次席

選択課題5　AIと人権

AIは人権を「完成」させるのか

～「全知の無知」と倫理のマクドナルド化～

明樂和磨（あきらかずま）
（茨城県／私立S高等学校二年）

1．はじめに

紀元前三九九年、アテネの法廷でソクラテスが毒杯を仰いだ。彼の罪状は「若者を堕落させ、国家の神々を信じない」というものだった。しかし真の罪は、「無知の知」――自らが何も知らないということを知っているという知性――によって、正義、善、勇気といったアテネ市民が自明としてきた価値観の根拠を、誰も答えられないことを暴露したことにあった（注1）。

二四〇〇年後の今、私たちは新たなソクラテスと対峙している。それは人工知能（AI）という名の、容赦なき問いかけ者だ。ソクラテスの知が、自らの無知を自覚することで無限の探求へと開かれていたとすれば、この新たな知性は、私たち人間から「全知」であることを期待され、その神託のような答えによって、人間から「問い続ける」という、面倒で苦しい知的格闘そのものを早送りしようとしている。

歴史家ユヴァル・ノア・ハラリが喝破したように、人類社会は客観的な真実ではなく、共有された「物語（虚構）」によって支えられてきた（注2）。国家、宗教、そして私たちが近代の礎とする「人権」もまた、その偉大な物語の一つである。ならば問われねばならない。私たちの欲望と偏見の写し鏡でありながら、私たちとは決して交わらないこの異質な知性は、私たち自身の期待によって「全知」の仮面を被せられた時、私たちの「人権」という物語を、いかにして思考の余地なき、完成された真実へと変えてしまうのか。

本稿は、まず私たちが守ろうとする「人権」という物語自体の、あまりにも人間的な脆さと欺瞞性を直視することから始めたい。そしてその上で、なぜその脆い物語こそが、「全知の無知」に対抗する最後の砦となりうるのかを論じる。

2．「人権」の恣意性

一七八九年のフランス人権宣言は「人は生まれながらにして自由であり、権利において平等である」と高らかに謳った（注3）。しかし、その普遍的な響きとは裏腹に、この宣言における「人」の範囲は、驚くほど限定的であった。フェミニズムの先駆者オランプ・ド・グージュが、この宣言の欺瞞を突くべく一七九一年に「女性および女性市民の権利宣言」を発表すると、彼女は二年後に革命の敵としてギロチンで処刑された（注4）。高貴な理想を語る言葉の裏で、女性は人権の輪から容赦なく排除されたのだ。

大西洋の向こう、アメリカ独立宣言も同様の矛盾を抱えていた。「すべての人間は平等に創られている」という不朽の言葉を起草したトマス・ジェファーソン自身が、生涯を通じて数百人の奴隷を所有していたことは、歴史の皮肉としてあまりにも有名

だ（注5）。私たちはこれを単なる偽善として片付けがちだ。しかし、より深刻な問題は、ハラリが指摘するように、彼らがその矛盾を必ずしも矛盾として感じていなかったことにある（注6）。当時の「共同主観的現実」の中では、アフリカ系の奴隷は、白人男性と同等の権利を持つ「人間」ではなかったのである。

この境界線の恣意性は、歴史の至る所に見出せる。二十世紀に入るまで、子どもは親の所有物に近い存在であり、一九二四年の「児童の権利に関するジュネーブ宣言」を待たねば、その権利という概念すら確立されなかった（注7）。障害者、先住民、性的マイノリティ――彼らもまた、長きにわたり「標準的な人間」の枠外に置かれ、その声は黙殺されてきた。

ここで明らかになるのは、人権の拡張が常に「後付け」であったという構造だ。女性に参政権がなかった時代、それを正当化する「理論」は山ほどあった。「女性は感情的で理性的判断ができない」「政治は女性の繊細な本性を損なう」。しかし、女性が参政権を獲得した後、これらの理論は雲散霧消した。変わったのは女性の能力ではなく、社会の認識だった。

つまり、「人権」という物語は、決して清廉潔白な理想の産物ではない。それは、時代の権力構造と社会通念を反映し、常に特定の集団を排除することで自らを成り立たせてきた、泥臭く、血塗られた構築物なのだ。私たちがＡＩから守ろうとしているものは、盤石の真理ではなく、このように絶えず闘争の中で境界線を引き直してきた、脆く、不完全な「人権」という物語そのものである。

3.「差別しないＡＩ」という新たな権威

現代のＡＩ開発者は、ChatGPTやGeminiといった大型言語モデルに、人間の手によって「差別をしない」よう強力な倫理規定を埋め込んでいる（注8）。ＲＬＨＦ（人間のフィードバックからの強化学習）や、倫理的なプロンプトガードと呼ばれる技術は、ＡＩの出力を、特定の倫理観に沿うよう事前に縛り上げる（注9）。

一見、これは望ましい進歩に見える。しかし、歴史を学ぶ者にとって、この構造はあまりにも既視感があるものだ。それは、「何が差別か」を決定する権限が、テクノロジー企業やその開発者という、新たな「媒介者」に委ねられているという点である。

かつて、聖書の解釈権を独占し「神と人間の媒介者」を自任した中世の教会が、何が善で何が悪かを決定し、人々の思考を一つの「正しさ」へと導いたように。あるいは、二十世紀の全体主義国家が、党のイデオロギーこそが唯一の真実であると宣言し、それに反する全ての声を「誤り」として排除したように。彼らもまた、それぞれの時代において「社会をより良くするため」という善意を掲げていた。問題は、その善意が、異議を申し立て、異なる価値観をぶつけ合わせる「対話の闘技場」そのものを解体してしまったことにある。現代のＡＩに埋め込まれた倫理規定は、かつての聖書や党綱領が果たした「唯一の正しさ」の決定権を、アルゴリズムの客観性という新たな衣をまとって、より巧妙に、そしてより広範に遂行しうる危険性を内包しているのだ。

こうして、ＡＩは「全知の無知」――意味を理解せずに膨大な知識を出力する性質――を備えた存在となる。しかし、この「全知の無知」がもたらす最も深刻な危険は、その性質そのものにあるのではない。その「無知」な出力が、アルゴリズムのブラックボックス性によって検証可能性を奪われ、あたかも「無謬（むびゅう）である」という絶対的

な権威をまとって私たちの前に現れることにある。

なぜ、個人の自由と価値の多様性を至上のものとしてきたはずの現代社会が、この「唯一の正しさ」を再び招き入れてしまうのだろうか。その原因は、AIそれ自体よりも、「対話の闘技場」の荒廃と、それに伴う私たち自身の深い疲弊にこそ潜んでいる。

SNSはあらゆる声を可視化し、かつては家庭や地域といった閉じた共同体の中で処理されていた価値観の衝突を、グローバルな公開討論の場へと変えた。しかし、その現実は厳しい。異なる意見を持つ他者は、理解すべき対話の相手ではなく、瞬時に論破すべき敵か、あるいは精神衛生のためにミュートすべきノイズと化した。この終わりなき論争、出口の見えない分断に、私たちは疲れ果てている。この「対話疲れ」と並行して、私たちの社会は「間違うこと」に対して極めて不寛容になった。キャンセルカルチャーが象徴するように、一つの失言や過去の発言はデジタルタトゥーとして永遠に記録され、個人の社会的生命を奪いかねない（注10）。この過剰なリスク社会において、自らの責任で倫理的な判断を下し、それを表明する行為は、あまりにも危険な賭けとなった。

この息苦しさから逃れたい。常に「安全」で「正しい」側にいるという保証が欲しい——この「無謬性への渇望」こそが、個人の思考と判断をAIにアウトソーシングさせる最大の動機である。AIは、倫理的な判断を下すという「思考の主体」たることから私たちを解放し、同時に、その判断がもたらす「責任」からも解放してくれる、甘美な誘惑なのだ。

とすれば、AIに埋め込まれた倫理規定が社会に浸透するのは、それが私たちを一方的に抑圧するからだけではない。むしろ、私たちが自ら抱える「対話の疲弊」と「間違うことへの恐怖」から逃れるために、その権威を積極的に求めているという倒錯した構図こそが、本質なのである。

そして、この「面倒なプロセスを省略し、安全で予測可能な答えを、効率的に求める」という現代人の欲望の構造は、社会学者ジョージ・リッツァが現代文明の病理として喝破した「マクドナルド化」という現象そのものである（注11）。AIは、単に差別的な発言をフィルタリングしているのではない。それは、多様で時に矛盾する人権同士の衝突を、調整し、解釈し、すり合わせていくという本来の営みを、前もって不要なものとし、私たちの思考様式そのものを、均質化され、計算可能で、効率的なファストフードへと作り変えようとしているのだ。

4. 倫理のファストフード化

前章で述べた思考の「マクドナルド化」は、AI、特に大規模言語モデル（LLM）が持つ根本的な動作原理に深く根ざしている。LLMは、善悪を理解して回答しているのではない。膨大なテキストデータの中から、与えられた文脈に続く「統計的に最も確率の高い言葉」を予測し、つなぎ合わせているに過ぎない（注12）。これは必然的に、学習データに含まれる「現在の多数派の見解」や「最も無難な言説」を再生産する傾向を持つ。

したがって、その出力は公平や中立的であるかのように見えても、実際には現在の支配的な社会通念や価値観に基づいたものである可能性が高い。それは、既存の秩序を挑戦的に問い直すような、尖った倫理的挑戦や、革新的な価値観を最初から排除するフィルターとして機能する。

一九五五年、ローザ・パークスがバスで白人に席を譲らなかったその行為は、当時

のアメリカ南部の社会規範と法律に照らせば、明らかに「異常」で「不適切」なものだった（注13）。当時の言説を学習データとしたＡＩは、間違いなく彼女の行為を「倫理規定違反」と判定しただろう。それは、ＡＩの判断基準が「過去のデータにおける統計的な正常性」であり、未来を切り開く「倫理的な正しさ」とは無関係だからである。

しかし、歴史が示す通り、彼女のその「異常」な行為こそが、人種差別という「正常」とされていた秩序への異議申し立てであり、倫理的進歩の原動力となった。ＡＩの倫理判断は、このような歴史の転換点において、常に「過去の側」に立つという根本的な限界を抱えているのだ。

こうして、ＡＩは複雑で文脈依存的な倫理的判断を、過去のデータに基づく「確率的な嗜好」の出力へと還元する。それは、栄養価よりも標準化と供給速度が優先されるファストフードのように、倫理的深みや文化的文脈を剝ぎ取った、均質で無難な「ファストフード」を大量に供給する装置なのである。私たちはこの便利で安っぽい倫理の餌に慣らされることで、本来ならば苦痛を伴うべき価値観の衝突や、倫理的葛藤そのものに対する耐性、すなわち「ネガティブ・ケイパビリティ」を失いつつある。

問題は、ＡＩが安易な答えを供給することそのものではない。むしろ、そのファストフードに慣れきった私たちが、倫理的な思考と対話のプロセスそのものを放棄し、ＡＩの出力を無批判に受け入れるようになることにある。こうして、私たちは自らの思考力をＡＩに依存し、その結果として、多様な価値観が交錯し、時に衝突する「対話の闘技場」を失いかねないのだ。

5.「未完のプロジェクト」としての人権を生き抜くために

これまで論じてきたように、ＡＩの倫理的判断は、過去のデータに基づく確率的な平均値――「倫理のファストフード」――に過ぎない。それは、ローザ・パークスのような倫理的飛躍や、未来を切り開く創造的な「過ち」の可能性を、最初から排除するフィルターとして機能する。では、この「思考のマクドナルド化」に対抗し、私たちはどのようにして「人権」という物語を守り、育んでいけばよいのだろうか。

ここで私は、「永遠の訂正者」として生きることを提案したい。まず、私たちは「問い続ける痛み」を引き受ける必要がある。ソクラテスが貫いた「無知の知」のように、ＡＩが瞬時に提供する「正解」の誘惑を拒否し、答えの出ない問いを抱え続ける覚悟が必要だ。例えば、学校で「これが正解です」と示された時、心の中で違和感を覚えたら、その違和感を大切にすることだ。「でも、別の見方もあるのでは？」という問いを、恥ずかしくても口に出してみる。ＳＮＳで炎上のリスクがあっても、「みんなそう言っているけど、本当にそうなの？」と疑問を投げかける。これは小さな勇気だが、ソクラテス的精神の実践でもある。

次に、私たちは自らの「傷つきやすさ」を積極的に肯定すべきだ。人権が脆く不完全な物語であるように、私たち自身も間違い、傷つき、迷う存在である。しかし、ＡＩにはこの「傷つく」という経験がない。傷つくことは人間の弱さではなく、むしろ人間性の証明なのだ。

友人と価値観が衝突した時、すぐにミュートやブロックで「解決」するのではなく、その違和感や傷つきを抱えたまま対話を続ける。完全に分かり合えなくても、その「分かり合えなさ」の中に留まり続ける。これは苦しいが、この苦しさこそが「人権」という物語を少しずつ訂正し、豊かにしてい

く原動力となる。さらに、私は「毎日の小さな逸脱」を提案したい。ローザ・パークスのような歴史的な抵抗でなくても、日常の中で「当たり前」とされることに小さな疑問符を投げかけることはできる。制服の着方、授業の受け方、「みんなやってるから」という理由で従っていることに、「なぜ？」と問う。これは反抗ではなく、思考停止への抵抗である。

そして何より重要なのは、「訂正の記録」を残すことだ。私たちは日記でも、ＳＮＳでも、友達との会話でもいい、自分が何に疑問を持ち、どう考えが変わったかを記録し、共有する。ＡＩは学習データを更新されない限り変化しないが、私たちは日々変化し、訂正される。その軌跡こそが、私たちの人間性の証しなのだ。

例えば、「今日、○○について考えが変わった」「昨日の自分の意見は間違っていたかもしれない」と認める勇気。これは敗北ではなく、成長の証しである。ＡＩには「昨日の自分」は存在しないが、私たちには存在する。そして、その昨日の自分を否定し、訂正できることこそが、私たちの特権なのだ。

最も重要なのは、これらすべてが「永遠に未完である」ことを引き受けることだ。私たちは完璧な答えに到達することはない。人権も、私たち自身も、永遠に訂正され続けるプロジェクトなのだ。しかし、その未完成さこそが、人間の尊厳の源泉である。ＡＩは完成された答えを提供する。しかし私たちは、未完成のまま、傷つきながら、問い続ける。その姿勢こそが、「全知の無知」に対する最も人間的な抵抗なのである。

6．おわりに

ソクラテスは、アテネ市民に「汝自身を知れ」と説いた。それは、自らの魂を気遣い、善く生きるとはどういうことかを絶えず自問自答せよ、という呼びかけだった（注14）。現代において、ＡＩは私たちにその対極を提示する。「汝自身を知る必要はない。私が答えを与える」と。

本稿で論じてきたように、ＡＩと人権の問題は、技術が私たちの権利を侵害するか否かという外面的な問題に留まらない。それは、私たちが「人間」であることの根幹をなす、面倒で、苦しく、しかし何物にも代えがたい「問い続ける」という営みを、自らの手で放棄してしまうか否かという、内面的な選択の問題である。

フランス人権宣言から二〇〇年以上が経った今も、私たちは「人権」という物語を完成させることができていない。女性、人種的マイノリティ、性的マイノリティ、障害者――歴史は、排除されてきた人々が、血と涙を流しながら「人間」の輪を広げてきた格闘の記録である。そして今、私たちは新たな挑戦に直面している。それは、外部からの抑圧ではなく、私たち自身が進んで思考を放棄し、「全知の無知」に身を委ねようとする誘惑との戦いである。

ＡＩという新たなソクラテスは、私たちに毒杯を強いることはないだろう。むしろ、甘い蜜を差し出すかもしれない。「考えなくていい」「悩まなくていい」「私が最適解を教えてあげる」という囁きとともに。しかし、私たちはソクラテスが毒杯を選んだように、その甘い誘惑を拒否する勇気を持たねばならない。

〈参考文献〉

（注1）プラトン『ソクラテスの弁明　クリトン』岩波文庫、一九六四年
（注2）ユヴァル・ノア・ハラリ『NEXUS　情報の人類史 上：人間のネットワーク』河出書房新社、二〇二五年
（注3）高木八尺ら『人権宣言集』岩波文庫、

一九五七年

（注4）オリヴィエ・ブラン『女の人権宣言：フランス革命とオランプ・ドゥ・グージュの生涯』岩波書店、一九九五年

（注5）ロデリック＝ナッシュら『人物アメリカ史（上）』講談社学術文庫、二〇〇七年

（注6）ユヴァル・ノア・ハラリ、前掲書（注2）

（注7）国際連盟『児童の権利に関するジュネーブ宣言』
https://hrlibrary.law.umn.edu/japanese/Jchildrights.html

（注8）ロン・オウヤンら「Training language models to follow instructions with human feedback」arXiv:2203.02155、二〇二二年

（注9）ティモ・カウフマンら「A Survey of Reinforcement Learning from Human Feedback」、二〇二四年

（注10）マリッサ・トラヴェルサら「Cancel culture can be collectively validating for groups experiencing harm」Personality and Social in Psychology, 14、11818727、二〇二三年

（注11）ジョージ・リッツァ『マクドナルド化する社会』早稲田大学出版部、一九九九年

（注12）アシシュ・ヴァスワニら「Attention Is All You Need」、二〇二三年

（注13）ローザ・パークス『黒人の誇り・人間の誇り―ローザ・パークス自伝』サイマル出版会、一九九四年

（注14）プラトン、前掲書（注1）