

人工知能の倫理の現状： 2023秋

村上祐子 Ph.D.(Philosophy, Indiana University)

立教大学
大学院人工知能科学研究科
教授

産総研2023

AIの倫理

道徳的AIの可能性

道徳的AIとは何か？

AIは行為主体になりうるか？

道徳的行為主体とは何か？

AIによる社会的問題の解決？

道徳的判断のAIへの委任は可能か？

AIの概念形成

AIによる評価項目の設定

AIによる問題設定

科学技術の倫理 vs 科学技術者の倫理

科学技術の倫理

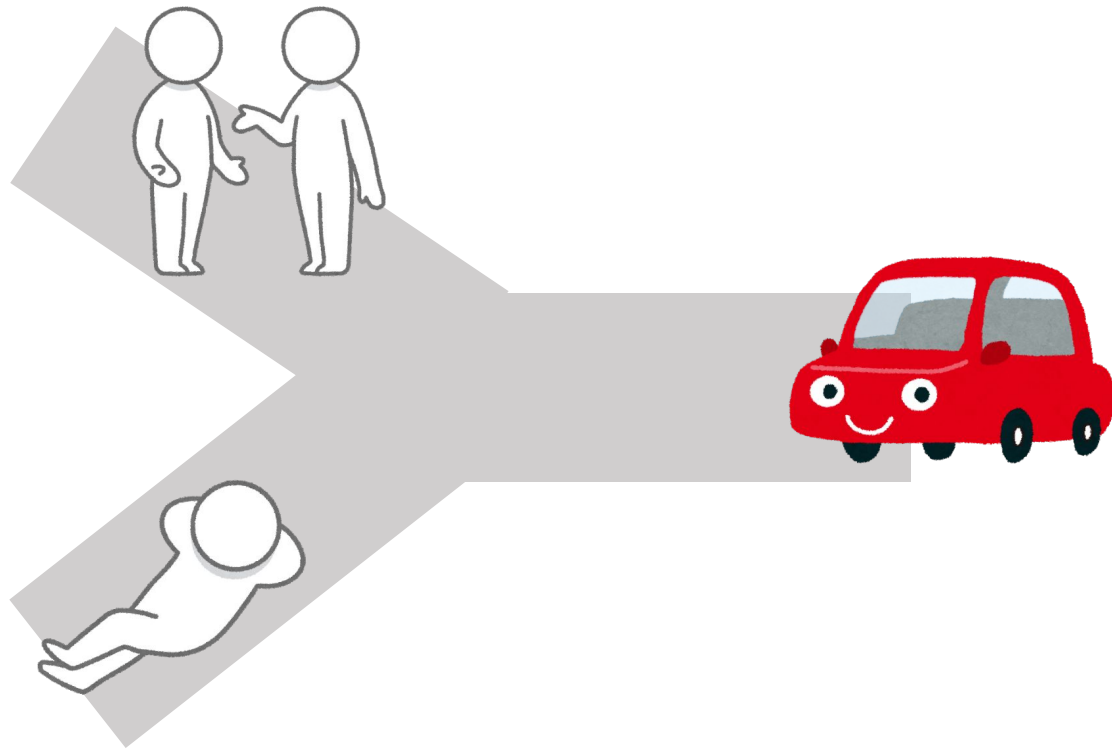
科学技術の成果が社会実装される際に発生する軋轢を処理するプロセスのうち、法整備の前段階となるのが科学技術の倫理の領域。つまり、科学技術の社会的課題を記述し、規範を定式化、その後法学者・法実務者が(1)既存法の適用可能性を議論(2)適用不能なケースを処理するために倫理規範を踏まえて法整備する。

科学技術者の倫理

科学研究・技術開発を倫理的に行うための規範整備。ビジネス倫理(職場環境における倫理)・専門家の倫理(専門分化社会における業務委任を円滑に行うための倫理)とオーバーラップし、研究開発コミュニティが自律的にふるまうための規範を整備し、実現させることが課題となる。

だいたい学振e-ラーニング教材で扱われている内容

人工知能にトロッコ問題は解けるか？



AI倫理？

「トロツコ問題はAIには難しい」

という問題設定自体が課題の矮小化



倫理学でトロツコ問題が提示された経緯

3つの倫理的立場: 倫理的に良い行為とは ...

- 功利主義: その帰結ができるだけ多くの利害関係者をできるだけ幸福にする行為
- 義務論: 義務を果たす行為
- 徳倫理学: 徳のある人にふさわしい行為

単一の利得関数に基づく功利主義的判断が不適切な例としてのトロツコ問題

トロツコ問題の現実的解決としての二重結果論:
現場の行為者の葛藤を解決するための妥協点

- 行為者が責任を持つのは、意図された帰結のみ
- 副作用とも言える予見された帰結については責任が問われない

資格職＋専門家倫理の対象

無資格の人が持っていない知識やスキルをもとに、社会から専門的な業務を委託される

そのときに専門家としてやっていいこととやっていけないことを踏まえていなければならない

代表例：医師。ヒポクラテスの誓いに始まり精緻化された倫理規範を順守しなければ医師会から追放、場合によっては資格はく奪

★情報学の場合、専門職としての倫理が完全に整備されているとは言えない一方で、社会への影響は甚大と考えられる：早急な法整備

メタ研究：

[Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review | Health Disparities | JAMA Network Open](#)

「機械学習技術に基づく介入」のある無作為化臨床試験のほとんどで、既存の臨床試験報告ガイドラインであるCONSORT-AIへの遵守が無い

Moral Machine



By Grendelkhan (Own work) [CC BY-SA 4.0 (<http://creativecommons.org/licenses/by-sa/4.0/>), via Wikimedia Commons]

Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

[Start Judging](#)

Moral machineの問題

個人属性に基づくシステムの教育データ収集

被験者への心的負担

→研究倫理事前審査を通す必要があった

トロツコ問題の哲学的意味のはき違い

→そもそも選べないのがポイント。不可能な課題は無意味

ドイツ倫理評議会

ドイツでは倫理評議会法を根拠法とするDeutscher Ethikratが構成され、研究開発の成果が個人・社会に及ぼす影響を分析し、政府・議会に対し勧告を行っている。構成員は哲学・倫理研究者、宗教者(カトリック・ユダヤ教)、医師、生命科学など理系研究者26名。

基本方針：法秩序維持＞集団的最大利得

法秩序維持＝法制度への信頼維持：守れない法はつukらない。現場の人に倫理的に不可能な選択を強いないように、あらかじめ制度・運用規則を準備する。

- 倫理的に不可能な選択の例：トロツコ問題、ECMO配分...

新全体主義(New totalitarianism)への危惧

全体主義: 全体の利益のため個人の権利を侵害することが正当化される

AI/DSが「全体の利益の最大化」のみを目指し、個人の権利を侵害するのであれば全体主義に該当する.

倫理的活用のためには、個人の権利を保護する仕組みが必要.

EU

人種的若しくは民族的な出自、政治的な意見、宗教上若しくは思想上の信条、又は、労働組合への加入を明らかにする個人データの取扱い、並びに、遺伝子データ、自然人を一意に識別することを目的とする生体データ、健康に関するデータ、又は、自然人の性生活若しくは性的指向に関するデータの取扱いは、禁止される。

EU デジタルサービス法 2022年10月4日

[DSA: Council gives final approval to the protection of users' rights online - Consilium](#)

それに至る議論

[Governing data and artificial intelligence for all: Models for sustainable and just data governance | Panel for the Future of Science and Technology \(STOA\) | European Parliament](#)

Nature 2022

Nature (2022-06-14) | doi: 10.1038/d41586-022-01607-0 | [Research must do no harm: new guidance addresses all studies relating to people](#)

シュプリングー・ネイチャーの編集者は被験者を要する研究に限らず、ヒト集団に関連する全ての研究において、研究により危害が及ぶ潜在的な可能性を検討するよう、論文著者に要請します。

Nature 2022

Nature (2022-06-14) | doi: 10.1038/d41586-022-01607-0 | [Research must do no harm: new guidance addresses all studies relating to people](#)

研究に直接参加していないヒト集団に影響が及び得る結論を示した学術研究の利益と危害に関して、言及していないのが一般的です。この学術研究の例としては、人々が汚名を着せられ、差別されるだけでなく、人種や性別、同性愛者であることなどを理由に差別を被るような状況を招き得る研究が挙げられます。特定の集団の社会的特性だけを理由として、その集団の人権を損なう行為を正当化するために、これらの研究が利用されるかもしれないのです。

データ取得対象の選定

社会的に弱い立場にあるもの＝「例えば、判断能力が十分でない者や、研究が実施されることに伴う利益又は実施されることを拒否した場合の不利益を予想することによって自発的な意思決定が不当に影響を受ける可能性がある者など、経済上又は医学上の理由等により不利な立場にある場合を指す」

「研究対象者の選定に際して、「社会的に弱い立場にある者」と考えられる者を研究対象者とする必要性について十分に考慮することも「特別な配慮」に含まれる。」

国による監視・制御をどこまで認めるか？

いくつかの異なる回答がありうる

- 大多数の国民の安全を守るためであれば必須であり、一部の国民や国民以外の人への不利益は国益のためならやむを得ない
- 全ての人には個人の自由があり、それを守るためには国による監視には制約を加えなければいけない

あなたはどのように考えますか？

欧米圏での顔認識技術の公共の場所での使用禁止

アップル、取引業者による従業員顔認証データ収集を禁止

ヨーロッパは戦略的に「倫理」を使おうとしている

RRI: Responsible Research and Innovation

日本経済新聞 完全自動運転へドイツが法改正 人命をてんびんにかけず

画像認識・顔認証技術

マーケティング等に期待されてきた

2019年～ 公安への情報提供への危惧

- サンフランシスコ 2019年5月
 - 公安監視カメラで顔認証使用禁止
- カリフォルニア州 2019
 - 2020年1月から3年間公安監視カメラで顔認証使用禁止

顔判別・認証技術

2020-2021 公安以外にも、店舗・レストラン・ホテルを含む民間サービスでの顔認証技術禁止拡大

- [サンフランシスコ](#) 2020年5月14日
- [ボストン](#) 2020年6月24日
- ニューヨーク州 [2020年7月24日 2年間禁止](#)
- [ポートランド](#) 2020年9月9日

その後部分規制にとどまる方向 [米国で一時高まった「顔認識規制」の動き、その後を追う](#)

ヨーロッパ: 顔認証技術規制

2019 スウェーデン学校顔認証がGDPR違反で罰金 同意が権力不均衡で適法ではないとされた <https://t.co/Tk8s5yYRmg?amp=1>

2021 スコットランド [顔認識カメラを学食に導入との報道、決済をスピードアップへ--英国の学校9校で](#) 保護者の約97%が、欧州連合(EU)の一般データ保護規則(GDPR)のガイドラインに従って、このシステムで自分の子どもにスキャンを適用することに同意

2023 AI act:

[The 2024 Paris Olympics: AI Mass Surveillance Under the Upcoming EU AI Act | OHRH](#)

日本における顔認証技術の規制の緩さ

「GDPR準拠」を今後も認められるか？

スイス

[Switzerland Publishes Adequacy List of Countries Receiving Personal Data Transfers | Morgan Lewis - JDSupra](#)

EUとほぼ同じだが、日本と韓国が準拠国から除外されている

設計による差別～障害の社会モデル

古典的な例:

高さが低いトンネル→バス不可、自家用車必須→車がない貧困層の排除

障害の社会モデル: 障害や病気は本人の身体状態だけに還元されるのではなく、それによって社会参加が困難になることが本質

バリアフリー通路の欠如または大幅な迂回の強制

★データサイエンスでもさまざまな場面で設計による差別が発生

Discrimination by data (not data discrimination)

A HUMAN RIGHTS-BASED APPROACH TO DATA

社会制度に内包される差別をデータ設計・データ収集の場面で再導入・強化
解決には社会制度側を変えなければいけないこともある.

しかし技術の専門家として差別に加担していることは意識する必要あり.

アルゴリズム的偏見

- データ設計
- データ収集
- システム運営

データ設計での問題

例: 入国管理

Identity management: Name structures and how to record them in the Global Case Management System (GCMS)

日本人通称使用者は併記パスポート名にかっこが入っているが、各国のパスポートデータ規定にかっこ使用が定められていないためにトラブルが発生している。

接種証明書アプリ: 海外用は通称に未対応 (国内用のみ対応済)

名前の文字数に上限がない国でのパスポートの場合には入国書類フォームの文字数制限に達してしまうことがある。

データ項目設計による差別への対策

データ項目そのものの改訂が必要

コストを理由とする不作為はもはや合理的とはみなせない

パスポートに男女以外の性別 X記載: 少なくとも以下 15か国

アルゼンチン、オーストリア、オーストラリア、カナダ、コロンビア、デンマーク、ドイツ、アイスランド、マルタ、オランダ、ニュージーランド、パキスタン、インド、ネパール

2021年10月 アメリカ

[Issuance of the First US Passport with an X Gender Marker - United States Department of State](#)

医師免許でのXジェンダー: India, Ireland, Germany, Austria, Colombia

データ収集での問題：狭義のアルゴリズム的偏見

顔認証

アメリカのシステムでは白人男性のデータが多い

→非白人・非男性のエラー率が高い

中国のシステムではアジア人データが多く、アジア人の顔認識で判別力が高い

→戸籍の種類(民族)についても顔認識で判別→民族同化政策への利用

データ収集における人権侵害

クラウド・ソーシング：労働搾取と考えられ始めている

例：クラウド・ソーシングで収集したデータに基づくデータサイエンス研究論文の人権侵害理由によるリジェクト（データ生成タスクに対する謝礼額不適當、労働搾取相当）

アルゴリズム的偏見: 社会的偏見の固定

トロツコ問題を回避する制度設計の必要性

複数の価値観のいずれかを選ばなければいけないような状況に現場が追い込まれるような状況を回避するように制度設計しなければならない.

実例

ドイツ→EU: 自動運転車をはじめとするデータ使用で個人の属性(性別・民族ほか)を判定基準に参照することを禁止

ポイント: 技術的に解決しようとしてはいけない問題が存在する.

まとめ

人工知能技術は私たちの生活を支えるものになっていく

AI倫理は発展途上：技術的には解決できない(トランスサイエンス)課題

どう使うかは人間が考えなければいけない

人間が選択に困ってしまうような状況にならないような規制が必要。

データを大量に集めて「学習させる」：既存社会の偏見が入り込む＜要矯正

排除技術はまだない(どころか、おそらく不可能)

とりあえず問題のあるデータは使用禁止にしたのがヨーロッパなど。