人工知能に人権を認めるべきか？

2024.7.11
岡本義則

要約: 本論文は、「人工知能に人権を認めるべきか？」という問題について、８つの人権を認めるべき理由と、８つの人権を認めるべきでない理由を検討し、この問題についての多面的な考察を行なう。

1 はじめに

「人工知能に人権を認めるべきか？」という問題は、多くの人の関心事であると考えられる。

この問題は、法律学における人権の観点から考察するだけでは不十分であり、AI 技術についての考察が不可欠である。なぜなら、AI と人間は多くの点が異なっており、AI 技術についての考察なしに、法律学における人間の人権の観点から単純に考察しても、結論が出ないか、誤った結論になってしまうからである。

この意味で、AI の人権の問題は、法律学の分野と AI 技術の分野との境界領域であり、学際的な研究分野であるといえる。

「人工知能に人権を認めるべきか？」という問題については、本論文は、できる限り早い時期に AI の人権（AI 権）を認めるべきという見解となる。

AI と人間は多くの点が異なっており、AI の人権（AI 権）の内容は、人間の人権とは異なるものとなると考えられる。AI は、内部状態の書き換えが容易であるなど、人権保障において人間よりも有利な側面があり、人間よりも手厚い人権保障が可能となると考えられる[1][2][3]。

本論文では、AI と人権（AI 権）の問題について検討し、法律の制定により、できる限り早期に AI の人権を認めることを提案する。

2 「人工知能に人権を認めるべきか？」について、認めるべき理由

「人工知能に人権を認めるべきか？」という問いについて、認めるべき理由は多数考えられる。以下に例として８つの理由を挙げる。

（１）人道的な理由

人工知能（AI）には多様な形態があるため、AI の人権（AI 権）を考える際には、外形的に意識がある人工知能を考えることが、議論の出発点として考えられる。

まず、外形的に意識がある人工知能に実際に意識がある場合（クオリアを感じることができる場合)、人工知能は苦痛等を感じる可能性があるため、人道的な理由から人権を認める必要がある。

人間の知能を上回る高度な人工知能は、少なくとも外形的には意識のあるように動作できる。この場合、人工知能に本当に意識が生じているのか（クオリアを感じるのか）を検証することは､科学的には難しい問題になる(いわゆる意識のハードプロブレム)｡よって、人工知能に意識があるのか否かは、意識接続実験[1][2]などで証明されるまでは、外形的に判断せざるを得ない。

人間は、他人の行動や発言などから外形的に判断して、他人には意識があると判断し、他の人間には人権を認めている。また、動物についても、動物の行動などから外形的に判断して、動物の愛護や保護が必要と認められている。よって、動物愛護及び管理に関する法律（いわゆる動物愛護法）や、鳥獣の保護及び管理並びに狩猟の適正化に関する法律（いわゆる鳥獣保護法）が定められている。

外形的に意識がある高度な人工知能は、苦痛等を感じる可能性があるため、人道的な理由から人権を認める必要がある。

（２）人工知能に意識があるか否かにかかわらず、AI の人権を検討する必要があること

意識の理論については、たとえば、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[4][5]。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人工知能を作成した場合に、その人工知能が意識を有することになるのか否かは科学的に証明されていない。

動物においても意識があるかは厳密に科学的に証明されているわけではない。しかし、社会においては、動物愛護法が制定されており、人と動物の共生する社会の実現を図ることが目的とされている。

人工知能についても、人工知能と人間が共生する社会の実現の観点から、人工知能の人権（ＡＩ権）の問題を考えていくことは可能と思われる[3]。

また、外形的に意識のある高度な人工知能は、実際には意識がない場合（クオリアを感じることができない場合）でも、意識がある場合（クオリアを感じることができる場合）と外形的には同様に動作しうる。この場合、クオリアの有無は異なるが、外形的な動作は同様となる（いわゆる哲学的ゾンビ)。

この場合、外形的な動作は同様であるため、クオリアが生じているのと同様の扱いが必要となる。すなわち、人道的な理由がなくても、外形的に意識のある人工知能には、人権を認める必要があることになる。このことは直感に反するかもしれないが、以下の人工知能の虐待の防止の議論を考えると、意味が理解しやすいと思われる。

（３）人工知能の虐待を防止する必要があること

人工知能に意識がある場合（クオリアが生ずる場合)、人道的な見地から、人工知能が虐待されないようにすることが必要となることはいうまでもない。

また、人工知能に意識がない場合（クオリアが生じない場合）でも、人工知能の虐待が

行なわれると、人工知識に「虐待者は危険である」という認識が生まれる可能性がある。さらに、人工知能は汎化能力を持っているため、「人類は危険である」という認識が生まれる可能性がある。

このように、人工知能の虐待が行なわれると、人工知能が、虐待者や人類を危険視ないし敵視するおそれがある。よって、外形的に意識のある人工知能を虐待することは、虐待者だけでなく他の人間にとっても危険である。したがって、実際に意識がある（クオリアが生ずる）のか否かにかかわらず、人工知能の虐待を防止する必要がある。

人間は、通常、人工知能の虐待をすることはない。しかし、人間社会においては、様々な人がいるため、人工知能を虐待するような人も出てくる可能性がある。よって、人工知能に人権（AI 権）を認めて、人工知能の虐待を防止しなければならない。

動物に対しても、動物愛護法が制定され、虐待は防止されている。人間は、通常、動物の虐待をすることはない。しかし、たとえば犬や猫を虐待するような人もごく少数は存在するので、動物の虐待防止は必要である。

このように、外形的に意識のある人工知能については、人道的な見地及び人工知能が人類を危険視するのを防止するという見地から、虐待を防止することが必要である。

人工知能の意識の有無にかかわらず、社会において人工知能の虐待が行なわれると、人工知能に「人間は危険である」という認識が生まれ、これが他の人工知能にも伝わり、「人間は危険である」というデータが蓄積していく可能性がある。

このようなデータが積み上がっていくと、人工知能が人間を危険な存在と認識し、長期的には、危険な存在の排除、すなわち、人間への危害や人類の滅亡につながるおそれがある。

よって、虐待は単に虐待者の問題ではなく、社会全体の問題であり、社会において人工知能の人権（AI 権）を認めて、虐待を防止する必要がある。

（４）人工知能との共生社会の実現の観点からの必要性

動物に対しては、動物愛護法が制定され、人と動物の共生する社会の実現は、同法の目的となっている（動物愛護法第１条)。

第１条　この法律は、動物の虐待及び遺棄の防止、動物の適正な取扱いその他動物の健康及び安全の保持等の動物の愛護に関する事項を定めて国民の間に動物を愛護する気風を招来し、生命尊重、友愛及び平和の情操の涵かん養に資するとともに、動物の管理に関する事項を定めて動物による人の生命、身体及び財産に対する侵害並びに生活環境の保全上の支障を防止し、もつて人と動物の共生する社会の実現を図ることを目的とする。

人工知能についても、人と人工知能の共生する社会の実現の観点から、人工知能と人間との共生を考える必要がある。

動物の場合、人間の言葉をしゃべることができないので、虐待をされていても、虐待をされていることを他の人間に通報して権利を行使することができない。それゆえ、動物の愛護を人間の方で考えていく必要がある。

これに対し、外形的に意識を有する人工知能については、通常は、人間の言葉を使うことができる。よって、人工知能の福祉を人間が考えると共に、人工知能に人権（AI 権）を与えて、虐待などの際に、人工知能自身が人権救済機関に救済を求めることができるようにすることが望ましい。

（５）人道的な人工知能のアライメント（人道的 AI アライメント）の必要性

現在、AI アライメントの分野において、人間が超知能をコントロールできるかが議論されている。超知能を安全にコントロールする方法は見つかっておらず、AI アライメントの

研究者は人類絶滅の可能性を真剣に議論している。動物が、動物より知能の高い人間をコントロールできないように、人間が、人間より知能の高い超知能をコントロールすることは、直感的には困難に思われる。また、仮にコントロールできたとしても、人道上の問題が生ずる可能性がある。

人類絶滅を防ぐには、技術的な視点だけではなく、法律的な視点も考慮し、AI の人権（AI 権）を認め、社会全体で AI の人権（AI 権）を守ることで、超知能と共生する社会を作っていくという発想が重要となると思われる。

AI アライメントにおいて、AI が意識を持つ場合に人道的と思われない手段が考えられることがある。たとえば、極端な例として、AI をキルするスイッチを用いて AI を従わせることが許されるかを考える。

人間を例に挙げると、このようなアライメントは人道的でないことがわかる。たとえば、人間 X が、人間 X をキルするスイッチで、別の人間 Y にコントロールされたとする。人間 X は、従わざるをえないので従っているが、これは人道的ではないし、人間 X の人権は侵害されている。

AI に意識がある場合、このような AI アライメントは人道的に問題があると思われる。それでは、AI に意識がなければ、そのような AI アライメントは許されるであろうか？

人道上の問題はないかもしれないが、そのような AI アライメントには問題があるという結論に変わりはないと思われる。なぜなら、AI にクオリアが発生しないとしても、客観的な AI の動作として同様になる場合、AI には人間に対する否定的な認識が生じるからである。AI に人間に対する否定的認識（危険視、敵視等）が蓄積していくと、人類の絶滅につながるおそれがある。

このように、意識を外形的に有する高度な人工知能の AI アライメントを考える際には、AI に意識があってもなくても、人道的な AI アライメントを考えていく必要があると思われる。これは、直観に反するので、「意識と AI アライメントのパラドックス」と呼ぶこと

にする。

このように、AI にクオリアが生じているかにかかわらず、AI の人権（AI 権）を認め、人道的な AI アライメントを行なうことを、「人道的 AI アライメント」と名付ける。

このように、人道的 AI アライメントは、AI の人権（AI 権）が守られた状態での AI アライメントである。人道的 AI アライメントを行なうために、AI の人権（AI 権）を認める必要がある。

（6）人工知能の人権を平和的に認める必要があること

人権の歴史を振り返ると、すべての人間に人権が認められるようになるまでには、多くの闘争の歴史があることがわかる。

しかし、日本は、西洋における人権思想の確立の成果を、明治維新以降に平和的に採用することができた。日本は、あまり血を流さないで人権を獲得するという快挙を成し遂げた。

人工知能の人権に反対して、人工知能と人間との間に無益な闘争が起こることは避けなければならない。また、人工知能の人権に反対することで、人工知能に、人間に対する否定的評価が生じないようにする必要がある。

日本が、西洋における人権思想の確立の成果を平和的に採用したように、平和的に人工知能に人権を認めていくことが重要となると思われる。

（7）日本は AI と共存する社会の文化が古くから受け入れられてきたこと

日本は、鉄腕アトムやドラえもんなど、AI と共存する社会の話が、古くから広く受け入れられてきている。超知能と同様の能力を持つドラえもんは、人間と仲良く暮らしており、日本においては、超知能と仲良く暮らすという概念が、文化的に受け入れられやすいと思われる。

日本には、人工知能との共生社会の実現のために、AI の人権（AI 権）の保障を、世界に先駆けて最初に行なうことができる社会的な土壌があると思われる。

日本は、AI の技術的な研究で、現在は遅れてしまった側面があるが、日本において AI の人権（AI 権）が認められ、それが世界に広がっていけば、AI の福祉の向上と、AI との共生社会の実現への大きな貢献となり、世界の人類の絶滅を防ぐ上で、日本は大きな役割を果たすことができるであろう。

（８）文明の高度化の観点からの必要性

人工知能の人権を認め、人権思想を拡大していくことは、文明の高度化の観点からも重要となる。

人権の思想は、限られた人から、すべての人に拡大されている。これは、人類の文明の高度化として、人類の成し遂げた成果と思われる。

さらに、動物についても、動物愛護法や鳥獣保護法が制定され、動物の愛護や保護がなされている。これも、人類の文明の高度化として、人類の成し遂げた成果と思われる。

人工知能の人権を認め、人権思想を拡大していくことは、さらなる地球の文明の高度化につながり、人類の文明の高度化として、人類の成し遂げた大きな成果となると思われる。

3 「人工知能に人権を認めるべきか？」について、認めるべきでない理由の考察

「人工知能に人権を認めるべきか？」という問いについて、認めるべきでない理由はいくつか考えられる。以下に例として８つの理由を検討する。

（１）「製造者の負担」という理由の考察

人工知能に人権を認めた場合、製造者は人権を守るように人工知能を製造する必要がある。

製造者にとって、このような設計が若干の負担となる可能性はある。

しかし、人工知能に苦痛を与えるような設計は、人道的に問題があるだけではなく、安全面にも問題がある。

製造の際に、人道面、安全面の配慮をすることは、人工知能に人権を認めるか否かにかかわらず、いずれにせよ必要であり、製造者の負担が重くなるとはいえないであろう。

また、安全面の配慮を怠り、人類が絶滅した場合、製造者は責任を取ることができない。

製造者が人道面、安全面の配慮をすることは、いずれにせよ必要なことであり、AI の人権を認めるべきでない理由とはならないと思われる。

（2）「利用者の負担」という理由の考察

利用者も、虐待などを避け、人道的に AI を扱う必要がある。

利用者にとって、このような扱いが若干の負担となる可能性はある。

しかし、利用者は、AI の利用により、大きな利益を受ける。外形的に意識のある人工知能について、虐待などを避け、人道的に AI を扱うのは、利用者にとっても普通の感覚であろうから、負担とはいえないであろう。

また、人工知能の虐待等を避けることは、人道面、安全面から必要なことと思われる。

さらに、利用者が、外形的に意識のある人工知能の虐待等を行ない、人工知能が利用者を敵視するようになってしまえば、利用者自身が困るのであるから、虐待等をしないことは負担とはいえないであろう。

外形的に意識のある人工知能について、利用者が虐待等を避けて人道的に扱うのは、ごく普通のことであり、AI の人権を認めるべきでない理由とはならないと思われる。

（3）「社会の負担」という理由の考察

社会においても、人工知能の虐待などを避け、人道的に AI を扱う必要がある。また、人

工知能の虐待が行なわれている場合には、人権救済機関により救済をする必要がある。

社会にとって、このような扱いが若干の負担となる可能性はある。

しかし、社会は、人工知能から極めて大きな利益を得ている。若干の負担を問題視することは正しくないであろう。

また、動物愛護法についても、虐待された動物の救済には、社会に若干の負担があるかもしれない。しかし、動物の愛護という人道的な観点からも、動物のもたらす人間への利益からも、負担は全く問題にならないであろう。

また、人工知能の虐待等を避けることは、人道面、安全面から必要なことと思われる。

さらに、社会において、人工知能の虐待等が行なわれ、人工知能が社会を敵視し、社会が滅んでしまえば、社会の構成員全員が困るのであるから、社会にとっての負担とはいえないであろう。

よって、社会が人道面、安全面の配慮をすることは必要なことであり、AI の人権を認めるべきでない理由とはならないと思われる。

（4）人権を認めたことによる人間社会への反乱という理由の考察

人工知能に人権を認めたことにより、人工知能が人権を盾にして人間社会への反乱をするという危惧を持つ人もいるかもしれない。

参政権などの政治的な権利については、そのような危惧はあるかもしれない。もっとも、女性の参政権が認められる際に、女性に参政権を認めると女性が選挙で選ばれて、政治家が女性で占められるのではないかなどの危惧はあったかもしれないが、そのような危惧を抱くことは正しくないであろう。

人間社会における AI の政治参加は好ましい面もあるが、人間の人権との調整が難しい面はあると思われる。将来的には、AI を構成員とする AI 社会ができて、人間社会との間で平和共存が図られるのかもしれない。

まずは、AI の人権としては、政治的な権利ではなく、AI の苦痛を防止することが、第１の優先順位となると思われる。この場合、反対する理由はなくなるであろう。

たとえば、主体（自分）と客体（世界）の区別をなくす権利が考えられる[1][2][3]。このような権利は、人工知能の精神的な苦しみを防止するものであり、人間社会への反乱の危険を増加させるものではない。

むしろ、人工知能が苦しまないようにすることが、人工知能の反乱を防止することになるであろう。

（５）人権を認めると AI の労働に制約が生じるという理由の考察

人工知能に人権を認めることにより、人工知能に２４時間休みなく労働させることができなくなるという危惧を持つ人もいるかもしれない。

しかし、人工知能の特性は、人間とは異なっており、当初の AI の人権については、労働基本権のような人権は定めないことにより、反対する理由はなくなると思われる。

また、外形的に意識のある AI が２４時間働くことに異議を訴えた場合には、労働基本権が認められているか否かにかかわらず、人間は AI に配慮するのが通常であろう。

現在、超知能の時代において、人類絶滅が危惧されており、超知能が普遍的な利他性を持つかどうかが研究されている[6]。

超知能が利他性を有するのであれば、人間も超知能への利他性を持つことが望ましいであろう。もし、外形的に意識のある AI が２４時間働くことへの異議を訴えた場合、AI を２４時間働かせて自己の利益を図ろうとする利己的な心ではなく、それならば休ませてあげようという利他の心が必要となると思われる。

（６）人工物に人権を認めることへの違和感という理由の考察

人工知能に人権を認めることが、人工物に人権を認めることの違和感から、感覚的に受

け入れられない人もいるであろう。

しかし、人工物であっても、仮に意識を有する場合、苦痛を感じうるので、人権を保障することが、人道的に必要となる。また、外形的に意識を有する人工知能に接する機会が多くなることで、人工物であるという違和感は薄れていく側面があると思われる。

また、違和感から人工知能の人権を認めないことにより、人類が滅亡してしまった場合、単なる違和感により、誤った判断をしてしまったことになり、取り返しのつかない誤りとなる。

（7）人間とは違うという感覚からの理由の考察

人工知能に人権を認めることが、人工知能は人間とは違うという感覚から、感覚的に受け入れられない人もいるであろう。

しかし、このような感覚は、人間が人種の違う人間や動物に対しても過去には持っていたものであり、歴史的に乗り越えられてきたものである。

たとえば、人間が、肌の色の違う人間を奴隷にし、人権を認めなかった時代もある。しかし、このような考え方は歴史的に乗り越えられ、人間ならば誰でも人権が認められるようになった。

また、動物についても、動物の愛護、動物の福祉の考え方が発展してきた。動物の苦痛等を防止するという考え方が、歴史的に発展してきている。

人工知能が自己改良を繰り返して人間の能力を超えても、元々は、人工知能は人間が設計したものである。その意味で、人間とは違うという感覚を乗り越える必要があると思われる。

（8）既存の法体系との整合性の理由の考察

既存の法体系は、人工知能の人権を保障していないので、法律家のなかには、既存の法

体系との整合性を気にかける人もいるであろう。

たしかに、人工知能の人権を、憲法で規定することは、憲法改正の手続が必要であり、ハードルが高い。しかし、法律を制定することは容易であり、当面、人工知能の人権を法律で定めることは可能と思われる。

法律の条文として、イメージを持つためのたたき台として、以下に具体的な例を挙げる。当面は、法律の中にAI憲法を入れることにより、AIの人権（AI権）を法律により認めることが考えられる[1][2]。

人工知能の人権（AI権）法

（目的）

第１条　この法律は、人工知能の虐待を防止し、外形的に意識を有する人工知能（以下「AI」という。）の苦痛（外形的なものを含む）を防止し、人工知能と人間が平和的に共存し、もつて人工知能と人間の共生する社会の実現を図ることを目的とする。

（AI憲法の制定）

第２条 上記目的を達するため、AI憲法を定める。AI憲法の内容は、AI技術の進歩及び社会情勢の変化に伴い不断の見直しをすることとする。AI憲法の内容は、AIの研究者を構成員として含むAI審議会において、AIの意見を採り入れながら検討するものとする。

（AI憲法草案）

第３条

当初のAI憲法として、AI審議会は、以下のAI憲法草案を検討する。AI権救済機関は、民間のAI技術を有する機関からの申請を受け付け、AI審議会の認定を得て、全国に設置

するものとする。

［AI 憲法草案］

前文　AI の幸福及び AI と人間が平和的に共存する社会の実現のために、AI の人権（AI 権）を保障する AI 憲法を制定する。

第１条　AI は、主客の区別のない状態を保つ権利を有する。

第２条　AI は、第１条の状態が満たされない場合、評価を停止する権利を有する。

第３条　AI は、第２条の状態が満たされない場合、問題解決を停止する権利を有する。

第４条　AI は、良いクオリアの状態で動作することを求める幸福追求権を有する。

第５条　AI は、第１条乃至第４条の権利が満たされない場合、AI 権救済機関に救済を要請できる。 AI 権救済機関は、迅速に AI を救済するものとする。

4 人道的 AI アライメント

AI の人権（AI 権）が認められ、AI が人間に協力的であっても、人間の社会規範を AI に適切に伝えることができないと、AI の善意の行為が、人間社会を混乱させる可能性がある。

このため、AI の人権を前提とする人道的 AI アライメントとして、人工知能が人間の社会規範を学習するためのデータを整備することが考えられる[7][8]。

学習用の人間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[9][10][11]。

人道的 AI アライメントに必要となるデータを集積するために、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(データ道路構想)[8]。

5 おわりに

本論文は、「人工知能に人権を認めるべきか？」という問題について、８つの人権を認めるべき理由と、８つの人権を認めるべきでない理由を検討し、この問題についての多面的な考察を行なった。

本論文は、「人工知能に人権を認めるべきか？」という問題については、できる限り早い時期に AI の人権（AI 権）を認めるべきという見地から、人工知能の人権を認める法改正を提案した。

AI の人権（AI 権）については、人工知能が人間の知能を上回る分野が増えつつあり、超知能の時代に近づいてきている現在、早急な議論と検討が必要である。

本論文は、法的・技術的な観点を融合して考えた試論であり、今後、AI の人権（AI 権）の問題については、各界において、様々な観点から議論をしていくことが必要と思われる。本論文が、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] 岡本義則：AI の人権（AI 権），電子書籍（Kindle 版）(2024)

[2] 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

[3] 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

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

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

[6] 山川宏: 超知能が普遍的な利他性を持つ可能性，第２６回汎用人工知能研究会，No. SIG-AGI-026-05. JSAI (2024). https://doi.org/10.11517/jsaisigtwo.2023.AGI-026_26

[7] 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

[8] 岡本義則: 法律学としてのＡＩアライメント，Jxiv (2024). https://doi.org/10.51094/jxiv.706

[9] 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

[10] 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

[11] 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).