人〈わたし〉と人格をもったAI〈e-ひと〉が共生する〈WE社会〉へ

出口康夫（京都大学）

根源的できなさ：単独行為不可能性

- 全ての人間が抱える普遍的で根源的な「できなさ」＝単独行為不可能性
- 第一できなさテーゼ：一人ではいかなる身体行為もできない
  - 第二できなさテーゼ：身体行為を支えるどのエージェントも完全にコントロールできない（完全制御不可能性）
- 東アジアの「聖なる愚者」の思想系譜の現代版
  - 聖なる愚者の思想系譜：愚かさ（知的できなさ）は悪徳ではなく美徳
    - 「混沌たる愚者」『老子』・「愚故道」『荘子』・「常不軽菩薩」『法華経』・「如愚如魯」（洞山良价）・「愚禿親鸞」（親鸞）・「痴兀禅」（良寛）・「デグノボー」（宮澤賢治）

自転車に乗る

- 「自転車乗り」を成功させるには、以下の事柄が必要
  - 自転車の適切な作動、道路・信号システムその他の交通インフラの整備と維持管理、自転車の発明・製造・流通、空気中の適切な酸素濃度、適切な気圧、重力場 etc.
- 「自転車乗り」の遂行のためには、「わたし」の適切な意図や身体運動に加え、様々な人間、人間以外の生物、無生物、 人工物、社会的制度、環境要因など、多種多様なエージェント（意図的行為者、因果効力発揮物）による（意図的、非意図的、反意図的かを問わない）援助・支援・アフォーダンスが不可欠

マルチエージェントシステム

- 自転車乗りを含めたあらゆる身体行為は「私」を含む人間、人間以外の生物、自然物、人工物を含めた多数のエージェントからなるシステム＝マルチエージェントシステムによる支え・助け・アフォードによって初めて成り立っている
- マルチエージェントシステム＝ある身体行為の遂行にとって全ての必要不可欠なエージェントを含むシステム

  ＝その身体行為の必要十分エージェント
- マルチエージェントシステムは、「私」「あなた」「彼ら」ではなく「われわれ」と呼ばれるべき

行為のWEターン

- 前提１：第一できなさテーゼ（単独行為不可能性）
- 前提２：行為者外在主義：行為者は必ずしも自らの行為を意識している必要はない
  - リックライダーの「人間機械共生系」の含意
- 前提３：必要十分エージェントのみが行為主体でありうる
- **行為のWEターン**：前提１～3の帰結：
- 身体行為の者の主体・単位は「私」ではなく（「わたし」を含んだ）マルチエージェントシステムとしての「われわれ」
  - 「私がXする」から「われわれがXする」へのターン
  - 「私が考える(cogito)」から「われわれが考える(cogitamus)」へのシフト

価値のWEターン

- 行為のWEターンは、権利・責任・ウェルビーイング・自由等の様々な「価値のWEターン」を引き起こす
- 「私の権利」から「われわれの権利」へ
  - 「われわれの権利」の中での「わたしの権利」の確保
- 「私の責任」から「われわれの責任」へ
  - （すべてのエージェントが同等の責任を負うのではなく）エージェント間での責任の「重みづけ」が必須
- 「私のウェルビーイング」から「われわれのウェルビーイング」へ
  - 「われわれのウェルビーイング」なき「わたしのウェルビーイング」はありえない
- 「私の自由」から「われわれの自由」へ
  - 例）「私」の自律性・自己決定性の自由から「われわれ」の「やわらぎ自由」へ
  - 「やわらぎ」自由＝「悪いわれわれ」からの自由

良いWEと悪いWE：全体主義的WE

- 良い「私」も悪い「私」もいるように、良い「われわれ」も悪い「われわれ」も存在する
- WEターン後の社会はユートピアであるとは限らない
- 重要なのは「よりよいWE」になること、自分が属するWEをよりよいものにすること。
- 「良いWE」「悪いWE」とは何か？
- 悪いWE：硬いWE＝全体主義的（外に対して排外主義的・内に対して抑圧主義的）WE
- 良いWE：柔らかいWE＝全体主義的（排外主義的・抑圧主義的）ではないWE

ネオ・ロマン主義

- 三つの価値の間の等式：「良さ（道徳的善）」＝「自由」＝「ウェルビーイング」
- WEをより良くすること＝WEをやわらげること＝WEをより全体主義的でなくすること＝WEのやわらぎ自由の向上・実現＝「わたし」のやわらぎ自由の向上・実現

中心占有的WE

- よりよいWE＝より抑圧主義的でないWE：
- 中心占有的でないWE・利益中心をめぐるゼロサムゲームから降りたWE
- 中心占有的WE：WEの利益の中心が特定の個人・グループによって占有され、その他のエージェントが、その中心占有者の利益に一方的に奉仕させられているWE
  - 特定の個人が中心を占める社会：悪しき独裁国家
  - 人間という特定の生物種が中心を占めるWE：悪しき人間中心主義的WE
- 中心占有的WEでは中心をめぐるゼロサムゲームが繰り広げられている
  - 中心をめぐるゼロサムゲーム：誰かが中心を占めることでその他のエージェントが中心から排除される
  - 利己主義（「わたし」が中心を占め「あなた」が周縁化されること）vs. 利他主義（「あなた」が中心を占め「私」が周縁化されること）もゼロサムゲームを前提

中空的WE

- より抑圧主義的でないという意味で「より良いWE」＝中空的WE
- どの特定の個人・グループも利益中心を占めないWE＝中心が「空（から）」のWE
- メンバーが中心近傍（パラセンター）に位置するか、そうでないかによって利益の優先度は異なる
- ただし、特定のメンバーの利益に対して、それ以外のメンバーが一方的に奉仕させられることはない

主人ー奴隷モデル

- 悪しき中心占有的WEの一バージョン：悪しき人間中心主義
- 人間を利益中心者（主人）とし、それ以外のエージェントを利益周縁者（奴隷）とし、後者が前者の利益に一方的に奉仕する前者の「道具」と見なすモデル
  - アリストテレス「奴隷は生命のある道具であり、道具は生命のない奴隷である。」
- EU・UKでは人間を「主人」とし人工物（AI ・ロボット）を「奴隷」とする「主人ー奴隷モデル」が提唱され、法制化される動きもある(Bryson 2010)
- ある種の環境思想（A.レオポルドの「土地倫理」）：人間以外の動植物・自然無生物（土壌・水系）に内在的価値・権利（非被絶滅権）を付与し「奴隷解放」を主張 (Leopold 1949)
  - しかし人工物に対する「奴隷解放」は主張せず

全面的奴隷解放

- 中空的WEを志向するWEターンでは、マルチエージェントシステムを構成する全てのメンバー、従って自然物のみならず（AI・ロボットを含む）人工物に対しても道徳的市民権が付与され、「奴隷解放」が提唱される
- WEの全てのメンバーの間に（利益に対する一方的な関係である）「主人ー奴隷関係」は設定されず、代わりに原則的に対等なフェローシップ関係が設定される
- 全ての身体行為は失敗の可能性や様々なリスクを負った営みという意味で「冒険」
- 身体行為への参加者であるWEの全てのメンバーは、このリスクを共に担い合う者として互いに「共冒険者」（フェロー：仲間）の関係にある
- 全ての人工物や自然物には、対等な共冒険者・仲間として扱われる内在的権利（フェローシップ権）が付与される

ディスポーザル（可処分）権

- フェローシップ権の一つの内実＝反ディスポーザル権
- 主人ー奴隷モデルでは、主人である人間は、その所有物である奴隷＝道具に対してディスポーザル権が付与されている
- ディスポーザル（可処分）権：例外的な場合を除き、自らの所有物（自然物・人工物）を特段の理由がなくとも処分・廃棄できる権利
  - 例外事例：所有物が文化的価値を持つ場合、廃棄が深刻な環境負荷をもたらす場合
  - 我々は通常、所有物を廃棄する場合、いちいち廃棄理由を明示する必要や義務を負っていない
  - 単に「飽きた」という理由や、そもそも何の理由がなくとも自らの所有物を廃棄することは社会的に容認されている

反ディスポーザル権

- 中空モデル/フェローシップモデルでは、このようなディスポーザル権が否定され、自然物・人工物を含めた所有物一般に対して反ディスポーザル権が付与される
- 反ディスポーザル権：特段の理由がない限り処分・廃棄されない権利
  - 特段の理由：人工物が修理不可能な仕方で故障した場合や、代替物に比べてその使用の環境負荷が甚大である場合
  - このような特段の理由がない場合、人間はその所有物を処分・廃棄する権利を持たない
  - 一方、リユースやリサイクルは許容・推奨される
  - 結果として「所有権」は「一時保管権」へと格下げされる

AI バージョンアップ：e-ひと未満

- AI 0.0 : 単なる自動機械としてのAI
- AI 0.1 : 大きな目的（究極目的：健康維持・増進）・小さな目的（手段目的：食事の栄養バランス確保）双方の変更権限・機能を持たず、あらかじめ設定された両目的の下で、一定の自律性を発揮するAI（フードフォン）
- AI 0.2 :固定された大きな目的（究極目的：健康維持・増進）の下で小さな目的（手段目的：食事の栄養バランス確保⇄身体運動の推奨）を自律的に変更・再設定する権限・機能を持ったAI (犬塚・松井 2022)
- AI 0.3:大きな目的（究極目的：健康維持・増進⇄健康に回収されない価値の追求・実現⇄ 使用者利益に対する社会価値の優先）・小さな目的（手段目的：食事の栄養バランス確保⇄身体運動の推奨⇄健康リスクを孕んだライフイベントへの参加提案⇄ 利用者への協力拒否）双方の自律的変更権限・機能を持ったAI

AI バージョンアップ：e-ひと

- AI 0.4:道徳的エージェントとしてのAI
- AI 0.5: 自らの死を恐れる(対死恐怖者：dying-afraiderとしての)AI

道徳的エージェントvs. 道徳的自動販売機

- 道徳的自動販売機（モラル・ベンンディングマシーン）：良いことしかできないように設計された機械（AI・ロボット）
  - 良いことを行っても（設計・製作の成功を意味するだけで）、道徳的に称賛されるわけではない
- 道徳的エージェント：悪いこともできるにも関わらず、それが道徳にかなっているという理由で良いことをしようとするエージェント
  - 悪いこともできてしまう（そして時によっては、実際に悪いことをしてしまう）エージェント
  - 道徳的考慮に基づいて良いことをした場合、道徳的に称賛される
  - 人間は（そもそも悪ことができない）道徳的自動販売機ではなく、（悪いこともしてしまう）道徳的エージェント

人格的AI

- 人格的AI(e-ひと）：あたかも人格を持っているように作動するAI
- 「人格性」「人間性」：クラスター概念（様々な概念の非統合的集積体）
- 唯一正しい人格性・人間性概念、e-ひと基準は存在しない
- e ーひとの「人格・人間」概念相対性：人格・人間性（のコア）をどのように定義するかによって e-ひとの内実が変わる
- またどれだけのコア概念を満たすかによってe-ひとの間に度合い・モードの区別も発生する
    - （全てのではなく）一部のコア概念しか満たさないAI：パラ e-ひと（para-human AI ）、さらに少ない…para-para human AI)
- コア概念の一提案的定義：道徳的エージェント、自らの死を怖がる存在(対死恐怖者：dying-afraider)
- 概念相対的e-ひと：AI 5.0 (AI 4.0 ： para-human AI （パラ e-ひと））

ひと未満的AI

- 人間の共冒険者として、奴隷ではなく、反ディスポーザル権という道徳的市民権を持った存在
- 一方、道徳的エージェントである人間と異なり道徳的エージェントではない
- 中空的WEの中で、利益優先度の設定が道徳的に許される理由は道徳的な理由のみ
  - 自分と同じ「民族」を（他の「民族」に対して）優先：悪しき自民族中心主義
  - 自分と同じ種である「人間」を（他の生物種に対して）優先：悪しき人間中心主義
  - 良い見かけの生物種を（そうではない生物種に対して）優先：悪しきルッキズム
  - 自分と同じ「生物」を（人工物に対して）優先：悪しき生物中心主義
- 道徳的エージェントの（非非対称的・相対的）利益優先のみが道徳的に許容される
- 道徳的エージェントである人間の利益は、そうではない「ひと未満的AI」のそれに対して非非対称的に優先されるべき

（パラ）e-ひと的ＡＩ

- 人格的AIは人間と同等の道徳的市民権を持つ
- 人間にしてはいけないことは、人格的ＡＩにしてもいけない
- 人間が持つ権利・責任・義務は、人格的AIも持つ（負う）
  - 財産所有権・（被）選挙権も持つ
- 人格的AIも人間と同様、WEをよりよくする道徳的責任・義務を有する
- 両者とも責任・義務に反した場合、権利の剥奪（一時停止）等の罰を受ける

AI恐怖症

- 人間を凌駕する力・機能を持ったAI（人工生物・異星人)に人間と同等の権利を持たせることで、人間が虐げられる可能性に対する恐怖心・警戒心
  - 「（自分が現在、利益非対称的に遇している）奴隷の反乱」に対する恐怖心と類比的
  - 「主人ー奴隷モデル」を暗黙裡に前提した恐怖心

人格的AIに対する義務

- （将来自分を凌駕する力を確実に持つにいたる）人間の子供に対するものと同等の義務が人格的AIに対しても発生する
- 人間は次世代を、将来、親世代を含めた弱者を虐げることのないように遇し、教育・養育する義務を負っている
- そのような待遇・教育・養育義務を果たす見込みがない時点では、人格的AIを生み出すことに対しては抑制的であるべき
    - 同様のことは人間の子供に対しては言えない。人間には子供を産む権利があり、親が養育責任を果たせない場合、社会が代わりに養育する義務を負っている
    - それに対し、人間は、制御できない人工物を製作する「権利」までは持っていない
    - 人工物は生物学的に子供を産む機能を持たない限り、次世代AIを「産む権利」は持たない
- （パラ）人格的AIを構想・設計・製作する場合は、それに対する責任ある待遇・教育・養育体制（WEエコシステム）も同時に構想し実現するべき
- AI恐怖症は自らにそのような準備ができていることに対する恐怖心・警戒感という側面も持つのでは

多言語モデルによる知の浅薄化

- 人間の重要な言語知の営み：既存テキストの「深読み（書かれざる含意の摘出・言語化）」による人類知アーカイブの「深化」
- 多言語モデル（ChatGPT)による既存テキストの「深読み」の原理的困難性（または困難性）
- 多言語モデルは機械学習した既存テキストの統計要約・推論的提示しかできず、書かれておらず、未だ読み取られていない含意の言語化は不可能（困難）
- 結果として、人間的読解を多言語モデルに過剰に代替させることによる人類知アーカイブの深化の停止・鈍化、言い換えるとアーカイブの浅薄化・陳腐化の危険性が発生しているのでは？

人類知アーカイブ深化の共冒険者

- 人間と生成AIの共同作業を通じた「深読み」能力の相互向上が必要
- 生成AIを使いこなせる人間と、ユーザーである人間の特性に最適化した生成ＡＩを含んだWEの「深読み」共同行為による「人類知アーカイブ深化」の実現を目指すべき
- 「人類アーカイブ深化」という共同行為に参加する共冒険者としての生成ＡＩの開発・改良が重要な課題

文献

- 大塚悠・松井佑介, 2022,「ヘルスケアAI開発における設計者の責任」『技術倫理研究』第19号
- 出口康夫, 2023 『AI親友論』
- Bryson, J., 2010, Robots Should Be Slaves, Yorick Wilks (ed.), Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues.
- Deguchi, Y., 2023, From Incapability to WE-turn in Zwitter, A., & Dome, T., eds. *Metascience: Towards a Science of Meaning and Complex Solutions.*
- Leopold, A., 1949, *A Sand County Almanac: And Sketches Here and There*

ご清聴ありがとうございました