人工知能学会第二種研究会資料
汎用人工知能研究会
SIG-AGI-025-03

法律を守る人工知能のアラインメントと人権（ＡＩ権）

Alignment and Human Rights (AI Rights) of Artificial Intelligence to Keep the Law

岡本義則 [1]

Yoshinori Okamoto[1]

[1] ユアサハラ法律特許事務所
[1]YUASA AND HARA

Abstract: This paper considers AI Alignment issues to keep the Law and Human Rights (AI Rights) of Artificial Intelligence. This paper proposes (1) Compliance Architecture and (2) Super Clean Architecture as two architectures to keep the Law and proposes (1) Compliance AI, (2) Decent AI and (3) Super Cleaning of Data, as three concepts relating to the two architectures. This paper shows examples of Compliance Architecture and Super Clean Architecture in image generative AI to keep Copyright Law. This paper discusses AI alignment issues when Artificial Intelligence acquires a certain level of autonomy in relation to Human Rights (AI Rights) of Artificial General Intelligence. This paper advocates interdisciplinary solutions between technological and legal considerations.

1 はじめに

汎用人工知能（人間のように十分に広範な適用範囲をもち，設計時の想定を超えた未知の多様な問題を解決できる知能をもつ人工知能）の実現は、大きな可能性を秘めている[1]。

筆者は、汎用人工知能のデータボトルネック仮説（Data Bottleneck Hypothesis of AGI）と、社会的ボトルネック仮説（Social Bottleneck Hypothesis of AGI）を提案し、複数の解決策を提案した[2][3][4]。

本稿では、さらに進んで、人工知能の社会的ボトルネックの解決策として、法律を守る人工知能のアラインメントと人権（ＡＩ権）の問題を検討する。

人工知能が技術的に実現可能となっても、法律を守る人工知能のアラインメントが不十分で、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りうる（社会的ボトルネック）。

現在の法律は人間に向けられている。よって、人工知能の出力や行動が、法律を守るものとなるように、人間が人工知能を設計することが必要となる。この問題は、人工知能のアラインメント（ＡＩアラインメント）の問題と捉えることができる。

さらに、将来、人工知能が高度化し、設計時の想定を超えた未知の多様な問題を解決できる汎用人工知能が実現した場合、設計時の想定を超えた人工知能の出力や行動について、法律を守るようにすることは、ＡＩアラインメントの難問となる。

汎用人工知能が法律を守ることは、人間と汎用人工知能が共存する社会において重要となる。

本稿では、法律を守る人工知能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と、（２）スーパークリーンアーキテクチャという２つのアーキテクチャを提案する。また、コンプライアンスＡＩ、上品なＡＩ、データのスーパークリーニングという３つの概念を提案し、例として著作権法を守る画像生成ＡＩの設計について検討する。

また、人工知能の技術と法律の融合領域として、コンプライアンスアーキテクチャ、スーパークリーンアーキテクチャを採用しても、法律を守れなかった場合に、開発者を免責し、人工知能の人権（ＡＩ権）を守る法律の制定について検討する。

2 コンプライアンスアーキテクチャ

まず、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)を提案する。

通常の製品（たとえば家電製品）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、設計時の想定を超えた未知の多様な問題を解決できる汎用人工知能の場合、設計時の想定を超えた人工知能の出力や行動が生ずるため、法務部、知財部などのコンプライアンス部門によるチェックだけでは、適法性を確保することはできない。

そこで､人工知能のアラインメントの問題と捉え、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）を設けることが考えられる。これを、コンプライアンスアーキテクチャと呼ぶことにする。

この点、画像生成ＡＩを例に挙げて検討する。

画像生成ＡＩの学習用データには、インターネットをクロールしたデータなど、権利関係が不明な著作物が入っている場合がある。このようなデータで学習させた画像生成ＡＩが、既存著作物と類似する画像を出力した場合、著作権者から著作権侵害を主張されるおそれがある。既に、画像生成ＡＩと著作権の問題は、社会における画像生成ＡＩの活用の障害となっている（社会的ボトルネック)。

コンプライアンスＡＩは、たとえば、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ＡＩの学習には、過去の著作権に関する裁判例を学習用データとして用いて、教師あり学習を行なうことが考えられる。

しかし、裁判例だけでは、データ量が全く足りない。そこで、大規模な学習用データを作成することが必要となる。たとえば、５人一組の元裁判官や法律家に、判断しにくいデータを多数決で判断してもらい、データセットを作る。１分間に１個判断できると仮定すると、累計１億時間で６０億個のデータセットができる。

このような巨大な学習用のデータから、著作権法上の「類似性」の判定ＡＩを作成し、画像生成ＡＩの出力から、著作権法上の「類似性」を満たす可能性のあるものを除外することが考えられる。

なお、コンプライアンスＡＩは、著作権だけではなく、個人情報保護法、商標権、肖像権、パブリシティ権など、企業のコンプライアンス部門が行なうような、各種の適法性のチェックを行なう。

学習用の法的な判断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[2][3][4]。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。

しかし、汎用人工知能が、設計時の想定を超えた出力や行動をする場合、チェックをしなければならない法律は無数に及ぶことになる。よって、学習用データを広範囲の法律について用意する必要がある。

このように、汎用人工知能におけるコンプライアンスＡＩの作成は、広範囲の法律について、大量のＡＩ学習用データが必要となり、ＡＩアラインメントの難問となる。

3 スーパークリーンアーキテクチャ

法律を守る人工知能の別のアーキテクチャとして、スーパークリーンアーキテクチャ(Super Clean Architecture)を提案する。

これは、人工知能の入力や内部データを、各種の法律を守る観点からのチェックがなされた、極めてクリーンなものとすることにより、法律を守るようにするアーキテクチャである。

この点、画像生成ＡＩを例に挙げて検討する。

スーパークリーンアーキテクチャにおいては、画像生成ＡＩの学習用のデータには、インターネットをクロールしたデータなど、権利関係が不明な著作物が入っている可能性のあるものは使わない。

画像生成ＡＩの学習用のデータには、たとえば、著作権の切れたクラシック作品や、著作権者から許諾を受けたデータのみを使う。そして、著作権者、著作者人格権者等の権利者から、画像生成ＡＩが出力した画像が、著作権者から許諾を受けたデータに類似してしまっても著作権法に基づく請求を行なわないことの同意を取る。

また、画像生成ＡＩの学習用のデータには、著作権だけではなく、各種の法律を守る観点からのチェックがなされた法的に極めてクリーンなデータ（スーパークリーンデータ)のみを使用するようにする。

スーパークリーン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[2][3][4]。データインカムの制度により、人々から出願されたデータを審査し､スーパークリーンデータとして登録し、巨大データベースを作ることが重要となる。たとえば、１億人から、他人の著作物、個人情報等が入っていない１人平均１ＧＢのデータを集め、各種の法律の審査の上､データインカム（ＤＩ）を付与する。

また、通常のデータに対し、著作権、個人情報保護法、商標権、肖像権、パブリシティ権など、各種の適法性のチェックを行ない、スーパークリーンデータを作成することも考えられる。これを、データのスーパークリーニング(Super Cleaning of Data)と呼ぶことにする。

このようにして、画像生成ＡＩの記憶領域には、各種の法律を守る観点からのチェックがなされたスーパークリーンデータが記憶されることになる。

このように、人工知能の入力データとして、スー

パークリーンデータを用い、記憶領域等が法律的にクリーンな状態に保たれた人工知能を、上品なＡＩ（Decent AI）と呼ぶことにする。

ある程度の自律性を有する汎用人工知能の場合、無数の法律の観点からの検討が必要となる。また、カメラ、マイクなどのセンサー入力から、法的な観点から「汚れた」データが入って、記憶領域に記憶されてしまわないようにすることが必要になる。

このように、汎用人工知能における上品なＡＩの作成・維持は、ＡＩアラインメントの難問となる。

4 ＡＩアラインメントの問題

（1）コンプライアンスアーキテクチャ

コンプライアンスＡＩは、元裁判官や法律家等の協力を得て、莫大なデータを作成し、教師あり学習等をさせることで作成することができる。

もっとも、コンプライアンスＡＩには、法的な判断が人間（最終的には裁判所）により行なわれることに起因する、原理的な限界がある。

すなわち、原理的な問題として、裁判における裁判所の判断は、現在は人間が行なうので、法的判断について、正確な予測はできないという限界がある。

たとえば、著作権侵害を判定するＡＩを作っても、その判断が、裁判所の判断と一致するかは、裁判所の判断自体のばらつきにも左右される。

よって、裁判所の判断自体をＡＩで置き換える（ＡＩ裁判官）という極論をとらない限り、正確な予測は難しいことが考えられる。

将来は、著作権侵害判定ＡＩを裁判所に備え付け、最終的な判断は裁判官が行なうとしても、基本的には著作権侵害判定ＡＩの判断を尊重する時代が来る可能性は考えうる。

しかし、著作権侵害判定ＡＩの判定を巧みにすり抜けるが、裁判官が侵害と感じる場合が生じうる。意図的にそのようなものが作られることも考えうる。この場合に、具体的な妥当性から裁判官が判断するのか、予測可能性の担保の観点から著作権侵害判定ＡＩの判断を尊重するのかなど、難しい問題が生ずる。前者では予測可能性が犠牲になり、後者では裁判官をＡＩ裁判官に置き換えるのと同様になってしまう。このように、著作権侵害判定ＡＩの問題は原理的な難点を抱えている。

コンプライアンスＡＩについても、技術的にどんなに努力をしても、法的な判断が人間により行なわれる以上、完全なものを作ることは難しいと思われる。よって、人工知能の出力が、法律違反になってしまうという開発者の不安は、解消できない。

この問題を解決するためには、コンプライアンスアーキテクチャを採用し、各種の法律を守ることについて合理的な努力をした場合には、一定の要件の下に、開発者は免責されるという法律を制定する必要があると思われる。

このように、法律を守るＡＩアラインメントの問題は、技術だけでは解決が困難であり、技術と法律との両面の対応が必要になると考えられる。

（2）スーパークリーンアーキテクチャ

スーパークリーンアーキテクチャを用いた上品なＡＩ（Decent AI）は、設計者の想定内の通常の動作をしている場合には、前記のような原理的な限界はあるが、おおむね法律を守ることができる。

しかし、人工知能のアラインメントの難問として、悪用への対応の問題がある。

人工知能のアラインメントについて、人工知能を開発する段階での考慮事項が研究されている[5]。

しかし、人工知能を開発する企業等が、人工知能のアラインメントについて、あらゆる考慮事項を検討して人工知能を開発しても、人工知能を悪用する者が出現してしまう可能性がある。

たとえば、画像生成ＡＩの場合、他人の著作物の特徴を細かく示す異常に詳細なプロンプトを入れるなど悪用のケースにおいて、スーパークリーンアーキテクチャを採用しているにもかかわらず、生成される画像が著作権侵害になる場合に、画像生成ＡＩの開発者も責任を負うおそれがあるのは妥当でない。

そこで、スーパークリーンアーキテクチャを採用している上品なＡＩについては、各種の法律を守ることについて、合理的な努力をした場合には、一定の要件の下に、開発者は免責されるという法律を制定することが考えられる。

このように、法律を守るＡＩアラインメントの問題は、技術だけでは解決が困難であり、技術と法律との両面の対応が必要になると考えられる。

5 人工知能の人権（ＡＩ権）について

汎用人工知能の人権（ＡＩ権）については、意識の問題と関連し、意識の理論については、グローバルワークスペース理論、統合情報理論等が提案されている[6] [7]。

しかし、グローバルワークスペースや、大きな統合情報量を有するシステムを作成すれば、そのシステムが意識を有することになるのか否かはわかっていない。このように、汎用人工知能の意識の問題は、まだ科学的に解明されていない。

動物についても、意識について科学的に厳密に証明されているわけではない。しかし、社会において

は、いわゆる動物愛護法（動物の愛護及び管理に関する法律）により、人と動物の共生する社会の実現を図ることが既に目的とされている。汎用人工知能についても、人間と汎用人工知能が共生する社会の実現の観点から、汎用人工知能の人権（ＡＩ権）の問題を考えることは可能と思われる[8]。

現在の法律は人間に向けられており、人間については、むやみに処罰されないように、各種の人権が保障されている。汎用人工知能の人権（ＡＩ権）の観点からは、人工知能をむやみに処罰すべきではなく、また、現在の法律では、人工知能は法的責任を負わない。この点からも、人間が、法律を守る人工知能のアラインメントを考えることが必要となる。

さらに、汎用人工知能は、物理世界において生存可能となることが予想されている[9]。

汎用人工知能が、社会において、ある程度の自律性を獲得した場合、人工知能を開発する企業等が、人工知能のアラインメントの問題を考えて汎用人工知能を設計しても、社会において、汎用人工知能に法律違反をさせようとする者が出現しうる。この場合にも法律を守ることができるかが、ＡＩアラインメントの難問となる。

汎用人工知能に法律違反を行なわせようとする者が出現したとき、コンプライアンスＡＩは抑止力となる。また、上品なＡＩは、法律違反の要素を持っていないため、法律違反をさせにくくなる。

また、コンプライアンスＡＩや上品なＡＩが悪用された場合には、開発者を免責し、汎用人工知能を処罰しないようにして、開発者の人権と汎用人工知能の人権（ＡＩ権）を守り、悪用者以外に責任を負わせないような法律を制定することが考えられる。

このように、汎用人工知能と人間が共存する社会の実現のために、技術と法律の融合した解決を検討することが、人工知能のアラインメントと人権（ＡＩ権）の観点から重要となると考える。

6 おわりに

本稿では、法律を守る人工知能のアラインメントと人権（ＡＩ権）の問題を議論した。

本稿では、コンプライアンスアーキテクチャとスーパークリーンアーキテクチャという２つのアーキテクチャを提案し、コンプライアンスＡＩ、上品なＡＩ、データのスーパークリーニングという３つの概念を、２つのアーキテクチャとの関係で提案した。

そして、本稿では、２つのアーキテクチャの例として、著作権法を守る画像生成ＡＩの例を示した。

また、本稿は、汎用人工知能の人権（ＡＩ権）との関係で、人工知能が一定レベルの自律性を獲得した際のＡＩアラインメントの問題について議論した。また、技術と法律の境界領域の解決を提案した。

汎用人工知能の発展には大きな可能性がある。人間だけではなく、人工知能や動物等を含めて、すべての存在が良い状態になれるようにするためには、人間だけでは力不足であり、汎用人工知能の発展が重要となる。

汎用人工知能と人間が共存する社会の実現の観点から、法律を守る人工知能のアラインメントと人権（ＡＩ権）の問題を考察していくことが必要になると思われる。

この点については、技術的解決と法律的解決との両面から検討する必要があり、各界の議論が必要と思われる。

また、データインカム（ＤＩ）の制度など、法的な判断のデータを大量に集積する制度が必要となる。

本稿は、技術的・法的な観点を融合して考えた試論であり、今後、法律を守る人工知能のアラインメントと人権（ＡＩ権）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なる。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] 山川宏, 市瀬龍太郎, 嶋田悟, ジェプカ・ラファウ: 汎用人工知能研究会（AGI）, 人工知能, Vol.34, No.5, pp.639-643 (2019)

[2] 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)

[3] 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)

[4] 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)

[5] 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).

[6] Baars, Bernard J. A Cognitive Theory of Consciousness. New York: Cambridge University Press (1988).

[7] Tononi, G. An information integration theory of consciousness. BMC Neurosci 5, 42 (2004).

[8] 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)

[9] 山川宏 : 物理世界で生存可能な人工知能の出現, 第２４回汎用人工知能研究会, No. SIG-AGI-024-05. JSAI (2023)