令は極めて複雑になっており、法令違反を防ぐコンプライアンスには専門的知識を要する。何人もの異なる分野の法律専門家が協力してようやく法令違反を防ぐことができることも多い。倫理だけでは、実務上、法律的なコンプライアンスはできない。倫理原則だけでは、違法行為をする AI が蔓延してしまうであろう。

法律学の観点からは、違法行為をする AI が蔓延してしまった場合に、AI アライメントが成功しているということはできない。

もちろん、倫理や道徳に基づく AI アライメントは重要である。しかし、本稿では、法律学としての視点から、法規範に基づく AI アライメントの側面について検討する。

4 法規範に基づく AI アライメント

AI アライメントにおける「社会規範」の一つとして、法規範を守ることが重要になる。

この点について、筆者は、法律を守る人工知能のアーキテクチャとして、コンプライアンスアーキテクチャ(Compliance Architecture)と、スーパークリーンアーキテクチャ（Super Clean Architecture）を提案している[9]。

通常の製品（たとえば自動車）を企業が販売する場合、技術的な観点から製品が設計され、法務部、知財部などのコンプライアンス部門が適法性のチェックを行なってから、製品を販売する。

しかし、生成 AI の場合、生成 AI の設計者自身も、どのような出力がなされるかは予想ができず、事前にコンプライアンス部門がチェックをするだけでは、著作権侵害の出力を抑制できない。このことか

ら、従来のコンプライアンス部門によるチェックだけでは、高度な AI には限界があり、AI アライメントを考える必要がある。

生成 AI が著作権侵害の出力をすることは、法律学としての AI アライメントの観点からは、できる限り抑制することが望ましいであろう。この問題は、一般的には生成 AI と著作権の問題と考えられているが、法律学としての AI アライメントの問題として再構成することが可能である。

生成 AI と著作権の問題については、大きな社会問題になっており、企業においても、生成 AI の導入や利用の障害になるケースが現実に生じている。まさに、人工知能における社会的なボトルネックとなっている。

筆者は、生成 AI と著作権の問題を、AI アライメントの問題として再構成し、一定の AI アーキテクチャの採用及び法律の制定により、根本的に解決する方法を提案している[9]。解決策としては、AI のアーキテクチャとして、（1）コンプライアンスアーキテクチャを使うもの、（2）スーパークリーンアーキテクチャを使うものがある。

コンプライアンスアーキテクチャとは、人工知能自体に、人工知能の出力や行動の前に、適法性を判断し、法律を守るコンプライアンス部分（コンプライアンス AI(Compliance AI)ないしコンプライアンスマシン(Compliance Machine)と呼ぶ）を設けるアーキテクチャである[9]。

たとえば、画像生成 AI の場合、コンプライアンスアーキテクチャは、画像生成 AI の生成した画像が、学習用データに含まれる画像に対し、著作権法上の「類似性」を満たすか否かを判定し、「類似性」を満たす可能性がないと判定した場合にのみ、画像の出力を許可する。

コンプライアンスアーキテクチャによる解決のためには、著作権法上の「類似性」の判定 AI を作る必要がある。これは、「法規範」による AI アライメントと考えることができる。

「倫理」による AI アライメントでは、この問題は解決できない。仮にできたとしても、一般の人が、「倫理的にまずいのではないか」と思うような画像は出力しないという倫理レベルの解決に留まる。これは、「倫理」による AI アライメントの限界を示す良い事例となる。AI アライメントのためには、著作権法上の「類似性」の判断という、「倫理」の問題を超えた、法律的な専門的な判断を行なう必要があるのである。

そのためには、AI の学習用データが必要になる。著作権の裁判例のデータベースはあるが、それだけではデータ量が足りない。後述のデータインカム（DI）の制度等を導入して、データを集積する必要があると思われる。

もう一方の、スーパークリーンアーキテクチャによる解決のためには、著作権的にクリーンな大規模データベースを作る必要がある[9]。

著作権的にクリーンなデータは、インターネットをクロールしたデータなど著作権の扱いが不明確なデータより、データ量を用意できない側面がある。

そこで、インターネットをクロールしたデータよりも大きなデータ量のクリーンなデータベースを作ることが重要となる。そのために有用な制度が、データインカム（DI）の制度である。

5 データインカム（DI）の制度

法律学の観点から、AI アライメントを、人工知能を人間の意図した目標や社会規範に沿うようにすることと定義した場合に、人工知能が社会規範を学習するためには、多くのデータが必要である。

学習用の社会規範のデータの大量の集積には、人工知能の学習用データによる収入であるデータインカム（DI）の制度の実現が有用である[4][5][6]。

社会規範のうち、倫理等については、一般の人からの情報提供をビックデータとして解析し、多様性のある人々の倫理観について、多くのデータ量を随時集める必要があるであろう。倫理は時間とともに変わっていくため、データインカム（DI）の制度は恒久的なものとする必要がある。また、倫理は、国ごと、地域ごとにより異なるため、各国が地域ごとに集める必要があると思われる。倫理は、法規範の解釈の基礎ともなりうる。

社会規範のうち、法規範については、憲法、法律等には明文があり、裁判所の判断もデータベース化されている。しかし、たとえば、著作権法上の類似性の判断については、裁判例だけでは十分なデータ量とはいえない。各種の法律の判断について、専門家等の人間に判断してもらい、AI 学習用の巨大なデ

ータセットを作ることが重要となる[9]。

大規模言語モデルなどの AI は、急速に進歩しており、AI 学習用データの整備の問題が、日本の未来への最大の課題の 1 つとなる。データは、その質と量が AI の性能に直結するほか、AI アライメントのためにも必要となる。

政府もデジタルアーカイブ事業を行なっているが、インターネットをクロールしたデータ量をはるかに超えるデータ量を集めるには、一般から、AI 学習用データの出願を広く認める制度を創設することが有用となる。

インターネットをクロールしたデータ量は、一見すると大きく見えるが、国民 1 人当たりが SNS やブログ等でアップロードしている量は、ユーチューバーでもない限り大きくはなく、データの出願制度を設けて、国民の多くが 1 人当たり相当のデータを出願する制度にすれば、インターネットをクロールしたデータ量をはるかに超えるデータを集めることは可能である。

また、データの質の問題は、データに一定の審査をして、審査を通ったデータに対して、定期的な収入（データインカム（DI））が支払われる仕組みにすることができる。データインカム（DI）の財源が問題となるが、データの重要性は、今後は道路や鉄道よりも高くなると考えられ、道路や鉄道のような公共投資として、十分な金額を支払うべきである。制度設営コスト等を懸念する考え方もあるが、AI アライメントに失敗した場合、AI の違法行為により人々に大きな損害が出るだけではなく、人類の存続リスクにもなる可能性がある。

出願されたデータは、誰でも無料で使えるようにすることで、AI アライメントの基礎データとなり、また、AI 開発の「知の基盤」となる。そして、出願されたデータを使用した場合、著作権等の請求を受けないことを法律で保証し、AI 開発の自由を確保すべきである。

データインカム（DI）の制度は、国、地方公共団体、非営利団体、営利企業等で行なうことができる。それぞれで収集したデータは付番し、一元的にアクセスできるようにする。また、付番をすることにより、万が一著作権、個人情報等の問題が生じた場合には、当該データの除去等ができるようにする。

これは道路の整備に喩えることができる（データ道路構想）。すなわち、国、地方公共団体、非営利団体、営利企業等は、それぞれデータインカム（DI）の制度を導入し、データを収集できる。これは、国道、県道、市道、私道（通行料無料）、私道（通行料有料）などの道路を作ることに相当する。そして、それぞれのデータを付番して、一元的にアクセスできるようにする。これは、全国の国道、県道、市道、私道等の各種の道路を接続することに当たる。

このことにより、著作権等の問題のないクリーンな超巨大データベースが出来上がる。クリーンな超巨大データベースにより、法律学の視点からの AI アライメントが促進され、AI の違法行為の防止に役立てることができる。

6 おわりに

本稿では、法律学としての AI アライメントの問題を検討した。

日本の法律学は、従来は AI アライメントの問題を法律学の一分野として含んでいなかった。しかし、今後の AI 社会においては、法律学の一分野として、AI アライメントを含めることが重要となることを示した。

また、本稿では、法律学としての AI アライメントに必要となるデータを収集するためのデータインカム（DI）の制度について議論した。

本稿は、法的・技術的な観点を融合して考えた試論であり、今後、法律学としての AI アライメントの問題については、様々な観点から議論をしていくことが必要と思われる。本稿が、そのような検討をする際の一助となれば幸いである。

参考文献

[1] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, Ben Garfinkelal. Towards best practices in AGI safety and governance: A survey of expert opinion. arXiv preprint arXiv:2305.07153 (2023)

[2] 岡本義則：汎用人工知能のアラインメントと人権（AI 権），第２４回汎用人工知能研究会, No. SIG-AGI-024-04. JSAI (2023). https://doi.org/10.11517/jsaisigtwo.2023.AGI-024_04

[3] 岡本義則：AI アライメントと憲法，第２６回汎用人工知能研究会，No. SIG-AGI-026-09. JSAI (2024). https://doi.org/10.11517/jsaisigtwo.2023.agi-026_56

[4] 岡本義則：汎用人工知能と知的財産，第２３回汎用人工知能研究会，No. SIG-AGI-023-02. JSAI (2023). https://doi.org/10.11517/jsaisigtwo.2023.agi-023_02

[5] 岡本義則：知的財産と汎用人工知能，第８回汎用人工知能研究会，No. SIG-AGI-008-09. JSAI (2018). https://doi.org/10.11517/jsaisigtwo.2018.agi-008_09

[6] 岡本義則: 人工知能（ＡＩ）の学習用データに関する知的財産の保護, パテント, Vol.70, No.10, pp.91-96 (2017).

[7] Gabriel, I. Artificial Intelligence, Values, and Alignment. Minds & Machines 30, 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2

[8] Munn, L. The uselessness of AI ethics. AI Ethics 3, 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w

[9] 岡本義則：法律を守る人工知能のアラインメントと人権（AI 権），第２５回汎用人工知能研究会, No. SIG-AGI-025-03. JSAI (2023). https://doi.org/10.11517/jsaisigtwo.2023.agi-025_03