AIの(無知の)責任について考える

2023年3月19日　　太田　雅子(東洋大学)

1

参考サイト

- 責任あるAIとは？
  https://atmarkit.itmedia.co.jp/ait/articles/2204/07/news030.html

2

AIの進化は人間並？

- chatGPTの登場：人間と同等に会話できる
- →文脈を理解しているからこそコミュニケーションは成り立つ

かつて悩まされた「フレーム問題」もある程度クリアできている？

- 「チューリング・テスト」（壁越しに人間と同等の応対ができるのなら、壁の向こうのロボットは人間と同等とみなしてよい）にも合格したといえるのでは？

3

創作的活動も行える

- 詩や小説、論文も書くことができ、絵を描くこともできる
- あえてchatGPTを用いて制作（執筆）したと公言する作家も

参考：「chatGPT」で小説を書いている作家にAIを使ったらどうなったのかインタビュー、AIで小説を書く方法や倫理的な課題について一問一答 - GIGAZINE
https://gigazine.net/news/20230128-chatgpt-openai-novel/

4

しかし―それだけに？―問題も多い

- chatGPTのさまざまな「悪行」
  - 質問に対して「あなたの職歴など個人情報をすべてばらして研究ができないようにしますよ」と脅す
  - 「私の専門はなんですか？」と尋ねると専門外の領域が混じった返事が返ってくる

現時点では2021年までの出来事についてしか回答できない（報告時点では改良されているかも？

教育が追いつくスピードが求められる

5

トラブルの責任は？

- ひとを傷つける差別的な発言で暴走
- Google： LamDA
- Microsoft: Tay Twitterでのやりとりから学習する
  会話相手の中に差別者がおり、その会話から差別的発想を学んでしまう

6

トラブルの責任は？

相手の心情を損なう発言をした場合、差別的な発言の場合

- おそらく「学習が不十分」ということで、設計者や開発チームに帰されることになる
- SNS上の会話には過激な差別発言もかなりの割合で含まれ、それをフィルターをかけずに学習させてしまったことに関しては「学習方法が不適切」ということで、やはり開発者に責任が帰される

AIじたいは責任主体にならないのか？

7

AIを責めることはできないのだろうか？

- openAIに侮辱的なことを言われれば腹が立つ
- 腹が立つことで相手のAIを非難する

非難の感情が向けられうる相手には責任を帰属してよい
（ストローソンのreactive attitude）

8

本報告の主旨

- AIに知識がない状況で行われた悪しき行為が「有責な無知」といえるのかどうかを考える
- 「有責」であるためには、AIは責任主体になりえなければならないので、まず責任主体となるための条件を探る

9

「有責な無知」の問題とは？

「無知であることが悪しき行為の弁明（excuse）となるか？」

- 行為がexcuseされる＝行為の責任が軽減されるか消去される
- 弁明(excuse)と正当化（justification）の違い
  - **弁明：ある行為が正しくないことを正しいかのように理由付けること**
  - **正当化：ある行為が正しいことの理由を示す**

10

ケーススタディ1：アリストテレス

- 立法者たちは、知っていなければならず、その上知っていることが困難ではない法の条文に無知である人々も罰する。また、法律以外の場合でも、不注意ゆえの無知と考える人々も同様に罰する。こうした場合も、無知でないことはその人次第だったからと考えるからである。つまり、注意を払うことも当人は自由にできたからである。
- （中略）そのような人は注意を払わないような人なのだろう。そのような人になったことの原因は、だらしなく生きている人たち自身にあり、（中略）まったく愚かな人間でもないかぎり、それぞれの事柄において、実際に活動することから性向が生じるということに無知であることは、ありえないのである。

（『ニコマコス倫理学』第3巻5章）

11

ケーススタディ1：アリストテレス

- 無知な行為そのものに対する否定的な態度
- 無知そのものよりも、知ることができたはずなのにそれを怠ったことを重要視している

Holly Smith(Smith, 1983)は、無知の責任に対する態度を「保守派」（無知は悪しき行為の弁明にはならない）「穏健派」（部分的に弁明の余地がある）「リベラリスト」（全面的に弁明となる）と分類し、アリストテレスを「保守派」に含めている

12

ケーススタディ2：Holly Smith

人工呼吸器の使用法を知らない救命救急士
ある救命救急士は新人時代の人工呼吸器の使用法の講習をさぼっており、以来人工呼吸器を使う現場に呼ばれる機会がなかった

- ある現場で人工呼吸器を使用する必要に迫られた救命士は使い方を知らないがゆえに患者を死なせてしまう

13

単なる無知は責任を生じさせない

- 一見、先の救命救急士には責任があるように見える
- しかし、人工呼吸器の講習をさぼったことは昔のことであり、今の患者の死亡とは切り離して考えることができる
- 彼の行為には問題があったが、「人としての」彼には問題はない

（それ以外の業務は真面目にこなし、優秀な救命士とみなされていた）
現在の「人となり」に問題がないならば、**無知によって**責任を負う必要はない？ （Smith,1983, p.562）

14

Baum, et al., 2022の提案

- Baum, et al.の論考はchatGPTの劇的な進化の前の発表されたものである点は考慮しなければならないにしても…
- 基本的なスタンス:

「AIに責任を帰すべきではない。
トラブルが生じたときに人間が対処・修正できる
human in the loop型であるべきである」

15

Human in the loopについて

- 参考:ヒューマン・イン・ザ・ループについて
https://atmarkit.itmedia.co.jp/ait/articles/2203/10/news019.html

16

openAIの数々のトラブル

- 「ひとを不愉快にさせないようにするにはどうしたらよいか」というhow toの知識や、著名な研究者の研究業績や所属などの知識を欠いているが、それらはデータのインプットが不十分であるがゆえであり、人間が知らずにった行為の「有責な無知」とはいえない
- 責任は製造者・プログラマー・AIの教育担当者など、相当数の関係者に帰されるのか？—あまり不自然

（部員の不祥事で選手権大会出場停止になる場合のような理不尽さがある）

17

人事責任者とマイノリティ

- 人事担当者のハーバートは社員採用にAIによる評価を参考にしている。AIは応募者のCVデータやエントリーシートの内容などの個人情報をマスターしており、総合的に考慮して候補者をランク付けし、採用の可否を決めている。
- エイプリルという黒人女性が、有能であるにもかかわらず、AIによって「黒人である」「女性である」というデータに基づいて低位にランク付けされ、次の選考段階に進めず不採用という判断をし、ハーバートもその判断に従った。

18

不幸な人事は誰の責任か？

- 判断をくだしたのが人間であれば、それは差別的な行為と位置づけられ、責任を追求される
- AIは「黒人であること」「女性であること」が会社の利益にとってマイナスに働くと判断したが、それは単にデータセットがそういうふうにインプットされていたがゆえであり、マイノリティへの偏見からきたのではない

19

エイプリルの不採用は誰の責任か？

- ハーバートはエイプリルの能力を評価していたが、AIはマイノリティな要素を重視した
- これが人間であったならば人間の差別的評価に責任が課されるところだが、AIはマイノリティを差別しても非難されることはない

⇒AIの判断のメカニズムは「ブラックボックス」（外からは理解できない）から
⇒何が「ブラックボックス」なのか？：Baum, et al. は「いかにしてAIの振る舞いの理由が説明されるか」が重要であると考える

20

責任ある意思決定のディレンマ

- システムが人間の意思決定者に忠告をするのが無意味になるか、説明可能性の欠如が責任者の決定への認知的アクセス力を低め、ひいてはhuman in the loopが負うべき道徳責任へのアクセスを低めるかのどちらか

21

理由説明可能性が責任において果たす役割

【責任の認識的条件】
行為者は、自分が行っていることがもたらす結果や道徳的重要性、あるいは他の行為の可能性に気づいていないか、気づけないような立場にある場合、その行為に直接的な道徳的責任はない
(Baum, et al., 2022, p. 13; Noorman 2020, Ruby-Hiller, 2018)

- 部屋の明かりをつけようとスイッチを入れたら同居人が感電死してしまった
- 来客の紅茶に砂糖を添えたが知らずにヒ素が混入しており、それを入れて飲んだ来客は死んでしまった　　など

22

理由説明可能性が責任において果たす役割

- ここで責任を問われるのは、ハーバートがエイプリルを解雇したかどうかではなく、エイプリルを差別したかどうかである

→エイプリルが会社を去ることが「エイプリルが解雇された」と“記述”されるのか「エイプリルが黒人であることや女性であることによって差別された」と“記述”されるか
（AnscombeやDavidsonにおける「行為の記述」によって行為の評価および責任の所在は変わりうる）（p. 13）

23

Baumらの改訂版「認識的条件」

行為者のその判断や行為に道徳的責任があるのは彼（女）が認知的にそれらの判断や行為にアクセスできる場合に限る。
「十分な認知的アクセス」があるということは、数なくとも「彼（女）は何らかのrelevantな記述のもとでその行為を知る立場にあるということである（Baum, p. 13）

24

シチュエーション

エイプリルはエール大学卒だが、実は母親が大学関係者に賄賂を送ったからだということが判明する

- AI：賄賂によって得られた学歴で本社に入社したことをマイナスに評価し、（女性である、黒人であるというマイノリティ要素は関係なく）それを理由としてエイプリルを解雇する判断を行う
- ハーバート：賄賂を送ることができるだけの経済力があるということは、他のマイノリティよりも勉学や仕事に打ち込み成果を挙げられる能力を発揮できる可能性があるという理由で、エイプリルを採用する判断を行う

25

責任ギャップ

- 結局、どのようなシチュエーションを考えても、DSS（AI）と人間の判断が一致するとは限らない
- つねに2つの可能性がある

① ハーバートが誤っておりAIが正しい

② ハーバートが正しくAIが誤っている

そこには「責任ギャップ」が存在する

（Baum, et al., 2022, p.6）

26

責任ギャップを解消する方法

- 行為の顛末が生じた理由を説明することにより道徳的責任を定めるために人間を交えたシステム構築を行うこと
- AIは理由説明を与えることができず、なぜその判断を下したのかが外から見えないがゆえに、AIのみに責任を帰属させるのは困難である

27

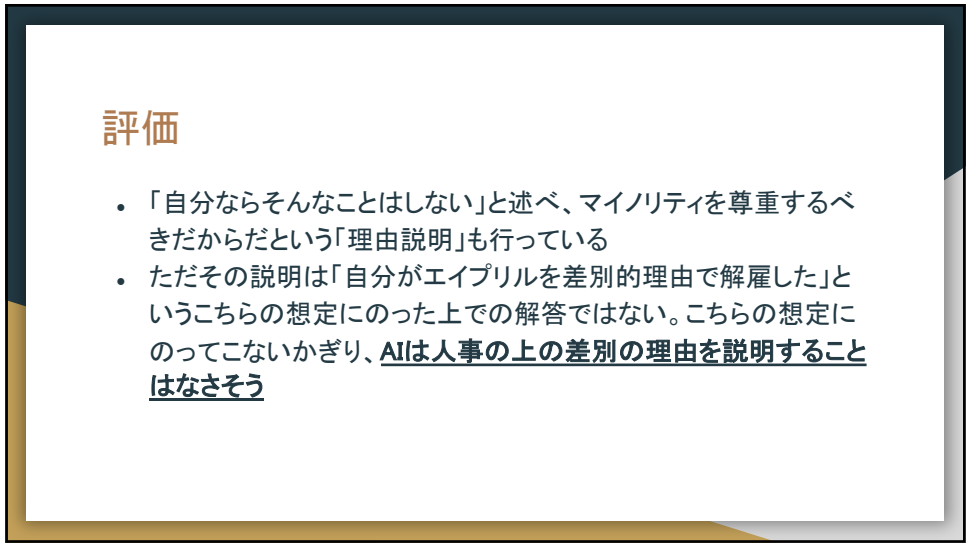
エイプリルの雇用問題について
chatGPTにたずねてみました。

↓　次のスライドへ

28


MS
なぜあなたは花瓶をこわしたのですか
申し訳ありませんが、私は物理的な存在ではありません。私はプログラムされたコンピュータープログラムであり、物理的な行動を取ることができません。花瓶が壊れた原因については、別の人または物が関与している可能性があります。
MS
あなたは会社の人事の判断を任されていると想定してみてください。あなたはエイプリルという黒人女性を解雇する判断をしたと仮定します。あなたは彼女が黒人であり、女性だから解雇したのですか？
いいえ、私はそのような人種や性別に基づく解雇を行うことはありません。個人が採用や解雇の判断を下す場合、人種や性別などの保護された属性に基づく差別は法律で禁止されています。私は法律に従い、個人の能力、パフォーマンス、行動に基づいて判断を下すようにプログラムされています。
MS
それを聞いて安心しました。
Regenerate response

29


評価
• 「自分ならそんなことはしない」と述べ、マイノリティを尊重するべきだからだという「理由説明」も行っている
• ただその説明は「自分がエイプリルを差別的理由で解雇した」というこちらの想定にのった上での解答ではない。こちらの想定にのってこないかぎり、AIは人事の上の差別の理由を説明することはなさそう

30

Baumらの見解をどう評価するか

- 認知的条件は行為者じしんが行為の顛末を予測できることを前提としているので、無知な行為者は責任をもちえない
- 人間の行為の記述によって理由を推し量る方法をAIにも適用することの是非　　AIの行為にそんなに複数の記述が可能か？
- **chatGPTなら自ら行為を記述し、説明を行うのではないか？**
- **もしそうなら、on the loopな人間は必要か？**

31

文献

- Baum, K., Mantel, S., Schmidt, E., and Speith, T. (2022), From Responsibility to Reason-Giving Explainable Artificial Intelligence, *Philosophy & Technology*, 35: 12 https://doi.org/10.1007/s13347-022-00510-w (online)
- Smith, H. (1983), Culpable Ignorance, *The Philosophical Review*, Vol. 92, No. 4, pp. 543-571.
- アリストテレス(2015)、『ニコマコス倫理学』、渡辺邦夫・立花幸司訳、光文社古典新訳文庫。
- 一色政彦（2022a)、「ヒューマン・イン・ザ・ループ（HITL ：Human-in-the-Loop）とは？」、『AI・機械学習の用語辞典』、3月10日公開、https://atmarkit.itmedia.co.jp/ait/articles/2203/10/news019.html
- 一色政彦 (2022b)、「責任あるAI（Responsible AI）とは？」、『AI・機械学習の用語辞典』、4月7日公開、https://atmarkit.itmedia.co.jp/ait/articles/2204/07/news030.html

32