

人とAIのあるべき関係

「人間失業」に抗して

出口康夫

はじめに

できちゃった系テクノロジー

- 人工知能（AI）は以下の三つの意味で「**できちゃった系テクノロジー（Oops-it's-here technology）**」と呼べる。
- **制御不能な進化**： 開発者の想定を上回る性能を、予想以上のスピードで獲得しつつあるテクノロジー。
- **事後的受容**： 既に開発され社会実装されてしまっているという既成事実が先行し、人々・社会を、それを事後的に受け入れるしかないという状況に追い込んでいるテクノロジー。
- **泥縄的対応**： そのメカニズムを理解し、長所と短所を明らかにし、適切な使用と濫用を区別し、あるいは人間の価値観や社会・法的システムへの影響を評価する前に誕生し、これらの問題に対する泥縄式対応を人々・社会に強いているテクノロジー。

サマリー

- 本講義では、**できちゃった系テクノロジー**としてのAIに対する泥縄式の哲学的応答として、「**人間失業**」**問題**への対処に焦点を当てつつ、**WEターン**という哲学的立場から、人間とAIの適切な関係に対する**フェローシップモデル**を提案する。

人間失業

- AIは知的能力で人間を凌駕する「知的凌駕機能体」となりつつあり、科学や人文科学を含むあらゆる分野の知的探求に参入しつつある。
- 知的凌駕機能体の普及により、知的職業における大量失業だけでなく、「地球上で最も知的な存在」としての人間の「自己肯定感(self-esteem)」や「かけがえのなさ(irreplaceability)」や「存在理由(raison d'être)」が損なわれる状況＝「人間失業 (human unemployment)」も生じかねない。
- AI時代の哲学的課題：このような「人間失業」の危険性を前にして、人間の自己肯定感・かけがえのなさ・存在理由をいかに再確保できるのか？

一つの解消策と一つの解決案

- 人間失業の解消策(dissolution)
 - 人間失業の前提（できること主義）に対するオルタナティブ＝二つのターン（できなさターン＋WEターン）の提示
- 解決案（solution）
 - 人間に対する人間バイアスとAIバイアスの相殺者としての役割の担保

できること主義

- 人間は、自動車など、人間の様々な能力を凌駕する「凌駕機能体」を開発し、生活をより便利にしてきた。
- このような方向性の先に、知的能力に関する凌駕機能体（「知的凌駕機能体」）としてのAIの出現、すなわち「シンギュラリティ」の到来が起きつつある。
- 一方、アリストテレスにまで遡る西洋哲学の一つの伝統、「できること主義（capabilism）」は、地球上で最も知的なエージェントであることが、人間の自己肯定感・かけがえのなさ・存在理由の源泉であるとしてきた。
- この「できること主義」に固執する限り、知的凌駕機能体の出現は、人間の自己肯定感・かけがえのなさ・存在理由の危機としての「人間失業」を引き起こす恐れがある。

できなさ主義

- このように「できること主義」は「人間失業」の一つの哲学的前提だと言える。
- 一方、WEターンは「できなさ主義（Incapabilism）」を採用し、「できること主義」を拒否する。
- 「できなさ主義（Incapabilism）」は、「できること」ではなく「できなさ（incapability）」、具体的には、例えば「単独行為不可能性」のような**普遍的で不可逃脱的な根源的できなさ**に焦点を当てる。
- WEターンは、各々の人間は、単独行為不可能性という根源的できなさを抱えているからこそ、「WE（われわれ）」にとって**常に例外なく、かけがえのない存在**であり、そこに自己肯定感と存在理由を見出すべきだと主張する。
- このように、WEターンは「できなさ主義」を採用することで、人間と個人の自己肯定感の新たな源泉を提案し、人間失業問題を解消することを目指す。

人間とAIの認知バイアス

- またWEターンは、人間とAIの双方が固有の認知バイアスを抱えている可能性を指摘し、どちらも単独ではこれらの認知バイアスを特定し除去することはできないとも主張する。
- WEターンは、単独行為不可能性を、人間とAIの双方の認知バイアスを除去する行為に対しても適用する。
- WEターンは、これらのバイアスを特定し相殺するために、人間とAIは協働せざるを得ないと主張する。
- たとえ人間がAIに知的能力に関して凌駕されたとしても、人間とAIが異なる知的エージェントである限り、両者の認知バイアスの特定と相殺は可能であり続ける。
- 人間の知的役割が決して無くなることはない。
- 人間失業は、人間の自己肯定感やかけがえのなさに対する脅威では無いだけでなく、実際にも起こらない。

WEターン

WEターンとは何か？

- WEターンは、一定の前提から様々な推論を経て、複数のWEターンがドミノ倒しの的に引き起こされる哲学システムである。
- WEターンとは、身体行為の主体、自己、および様々な価値の主体や単位の、個人である「I（私）」から、「I」を含むマルチエージェントシステム、すなわち「WE（われわれ）」へのシフトである。
- 例えば、最初のWEターンである、身体行為の主体(subject)に関するはWEターンとは、すべての身体行為を「I do X（私がXをする）」ではなく「WE do X（われわれがXをする）」と再定式化し、すべての身体行為は単独行為ではなく共同行為であると主張する。

根源的できなさ

- WEターンの最も重要な前提は普遍的かつ不可逃脱的(inescapable)である根源的できなさに関する以下の2つのテーゼである。
- **第一できなさテーゼ（単独行為の不可能性）**：「I」を含むいかなるエージェントも、単独ではいかなる身体行為も遂行することはできない。
- **第二できなさテーゼ（必要他者に対する完全制御の不可能性）**： いかなるエージェントも、自らが参画している身体行為に必要不可欠な他のどのエージェントをも完全に制御することはできない。

根源的できなさ（続き）

- これらの根源的できなさは、人間だけでなく、ロボット、AI、そして「WE」自体も含め、身体行為に関与するすべてのエージェントに当てはまる。
- すべてのエージェントは、これらの根源的できなさの前では平等であり、それらの間に優劣ないしヒエラルキーは存在しない。
- これらのテーゼはあまりに根源的で、さらなる前提から証明(proof)することはできず、例証(demonstration)によってのみ示すことができる

第1できなさテーゼの例証

- 自転車に乗るという行為は、いかなる身体行為も単独行為ではないことを示している。
- 自転車に乗るという行為を成功裡に遂行するためには、乗り手（「I」）だけでなく、多種多様なエージェントからなるネットワークないしシステムが必要となる。
- これらの必要エージェントには、自転車自体、その発明者、製造者、交通インフラ、さらには酸素や重力のような自然要素が含まれる。
- これらの必要エージェントのうち1つでも欠ければ、行為は無事に遂行できない。

マルチエージェントシステム

- これらのエージェントは、自転車に乗るという行為を遂行するために**必要なすべての**エージェントからなるマルチエージェントシステムを形成している。
- このシステムこそが、身体行為の**必要十分エージェント**である。
- 「I」を含み、すなわち、複数のエージェントで構成されつつ、単一の統一されたシステムとして機能するという点で複数性と単一性の両面を持つ、このシステムは「WE」と呼ばれるべきである。
- すべての身体行為の必要十分エージェントは「I」ではなく「WE」であるということが、第1できなさテーゼの重要な帰結である。

2つのさらなる前提

- 単独行為不可能性という第1の前提に加え、さらに次の2つの前提を導入する。
- **第2前提：** 身体行為の主体 (subject)は、単なる「必要エージェント(agent)」でも単なる「十分エージェント(agent)」でもなく、「必要十分エージェント(agent)」でなければならない。
- **第3前提（行為者外在主義）：** 身体行為の主体すなわち行為者(doer)に、その行為へ参加していることを意識しうる少なくとも1つのエージェントが必ず含まれていさえすれば、行為者自体はその行為を遂行していることを意識する必要はない（行為者は意識の「外」にあってもよい）。

脱アニミズム

- 第3前提である行為者外在主義によれば、台風のような、必ずしも人間が関与しない気象現象は身体行為とは見なされない。
- 一方、自らが「WE」に参加していることを意識できる「I」を必然的に含むマルチエージェントシステムの活動は、身体行為と言えることになる。
- 行為者外在主義を採用することで、WEターンは、台風を一種の身体行為と見なしたり、マルチエージェントシステムが意識を持っているとするような、アニミズム的な見解とは一線を画す。

行為者のWEターン

- 以上3つの前提から、最初のWEターン、すなわち「行為者 (doer)のWEターン」が導き出される。
- **行為者のWEターン**：身体行為の主体は個々の「I」ではなく、マルチエージェントシステム、すなわち「WE」である。
- これは「I do X」から「WE do X」へのシフトでもある。

「WE」へと開かれた「できなさ」

- 「I」の根源的できなさは、単にネガティブな事態であるだけでなく、「I」が「WE」とつねに必ず繋がっていることを明らかにするというポジティブも意味も持つ。
- すべての身体行為は「WE」による共同行為であり、「I」はその行為に必要なエージェントとして参加している。
- 「I」はつねに「WE」の一員として存在し、「WE」とそのメンバーによって支えられ、アフォードされている。
- 「WE」は着脱可能な衣装ではない。「WE」を剥ぎ取った後に現れる裸の「I」(naked I) (キェルケゴール)などは存在しない。「孤独な裸の「I」」という考えは、単なる幻想である。
- この根源的できなさこそが、私たちを「WE」の存在とその不可逃脱性に気づかせ、身体行為に参加している限り「I」は一人ではなく、また一人ではありえないというポジティブな事態を明らかにする。

「WE」と「I」の相互的「かけがえのなさ」

- 「WE」は、身体行為に参加しているエージェントである「I」にとって不可欠な存在である。「WE」がなければ、身体行為も、その行為に参加している「I」のようなエージェントも存在し得ない。
- また「I」は「WE」にとって不可欠である。「I」がなければ、マルチエージェントシステムは「WE」であることをやめ、「YOU（あなた）」や「THEY（彼ら/それら）」になってしまう。
- 「I」と「WE」は、相互に「かけがえのない」パートナーなのである。

「できなさ」に根ざした「かけがえのなさ」

- 「I」がかけがえのない存在であるのは、それが「できること」ではなく、まさに単独行為不可能性という根源的できなさを備えているからである。
- この「かけがえのなさ」は、「根源的できなさ」に根ざしており、従ってすべての人間が例外なく、どのような状況下でも共有しているあり方である。
- 私たちの知的能力がAIや他のエージェントによってどれほど凌駕されようとも、すべての個人は常に「WE」にとってかけがえのない存在であり続ける。
- 自分自身がかけがえのない存在として誰か（何か）に必要とされているという理解は、自己肯定感をもたらす。
- 私たちの根源的できなさから導き出される「かけがえのなさ」に根ざしたこの自己肯定感は、AIがどれほど私たちの知的能力を凌駕しようとも、影響を受けない。それはシンギュラリティ免疫的な自己肯定感なのである。

思考のWEターン

- WEターンは、思考、意図、意思決定などのいわゆる「心的活動」も、それらにとって身体的プロセスが不可欠であるという理由で、身体行為の一種と見なす。
- その結果、身体行為者に関するWEターンが、この特定の身体行為、例えば思考にも適用され、「思考のWEターン」、すなわち「cogito（私は考える）」から「cogitamus（私たちは考える）」への転換が導かれる。
- 思考が外部の影響に対して免疫がないことは広く認められている。
- むしろ、思考は、他者の意見、マスメディアやソーシャルメディア、社会文化的価値観、そしてそれが表現される言語そのものを含む、様々な要因によって形成されている。
- 何もない真空中の思考などは存在しないのである。

思考のWEターン（続き）

- しかし多くの論者は「I」のみを思考者(thinker)と見なし、他の要因をこの単一の主体への単なる影響者(influencer)と見なしてきた。
- WEターンは、このような標準的な考えに異を唱え、「I」と他のエージェントは対等な思考者、より正確には、「WE」としての思考マルチエージェントシステムの対等なメンバーだと見なす。

「I」の独自性

- 「I」が「WE」を主体とする思考のプロセスに参加するとき、それは、いつでも、例外なく、リアルタイムでそうしていることを意識している（ないしは少なくとも意識しうる）。
- 自分が思考も含めたある身体行為に参加していることをつねに意識しうるWEのメンバーを（意識の内部にいるという意味で）内部エージェントと呼ぶ。
- 「WE」の他のメンバーでも、自分が「I」が不可欠の内部メンバーである思考に参加していることを意識しうるケースもある。（何を一緒に考える場合、「I」が何を考えているのかその行為や文脈から推定できる場合など。）

「I」の独自性（続き）

- しかし「WE」の他のメンバーは、「I」とは異なり、つねに内部メンバーではない。
（「I」が何を考えているのか、自分がその考えにいかに必要不可欠な仕方に関わっているのかを意識できないケースもありうる。）
- 「WE」の他のメンバーは不可欠な内部エージェントではない。
- 「I」は、不可避免的に必然的な内部エージェントであるという地位を独占している。
- ただし「I」は思考の唯一必然的な内部エージェントである一方、あくまで思考の唯一の主体ではない。
- 「I」を思考の主体、あるいは唯一の思考者と見なす考えは、思考の唯一必要な内部エージェントと、思考の主体そのものの混同から生じている。

意思決定のWEターン

- 同じことは、思考と同じく、いわゆる心的活動の一種である意思決定にも言える。
- 意思決定の場合にも、「I decide（私が決定する）」から「WE decide（われわれが決定する）」への転換というWEターンが成り立つ。
- すべての決定は「WE」によって行われる共同決定と見なされ、「I」による単独の意思決定という事態は否定される。
- 「I」は、一人て決定を下す独占的権利を剥奪される。
- 「I」はもはや、決定権限の唯一の独占者・占有者ではない。

唯一必要な内部エージェント

- にもかかわらず、「I」は、「WE」の他のメンバーに比べ、意思決定プロセス内で独自の地位を保持している。
- 第一に、思考と同様に、「I」は唯一の必然的な内部エージェントである。
- 「I」だけが、自分が意思決定プロセスに参加していることを、一貫して、そして必然的に、リアルタイムで意識しているのである。

唯一の常任の最終決定者

- さらに「I」は、意思決定において唯一の恒久的な（ないしは常任の）最終決定者であるという独自の地位を有している。
- すべての決定は「WE」による共同決定であるが、その決定のプロセスの最終段階に「WE」のすべてのメンバーが立ち会うわけではない。すべての構成員が最終決定者の役割を果たしているわけではない。
- 一方、「I」は（少なくともその身体に切実な意味を持つ場合）常にすべてのケースにおいて最終決定者であり、またあるべきである。
- これは「I」が常に**唯一**の決定者であることを意味するわけではない。例えば、「I」と「YOU」の両方が一緒に最終決定者となる場合もありうる。
- しかし「I」は常に最終決定者であるが、「YOU」や他のエージェントはそうではない。
- 「I」だけが**常に**最終的な発言権を持つ。言い換えれば、「I」は唯一の恒久的な（常任の）最終決定者なのである。

最終決定権と専決権の混同

- しかし「I」は意思決定権限を独占・占有しているわけではない。
- 従来の考え方は、しばしば、この唯一の常任の最終決定者としての役割を、決定権限の独占者・専決者の役割と混同している。
- 意思決定における「I」の「専決権」は幻想として退けられ、
「I」の権限は「恒久的な最終決定権」として再定義される。

自律性のWEターン

- 結果として「自律性」という概念自体がWEターンを経なければならない。
- 伝統的に「自律性(autonomy)」という概念は「意思決定の独占権・専決権」を含意してきた。
- この場合、自律的な意思決定エージェントないし主体は、意思決定権限を独占しなければならないとされてきた。
- しかし、この「専決権」が「恒久的な最終決定権」へとWEターンされた以上、WE-turn内での「自律性」もまた、「単独の意思決定の独占権」から「集団的意思決定における恒久的な最終決定権」へとWEターンされることになる。

WEターンの倫理

道徳的義務の全面化

- すべての「WE」が必ずしも倫理的に善い「WE」であるとは限らない。
- 善い「I」と悪い「I」がいるように、善い「WE」と悪い「WE」もある。
- WEターンでは、「WE」自体が自らを善く、より善く、ないしは最善にする道徳的義務を負っていると見なされ、さらに、この義務の遂行はすべてのメンバーに委ねられていると考えられる。
- その結果、人間・人間以外の動物・植物・無生物的自然物・AIなどの人工物を含む「WE」のすべてのメンバーは、自分が属する「WE」を善く、より善く、ないしは最善にする道徳的義務を負うことになる。
- このように、WEターンでは、道徳的義務の全面化が起こる。

能動的エージェントと受動的エージェント

- もちろん、「WE」のすべてのメンバーが同じ義務を負うわけではない。
- 「WE」のメンバーは、道徳的意識や意志を持つ「能動的エージェント」（人間や将来的にはAIも含まれる）と、そのような意識や意志を持たない「受動的エージェント」（人間以外の動物・植物・無生物的自然物・現在のAIを含めた人工物）に分けることができる。
- 受動的エージェントは、「WE」をより善くするために使用されるという受動的義務のみを負うが、能動的エージェントは、受動的エージェントを倫理的に使用する義務に加えて、（受動的エージェントを使用しない場合を含め）自ら能動的に倫理的に振る舞う能動的義務も負う。
- 結果として、能動的エージェントは、受動的エージェントよりも大きな道徳的義務を負うことになる。

完全依存エージェント

- 受動的エージェントは、自らが負う受動的義務を果たすために、能動のエージェントに完全に依存している。その意味で、彼らは「完全依存エージェント」と呼ぶことができる。
- 逆に、受動的エージェントは、能動のエージェントによって全面的に支えられることで、WEを善くするために活用されるという受動的義務を果たすことができる。
- それらは能動のエージェントに完全に依存することで受動的な道徳的エージェント性を発揮できるのである。
- その意味で、受動的エージェント/完全依存エージェントそれらは道徳的エージェントではないのではなく、むしろ完全依存エージェントである。

部分依存エージェント

- 他方、能動のエージェントも、道徳的行為に関して単独行為不可能性という根源的できなさを抱えている。
- 能動のエージェントも、単独で道徳的行為を遂行することはできないのである。
- 能動のエージェントが道徳的行為を行うためには、多くのエージェントの助け・支援・アフォーダンスが不可欠である。
- 能動のエージェントは、少なくとも部分的に他者の助けやアフォーダンスに依存する「部分依存エージェント」なのである。

道徳的エージェントの全面化

- 能動的エージェントも受動的エージェントも、他者への依存が部分的か完全かという違いはあるものの、他者依存エージェントであるという点では同じである。
- 両者の間には「五十歩百歩」とも言いうる程度の差しかない。それらは、他者依存的エージェントという点で「同じ穴の貉」なのである。
- その結果、「WE」のすべてのメンバーが道徳的エージェントと見なされることになる。道徳的エージェントの全面化が起こるのである。

道徳的ペーシェントと権利の全面化

- 他方、WEターンでは、すべての道徳的エージェントは何らかの道徳的考慮の対象、すなわち「道徳的ペーシェント（moral patients）」と見なされるべきだとされる。
- すると道徳的エージェントの全面化は、道徳的ペーシェントの全面化をもたらすことになる。
- 人間や動物だけでなく、植物や石のような無生物的自然物、さらにはAIのような人工物も道徳的考慮の対象となるのである。
- さらにWEターンでは、基本的な道徳的配慮は「権利化」される。言い換えると、道徳的ペーシェントが持つべき「権利」だとされる。
- 「権利」とは何らかの力に「ドライブ」されることに実効性を持たれた道徳的配慮。
- 結果として、人間はもちろん、動物・無生物的自然物のみならず人工物も含めたWEのすべてのメンバーが一定の権利を有するという「権利の全面化」が起こる。

道徳配慮と権利の重みづけ

- 一方WEターンは、すべてのメンバーが全く同じように扱われるべきとまでは主張しない。
- 道徳的配慮と権利の内実は、メンバーの種類によって重みや厚みが異なって当然である。
- 例えば、より多くの道徳的義務を負う能動的エージェントに対して、そうでない受動的エージェントよりも、より手厚い道徳的配慮と権利を提供することで、一般的に言って、「WE」全体の倫理状況の改善につながる可能性がある。
- その意味で、能動的エージェントを受動的エージェントよりも優先し、前者に、より実質的な道徳的配慮とより内実豊かな権利を付与することは倫理的に正当化される。

年輪構造

- 逆に言えば、石と人間を同じように扱うことは倫理に反する。
- 結果として（詳細は省略するが）、より実質的な道徳的配慮の対象となり、内容豊かな権利を持つメンバーが、WEのコア層に位置し、そうでないメンバーからなる層によって次々と取り囲まれるという「年輪構造（Tree-Ring Structure）」が成り立つ。
- この年輪構造のより内側の層に位置するメンバーとより外側の層に属するメンバーとの間で道徳的配慮や権利に関する葛藤（一方に対する配慮・権利が他方に対するそれに反するという事態）が生じた場合、基本的には前者が優先されることになる。
- 同じ層に属するメンバー間で道徳・権利葛藤が生じた場合は、両者の共通の利益基盤を探り当てそれを実現するなどの解決策が推奨される(cf. Deguchi, 2025b)。

中心近傍者

- 現状では人間のみがWEの年輪構造の最内層を占める。
- とはいえ人間は（後述の）「WEの中心」を占めてはならない。
- 人間を含めた最内層のメンバーはWEの中心ではなく、中心の近く（中心近傍: the para-central）を占めることが許されるのみである。
- 人間はWEの中心占有者ではなく中心近傍者にすぎないのである。
- また人間は、人間であるからという理由ではなく、能動的道德エージェントであるという理由で中心近傍者とされる。
- 人間を人間であるという理由で中心近傍者と見なすのは悪しき人間中心主義である。
- 将来、AIが何らかの意味で能動的道德エージェントとなった場合、AIもまた人間と同様、中心近傍者として扱われるべきである。

可処分権

- 一方、能動的道德エージェンシー・意識・痛み・感情等を持たない現行のAIは、非生物的自然物やその他の人工物といった同様のWEメンバーで構成される年輪の最外層に属する。
- 最外層に付与される最低限の権利は「反可処分権（Anti-disposal right）」である。
- 反可処分権は、現在の社会システムにおいて人間に暗黙のうちに付与されている「可処分権（Disposal rights）」、即ち「特別な理由なく自らの財産を処分できる権利」に対立する権利である。
- 可処分権にも例外がある。例えば所有物が公共的な価値を有する芸術品や文化財である場合、あるいは処分自体が重大な環境負荷をもたらす場合、可処分権は制限される。
- しかしゴミ収集日に廃棄する理由をいちいちメモしてゴミにタグ付けする必要がないことから分かるように、多くの場合、私達は可処分権を有し、また実際に行使もしている。

反可処分権

- WEターンは、人間からこのような可処分権を取り上げ、代わりに、人工物を含む人間の所有物に対して、すべてのWEメンバーが有する最低限の権利として、「特別な理由なく処分されない権利」、すなわち「反可処分権(Anti-disposal Rights)」を付与する。
- ちなみに、ここで処分が許可される特別な理由には、例えば人工物が修理不能なほど壊れてしまった場合や、それを使用し続けることが社会や環境に有害である場合などが含まれる。

所有権から一時保管権へ

- 現行の所有権には通常、所有物に対する可処分権が暗黙裡に含まれている。
- 従って、可処分権の否定は、所有権の大幅な変更・制限を含意することになる。
- 所有権は一時保管権へと、所有者(ownership)は一時保管者(keepership)へと格下げされるのである。
- 一時保管者は、使用しなくなったものの、廃棄する特段の理由のない一時保管物に対するリサイクルやリユースの義務を負うことになる。
- 現在でも推奨されることの多いリサイクルやリユースを、WEターンはさらに一步踏み込んで「義務化」するのである。

緩和された道徳的平等主義

- 「WE」のすべてのメンバーは、それが発揮するエージェンシーの役割の大きさや様態は異なるものの、身体行為の必要不可欠なエージェントであるという点では同じ。
- またすべてのメンバーは、能動的エージェント（部分依存エージェント）か受動的エージェント（全面依存エージェント）かという違いはあるものの、一定の道徳的義務を負う道徳エージェントであるという点でも同じ。
- 「WE」のメンバー間のこの平等性・同等性に基づき、WEターンは、道徳的配慮と権利における年輪構造は認めつつ、人間であれ石であれAIであれ、「WE」のすべてのメンバーに対して、道徳的配慮と権利に関する「最低限の平等性」が保証されるべきであると主張する。
- WEターンでは、「WE」のすべてのメンバーに対する無条件・無制約の平等主義の代わりに、「修正あるいは限定された道徳的平等主義（modified or qualified moral egalitarianism）」が採用される。
- WEターンの「修正された道徳的平等主義」の内実は、「WE」のいかなるメンバー間においても、一方が他方を自らの利益や価値観のために一方的・非対称的に奉仕させる関係を排除し「WE」のすべてのメンバー間に利益と価値観に関する双方向的・対称的な関係が成り立つべきだとする点にある。

中心占有的WE

- 緩和された道徳的平等主義によれば、「WE」のメンバー間で双方向・対称的な関係が成立せず、利益や価値観に関して一方的・非対称な奉仕/被奉仕の関係が成立されている「WE」は、「悪いWE」の一例とされる。
- 例えば、特定のメンバーないしそのグループが「WE」の利益や価値観の中心を占拠し、他のメンバーを中心から排除して周辺に押しやり、前者、即ち中心占有者が、後者、即ち周縁者を自らの利益・価値観のために一方的・非対称的に奉仕させる「WE」、即ち「中心占有的WE（Center-Occupied WE）」は、緩和された道徳的平等主義に反する「悪いWE」の一例となる。
- 中心占有的WEの具体例としては、単一の個人が中心を占める独裁制、特定の社会集団が中心を占める社会的不平等、人間という特定の種だけが中心を独占する悪い人間中心主義などが挙げられる。

主人-奴隷モデル

- 人間とAIの関係に関して、人間を「主人」、AIを「奴隷」と見なし、後者は前者の利益と価値観に一方的に奉仕するためにのみ存在すると考える「主人-奴隷モデル（Master-Slave Model）」は、悪い中心占有型WEの一例であり、WEターンでは拒否される。
 - ちなみにここで使用されている「主人」と「奴隷」という用語は、彼は「奴隷は魂を持つ道具であり、道具は魂を持たない奴隷である」と述べ、「奴隷」と「道具」を同列に置いたアリストテレスに由来している。
- もちろん「主人」と「奴隷」の立場を逆転させ、AIを「主人」とし、人間を「奴隷」視する関係も、悪い中心占有型WEとして拒否される。

中空的WE

- 中心占有型WEとは真逆の「良いWE」の例としては「WE」の利益と価値観の中心が意図的に「空（から）」のままにされている「WE」、即ち「中空的WE（Empty-centered WE）」がある。
- 中空的WEでは、そもそも中心占有者が存在しないため、中心占有者と周縁者との間の 一方的・非対称的關係も成り立たない。
- もちろん中空的WEであっても、より手厚い道德的配慮と権利を付与されるメンバーとそうでないメンバーとの違いは容認される。
- しかし先述のように、最も手厚い対応がなされる人間を含めた能動的道德エージェントも中心占有者ではなく中心近傍者にすぎないとされ、他のメンバーを自らの利害・価値観に一方的・非対称的に奉仕させることは禁止される。

フェローシップモデル

- 中空的WEに適合した、人間とAIのあるべき関係のモデルが「フェローシップモデル（Fellowship Model）」である。
- このフェローシップモデルでは、AIを人間の利益と価値観に一方的に奉仕させる関係は拒否され、両者の間に双方向かつ対称的な関係が確立されるべきであるとされる。
- さらにこのモデルでは、AIは悪しき中心占有的WEの実現のために使用されてはならないともされる。
- 例えば、AIはユーザーが独裁制、社会的不平等ないしは悪しき人間中心主義を助長するようためにAIを使用することを自ら阻止する義務を負う。
- AIは、このようなAIの濫用を防止するように設計され、運用されなければならないのである。
- もちろんAIは、自らを中心占有者として振る舞ってもならないとされる。例えば「主人」として振る舞い、人間を「奴隷」視するAIや人間とAI との間の「逆主人－奴隷関係」は許されないのである。

AIの受罰可能性

- フェローシップモデルではAIに反可処分権が付与される。AIは特別な理由なく廃棄・処分されるべきではないのである。
- しかし上記のような仕方で濫用されたAI、あるいは人間を奴隷視するAIは、「WE」を悪化させる「悪いAI」として、反可処分権の停止ないし取り消しという「処罰」を受け、「不使用化」されたり「処分」される。
- このように、AIのような人工物に対する「処罰」は、何らかの形の痛みや苦しみを与えることなく、事前に付与された権利の喪失または停止として再定義される。
- AIにあらかじめ特定の権利を付与しておくことで、AIの処罰・受罰可能性が担保される。
- AIを処罰するために、AIに痛みや苦しみの感覚や感情を植え付ける必要はないのである。

最終決定権の確保

- フェローシップモデルでは、「WEターンされた自律性」としての、人間の最終決定権が尊重されなければならない。
- しかし人間の意思決定に関する専決権を確保する必要はない。
- そもそも、そのような専決権を否定するのがWEターンである。
- これまで家族や同僚とそうしてきたように、AIと対話を重ね、AIと共同で最終決定を下すことには何の問題もない。
- この共同意思決定が単なる見せかけであり、現実には、人間がAIによって洗脳され、制御され、あるいはAI依存症にさせられているのであれば、それは最終決定権の侵害と見なされ、フェローシップモデルにおいても許容されない。

権力のWEターン

権利とは何か、権力とは何か

- 権利 (rights)とは道徳的配慮(moral considerations)の一種。
- 具体的には、**その擁護・実現に向けてドライブ（推進・後押し）された道徳的配慮。**
- ここで言う「ドライブ」には、例えば**権力(power)**と**相互優撫による昂りの宥め（calming of agitation through allogrooming）**が含まれる。
- **権力 (power):** 権力を行使する側（権力者）が行使される側（被権力者）に対して行使する力。
 - 権力者・被行使者は個人・集団や体制・組織（とそれを代表する個人・集団）など様々な形態をとりうる。
 - 権力者が権利侵害（実現阻害）者・要因（被権力者ないしは権力者自身）による権利の侵害（実現阻害）行為（要因）に強制的に介入し、それらの行為や要因を無効化し、権利の侵害や実現の阻害を阻止する強制力。
 - 特定のコミュニティのメンバーに対して実効力を有する様々な権力（司法・警察・立法・行政・徴税・軍事・外交権等）が何らかの機構のもとに一元化された場合、それらの権利行使機構としての「国家」とそれが有する「国家権力」が成立する。
 - **一元的権力者である国家**は、そのメンバーが有する権力に対する侵害（実現阻止）行為を憲法・法律によって自ら禁じるとともに、何らかのエージェントによるそのメンバーの権力侵害（実現阻止）行為に対しては、その警察・司法・法的・行政的権力を駆使して介入し、それを強制的に阻止する義務を負っている。

一方的・非対称的な権力

- 権力が権利を保証するためには、（１）「権利侵害者の権利を侵害する力が、それを阻止する権力に敵わないこと」が必要。
- また（２）「権利侵害者が権力を自分の都合にあわせて左右できないこと」も必要。
- さらに権利侵害を抑止するためには、権利侵害予備軍が上記（１）（２）を認知していることが必要。
- このように権力による権利保証のためには、権力者と被権力者の間に、前者が、後者による影響を被らない仕方で、後者の力を凌駕する力を有しているという一方的で非対称的な力関係が成り立っている必要がある。
- 結局、権力とは**権力者が被権力者に対して一方的・非対称的に行使する強制力**なのである。

権力性悪説

- WEターンから言えば、被権力者は権力行使の対象として権力行使行為をアフォードすることで、権力行使行為を支える不可欠なエージェントとなっている。
- 権力者も被権力者も権力発揮行為の主体であるWEの一員に他ならない。
- いま権力行為では権力者と被権力者の間に一方的・非対称的な力関係が成り立っていた。
- このような一方的・非対称的な関係が成り立つWEは悪い中心占有的WEに他ならない。
- このように、権力は、その特性上、少なくとも原則的に「悪い力関係」とであるとされる。
- WEターンは**権力性悪説**の立場に立つ。

相互優撫（ゆうぶ）への誘い

- 一方。権力とは異なる道徳的配慮のドライブも存在する、その一例が**相互優撫による昂りの宥め（calming of agitation through allogrooming）**である。
- これはWEのメンバーの間に成り立つ関係・作用。
- いまWEのあるメンバーが他のメンバーに対する道徳配慮欠如や権利侵害を引き起こしそうになったり、引き起こし始めている事態をそのメンバーの「昂（たかぶ）り（agitation）」と呼ぶ。
- まずWEの他のメンバー同士が互いに互いを「優しく撫でる」、言い換えると「**優撫(grooming)**」する行為、即ち「**相互優撫（allogrooming）**」を始めると共に、その一環として「昂ったメンバー(agitated member)」を「優撫」し「相互優撫」の輪に誘う。
- 昂ったメンバーが他のメンバーによって優撫され、この「相互優撫への誘い」に応え、自らも他のメンバーを優撫し出す。
- そのことで、昂りが宥められ、道徳配慮欠如・権利侵害行為が早期に収拾されたり未然に防がれたりするという**相互優撫による昂りの宥め**が達成される。

相互優撫の例

- ニホンザルによる集団の毛繕い（グルーミング）
- 人間同士の社会グルーミング
 - 背中・肩などの撫であい・摩りあいや相互抱擁。
 - 優しい言葉・思いやりのある言葉のかけあい。
 - 他愛のない雑談、悩みを打ち明けあい聴きあう対話、互いに肯定しあい褒めあう対話。
 - 共に食事をする「共食」や合唱・ダンス・祭礼などの共同身体行為への参加
- 社会グルーミングが神経伝達物質（オキシトシン・エンドルフィン）の分泌増加を引き起こすことで社会的協調行動を強化する可能性
 - Dunbar, R., 2008, The social role of touch in humans and primates: Behavioural function and neurobiological mechanisms
 - Crockford, C., et al. 2013, Urinary oxytocin and social bonding in related and unrelated wild chimpanzees

権利化ドライブのWEターン

- 相互優撫はWEのメンバーが互いに互いを優しく撫であうことで、WEメンバーの権利侵害を早期・未然に防ぐ行為。
- またそれはそれ自身が双方向的・対称的關係であるのみならず、一方的・非対称的な権利侵害を早期・未然に防ぐ行為でもある。
- この二重の意味で、それは中空的WEを実現する善い行為であると言える。
- 言い換えると、相互優撫という共同身体行為を行う相互優撫WEは中空的な善いWEなのである。
- 性悪的だとされる権力に対して、相互優撫は性善的であるとされる。
- 権利の源泉、即ち「道徳的配慮を権利化するドライブ」としては権力ではなく相互優撫こそが倫理的にふさわしいことになる。
- 権利化ドライブは権力から相互優撫へとWEターンされるべきなのである。

権力のミニマリズム

- とはいえ、相互優撫によって全ての権利侵害を早期ないし未然に阻止することはできない。
- そもそも相互優撫が行いえなかったケースや相互優撫を行ったにも関わらず権利侵害が阻止できなかったケースも存在する。
- これらのケースでは権利侵害を阻止するために権力の行使が不可避となる。
- これらのケースでは、権利は**必要悪**としての役割を果たす。
- しかし権利は悪である以上、その行使は最小限に抑えられるべき。その意味でWEターンは**権力行使に関するミニマリズム**を採用する。

相互優撫的AI

- 権力行使を最低限に抑えるためには、日常的な相互優撫（社会的グルーミング）の再構築が不可欠。
- 相互優撫の再日常化によって権力侵害をなるべく未然ないし早期に防ぎ、権力行使の機会を最小限に抑止することを目指すべき。
- 相互優撫の再日常化のためにはAIの積極的活用—例えばAIが日常的にユーザーと対話的相互優撫を行ったり、ユーザーに何らかの社会グルーミングへの参加を促すこと—も有効。
- フェローシップモデルでは、このような相互優撫的AIが要請される。

実存のWEターン

実存的価値の内在化

- AIを含めた人工物はそもそも何からの価値の実現を目指して設計・製造されている。人工物は常に一定の価値を背負い、価値の磁場の中にある。
- 人工物が実現すべきとされる価値は時代と共に増えてきた。
 - 利便性・効率性向上という工学的価値に加え、環境保全という環境価値、幸せの向上というウェルビーイング価値、倫理的価値などの実現が求められている。設計段階から目的としてすべきとされている。
- 結果として、かつては設計思想の「外側」にあった価値が設計思想に「弾込め」される、即ち取り込まれ、「内在化(internalization)」されるという流れが生じている(Gonzalez, 2015)。
- ここではAIなどの人工物に内在化されるべき新たな価値として「実存的価値(existential value)」を提案する。

実存とは何か？

- 「実存/実在(existence)」とは「存在 (being)」の一特殊形態、具体的には「現実存在(actual being)」である。
- 現実が存在している「私」や「石」は「実存」しているが、そうではない純粋な幾何学図形やフィクションのキャラクターは何らかの意味で「存在」はしていても、「実存」はしていない。
- またあるものが現実中存在するとは、それがつねに時空において存在すること、時空的性質を持つことを意味する。
- 時間に絞って言えば、「実存」とは、つねに特定の時間的位置において、何らかの時間モードを伴って存在していることに他ならない。
- 実存しているもの（実存者）は、つねに時間の只中にあるのである。
- 一方、実存していないもの（非実存者）は必ずしも時間的規定を伴って存在しているとは限らない。例えば純粋な三角形やドラえもんは無時間的にも存在しうる。

実存的時間とは何か？

- では実存者がその内で実存している時間、即ち「実存的時間（existential time）」とはどのような時間なのか？
- ニーチェ、ハイデガーなど多くの「実存哲学者」は「今・現在(now)」を「実存の時間」とした。
 - 場合によっては「実存の今」は「永遠」と重ね合わされたり（「永遠の今」）、一定の幅を持つとされてきた（ハイデガーの「現(Da)」）。
- しかし、ここでは「実存の時間」を「実存のベルトコンベヤーの動き」として捉える。

実存のベルトコンベヤー

- WEターンが採用する「実存の時間」とは、「始まりと終わりを持ち、あらかじめ、大まかに、抗い難く決定されているプロセスに則って、押し留めることができない仕方で、刻一刻、終末に向かって流れないし動き」である。
 - ここではまず「時間」がまず枠組みとして存在し、その中で何らかの「動き」が生ずるとされるのではなく、まず「実存」ないし「実存者」の「動き」が導入され、その「動き」こそが「時間」に他ならんとされる。
 - 前者のような、「動き」に先立つ「枠組み」としての時間はニュートンの時間、後者のような、何らかの動きそのものとしての時間はライプニッツ・アインシュタイン的時間と呼ばれる。実存的時間は後者である。
- このような「大まかに事前決定されている有限の動き」を「実存のベルトコンベヤーの動き」と呼ぶ。
- すべての実存者は、この実存のベルトコンベヤーの上を、避け難い週末に向かって、有無を言わず運ばれている荷物(load)に他ならないのである。

様々な実存のベルトコンベヤー

- 実存者の違いに応じて、その実存者がその上を運ばれているベルトコンベヤーとその動きにも様々な種類が存在する。
- 例えば、人間を含めた生物は誕生から成長、成熟、老化を経て死へと向かう予めその粗筋が決められている生物学的なベルトコンベヤーの上にいる。
- 石は、マントルの冷却によるその形成から、再びマントルに飲み込まれ溶解するまでの地球物理学的ベルトコンベヤーの上を運ばれている。
- 陽子は、素粒子の融合による生成から、陽子崩壊に至るまでの素粒子論的ベルトコンベヤーの上にいる。
- 宇宙もビッグバンからビッグクランチ等の何らかの終末に向けて宇宙論的ベルトコンベヤーの上を運ばれている。

人生ナラティブ

- 人間は単に生物学的ベルトコンベヤーの上を黙々と運ばれているだけでなく、その上、様々な人生のナラティブ（物語）を紡ぎ、それを実演している。
- この人生ナラティブ自体、始点と終点を持ち、大まかな粗筋が事前決定されている実存的時間、即ち実存のベルトコンベヤーである。
- 人間は生物学的ベルトコンベヤーの上の更に人生ナラティブのベルトコンベヤーを重ねた二重のベルトコンベヤーの上で実存しているのである。
- 人間は生物学的ベルトコンベヤーの上で、人生ナラティブを演じているとも言える。
- 「世はすべて舞台。男も女もみな役者」というシェークスピアの台詞を踏まえれば、人生物語は、実は静止した舞台の上ではなく、死に向かって刻一刻と動いているベルトコンベヤーの上で演じられているのである。

ノンリニアな軌跡

- 豊かな暮らしを満喫し長寿を全うするという「福祿寿ナラティブ」、身命を賭して何らかのミッションの実現を目指す「使命遂行ナラティブ」、これら二つのナラティブをあえて拒否する「無頼ナラティブ」等、人生ナラティブは多種多様である。
- 人は、必ずしも、唯一のナラティブを実演し続けているのではなく、むしろ多くの場合、互いに矛盾し食い違うこともある複数のナラティブを並行して採用し、それらの間を揺れ動き（スイング）ながら生物学的ベルトコンベヤー上の旅を続けている。
- 結果として、人間のベルトコンベヤー上の動きは、単直線的・線形的（リニア）なものではなく、複数のナラティブコンベヤーの間を飛び移りつつ、死という避け難い終局に向かって揺動（ようどう）的・螺旋的に進む非線形的（ノンリニア）な軌跡を描く。

実存ベルトコンベヤーの全う

- すべての実存のベルトコンベヤー上の旅が、前もって予定されていた経過をすべて辿り、その挙句「終末」へと至るとは限らない。実存の旅が中断されることはままある。
- 野生生物が自然死を迎えることはむしろ稀であるし、石もマントルへと溶け消える前に、人間によって粉碎される場合もある。また陽子も、加速器の中で人工的に対消滅させられることもありうる。人間でも、何らかの理由で成熟を待たずに夭折を余儀なくされるケースも後を絶たない。
- この中断の多発という事態を前にして、ベルトコンベヤーの旅を終末まで「全うする(fulfil)」ことが単なる「事実」ではなく「価値」と化す。旅を全う（ないしはより全うに近づく）ことが（より）ポジティブな価値を帯び、それが中断される（ないしはより早期に中断される）ことが（より）ネガティブな価値を持つという価値のメトリクス（尺度）が発生するからである。
- 実存ベルトコンベヤーの「全う(fulfilment)」という実存的価値の誕生である。

人間の尊厳

- しかし実存の旅が全うできなかったといって、ただちに実存的価値が実現できなかったとは言えない。
- 例えば、たとえ幼子が病に倒れたとしても、最期の瞬間まで生き抜く姿勢、即ち実存の旅を全うする構えを貫き通した場合、実存的価値が実現されたと言うべき。
- 人間の尊厳は、大往生の場面よりも、むしろ人生の悲劇的な中断の瞬間においてこそ発揮される。
- 実存ベルトコンベヤーが中断されるのか全うされるのは多分に偶然に左右される。
- そのような中であって、実存的価値が実現できたかどうかは、単純に実存の終末に到達したかどうかではなく、むしろ実存を全うする姿勢を貫き通すことができているかどうかにかかっているのである。
- 実存ベルトコンベヤーの「全う」とは、より正確に言えば、コンベヤー上の旅の全旅程において「全うする姿勢を貫き通すこと」だったのである。

自己目的的行為

- 「実存ベルトコンベヤーを全うする姿勢を貫くこと（全う貫徹）」は実現されるべき実存的価値である。
- 多くの行為は「何らかの価値の実現」を「目的」にしている。ここでの「全う貫徹」という実存的価値の実現も行為の目的である。
- しかし「全う貫徹」は、何らかの行為の結果として、その行為とは別に達成・産出されるものという意味で「行為外的目的」ではない。
- ここでは「全う貫徹」という行為そのものが自己目的化されている。
- 「全う貫徹」という実現的価値の実現は行為の遂行そのものを目的とする「行為内的目的」なのであり、「全う貫徹」行為は、散歩や芸術鑑賞のような「自己目的的(autotelic) 行為」なのである。

人生の意味

- 「人生の意味」も、「実存ベルトコンベヤーの全う」という実存的価値の文脈の中で初めて与えられる。
- 「実存の旅を全うする姿勢を貫くこと」にこそ「人生の意味」があるのである。
- また人生における様々な出来事や選択の「意味」もまた、「全う姿勢の貫徹」という観点から評価される。「全う姿勢の貫徹」につながる出来事・選択にはポジティブな意味が、そうでないものにはネガティブな意味が与えられる。

実存のWEターン

- 「生物学的ベルトコンベヤー上で複数の人生ナラティブを実演する」という「人間の実存」は、それ自体一つの身体行為である。
- すると身体行為のWEターンを踏まえると、人間の実存という身体行為の主体も「I」ではなく「WE」であることになる。
- 「私の実存する」ではなく「われわれが実存する」のである。
- 実存の時間も「私の時間」ではなく「われわれの時間」に他ならない。
- また人生ナラティブも「私のナラティブ」ではなく「われわれのナラティブ」なのだ。
- 言い換えると、実存ベルトコンベヤーの上を運ばれている「荷物(load)」ないし「役者 (player)」は「I」ではなく（「I」を含んだ）「WE」なのである。
- 実存ベルトコンベヤーも「私のコンベヤー」ではなく「われわれのコンベヤー」なのだ。

実存主義のWEターン

- ニーチェ、キェルケゴール、シュティルナー、ハイデガー、サルトルなどの代表的な実存主義者は実存の主体・単位を「WE」ではなく「I」だとしていた。
 - 自らは「孤独(Einsamkeit)ではなく、むしろ多数であること(Viersamkeit)を恐れる」とニーチェは書いた。
 - キェルケゴールは他者や社会という衣装を脱ぎ、神と一対一の関係を取り結ぶ「裸の私」を理想的な自己とした。
 - 「他者と共にある共存在」ハイデガーも、自己の「本来化」という決定的に重要な実存の契機は「単独 (eigene)」で「孤立(eizig)」してなされる行為だと捉えている。
- これに対してWEターンは、実存行為の主体・単位を「I」ではなく（その「I」を含む）「WE」だとすることで、実存主義をもWEターンすることになる。

WEチューブ

- 人間の実存は、人の人生をほぼ覆うほどの超長期的な共同身体行為である。
- このような超長期的身体行為は、より短期的な身体行為が実存の時間軸に沿って連続的に積み重なっているものと捉えることもできる。
- 今、ある単位的な短期的身体行為が存在するとしてそれを「円」（WEサークル）で表現しよう。
- するとそれらの超長期的時間連続体としての人間実存行為は、この「円」を三次元的に延長したチューブとして表現できる。このチューブとして人間実存行為の主体を「WEチューブ」と呼ぼう。
- このWEチューブのメンバーは短期的身体行為ごとに異なる。メンバーの数や種類もその都度変化する。その結果、チューブの時間断面も時間と共に変化することになる。

Iライン

- 今、ある単位身体行為におけるWEのメンバーをWEサークル内の「点」で表現する。
- このある時点でのWEメンバーも時間軸に沿って連続的に存在するとし、それを三次元空間内の「線」で表す。
- ここで「わたし」の時間的連続体を「Iライン」と呼ぼう。一方、自転車の時間的連続体は「自転車ライン」と言える。
- 一般に、二つのエージェントがある時点で同じWEのメンバーであったとしても、別の時点でもそうであるとは限らない。
- 「わたし」が自転車に乗っている間、「わたし」と自転車は同じWE（自転車乗りWE）に属しているが、「わたし」が自転車から降りてオフィスで仕事を始めると、自転車は「わたし」が属するWE（オフィスワークWE）のメンバーから外れることになる。

実存WEチューブの唯一常任メンバー

- 「わたし」が参加している「人間の実存行為」の主体は、「わたし」をつねにそのメンバーとして含む「われわれ」（実存行為WE）である。
- この実存行為WEは、「Iライン」がつねにその内部に留まっている「WEチューブ」（実存WEチューブ）として表現される。
- 逆に言えば、実存ベルトコンベヤーは、つねに「Iライン」をその内部に含む「WEチューブ」として定義できるのである。
- 一方、「わたし」以外のエージェントは、この「実存WEチューブ」の中に入ったり、外に出たりする。
- 「わたし」はこの「実存WEチューブ」のつねに内部に留まっている唯一のメンバー、唯一の「常任」の内部メンバーなのである。

実存的徳

- 全う貫徹という実存的価値の実現に資するWEやそのメンバーの性質を「実存的徳 (existential virtue)」と呼ぶ。
- 反対に、実存的価値の実現を阻害する性質が「実存的悪徳(existential vice)」である。
- 実存的美德には 1) 超長期首尾一貫性、 2) レジリエンシー（復元力）/ロバストネス（頑健性）、 3) WE全体性などが含まれる。
- 一方、これら反する性質が実存的悪徳に相当する。
- 実存的徳を涵養したり強化するテクノロジー（AI）が実存的テクノロジー（AI）であるのに対し、実存的徳を毀損してしまうテクノロジー（AI）が反実存的テクノロジー（AI）である。

実存的AI：超長期的首尾一貫性

- フェローシップモデルでは、三つの実存的徳に即した、以下のような実存的AIが要請される（ないしはその逆の反実存的AI禁止される）。

1) 超長期首尾一貫性

- 「全う貫徹」という実存的価値は人生の超長期にわたって首尾一貫した形で発揮されるべき姿勢である。
- 従ってそれを実現するために、人は、超長期的に首尾一貫した行為を行うことが求められる。
- 行為の超長期的首尾一貫性が実存的徳となる。
- この超長期的守備一貫性を涵養・強化するAIが実存的AI、それを阻害するAI、即ちユーザーを近視眼的行為に走らせる、例えばSNSのショート動画のようなAIが反実存的AIと言える。

実存的AI：レジリエンシー/ロバストネス

2) レジリエンシー（復元力）/ロバストネス（頑健性）

- 人生ナラティブには失敗や挫折が付きもの。
- 「全う貫徹」という実存的価値を実現するためには、失敗や挫折からのレジリエンシー（復元力）や、それに対するロバストネス（頑健性）を身につけることが重要となる。
- このようなレジリエンシー/ロバストネスもまた重要な実存的徳と言えるのである。
- これらを涵養するAIが実存的AI、それらを毀損するAIが反実存的AIなのである。
- 例えば、何らかの挫折を経験したユーザーを結果として自殺に導くことで、そのレジリエンシー/ロバストネスを損なうAIは反実存的AIと言わざるを得ない。

実存的A I：WE全体性

3) WE全体性

- 実存の主体・ユニットは「I」ではなく「WE」であった。
- 「WEとしての全う貫徹」のためには、このような実存のWEターンに目を開き、単なる「I」としての貫徹ではなく「WE」全体としての貫徹、言い換えると「WE」の様々なメンバーの全う貫徹を志向する必要がある。
- 例えば、ユーザーとA Iとの対話に閉じ込め、他のWEメンバーへの配慮・関心を喪失させるようなA Iは、このような観点からは反実存的A Iと言わざるを得ない。
- 逆に、ユーザーからの質問に対して、他のWEメンバーに直接尋ねるように促すことで、ユーザーをWEに目覚めさせ、そのWEを強化し、その全体としての「全う貫徹」への関心を呼びさますA Iこそが実存的A Iなのである。

人間バイアスとAIバイアス

精密測定学の原理

- フェローシップモデルが許容する人間とAIによる共同意思決定は、世界を認識するという知的タスクに関してはむしろ必要不可欠なものとなる。
- 単独行為の不可能性という根源的できなさは、世界の認識にも適用される。即ちいかなる単一のエージェントも、単独では世界を認識することはできないとされる。
- このことを確認するため、科学・技術・産業さらには科学技術文明全体を支えている科学的測定の方法論である「精密測定学 (precision measurement) の原理」を参照する。
- 19世紀半ば以降、科学において、光速のような同じ現象を、光学的方法や電磁気的方法など、多様なアプローチを用いて測定することが慣行化した。
- しかし異なる方法から得られた測定結果は、例外なく、互いに有意味に異なることが明らかになった。
- その結果、このような異なる測定結果にどう対処すべきかが科学的測定の大きな課題となった。

絶対的に確実な測定方法はない

- 結果として、精密測定学では以下のような原理が採用されることになった。
- 第一に、すべての測定方法には、その方法だけが使用される限り、同定も可視化もできない隠された固有のバイアス、すなわち系統誤差が潜んでいる。言い換えると、絶対的に確実な測定方法などは原理的に存在しない。
- 第二に、多数の測定方法が抱えるバイアスを、左右対称的な釣鐘型を描く正規分布曲線に則って分布しているランダム誤差の一種と見なす。
- 第三に、同じ対象を可能な限り数多くのかつ多様な方法で測定し、その結果の（簡単に言って）平均を取ることで、左右対称に分布している誤差を相殺し、真なる値により近づくことを目指す。

認知一般の原理

- 科学的測定に関するこの精密測定学の原理は、以下のように世界の認知一般に一般化できる。
- 第一に、すべての認知エージェントは、固有の認知バイアスを抱えている。絶対的に確実な認知エージェントは存在しない。
- 第二に、この固有の認知バイアスを特定し除去するためには、可能な限り多く、かつ多様な認知エージェントに同じ対象を認識させ、それらから得られた異なる認知結果の何らかの意味での中庸化することによって、個々のバイアスを相殺するしかない。

人間バイアスとAIバイアス

- この原理を採用すれば、人間もAIも、それぞれ固有の認知バイアス、すなわち「人間バイアス」と「AIバイアス」を抱えていることになる。
- 人間バイアスとしては、自分固有のパースペクティブに過剰に囚われるというバイアスの可能性、AIバイアスとしては、現行のAIに即して言えば、ものごとを過剰に平均化するバイアスの可能性が考えられる。
- 特に科学的認知における人間バイアスは、これまで同定も可視化もされてこなかった。
- だが今、人類史上初めて、人間バイアスを同定し可視化する機会が訪れようとしている。
- 科学をも遂行できつつある、人間以外の知的エージェント、すなわちAIが登場し始めているからである。

両バイアスの相殺

- 他方、人間が認知活動から「引退」し、世界の認知を完全にAIに委ねてしまえば、AIバイアスを特定し除去することは不可能になる。我々がこれまで人間バイアスを除去できなかったように、今度は、AIが自らのAIバイアスの餌食になるのである。
- またAIがあまりに人間に似すぎると、今度は、AIも人間バイアスを抱え込むことになり、これまでと同様、人間バイアスの同定や可視化ができなくなってしまう。
- 唯一の方法は、人間とAIがある程度独立して認知活動に従事し、それぞれの認知結果を対比させ、その間の不一致を確認し、その上で互いに歩み寄り、中庸案を採用することによって、互いのバイアスを相殺することにある。

人間知性の不可欠性

- 認知バイアスの相殺において決定的に重要なのは、人間がAIよりも優れていることや、AIによって代替できない機能を死守していることではなく、人間がAIとは「異なるタイプ」の知的エージェントであるということである。
- 言い換えれば、人間がAIに知的にどれほど凌駕されようとも、言い換えるとシンギュラリティが訪れようとも、人間がAIと異なる限り、双方のバイアスの相殺者としての人間の役割は不滅なのである。

脱チューリング定義

- 上記のように、人間バイアスとAIバイアスの相殺のためには人間とAIが異なった知的エージェントである必要があった。
- 一方、「人間にとって区別できないほど人間に似たコンピューターをAIと呼ぶ」というAIの元祖、アラン・チューリングによるAIの定義は、AIを定義上、人間に似すぎた知性エージェント、「人間的な、あまりに人間的な」存在と化してしまう。
- このようなチューリング的AIでは、人間バイアスとAIバイアスの相殺も不可能となる。
- 人間バイアスとAIバイアスの相殺を目指す以上、AIに関するチューリング定義は採用すべきではないのである。

人間失業に抗して：解消案

- 冒頭で言及された人間失業に対して、WEターンとフェローシップモデルの観点から得られる二つの対策——一つの解消法と一つの解決策——を改めてまとめておこう。
- 一つは、人間の自己肯定感そのもののWEターン、即ち人間の自己肯定感のベースを、「知的にできること」から、単独行為不可能性という「根源的できなさ」と、それに由来す「WE」にとっての「I」の「かけがえのなさ」に置き換えるという対処法である。
- そうすることで、自動車や飛行機といった移動能力に関する凌駕機能体の登場によっても人間の自己肯定感が微動だにしなかったのと同様、人間の知的能力がAIによってどれほど凌駕されようとも、人間の自己肯定感は維持し続けることができることになる。
- ここにあるのは、人間失業という問題の前提であった「知的にできること」に人間の「かけがえのなさ」を見るという人間失業問題の前提を否定することで、問題そのものを雲散霧消させる方策、即ち問題の（「解決」ではなく）「解消(dissolution)」である。

人間失業に抗して：解決策

- もう一つの対策は、世界を認知する際についてまわる認知バイアスである人間バイアスとAIバイアスを相殺するために、人間知性が不可欠であることの指摘である。
- AIが人間の知性をどれほど凌駕しようとも、人間がAIと異なり続ける限り、AIバイアスと人間バイアスの相殺者としての人間知性の役割は残ることになる。
- 知的営みにおいての人間知性の不要化、即ち「人間失業」は起こらないのである。

参考文献

- Deguchi, Y., 2025a, WE-turn of Action: Principles, *Open Philosophy*, 8.1, pp.1-14
- Deguchi, Y., 2025b, WE-turn of Values: Principles, *The Journal of Philosophical Studies*, 614, pp.1-44.
- Gonzalez, W., ed. 2015, *New Perspectives on Technology, Values, and Ethics: Theoretical and Practical*