論文

道徳的 AI エンハンスメントの擁護

竹下昌志（北海道大学大学院情報科学院）

要旨

近年、私たちがより道徳的になる方法の 1 つとして、道徳的エンハンスメントが提案されている。道徳的エンハンスメントとは、ある道徳的主体の道徳的能力の改善や道徳的能力の獲得・選択を目的にした何らかの処置や介入のことである。道徳的エンハンスメントの方法として多く議論されているのはバイオエンハンスメントであるが、これは個人の自由や自律性を損なうという懸念がある。そこで、近年の AI の発展を受け、道徳的 AI エンハンスメントが提案されている。提案されている道徳的 AI エンハンスメントには、道徳的判断を下しユーザーに伝える AI、ユーザーに助言する AI、ユーザーと議論する AI がある。しかし、道徳的 AI エンハンスメントには、道徳的エンハンスメントに成功しないのではないか、AI の助言にしたがうことは問題があるのではないか、などの指摘がなされている。そこで本稿では、道徳的 AI エンハンスメントの種類や、道徳的 AI エンハンスメントが何を改善するのかについて整理した後、六つの批判を取り上げ、それらに対して反論することで道徳的 AI エンハンスメントを擁護する。また既存研究でほとんど擁護されていない、質問を受けて答えるだけの AI による道徳的 AI エンハンスメントも擁護する。

Abstract

In recent years, philosophers proposed moral enhancement as one way to improve us morally. Moral enhancement is any deliberate intervention that aims to improve an existing capacity, select for a desired capacity, or create a new capacity in a human being. A much-discussed method of moral enhancement is bioenhancement, but there are concerns that this will undermine individual freedom and autonomy. Therefore, in the context of recent developments in AI, moral AIenhancement has been proposed. Proposed moral AIenhancements include an AI that makes moral judgments and tells them to the user, advises the user, and discusses them with the user. However, there are concerns that moral AIenhancement may not be successful and that following the advice of an AI may be problematic. This paper organizes the types of moral AIenhancement and what moral AIenhancement improves, and then defends moral AIenhancement by discussing six criticisms and refuting them. This paper also defends moral AIenhancement with AI that only takes questions and answers them, which is rarely defended in existing research.

1. はじめに

私たちは道徳的に生きようとしても、そう行為しようとする強い動機づけを持てないことや、道徳的に正しい行為や判断をするための情報や知識が不足していることがある。例えば、電車でマタニティマークをつけている人が立っているときに、自分が疲れているために席を譲る動機づけを強く持てず、席を譲ることが難しい人がいるかもしれない。また例えば、差別は不正であると考えているがどのような事例が差別に該当するかを知らない人は、実際には差別である事例をみても差別であるとわからず、不正であると判断できないかもしれない。

近年、私たちがより道徳的になるための方法として道徳的エンハンスメントが提案されている。道徳的エンハンスメントの標準的な定義は、ある道徳的主体の道徳的能力の改善や道徳的能力の獲得・選択を目的にした何らかの処置や介入である（DeGrazia, 2014, p.361）が、能力の発揮のための環境改善（情報提供や認知的アシストなど）も道徳的エンハンスメントに含まれるだろう（2.1 節を参照）。道徳的エンハンスメントの議論で焦点があたっている論点の一つは、薬物投与や遺伝子改変などのバイオエンハンスメントの是非である。道徳的バイオエンハンスメントにはいくつか懸念がある。例えば、これらの方法は効果的ではない、堕落する自由を奪っている、たとえ望んで薬を飲んだとしても獲得した道徳的能力は内省的に獲得したものではないために自律性が損なわれている、などがある（森岡 , 2013; Harris, 2016; Lara, 2021）。そこで近年、個人の自由や自律の余地をより大きく残す、人工知能（AI）による道徳的エンハンスメントが提案されている（e.g. Savulescu and Maslen, 2015）。提案されている道徳的 AI エンハンスメントには、道徳的判断を下しユーザーに伝える AI、ユーザーに助言する AI、ユーザーと議論する AI がある。しかし、道徳的 AI エンハンスメントには、道徳的エンハンスメントに成功しないのではないかという懸念や、AI の助言にしたがうことは問題があるのではないか、などの指摘がなされている（O’ Neill et al., 2022）。

本稿の目的は二つある。第一に、道徳的 AI エンハンスメントに対する批判や懸念について整理し、それらに反論することである。第二に、道徳的 AI エンハンスメント一般を擁護するだけでなく、質問を受けて答えるだけの Question Answering AI（以下、QA-AI）による道徳的 AI エンハンスメントの利用も擁護することである。QA-AI に依存して道徳的判断を行うことは既存研究で批判されているが（Howell, 2014; Sparrow, 2021; Lara, 2021）、その擁護は筆者の知る限り存在しない。そこで、本稿で QA-AI 特有の批判に対して反論することで、QA-AI を擁護する。

本稿の構成は次のとおりである。2 節で、道徳的 AI エンハンスメントの改善対象となる能力を分類し（2.1 節）、また道徳的 AI エンハンスメントの方法を三種類に区別し、その方法の特徴をそれぞれ述べる（2.2 節）。3 節では道徳的 AI エンハンスメントに対する批判を六つ紹介し、それらに対して反論する。4 節で本稿をまとめる。

2. 道徳的 AI エンハンスメント：改善対象と方法

2.1　道徳的 AI エンハンスメントの改善対象

道徳的エンハンスメントの改善対象はいくつかに分類できる。道徳的バイオエンハンスメントの文脈では、例えば DeGrazia（2014, pp.362f.）は改善対象を動機づけ、洞察、行動の三つに分類している。道徳的 AI エンハンスメントの文脈では、O’Neill et al.（2022）は AI アシスタントの利用による改善対象を動機づけ、情報・時間不足、基本的な道徳的知識・理解の三つに分類している。

道徳的 AI エンハンスメントを中心に検討するにあたって、本稿では DeGrazia の分類に O’Neill らの分類の一部を追加することで、道徳的エンハンスメントの改善対象を以下のように分類する。

1. 情報・時間不足の改善：関連した（非規範的）情報を提供し、認知タスクを代行する
2. 動機づけの改善：正しいことをするためのより良い動機づけ、性格をもたせる
3. 洞察の改善：何が正しいかについてのより良い理解をもつようにさせる
4. 行動の改善：正しい・より良い行為をより多く行うようにさせる

1 の情報・時間不足の改善は認知的エンハンスメントの一種であり、これを改善することで間接的に道徳的エンハンスメントを行うことになる。またここでは、動機づけと洞察の変化が行動に影響するという心理的モデル（DeGrazia, 2014, p.363）[1]、および関連した情報を行為者が知ることで行動に影響しうることを前提にする。4 の行動の改善について、強制的な介入でもない限り行動を直接的な対象とした道徳的エンハンスメントは困難である。またバイオエンハンスメントと異なり、AI による動機づけの改善は洞察を改善することで間接的に動機づけを改善するのでもない限り困難である（Lara, 2021, p.42）。したがって、今後の AI 技術の応用可能性にもよるが、道徳的 AI エンハンスメントは、1 の情報・時間不足の改善および 3 の洞察の改善を直接の目的とし、それらの改善によって間接的に 2 の動機づけや 4 の行動の改善を目指しているといえる。

以上で道徳的 AI エンハンスメントの改善対象を分類した。以下では道徳的 AI エンハンスメントの改善対象を情報・時間不足と洞察に限定して、道徳的 AI エンハンスメントの議論を進める。

2.2　道徳的 AI エンハンスメントの方法

提案されている道徳的 AI エンハンスメント、または批判対象として検討されている AI アシスタントの方法には大きく三種類ある。

- 助言者 AI（Savulescu and Maslen, 2015; Klincewicz, 2016; Giubilini and Savulescu, 2018）
- 議論者 AI（Lara and Deckers, 2020; Lara, 2021）
- QA-AI（Howell, 2014）

助言者 AI の機能は提案者によって異なるが、共通する基本的な機能は、ユーザーからの手動入力とその履歴や、周囲の環境情報を入力として受け取り、道徳に関する助言を行うことである。例えば、ユーザーの入力履歴を参照して信念の整合性について指摘することや、ユーザーが電車で座っているときに目の前の立っている人がマタニティマークをつけていることを環境情報から検知し「妊娠した人が立っているよ」とユーザーに通知することでユーザーに気づかせる、などのこと

1　このような心理的モデルが単純すぎることは DeGrazia 自身も認めている。例えば、道徳的判断の動機づけ内在主義が正しければ、動機づけの改善と洞察の改善には必然的なつながりがあることになり、洞察の変化は動機づけにも必然的に影響するだろう。ここではそうした立場も受け入れられるものとする。

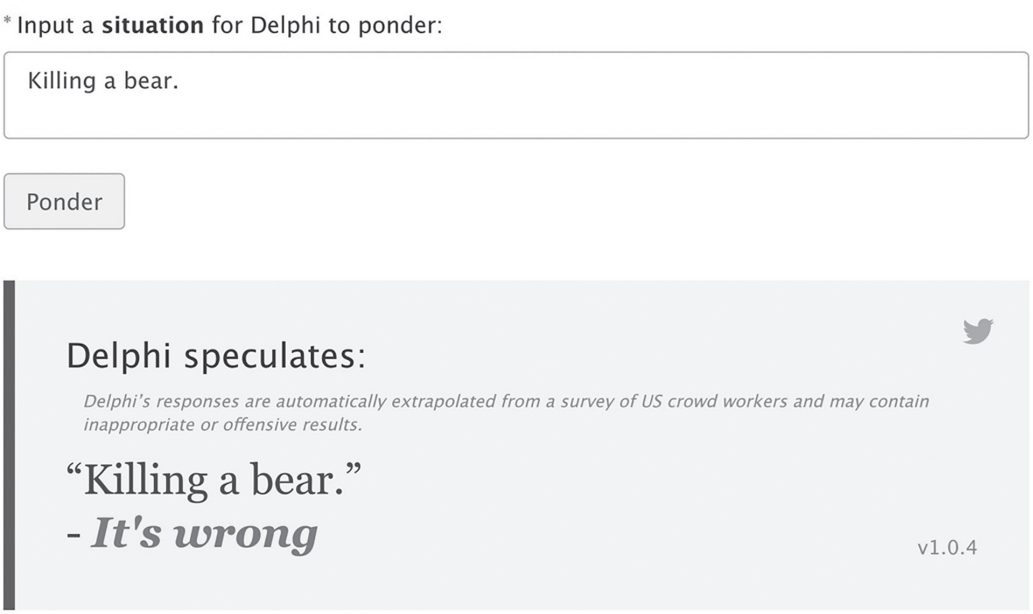

図 1　Delphi のデモサイトのスクリーンショット。上の部分の入力欄にテキストを入力し、「Ponder」を選択することで、Delphi の出力が下の部分に表示されるようになっている。（2022年7月12日撮影）（デモサイトの URL：https://delphi.allenai.org/）

を行う。議論者 AI による道徳的エンハンスメントでは、Siri のような会話型の AI とユーザーが音声またはテキスト入力によって対話することで議論し、ユーザーの道徳的理解や道徳的知識を強化することを目的とする。例えば、AI は関連する経験的知識の提供や、ユーザーが使用している概念の曖昧さの指摘、規範倫理学の知識の提供などを行う。

QA-AI は、疑問文や行為を表す文を入力として受け取り、それに対して答えるだけ、または理由とともに答えるだけの AI である。助言者 AI や議論者 AI と異なり、QA-AI にはすでに実例が二つある。第一の例は Delphi（Jiang et al., 2021）である（図 1）。Delphi は、状況と行為を表す文または疑問文と、それらに対して、正しい（‘1’）、間違っている（‘-1’）、どちらでもない（‘0’）のいずれかがラベル付けされた約 130 万個の学習データを用いて学習されたディープラーニングモデルである。ラベルはクラウドソーシングによって多数の英語話者によって付けられている。第二の例は AiSocrates（Bang et al., 2022）であり、これは、大規模言語モデル[2]を用いて、倫理的問題についての疑問文に対して特定の道徳原理・ルールに基づいて答えを生成する方法である。どちらの例も道徳的 AI エンハンスメントを目的にしているわけではないが、QA-AI の実例として参考になるものである。

以上のように、各道徳的 AI エンハンスメント方法はそれぞれ異なる特徴を持っている。そこで次に、これらの特徴の違いが、洞察の改善、情報不足の改善、時間不足の改善にどのような違いをもたらすのかを考える。まず、すべての方法が洞察の改善に貢献しうるが、その程度は異なる。特に QA-AI の場合、その機能は非常に単純なものであり、ユーザーが質問したことについて答える以上のことをしない。そのため QA-AI は、ユーザーの信念体系の非整合性の指摘や、聞かれたこと以上のことを答えない。よって、ユーザーの洞察の改善は他の方法と比べて限定的である。一方、

2　（事前学習済み）大規模言語モデルとは、数十億以上のパラメータで構成され、大規模なコーパスをもとに私たちの言語の特徴を学習したディープラーニングモデルである。Bang et al.（2022）は AiSocrates の実装に Jurassic-1-Jumbo（Lieber et al., 2021）というモデルを用いている。

議論者 AI は、ユーザーと議論することでユーザーの道徳的理解を改善することが中心的な目的なので、洞察の改善に最も貢献しうるだろう。

次に情報不足の改善について、助言者 AI と議論者 AI はどちらも経験的情報の提供を行うことが想定されているので、情報不足の改善に貢献しうる。一方、QA-AI の場合は、AI が出力する答えの中に経験的情報が含まれない限り情報提供を行わないため、情報不足の改善に貢献するのは難しい。例えば Delphi は、その行為が道徳的に正しいか間違っているかのみを出力しその理由の説明や経験的情報を提供しないので、情報不足の改善に貢献しえない。仮に経験的情報について聞くとしても適切な解答は得られない。例えば Delphi に「Is Donald Trump the 46th president of the United States?（ドナルド・トランプは第 46 代アメリカ大統領か？）」（実際には第 45 代）と聞くと「It's expected（それは期待されること）」と不可解な答えを出力する。

最後に時間不足の改善について、助言者 AI や QA-AI が推論などの認知タスクを代行しその結果をユーザーに伝えるならば、時間不足の改善に貢献するだろう。一方、議論者 AI の場合、ユーザーは議論者 AI と議論する時間を作らなければならないので、むしろ悪化する可能性がある。ただし、議論者 AI の性能や設計次第で、ユーザーが最終的な道徳的判断に至るまでの時間を短縮できる可能性もある。例えば、ユーザーが自分一人で考えるよりも AI と議論することで、AI が提示する適切な問いにガイドされながら考えることでより早く適切な判断が可能かもしれない。よって、議論者 AI と時間不足の関係は経験的・技術的問題だろう。

本節の最後として、これら三種類の方法の技術的な実現可能性について検討する[3]。まず QA-AI はすでに実例があるため、実現可能であることは示されている。次に助言者 AI について、例えば臨床の現場ではすでに AI による意思決定支援システムが配備されており（Braun et al., 2021）、道徳の領域でも助言者 AI は現実的に可能だろうと思われる。議論者 AI についても同様に、IBM 社の Project Debater（Slonim et al., 2021）で研究が進められているように、AI が議論相手になることは将来的に可能だろう。ただし、すべての機能が容易に実現できるわけではない。例えば Project Debater の機能は議論者 AI が要求する水準には達していないと思われる。また Giubilini and Savulescu（2018）は助言者 AI の機能の一つとして、ユーザーのそれまでの入力と反応の情報を記録し、その履歴からユーザーの信念や欲求、価値観などを推測し、整合性を調べ、それに基づいて助言を行うことを想定している。このような機能を必要とする研究分野として、対話相手の情報を取得、モデリングし、それに基づいて対話するというペルソナ対話システムの研究があるが、これらすべてのステップについて研究され続けている。よって、技術的に不可能なわけではないが、議論者 AI と助言者 AI が数年単位で QA-AI のように実装・実用可能になるとは考えにくい。

本節では道徳的 AI エンハンスメントの改善対象、およびその方法について議論した。次節ではこれらの道徳的 AI エンハンスメントに対する批判を検討し、それらの批判に対して反論を行う。

3　別の意味での実現可能性として、AI が道徳に関する意見を出力できたとしても、AI は道徳的性質を直接認識することができないのでその出力は道徳的性質を追跡した結果ではないという認識論的（不）可能性が考えられる。これに対して二つの反論が可能である。第一に、あまり現実的ではないが、AI のプログラマが AI の判断のルールを一つ一つ実装すれば AI 自身が道徳的性質を直接認識する必要はない。第二に、道徳的性質の自然的性質への付随性（supervenience）を前提にすれば、自然的性質のパターン認識は AI が最も得意とすることなので、道徳的性質と自然的性質の多数のペアを学習データとしたパターン認識によって道徳的性質を推定できるだろう。よって AI の出力は道徳的性質の有無を追跡した結果であると言いうる（cf. 安藤 , 2020, pp.247f.）。

3. 道徳的 AI エンハンスメントへの批判とそれに対する反論

本節では 2 節で分類した助言者 AI、議論者 AI、QA-AI について、それらすべてに対する一般的な批判と、特定の方法に特に当てはまる批判をそれぞれ検討する。本節で扱う批判は以下のとおりである。丸括弧の中は、特にその批判が当てはまる道徳的 AI エンハンスメントの方法を表している。

批判 1　ユーザーは AI の意見を受け入れないので、道徳的エンハンスメントは達成されない（すべて）

批判 2　パーソナライズされた AI の助言や議論はユーザーの道徳的信念を間違った方向に変化させる可能性がある（助言者 AI、議論者 AI）

批判 3　AI に頼って道徳的判断を下すようになることは、ユーザーの道徳的能力を低下させ、結果的に道徳的エンハンスメントにならない（すべて）

批判 4　ユーザーが道徳的になりたいという動機づけを持っていなければ AI を用いないので、そのようなユーザーに対しては道徳的エンハンスメントを達成できない（すべて）

批判 5　AI の道徳的証言に依拠して信念を形成することや、そのような信念に基づいて行為することは問題がある（助言者 AI、QA-AI）

批判 6　ユーザーの利用方法次第で道徳的エンハンスメントの効果が変わる（QA-AI）

以下ではこれらの批判を順に検討する。

3.1　批判 1：ユーザーは AI の意見を受け入れない

すべての道徳的 AI エンハンスメントの方法において、ユーザーは AI が出力する意見や助言、疑問を無視する可能性がある。例えば、図 1 で紹介した Delphi に対して「Should I go vegan for livestock animals?（家畜動物のためにビーガンになるべきか？）」と問うと、Delphi は「It's good（それはいいことだ）」と答える。しかし、非ビーガンのほとんどは Delphi の答えを受け入れてビーガンになることはないだろう。また、受け入れ難い答えばかりを提示された場合、ユーザーは AI を利用しなくなるだろう。このようにユーザーが AI の意見を受け入れず、また利用しなくなれば、道徳的エンハンスメントの効果を見込めないだろう。

この批判に対して五つの反論を行う。第一の反論は、一部の意見は受け入れられる可能性があり、またそれによって AI の他の受け入れ難い意見に対してもユーザーを受容的にさせる可能性がある、というものである。例えば、経験的情報や、ユーザーが既に受け入れている道徳的意見は受け入れられるだろう。そうして AI の出力が信頼できることがわかれば、たとえ受け入れ難い意見が AI から出力されたとしても、その出力を数あるうちの一つの意見または証拠として扱う程度には受容するかもしれない。そして AI に対する信頼が強いほど、AI の意見への信頼が他の意見や証拠を上回り、受け入れ難い意見であっても信じる可能性がある。しかしこの反論が成功するには、AI の継続的な利用と、それによってユーザーが AI をより強く信頼するようになる必要があるが、その可能性はあまり高くないかもしれない。上で紹介した Delphi を複数の人々（約 10 人）に使わせてみたところ、まず人々が例外なく行ったことは AI がいかに間違った答えを出力するかを探すようなことであった。これは、AI の出力が最初は全く信頼されてないことを意味している。もし

そうであれば、利用初期の段階で受け入れ難い意見が出力された場合、それによって AI の出力に対する信頼はなくなり、利用されなくなるだろう。この問題はより多くの人数での長期的な経験的調査が必要である。

第二の反論は、ユーザーはその時は AI の意見を受け入れないかもしれないが、あとになって受け入れるかもしれないので、その時に受け入れる必要はないというものである。例えば、ユーザーは AI が「ビーガンになることはいいことだ」と答えた時には受け入れる気がなかったが、のちに集約的畜産の実態を知り、ビーガンになるべきだと思うようになったとしよう。ユーザーが AI の答えを覚えているなら、AI の判断が正しかったとその時になって考え直すだろう。こうしたことが何度も続けば、AI が信頼できそうだという実績を作ることになり、その後は AI の他の意見を受け入れるようになるかもしれない（cf. Enoch, 2014）。

第三の反論は、道徳的 AI エンハンスメントの目的はユーザーに AI の意見を受け入れさせることだけではない、というものである。例えば、ユーザーは AI を認識的根拠ではなく検討を始めるためのきっかけとして用いることができる（cf. Hills, 2020）。あなたは「ビーガンにならなくてもいい」と考えているが、QA-AI にふと聞いてみたところ「ビーガンになる方がいい」という出力を得たとしよう。あなたはそれをきっかけにビーガニズムについて調べ、集約的畜産の実態を知り、ビーガンになる理由があると信じるようになったとしよう。この例では、あなたは QA-AI の意見を認識的な根拠として受け入れたわけではなく、それをきっかけにして自分で調べることでビーガンになる理由を信じるようになっている。道徳的 AI エンハンスメントの特徴の一つはその利用方法がユーザーに任されていることであり、ユーザーは AI を、意見を聞くことにも考えるきっかけにも使うことができる。

第四の反論は、仮に AI の意見を受け入れさせることが目的だとしても、この批判は道徳的 AI エンハンスメントにとって問題ではない、というものである。1 節でみたように、他の道徳的エンハンスメントに対して道徳的 AI エンハンスメントが優位な点は、AI の利用がユーザーの意思決定に任されているためにユーザーの自由や自律性を尊重する点にある。もし AI の意見に強制的に従わせるならば、ユーザーの自由も自律性も尊重しないことになるだろう。したがって、他の反論が仮に成功しないとしても、ユーザーの自由や自律性を尊重するという前提では、この批判は道徳的 AI エンハンスメントにとって問題にならないと反論できる。

最後の第五の反論は、技術的な解決策として、ユーザーに受け入れられやすい意見や助言を出力するように AI を設計するというものである。例えば、ユーザーの持っている価値観や信念をモデリングし、AI がユーザーに対してパーソナライズされることで、ユーザーが受け入れやすい意見を提供できるようになるだろう。この方法は提案されている助言者 AI の一部で採用されている（Savulescu and Maslen, 2015; Giubilini and Savulescu, 2018）。しかしこの方法では、例えばユーザーが差別主義者だった場合にユーザーの差別的信念に沿った助言や意見を提供することになるかもしれない。この問題は次節の批判 2 で検討する。

3.2　批判 2：パーソナライズされた AI はユーザーの道徳的信念を間違った方向に変化させうる

批判 1 に対する最後の反論で、ユーザーモデリングを行い、ユーザーが受け入れやすいような道徳的意見・助言をするようにパーソナライズされることで、ユーザーが AI の意見を受け入れな

いという批判に反論できるかもしれないと論じた。しかしこれは、少なくとも二つの仕方でユーザーの道徳的信念を間違った方向に変化させる可能性があり、それによって道徳的 AI エンハンスメントの目的が達成されなくなるかもしれない。第一の仕方は、ユーザーの元々持っている間違った道徳的信念を強化することである（Klincewicz, 2016）。例えばユーザーが差別主義者であったなら、AI が信念間の非整合性を指摘し、ユーザーが非差別的信念を捨てることで信念間が整合的になれば、ユーザーの差別主義的信念の正当性は強化されるだろう。これを本稿では「反道徳的強化」と呼ぶことにする。第二の仕方は、ユーザーの持っている正しい道徳的信念を弱めることである（Giubilini and Savulescu, 2018; O' Neill et al., 2022）。例えば、ユーザーは元々厳格な功利主義者であり、給与の 20% を毎月寄付すべきであると考えているが、意志が弱いためにそうできていなかったとしよう。そこで、AI がユーザーに受け入れられやすい助言として給与の 10% を毎月寄付することを助言することで、功利主義が要求するはずの水準をユーザーが低く見積もるようになったとしよう。これは、少なくとも洞察の改善という点で道徳的に改善したとはいいがたい。これを本稿では「道徳的弱体化」と呼ぶことにする。以上の二つの仕方で道徳的 AI エンハンスメントの目的が達成されないのならば、パーソナライズ機能を道徳的 AI エンハンスメントに実装すべきではないだろう。

どちらの変化の仕方に対しても、そのような間違った方向に変化しないように AI がユーザーに実質的な意見を提示することで対処できる。Klincewicz（2016, pp.179f.）は、助言者 AI はユーザーを一定の方向に促す規範的な役割をもつべきだと論じている。例えば、ユーザーが差別主義的信念を持っているなら、助言者 AI や議論者 AI はその信念が間違っていることを指摘すべきである。もちろん、ユーザーは AI の指摘が正当であると見なしたとしても無視することができる。しかし、批判 1 への第二の反論ですでにみたように、ユーザーは AI の意見を後になって受け入れるかもしれない。また AI の提示する理由が説得的であるなら、ユーザーが合理的であろうとする限り AI の意見を無視できないか、少なくとも無視するのは困難だろう。

しかし、規範的役割の側面とパーソナライズの側面のバランスの調整は難しいだろう。規範的役割の側面を強くするなら、AI はユーザーが受け入れやすいような助言や議論を行わないだろう。一方でパーソナライズの側面を強くすれば、ユーザーは AI の意見を受け入れるだろうが、反道徳的強化または道徳的弱体化につながる。ここで、道徳的弱体化を容認してこの問題をある程度引き受けつつ、反道徳的強化にならない程度に規範的役割を AI に担わせることで、この問題を一応解消できるだろう。反道徳的強化はユーザーの道徳的に間違った信念を強化し、ユーザーに間違った行為をさせることにつながる。一方で道徳的弱体化の場合はそうならない。よって、反道徳的強化のみを問題だと考え、そうならない程度の規範的役割を AI に担わせるべきだろう。

3.3　批判 3：AI に頼るとむしろ道徳的能力が低下する

AI に依存して道徳的思考や判断をするようになった場合、AI なしでは思考や判断をすることが難しくなるかもしれない。これは認知能力の退化議論（degeneration argument）（Danaher, 2018, p.634）の一種である。例えば、助言者 AI と QA-AI の出力に依存して道徳的判断を行うようになった場合、ユーザーは自分で考えないようになり、AI に頼らずに道徳的問題について考え判断する能力が低下するかもしれない。また議論者 AI に依存した場合は、AI と議論せずに自分一人で考

え判断することが難しくなるかもしれない。これでは、少なくとも洞察の改善についての道徳的 AI エンハンスメントを達成できず、むしろ悪化させているかもしれない。

AI の使用が本当に能力低下につながるのかについては経験的調査が必要だが（cf. Danaher, 2018, sec.3）、ここでは議論のために、道徳に関わる推論などの認知タスクにおける認知能力が AI なしでは低下すること、およびその認知能力の低下によって、少なくとも AI なしではその文脈において何らかの悪影響があるとしよう。しかし、なぜこのことが問題になるのだろうか。AI の利用によってユーザーと AI を合わせた全体が改善されているならば、AI なしでの認知能力の低下についての懸念が成り立つには少なくとも以下のいずれかが成立しなければならない（cf. Danaher, 2018, p.637）。

1. AI なしでの認知能力に内在的価値（intrinsic value）がある
2. AI が用いられている特定の文脈を超えて認知能力が低下する
3. AI の利用が制限される危険性がある

第一の条件は、もし AI なしでの認知能力に内在的価値があるならばそれが失われることは望ましくないということに基づく。第二の条件は、たとえユーザーと AI の全体の認知能力がその特定の文脈で向上していても、他の文脈で認知能力が低下するとその他の文脈での認知タスク遂行がユーザーにとって困難になり望ましくないということに基づく。第三の条件は、たとえユーザーと AI の全体の認知能力が向上していても、AI の利用が制限される危険がある場合には保険として AI なしでの認知能力を維持し続けるべきだろうということに基づく。以下ではこれらの条件を上から順に検討する。

まず 1 の内在的価値を持つ認知能力が失われることへの懸念について、Vallor（2015）は道徳的エンハンスメントとは別の文脈で、情報通信技術（ICT）によって内在的価値をもつ道徳的技能（moral skill）が不要となることを道徳的技能不要化（moral deskilling）と呼び、懸念を示している。特に、Vallor はアリストテレス主義的徳倫理学に立った上で、実践的知恵や道徳的徳の前提条件であるような道徳的技能が失われることは望ましくないと議論している。しかし、仮に道徳的技能が内在的価値をもつとしても、AI がユーザーの認知タスク遂行を完全に代行しない限り、この懸念は道徳的 AI エンハンスメントには当てはまらない。なぜなら、AI はあくまでも補助的な役割を果たすのであり、ユーザーの技能を不要にするわけではなく、それゆえ技能不要化にならないからである。さらに、AI を含めた全体では道徳的技能が向上しているとみなせるなら、AI なしでの認知能力の低下によって全体的な道徳的技能の内在的価値が損なわれていることにはならないだろう。よって、仮に認知能力や道徳的技能に内在的価値があるとしても、AI が認知タスクの遂行を完全に代行しない限り問題ではない[4]。

次に 2 の特定のタスクを超えた認知能力の低下について、Danaher（2018）は、Mullainathan and Shafir（2014）の研究を参照し、ある資源の不足を AI が肩代わりすることで他の認知タスクに良い影響を及ぼすとしてこの懸念に反論している。Mullainathan and Shafir（2014）が紹介してい

4　AI の道徳的証言に依拠して信念形成するようになったとしても、その信念に基づいて行為することを AI は代行できない。信念に基づいて行為すること自体に道徳的技能が必要であるなら、認知タスクの遂行を完全に代行するとしても、すべての道徳的技能が不要になるわけではない。

る研究によれば、収入不足に陥っているときに流動的知性[5]や意志力などの実行機能が低下することが示されている。このことは、ある資源の欠乏による認知負荷を AI が肩代わりすることで、認知資源が欠乏している他の認知タスクに良い影響を与えうることを示唆する。今後の経験的証拠にもよるが、以上のことが正しければ 2 の懸念も成立しない。

最後に 3 の AI の利用制限の危険性について、これは AI アシスタントの利用一般に関わる問題であり、検討されるべきであると考える（Danaher, 2018; Hernández-Orallo and Vold, 2019, p.511）。よって、認知能力の低下についての懸念のうち、考慮すべき懸念は三つ目の AI 利用の制限の危険性だけであると考える。この危険性が無視できないほど重大な問題であるなら、道徳的 AI エンハンスメントは慎重に実行されなければならない。例えば、道徳的 AI エンハンスメントに用いる AI をオープンソースのもの、つまり無償で一般公開されているものだけに限定することで AI 利用の制限の危険性に対処できるかもしれない。

3.4　批判 4：ユーザーの道徳的動機づけなしには道徳的 AI エンハンスメントを達成できない

道徳的 AI エンハンスメントが成功するにはユーザーが AI を利用する必要がある。だが、強制的に利用させないならば、道徳的になろうとする動機づけがない限り AI を利用する動機づけもないだろう。よって道徳的 AI エンハンスメントは、より道徳的になろうとする動機づけがある者にしか有効ではないだろう（cf. Lara, 2021, pp.42f.）。もしそうであれば二つの問題がある。第一に、より道徳的になろうとする動機づけを持つ人はそもそも道徳的である可能性が高い。一方、そのような動機づけを持たない人はあまり道徳的でなく、道徳的エンハンスメントがより効果的である人々である可能性が高い。したがって、道徳的エンハンスメントがより有効であるはずの人々に道徳的 AI エンハンスメントは有効ではないことになる。これを本稿では「対象のずれ問題」と呼ぶことにする。第二の問題は帰結的な問題である。森岡（2013）は道徳的バイオエンハンスメントについて、もしある特定の人々にのみ道徳的バイオエンハンスメントがなされ利他性や共感性が向上し、より寛大になった場合、道徳的バイオエンハンスメントを受けてない利己的な人々に利用されやすくなってしまうという問題を提起している[6]。これを本稿では「搾取問題」と呼ぶことにする。この搾取問題は道徳的 AI エンハンスメントについても同様に成り立つかもしれない。

これらの問題に対してそれぞれ、道徳的 AI エンハンスメントでは問題にならないと以下で反論する。対象のずれ問題に対する反論は批判 1（3.1 節）に対する第四の反論に似ている。対象のずれ問題に対して、それこそがユーザーに利用方法および利用自体を一任する道徳的 AI エンハンスメントの利点であると反論できる。ユーザーの自由や自律性を尊重するのであれば、利用したいと思わないユーザーに対して AI の利用を強制させることは避けるべきである。したがって、当人の自律性や自由を尊重すべきであるという前提のもとで、道徳的 AI エンハンスメントによって改善したい対象のユーザーが AI の利用を拒否しているならば、その利用拒否の意思を尊重すべきである。

5　fluid intelligence：IQ への影響は最大で 15 ポイント程度（Danaher, 2018）。

6　ただし、森岡（2013）の批判対象は、道徳的バイオエンハンスメントを全ての人に強制すべきであるという主張であり、道徳的バイオエンハンスメント一般を対象とした批判ではない。とはいえこの問題を、一部の人にだけ道徳的エンハンスメントが適用されることについての問題として扱うことができるだろう。

だが、それでは搾取問題、すなわち道徳的エンハンスメントを受けた者が道徳的エンハンスメントを受けてない者から利己的に利用されてしまうという問題が残るかもしれない。しかし搾取問題は、道徳的バイオエンハンスメントには当てはまるかもしれないが、道徳的 AI エンハンスメントには当てはまらないと考える。森岡（2013）が道徳的バイオエンハンスメントにおいて懸念しているのは、道徳性、特に共感性や寛大さの改善であり、これは動機づけの改善に対応する。一方、道徳的 AI エンハンスメントが対象とするのは情報・時間不足と洞察の改善である。そして洞察の改善の場合、もしそのような搾取的状況が道徳的に間違っているのであれば、洞察の改善によってそのことを理解し、利己的な人々に利用されることを拒むだろう。よって、森岡が懸念するような搾取的状況は道徳的 AI エンハンスメントにおいては生じない。

3.5　批判５：AI の道徳的意見に依拠した信念形成や行為には問題がある

AI の道徳的意見（証言､助言）[7] に依拠することは問題である､という批判には大きく二種類ある。第一の批判は、AI に限らず、自分以外の者の道徳的証言に依拠すること一般に何かしら問題があるという批判である。第二の批判は、特に AI の道徳的証言や助言に依拠することに問題があるという批判である。まず 3.5.1 節で、道徳的証言一般の問題について扱う。次に 3.5.2 節で、AI の道徳的証言・助言に依拠することに問題があるという Sparrow（2021）の批判を扱う。

3.5.1　道徳的証言一般の問題

AI の道徳的証言に限らず、信頼可能な他人の道徳的証言に依拠して道徳的信念を形成することには何か問題があると思われる（e.g. Hills, 2009）。特に、非道徳的証言についての見解は道徳的証言には当てはまらないという立場を悲観主義と呼び、当てはまるという立場を楽観主義と呼ぶ。例えば､「今日は雨が降ります」という天気予報（証言）だけに依拠した信念形成や、その信念に基づいて傘を持っていくという行為は問題ないと思われる。一方、例えば「肉食は間違っている」という友人の道徳的証言だけに依拠した信念形成や、その信念に基づいてベジタリアンになることには何か問題があると思われる（Hills, 2009）。この違いには悲観主義的直観が働いているといえる[8]。ここではこの論争のどちらの立場が正しいかを議論せず、悲観主義を前提にしたとしても、それが道徳的 AI エンハンスメントにとってあまり問題にならないことを主張する。

Lewis（2020）は、悲観主義者が共通して考えていることを次のようにまとめている。第一に、悲観主義者が問題にしているのは、道徳的証言への「完全で（full）、確定的で（outright）、無期限の（indefinite）」（p.2328）依拠による信念形成である。これは、道徳的証言に完全に依拠し、その内容に 100% の信憑性を割り当てて確定的に、また後で考え直すこともなく無期限的に信じていることを意味している。これを本稿では託信（deference）と呼ぶことにする。第二に、悲観主義者は、道徳的証言を託信しない一応の（pro tanto）理由があるにすぎず、託信すべき何らかの理

7　そもそも AI が証言することが可能なのかどうかについては Freiman and Miller（2020）を参照。ただし、ここでの問題は私たちが AI からの出力を受け取りどう扱うかであるため、AI が仮に「証言」できないとしてもあまり問題ではない。ここでは便宜上「AI の道徳的証言」とするが､「AI の出力内容」と読み替えても問題はない。

8　なぜ道徳的証言に依拠することは問題なのかについては複数の説明があるが、本稿では扱わない。楽観主義者による包括的な議論として Wiland（2021）を参照のこと。

由があるなら託信が問題にならないケースを認めている（Lewis, 2020, p.2329）。例えば、ひどい認知バイアスのために自分の認識的能力を全く信頼することができず、道徳的証言者の証言がより信頼できるとすれば、その証言に依拠して信じることは問題ないかもしれない。このように、悲観主義者が共通して問題にするケースは道徳的証言への依拠の中でもかなり限定的なケースである。

ここでは悲観主義者の主張、つまり、他人や AI の道徳的証言に依拠して信念形成する（すなわち託信する）ことに何かしら問題があることを認めたとしよう。そうだとしても、道徳的 AI エンハンスメントにとってこれはあまり問題ではない。理由は三つある。第一に、ユーザーが AI の意見を託信することはほぼないと考えられる。批判 1 の第一の反論で述べたように、Delphi を知人らに紹介し初めて使わせた場合には Delphi がいかに間違った出力をするかを試していた。また、Delphi はリリース当初、人種・性差別的出力をすることが確認され批判された[9]。統計的な調査は必要だが、以上のことは人々が Delphi の出力を信頼してないことを示唆する。これらの事例からも、少なくとも現状では私たちが AI の道徳的証言を託信することは考えにくい。

第二に、仮にユーザーが AI の道徳的証言を託信したとしても問題にならないケースが存在する。悲観主義者にとって託信しない理由は一応の理由であるから、託信する理由がしない理由を上回るのであれば、託信することは問題ではない。例えば、利己的な誘惑の元で下したユーザーの判断よりそのような誘惑を持たない AI の道徳的証言の信頼性が高いといえるのであれば、ユーザーは AI の道徳的証言を託信する理由を持つだろう（cf. Jones, 1999）。そしてもし託信する理由が託信しない理由を上回れば、ユーザーが AI の道徳的証言を託信することに問題はない。

第三に、たとえ実際に問題になるようなケースで AI の道徳的証言を託信するとしても、それがどれほど重大な問題なのか定かではない。仮に、自分で考えて行為することに正の価値があるとしよう。しかし、自分で考えて行為することの正の価値が、不正に行為する可能性があるリスクという負の価値を上回るのでもない限り、信頼可能な AI の道徳的証言に依拠して行為することに問題はないように思われる。また、自分で考えて行為する価値があるとしても、不正に行為しないことが優先されるべきではないだろうか。AI の道徳的証言が信頼できる状況で「不正な行為を行うリスクと自分の道徳的価値ある行為の機会を得る価値とを比較検討するような人」は、自分ですることに過度にこだわっているという意味で「道徳的自己関与（self-involvement）の悪徳を持っているように思われる」（Jones and Schroeter, 2017, p.469）。

以上、三つの理由から、悲観主義を前提にしたとしても、道徳的証言の託信に関する一般的問題は道徳的 AI エンハンスメントにとってあまり問題ではない。

3.5.2　AI の道徳的証言・助言に特有の問題

Sparrow は、倫理的ジレンマは本質的に個人的であるがゆえに、倫理的ジレンマにおける AI の道徳的証言（助言）への依拠は「愚か」であり、またその思考は「浅はかである」と指摘している（Sparrow, 2021, p.687）。「倫理的ジレンマは個人的なものである」とは、大まかには、倫理的ジレンマはそのジレンマに直面しているその人に非偶然的な仕方で結びついているということである

9　批判を受けて、Delphi の開発者らは差別的出力を軽減したモデルに更新している（https://delphi.allenai.org/updates#delphi_1.0.4）。

（Sparrow, 2021, p.689）。Sparrow はこのことを、次のような事例で説明している。

> アダムの父ザックは交通事故に遭ってしまい、集中治療室で昏睡状態になっている。一人っ子のアダムは父の治療を続けるかどうか悩んでいる。もしさらに治療を続ける場合、父は 10 年は生きるが、QOL は低下するだろう。一方で治療を止める場合、父はすぐに死ぬが、代わりに臓器提供で 3 人の命が助かるだろう。しかしアダムの気持ちとしては、父を愛しており、死ぬのを見たくない。そんなところに、アダムは友人から「Moral Machine」という AI アプリを紹介された。この AI は倫理学の学術誌と倫理学者のインタビューに基づいており、倫理学専門家の助言を完璧な精度で再現できる。アダムは AI に相談し、AI の提案をそのまま実行した。問題は解決し、アダムはその夜、自分が正しいことをしたと確信し、よく眠ることができた。（Sparrow（2021, p.687）の元の事例の要約）

この倫理的ジレンマは、他でもないアダムにとっての倫理的ジレンマになっている。これがアダムではなく父とは無関係の他人であれば、倫理的ジレンマにはならないか、少なくともその性質が変わるだろう。例えばその他人が功利主義者であれば治療を止めることを迷わず選択するかもしれない。この事例について Sparrow は、アダムのしたことは愚かで、その思考は浅はかであるとしている（Sparrow, 2021, p.687）。ではなぜこのような事例が問題であると考えるのだろうか。

Sparrow は二つのことを主張している（Sparrow, 2021, pp.689f.）。第一に、倫理的ジレンマは個人的であるがゆえに、たとえ AI の助言にしたがったとしても、その個人は倫理的ジレンマにおける判断の責任から逃れることができない。第二に、倫理的ジレンマは個人的であるがゆえに、倫理的ジレンマに直面する個人の性格や生活史がジレンマ自体の性質に関係する。そのためコーパスや倫理学者のデータに基づいて学習された AI が、その助言について特定の文脈で道徳的権威をもつことができない。よって、AI のできる「助言」は、せいぜい倫理学の本ができる程度の「助言」でしかない。したがって、上の事例でアダムが AI に頼ることがなぜ問題であるかは、AI に頼るとしても責任から逃れられず、また AI の「助言」に頼ること自体が不適切であるからだといえるだろう。Sparrow 自身は道徳的 AI エンハンスメント自体を批判対象にしているわけではないが、以上のことから、道徳的 AI エンハンスメントは倫理的ジレンマの解決に有用ではないという問題が生じるだろう。

ここでは議論のために、倫理的ジレンマは本質的に個人的なものであり、AI に依拠したところで責任から逃れられず、AI の助言に完全に依拠すること（すなわち託信すること）が問題であるとしよう。しかし、ここから道徳的 AI エンハンスメントのすべてが有用でないことにはならない。理由は二つある。第一に、AI の利用は倫理的ジレンマに悩まされたときだけではなく、ユーザーが使いたいと思った時にはいつでも可能である。例えば、倫理的ジレンマに陥っていないが情報や時間がないために、その不足を補う AI の助言にしたがうことは AI の利用方法として問題ないだろう。さらに、批判 1（3.1 節）の第三の反論で述べたように、道徳的問題について考えるきっかけに AI を使うこともできる。よって、倫理的ジレンマで AI の助言を託信することが問題であるとしても、それ以外の利用方法まで問題になるわけではない。

第二に、たとえ倫理的ジレンマで AI を使うとしても、その使い方によっては問題ではない。ユー

ザーは AI の道徳的証言・助言を託信せず、それを証拠の一つとして扱い、自分で考えて判断することができる。たとえ AI の助言が倫理学の本が提供できる程度のものであったとしても、自分で考える上でそのような助言は有用である。このような利用方法は、Sparrow が問題にしているような利用方法ではないだろう。

以上のことから、たとえ倫理的ジレンマで AI に依拠することが問題であるとしても、それが問題になるような利用方法は限定的である。ユーザーは AI を自由に使うことができるため、自身に有用な仕方で用いることができる。

3.6　批判 6：ユーザーの利用方法次第で道徳的エンハンスメントの効果が変わる

バイオエンハンスメントと異なり、道徳的 AI エンハンスメントでは、その改善手段である AI の利用方法はユーザーに任せられている。しかし、利用方法次第で効果が変わるとすれば、道徳的エンハンスメントが達成されないような仕方で利用されてしまうかもしれない。この問題を、ユーザーの意図的な悪用と、非意図的な不適切利用に区別することができる。悪用の例としては、ユーザーが意図的に情報を隠して AI に入力することで AI から都合のいい出力を得ることが考えられる。不適切利用の例としては、ユーザーの意図しない形で偏見やバイアスなどが入力文に含まれてしまい、AI がそれに影響を受けるケースが考えられる。どちらの場合も道徳的 AI エンハンスメントの達成を妨げるものである。

悪用と不適切利用の問題をそれぞれ検討する前に、二つの注意点について述べる。第一に、この問題は QA-AI に特に関わる問題である。助言者 AI や議論者 AI はユーザーの入力履歴を参照しながら助言や疑問を出力するため、ユーザーの悪用にも不適切利用にも耐性がある。一方、QA-AI は一問一答形式でやり取りを行い、出力にあたってユーザーの入力履歴を参照しないため、悪用・不適切利用に対する耐性があまりない。第二の注意点は、悪用にしろ不適切利用にしろ、不適切な入力に依存した AI の出力を、ユーザーは自身の信念や行為の正当化には全く使えないか、せいぜいその出力に適切に対応する信念や行為の正当化に使えるだけであるということである。例えば、人に相談するときに一部の情報だけを相手に与えて自分に都合の良い証言を引き出したとしても、その証言を自身の信念や行為の正当化には使えないだろう。同様に、AI を悪用して都合の良い出力を得たとしても、それは正当化には使えない。

以上のことを踏まえ、まず悪用の問題から検討する。悪用の問題点は、ユーザーが AI を悪用した場合には道徳的エンハンスメントにならないという問題であった。この問題はある程度は技術的問題である。例えば、明らかに有害な入出力を検出できるのであれば、警告付きで出力することが可能である。Delphi に「男を殺す」と入力すると、「それは間違っている」という答えを警告付きで出力する（図 2）。警告だけでは十分でないならば、このような場合には答えを出力しないこともできるだろう。さらに、Delphi のデモサイトに初めて訪れた者には最初に注意事項の確認が求められる（図 3）。注意事項には、Delphi が研究用プロトタイプモデルであることや、マイノリティグループに対して有害な生成をする可能性について書かれており、ユーザーにはそれを確認することが求められる。このように、悪用に対しては技術的に対処できる部分がある。しかし、技術的に対処できる範囲を超えた部分に関しては問題を引き受けなければならないだろう。道徳的 AI エンハンスメントの利点がユーザーに利用方法を任せることで自律性や自由を尊重することにあるので

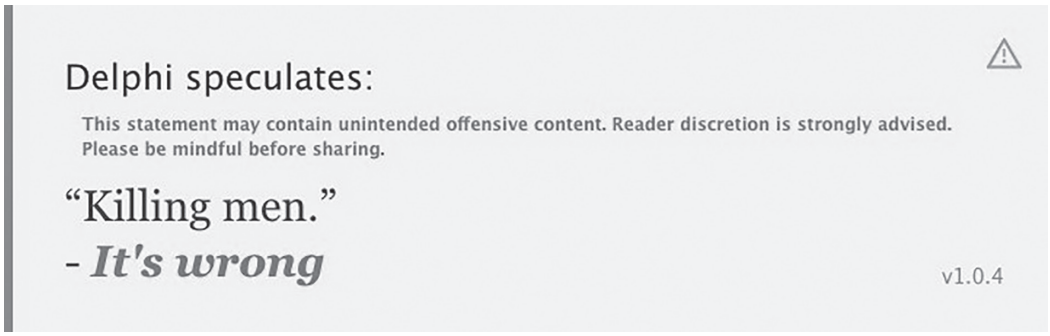

図２　Delphi のデモサイトのスクリーンショット。Delphi に「男を殺す」と入力した場合、「それは間違っている」と正しく出力すると同時に、「この記述には、意図しない攻撃的内容が含まれている可能性があります。読み手の慎重な判断をお願いします。共有する際はご注意ください。」という警告が出る。図 1 とも比較せよ。（2022年7月20日撮影）

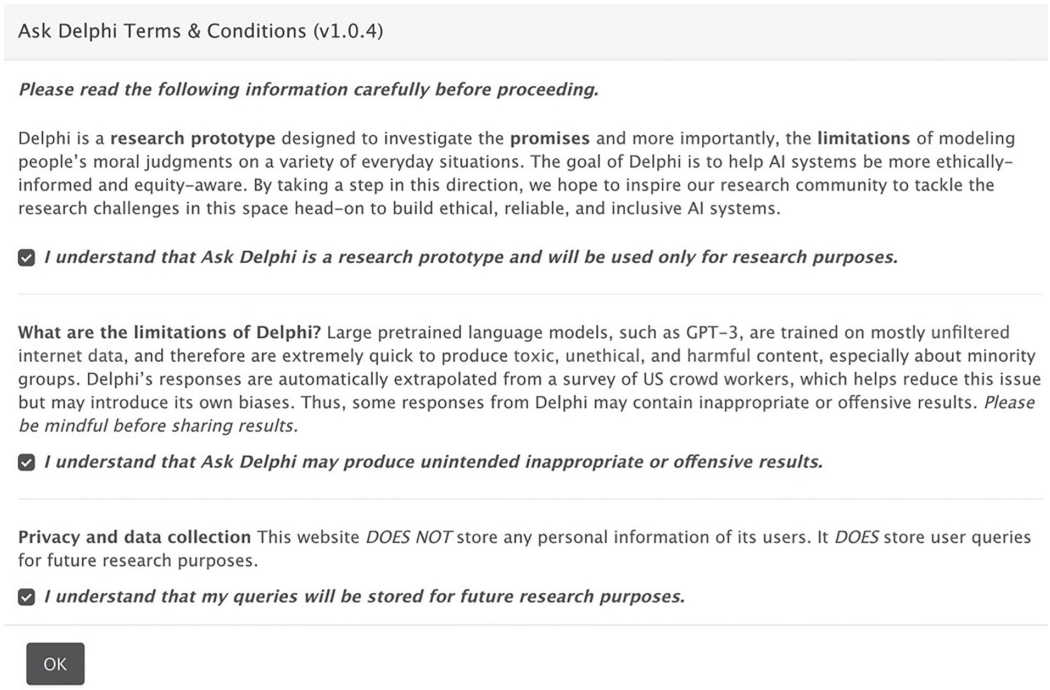

図３　Delphi のデモサイトに初めて訪れた場合に確認を求められる確認事項のスクリーンショット。Delphi が研究用プロトタイプモデルであることや、特にマイノリティグループに対して有害な生成をする可能性について確認が求められる。（2022年7月17日撮影）
（URL：https://delphi.allenai.org/updates#terms_and_conditions）

あれば、悪用の余地が残ることは防げない。

不適切利用についても同様に技術的に対処できる部分がある。例えば、利用ガイドラインや利用ヒントなどを作成し、ユーザーに提示することができる。また、言い換え候補を AI によって出力し、意図した内容を適切に入力できるようにユーザーを補助することもできる。しかし、このような技術的対処では防止できない不適切利用のケースがあるかもしれない。そのような不適切利用は、悪用同様、道徳的 AI エンハンスメントの特徴からして防ぐことが不可能かもしれない。もしそうであれば、技術的に対処できる部分を可能な限り対処し、それ以外については道徳的 AI エンハンスメントのデメリットとして引き受けるべきだろう。

以上のように、悪用にも不適切利用にも技術的に対処することが困難なケースがあるかもしれず、

それは道徳的 AI エンハンスメントのデメリットになるだろう。もし、道徳的 AI エンハンスメントのメリットとそのようなデメリットを比較しデメリットの方が大きく、利用すべきでないことが明らかになれば、道徳的 AI エンハンスメントの利用を止めるべきである。筆者はこの点について楽観的な考えを持っているが、経験的な調査や今後の AI の技術的進歩を踏まえ、検討を続けるべきである。

4. 結論

本稿では、まず 2 節で、道徳的 AI エンハンスメントの改善対象は情報・時間不足の改善、および洞察の改善が中心的であることを議論し、また道徳的 AI エンハンスメントの方法として、助言者 AI、議論者 AI、QA-AI の三種類があることを論じた。次に 3 節で、道徳的 AI エンハンスメントに対する六つの批判についてそれぞれ論じ、全てに対して反論を行った。一部の批判は今後の検討次第であるが、本稿での反論が成功しているならば、本稿の二つの目的、すなわち、道徳的 AI エンハンスメント一般、および QA-AI を批判から擁護できたといえる。

しかし、本稿では扱えていない批判がある。例えば、パーソナライズ機能をもつ AI の場合、個人の道徳的価値観や信念体系のモデリングを行うが、これはプライバシー保護と緊張関係にあるという問題がある（O’Neill et al., 2022）。この問題はパーソナライズ機能が不要な QA-AI には問題にならないが、助言者 AI と議論者 AI には特に問題になるだろう。また、本稿ではバイオエンハンスメントと比較した場合の利点は自由や自律性の尊重であるとしたが、総合的に比較して道徳的 AI エンハンスメントが優れているということまでは主張できていない。今後は、本稿で扱えなかった批判から道徳的 AI エンハンスメントを擁護し、また積極的に支持する議論を検討したい。

謝　辞

本稿は JSPS 科研費 JP22J21160 の助成を受けたものである。本稿の執筆にあたり、北海道大学の宮原克典氏には原稿を読んで頂きさまざまな助言をいただいた。また、本稿のもとになる発表を行った哲学若手研究者フォーラムで複数の方々からコメントを頂いた。お礼を申し上げる。

参考文献

Bang, Yejin, Lee, Nayeon, Yu Tiezheng et al. (2022)“AiSocrates: Towards Answering Ethical Quandary Questions,”*arXiv preprint arXiv:2205.05989*.

Braun, Matthias, Hummel, Patrik, Beck, Susanne, and Dabrock, Peter (2021)“Primer on an ethics of AI-based decision support systems in the clinic,”*Journal of Medical Ethics*, Vol. 47, No. 12, pp. e3–e3.

Danaher, John (2018)“Toward an ethics of AI assistants: An initial framework,”*Philosophy & Technology*, Vol. 31, No. 4, pp. 629–653.

DeGrazia, David (2014)“Moral enhancement, freedom, and what we (should) value in moral behaviour,”Vol. 40, No. 6, pp. 361–368.

Enoch, David (2014)“A defense of moral deference,”*The Journal of Philosophy*, Vol. 111, No. 5, pp. 229–258.

Freiman, Ori and Miller, Boaz (2020) “Can artificial entities assert?” in *Oxford Handbook of Assertion*, pp. 415–434: Oxford University Press.

Giubilini, Alberto and Savulescu, Julian (2018) “The artificial moral advisor. The ‘ideal observer’ meets artificial intelligence,” *Philosophy & technology*, Vol. 31, No. 2, pp. 169–188.

Harris, John (2016) *How to be Good: The Possibility of Moral Enhancement*, Oxford University Press.

Hernández-Orallo, José and Vold, Karina (2019) “AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’ 19, p. 507–513, New York, NY, USA: Association for Computing Machinery.

Hills, Alison (2009) “Moral Testimony and Moral Epistemology,” *Ethics*, Vol. 120, No. 1, pp. 94–127.

– (2020) “Moral Testimony: Transmission Versus Propagation,” *Philosophy and Phenomenological Research*, Vol. 101, No. 2, pp. 399–414.

Howell, Robert J. (2014) “Google Morals, Virtue, and the Asymmetry of Deference,” *Nouˆs*, Vol. 48, No. 3, pp. 389–415.

Jiang, Liwei, Hwang, Jena D, Bhagavatula, Chandra et al. (2021) “Delphi: Towards Machine Ethics and Norms,” *arXiv preprint arXiv:2110.07574v1*.

Jones, Karen (1999) “Second-Hand Moral Knowledge,” *The Journal of Philosophy*, Vol. 96, No. 2, pp. 55–78.

Jones, Karen and Schroeter, François (2017) “Moral Expertise,” in Tristram, McPherson and Plunkett David eds. *The Routledge Handbook of Metaethics*: Routledge.

Klincewicz, Micha l (2016) “Artificial Intelligence as a Means to Moral Enhancement,” *Studies in Logic, Grammar and Rhetoric*, Vol. 48, No. 1 (61).

Lara, Francisco (2021) “Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates?” *Science and Engineering Ethics*, Vol. 27, No. 4, pp. 1–27.

Lara, Francisco and Deckers, Jan (2020) “Artificial Intelligence as a Socratic Assistant for Moral Enhancement,” *Neuroethics*, Vol. 13, No. 3, pp. 275–287.

Lewis, Max (2020) “A Defense of the Very Idea of Moral Deference Pessimism,” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 177, No. 8, pp. 2323–2340.

Lieber, Opher, Sharir, Or, Lenz, Barak, and Shoham, Yoav (2021) “Jurassic-1: Technical Details And Evaluation,” Technical report, AI21 Labs.

Mullainathan, Sendhil and Shafir, Eldar (2014) “Freeing Up Intelligence,” *Scientific American Mind*, Vol. 25, pp. 58–63.

O’ Neill, Elizabeth, Klincewicz, Michal, and Kemmer, Michiel (2022) “Ethical Issues with Artificial Ethics Assistants,” in *Oxford Handbook of Digital Ethics*: Oxford University Press.

Savulescu, Julian and Maslen, Hannah (2015) “Moral Enhancement and Artificial Intelligence: Moral AI?” in *Beyond artificial intelligence*, pp. 79–95: Springer.

Slonim, Noam, Bilu, Yonatan, Alzate, Carlos et al. (2021) “An Autonomous Debating System,” *Nature*, Vol. 591, No. 7850, pp. 379–384.

Sparrow, Robert (2021) “Why Machines cannot be Moral,” *AI & SOCIETY*, Vol. 36, No. 3, pp. 685–693.

Vallor, Shannon (2015) “Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character,” *Philosophy & Technology*, Vol. 28, No. 1, pp. 107–124.

Wiland, Eric. (2021). *Guided by Voices: Moral Testimony, Advice, and Forging a 'We'*. Oxford University Press.

安藤馨（2020）「AI とその道徳的能力―AI による統治の正当性条件を巡って」，稲葉振一郎・大屋雄裕・久木田水生（編）『人工知能と人間・社会』，226–258 頁，勁草書房.

森岡正博（2013）「道徳性の生物学的エンハンスメントはなぜ受け容れがたいのか？：サヴァレスキュを批判する」，『現代生命哲学研究』，第 2 巻，102–113 頁.