特集
Special Feature

［デジタルアーキテクチャデザイン］

6 デジタル社会における AI ガバナンス

応用 —倫理と法制度—

中川裕志 | 理化学研究所・革新知能統合研究センター

AI ガバナンス

ガバナンスを統治，規制などというと難しいが，組織や技術に限定した場合，それを本来の目的に適応するように使う方法あるいは改訂していく方法と考えられる．AI は技術であり，その目的は人間社会の利便性や質の向上，さらには個人の幸福に資することだとすれば，AI ガバナンスの意味はかなり明確化する．とはいえ，AI は常に同じ動きをするツールとは言い切れない．その動きは学習に使ったデータに依存する．また，利用しつつ構造を変える強化学習による分類器や予測器の更新が起きるため，同じ入力に対して同じ結果が返されることが保証されているわけではない．つまり，常に利用者の予測通りに動くというわけではないので，AI ガバナンスの意味は複雑になる．AI ガバナンスには次の 2 種類があるので節を分けて説明する．

AI をガバナンス

AI をガバナンスするとは，人間が AI を道具として使える状態を目指すことであり，具体的行動としては道具としてコントロールしきれるような AI を設計，開発することである．この議論はシンギュラリティや汎用 AI（AGI）が出現し，人間がこれを制御できなくなるという危惧が広まり，人間が AI を完全に支配できる，つまりガバナンスできるように AI を設計すべきであるという議論が 2010 年代に起こったことによる．これは初期の AI 倫理指針[1), 2)]に色濃く反映されている．だが，この感情的とも受け取れる動きが収まると，後に述べるようにより現実的な議論が展開されるようになった．

AI でガバナンス

人間／組織がほかの人間／組織をガバナンスするために AI を利用するという AI の使い方がある．具体例としては，顔認証システムという AI 技術を用いるガバナンスがある．チェーン店のある店で捉えた万引き犯の顔を全チェーン店に送り，顔認証システムによって入店者の中にこの万引き犯を発見したら警戒をするという万引き防止に使うことは，組織が人間を AI でガバナンスする例である．

別の例としては，検索エンジンや SNS のプラットフォームにおける利用者の行動データを AI に与えて，その利用者にフィットする広告を推論して表示するターゲット広告がある．これは，プラットフォームが利用者をガバナンスするために AI を用いるケースである．

最近議論を呼んでいるケースとして公共の場所における顔認証システムの利用がある．顔認証システムはすでに多くの場所で使われ，たとえば空港ではパスポートの顔写真とカメラから入力した本人の顔が一致することを顔認証システムで行い，省力化と本人確認の正確さに寄与している．

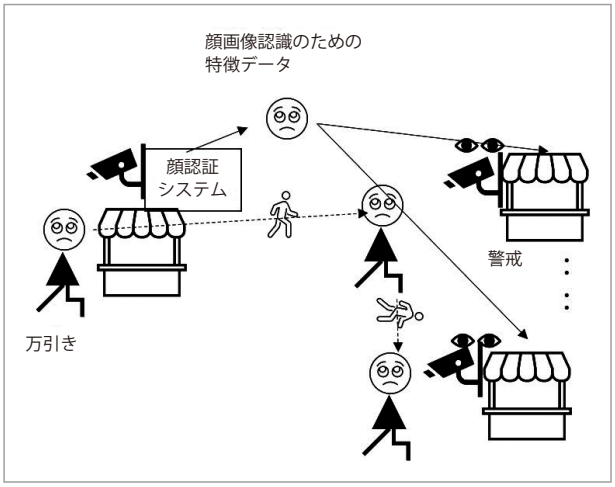
これは有益な利用法だが，上で述べたような問題のある使い方もある．すなわち，図 -1 に示すようなコンビニエンスストアのチェーンの全店舗で万引き防止のために店内カメラを設置して顔認証システムを導入した場合である．店内カメラをチェックした店員がある人物が万引きしたと認識し，顔認証システムに撮影されていた顔画

© 2021 Information Processing Society of Japan

像を用いて，この人物を識別できる特徴データを抽出する．この顔特徴データを同じ系列の他の全店舗に配布する．他の店舗では，入店する客の顔画像を常時撮影して顔認証システムに入力している．顔認証システムがすでに配布されている万引き犯だと識別した場合，すぐに店内の店員に警報を出し，その人物の行動に注意するようにしたとしよう．

この方法は万引きの再犯防止には効果的だと思われるが，出来心で１回だけ万引きした学生がどこの店舗でも常に店員に疑い深く見張られたり，店員に付きまとわれるのはやりすぎではないかという意見もあるだろう．もし最初の万引き認識が誤認だったとすれば，誤認された人にとってはとんでもない濡れ衣を着せられたことになってしまい，社会的にも精神的にも被害甚大である．顔認証システムの認識性能は現在でも 100% ではないため誤認識される可能性も残る．このような顔認証システムの利用法は果たして倫理的かどうかは問題視されている．

元来，顔画像はセンシティブな個人データだと考えられている．顔認証システムが多数の公共の場に仕掛けられていると，顔認証システムを仕掛けた側，たとえば国の機関は個人の居場所の履歴を把握できる．EU では居場所の履歴は個人データとして保護対象と考えられている．日本に住んでいると分かりにくいが，次のような例を考えるとよい．ある人が，しばしば特定の



■図 -1　顔認証システムによる万引き防止

宗教の施設に出入りすることが顔認証システムで分かったとしよう．すると，この人はその宗教の教徒であることが推定される．この宗教が嫌われている社会では，この宗教の教徒であることが分かると差別や迫害を受けかねない．これは信教の自由を阻害することになってしまうわけである．こういった事情があるため，EU では顔認証システム公共の場所での利用を制限する方向で議論が進んでいる．

もちろん，警察にとって顔認証システムは犯罪捜査には有力な手段である．つまり，顔認証システムには，プライバシ保護と社会に安全維持という相反する目的がある．したがって，顔認証システムで収集された顔画像データ，ないしは顔認証のための特徴データの使い方は法律によって厳格に規律されなければならない．社会を AI でガバナンスする場合に起きる典型的な問題点が顔認証システムに表れているので，AI の利用法について考える場合に思い出してほしい具体例である．

公平性

個人への影響が強いケースとしては人事における AI の利用がある．たとえば，会社が求人を行うときに採用部門が求職者を選考するための支援ツールとして AI を使用する場合である．さらに，社内の人事評価での AI 利用もある．

より大きなスケールでは，国民全員を至るところに張り巡らされた顔認証システムを使ったり，常時スマホの位置情報を追跡したりして個人データを収集する．収集したデータを使って個人をスコア付けして，スコアに応じた行政サービスを行う国もある．国民総スコア付けの是非は問わないにしても，その根拠とスコア付けのためのデータやアルゴリズムは少なくとも公平なものでなければならない．しかし，公平性とはそもそもどのように定義するのだろうか．

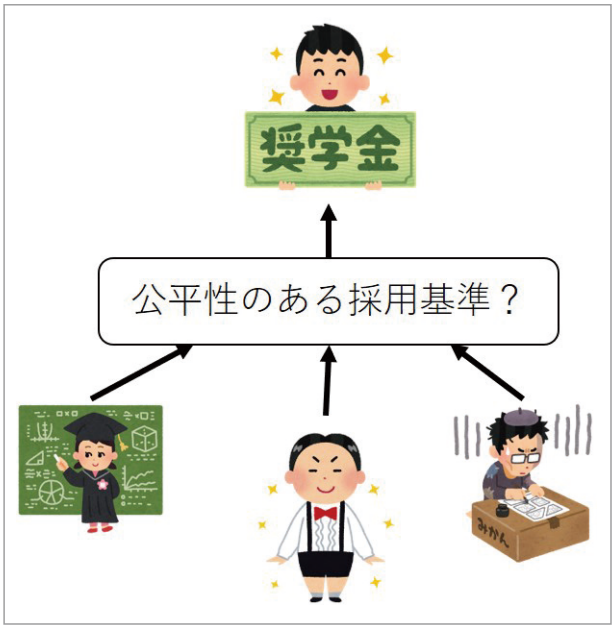
図 -2 に示すような学生に奨学金を与える方法を例にして考えてみよう．親の収入が低い順に与える方法は，経済的な平準化には寄与するだろう．しかし，才能ある学生やまじめに学業に励む学生には不利かもし

© 2021 Information Processing Society of Japan

れない．学業成績の高い順に与える方法もあるが，これでは経済的苦境にある学生を援助することにならない．また，性別や人種で差が大きいとき，これまで差別的に扱われていた性別，あるいは人種に特別枠を設けて公平化する，いわゆるアファーマティブアクションがある．しかし，これによって才能やスキルの高い学生が排除される逆差別も起こりがちである．

つまり，公平性とはどの要素を平等化したいのか，どういう社会を望むのか，などの外部から与えられた目標に依存して定義されるものである．もちろん，性別，経済状態などの公平化したい要素を定め，その要素に因果的に関連する要素も交絡因子として統計的に求めることができれば，公平性の数理モデルが確定できる．そのモデルに限った上で公平性を最大化することはできる．ただし，ガバナンスという観点からすれば，望むべき社会の在り方を議論しなければならない．よって公平性の定義はその時代における倫理観，価値観に依存するということを念頭に置かなければならない．

このような価値観，倫理観を AI という対象にどう適用するかというテーマが，最近の AI 倫理のトピックになってきている．



■図 -2　公平性のある採用基準とは

AI 倫理の概要

人間の行動を対象にした倫理と AI 倫理は大きく異なる．AI 倫理の重点項目は当初，人間に敵対する可能性のある汎用 AI を制御すべきというものだったが，2017 年から 2018 年あたりを境にして，より現実的な項目に重点が移った．すなわち，文献 2) 以降に公開された主要な倫理指針では以下のような項目が重視されている．

well being，人権，安全性，教育，AI 兵器，法的位置づけ，公平性，差別，透明性，アカウンタビリティ，トラスト，悪用，プライバシ，AI エージェント，独占禁止と国際協調，政策への反映

AI が人類の福祉や幸福に資するという否定しがたい目標を持つ以上，well being は当然の項目となる．個人単位での人権の保護も当然必要な項目である

また，人間に意識的に敵対しなくても，AI の行動が人間にとって安全であること，さらにマルウェアやボットに乗っ取られて不適切な動きをすることは絶対に避けなければならないので，安全性も当然確保されなければならない．

AI が身の回りに溢れ，社会活動の隅々まで影響するようになってくると，一般人であっても AI の機能などについては勉強する必要があるという意味で，AI に関連する内容を教育することは重視される．しかし，このことは専門家以外の方に AI 関連のプログラムの読み書きを求めるものではないし，実際それは不可能である．むしろ，AI が扱っている対象，現象の意味内容を理解するために必要な数学的基礎を身につけることを期待している．また，技術者や開発者の側に対しては，AI が使われる場面を想定して，開発方向の利害得失を事前評価できる知識として，法令，経済，地政学，文化などへの理解を期待するところである．

AI 兵器の禁止は道徳的に必要なことであるし，それ自体に反論も少ないと思われるが，文献 1) 以外では意外なことにあまり表立って扱われていない．文献 2) では，AI 兵器について許容できない AI 兵器の定義

© 2021 Information Processing Society of Japan

を議論している．政治的にセンシティブな項目である上に，仮に禁止を打ち出しても，CCW[☆1]の交渉の場では，図 -3 に雰囲気を示すような総論賛成，各論反対の状況で禁止へ向けての実効性が期待できないという事情であろう．

AI の法的位置づけもまた微妙な論点である．AI に自然人と同じ人格権を与えるという議論は非現実的であるにしても，部分的な人格や法人格を与える議論は AI の利用目的に応じて検討すべき論点である．たとえば，AI ツールを使った芸術作品の創作で，人間の寄与が「XX 風な絵を YY を題材にして創作せよ」程度に大雑把だったとき，結果として得られた作品に AI にいかなる権利を与えるか？　これは法的も難問だが，上記の法人格付与などを絡めて，いずれ議論されるであろう．

アカウンタビリティ

AI を組み込んだシステムを使って出力された結果が，そのシステムの利用者にとって不本意であったときの対処法が AI の実用化において重要である．この論点にかかわる一連の概念が透明性，アカウンタビリティ，トラストおよび悪用である．これらの間の関係を図 -4 に示す．

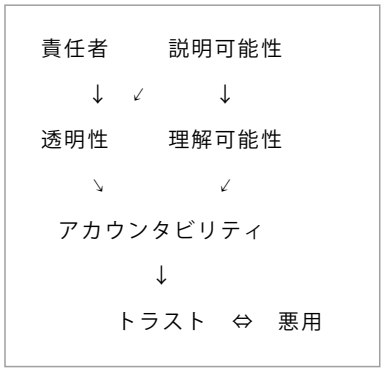
☆1 特定通常兵器使用禁止制限条約．

■図 -3　AI 兵器禁止のコンセンサスの困難さ

AI を組み込んだシステムから利用者にとって不本意な結果が出たとき，利用者はその結果が出てきた理由の説明を要求するだろう．そのときに AI が理由の説明を生成できることが説明可能性である．ただし，その説明は開発者や専門家以外にも分かる理解可能性のあるものでなければならない．利用者が不利益を被っている場合，説明を受ければ終わりというわけではなく，不利益の補償をする責任者と，説明を合わせて当該利用者に提示することが必要であり，これが AI システムの利用者から見た透明性に相当する．その上で，理由に納得し，補償も受けられれば，AI システム側では責任を果たしたことになるが，これがアカウンタビリティである．

形式的にはこれで十分に見えるが，一般の利用者にとってはアカウンタビリティの概念およびその背後関係は理解容易なものではないことが多いだろう．特に AI の説明可能性に技術はいまだに一般人を理解できる説明を作るという観点からは技術的に不十分である．結果として，アカウンタビリティを，内容を理解せずに信用するという意味のトラストで置き換える方向になる．トラストは心理学的ないし社会的概念である．すなわち，組織 W が出してきた Z という結果の正しさが証明されていないが，それまでの組織 W の行ってきた行動実績から，正しさが証明されていない Z を正しいとみなしてしまおうというメンタリティがここでいうトラストである．

ここでのトラストは従来の工学的ツールがトラストできるという技術的トラストとは違う．技術的なトラストは，



■図 -4
アカウンタビリティ等諸概念の関係

© 2021 Information Processing Society of Japan

ツールが同じ利用条件なら常に同じ結果を出し続けるということを意味する．また，このことは故障しないことも意味する．すでに述べたようにAIは正しい動作をしている場合であっても，学習や利用環境に依存して異なる結果を出してくることがある．よって，技術的なトラストとは違う意味でトラストを捉えなければならないことに留意されたい．

悪用

AIの誤用や悪用はAI倫理の点からはもちろん好ましくないが，法律に抵触しそうなケースも現れてきているので，それらについて説明する．

誤用は間違った利用であり，典型的には利用者がAIの性質を十分に理解せずに使っているケースである．たとえば，IT製品は適切に更新しないとセキュリティ上の危険性があるが，AIを含むIT製品でもまったく同様である．一方，悪用は恣意的に特定ないし不特定の人物や組織を攻撃する目的でAIを利用する場合である．たとえば，図-5に示すように，上司が部下の若い社員に無理難題を押し付けるとき，AIが出した結論だから従えというケースが考えられる．このようなケースに対して文献2)は以下のような指針を提案している．

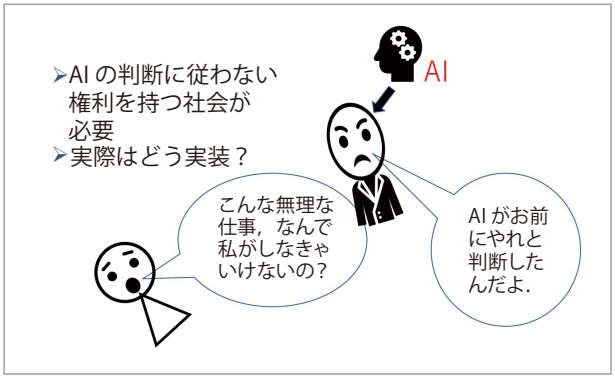
a) AIの出してきた結果に疑問があり納得できない場合，内部告発を制度的に保証する．これによって組織内のパワハラや不正行為を露見しやすくする．

b) AIが悪用された場合の救済策を立法化しておくことが必要である．つまり，パワハラ等で受けた被害を救済することは社内だけではうやむやにされがちなので，国の法律とすることによって強制力を持たせるべきである．

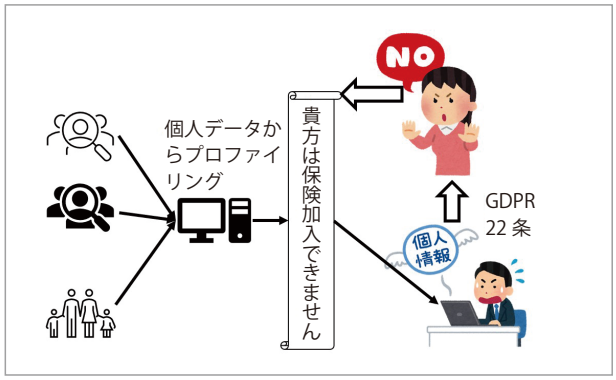
c) AIが悪用された場合の救済策としては，被害者や訴えた人が経済的に不利益を被るケースが起こりそうである．つまり泣き寝入りや萎縮して訴えないという状況が想定される．この状況を救うために保険制度を整備することが重要である．

プライバシ

プライバシの保護は多数のAI倫理指針で重視されている．多くの個人データがインターネット経由で流通してしまう現在においては，第一に個人データをインターネットに流さないという個人の行動規範に頼ることには無理がある．購買履歴，IoTから収集される個人の健康データ，家の状況データ，AIスピーカの入力など知らないうちに大量の個人データが企業に収集されている．したがって，そうやって収集された個人データが個人の不利益になるような使い方をされないという個人情報保護法，あるいはEUの一般データ保護規則（GDPR）のような法律による規制が重視されている．法律といっても，企業へのプライバシ保護順守だけを要請するだけではなく，個人データのトレースなどを行える権利，すなわち自己情報コントロール権，あるいは情報自己決定権という概念が重視されている．図-6に



■図-5　上司によるAIの悪用



■図-6　GDPR22条の概念

© 2021 Information Processing Society of Japan

概念を示す GDPR22 条 1 項の「個人データをプロファイリングした結果から機械的に得られた個人に関する決定に服さない権利」は，上記の権利の一例である．

とはいうものの，この権利の実現を機械的に支援する AI の説明可能性は技術的に不十分であることはすでに述べたとおりである．

個人の側としても，事業者側の倫理に期待する以外に，個人データの利用を自身で管理する AI エージェントの概念と実装が将来技術として期待される [2]．

政策

AI 倫理指針の技術的な指針が収束するにつれて，政策的な指針が強調されるようになってきた．独占禁止と協調は多くの指針で建前的に触れられているが，政策への反映に関しては国ごとに異なる主張が鮮明になってきている．以下で米国と EU の対比を紹介する．

文献 3) では AI 関連産業を EU 域内でどう扱うかという政治的方針が書かれ，EU 域内の AI 関連産業の保護育成を主張している．政治的方針だけに EU の保護主義的な主張も散見される．たとえば，長いサプライチェーンを持つ AI 製品は EU 域外の素材が内部で使われている場合であっても EU 基準で審査，また不具合は EU 域内で改善せよとしている．

これと対照をなすのが，文献 4) である．この指針の名宛人は AI 製品開発企業である．そして，便益より大きな害をもたらすという意味でのリスク評価と管理を重視している．リスク評価が甘ければ社会全般からの支持を失うので事業者はそれに留意して開発せよと書かれている．このようにできるだけ自由に AI 開発を行うという方向性に将来性があるのではないだろうか☆2．

政策的問題は国内，ないし域内で考えることは主権の範囲のようであるが，人やデータの移動が頻繁な今日，国境を越えて重要な意味を持つ．IT プラットフォームが収集している個人データをその企業が使っている分にはプライバシ保護の点からは問題にならないと考えられる☆3．しかし，巨大 IT プラットフォームが個人データを大量に保持している状態で，ターゲット広告などが効果的に行えるという立場を利用して独占的地位で他社に不公平な取引条件を課するという問題は，独占禁止法上の問題として俎上に上がることもある☆4．一方で，国家ぐるみで個人データを独占的に管理し，民族的差別すら行うケースもあり得る．独占禁止と協調というテーマは，法制度や地政学的問題を孕んでいることを技術者としても十分に心得ておくべき時代であろう．

参考文献

1) Future Life Institute : ASILOMAR AI PRINCIPLES (2017).
2) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems : Ethically Aligned Design Version2 (2018).
3) EUROPEAN COMMISSION : White Paper on Artificial Intelligence A European Approach to Excellence and Trust, Brussels, 19.2. (2020).
4) USA Whitehouse. Guidance for Regulation of Artificial Intelligence Applications : MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (2019) .

(2020 年 10 月 26 日受付)

☆2 AI ガバナンスの関係から政策的動向も含めて調査し俯瞰した資料『我が国の AI ガバナンスの在り方 ver. 1.0』https://www.meti.go.jp/press/2020/01/20210115003/20210115003-1.pdf はこの分野の動向把握に役立つ．

☆3 Facebook が収集した個人データをケンブリッジアナリティカに渡して，米国大統領選挙に使わせたことが問題視された．しかし，米国では個別訪問は禁止されておらず，Facebook が収集した個人の政治的行動が個別訪問して効果がありそうな家を調べるのに使われても違法ではないし，Facebook の利用規約には，収集した個人データを第 3 者に使わせる可能性があることが明記されている．米国議会で問題視されたのは，この個人データがロシアにわたり，米国大統領選挙に介入されたのではないかという点であった．

☆4 米国司法省が 2020 年 10 月に Google を訴えた．

■中川裕志（正会員） nakagawa3@gmail.com

1975 年東京大学工学部卒業．1980 年東京大学大学院工学系研究科修了（工学博士）．1980 ～ 1999 年横浜国立大学工学部勤務．1999 ～ 2018 年：東京大学情報基盤センター教授．2018 年～現在：理化学研究所・革新知能統合研究センター・チームリーダ．AI 倫理などの研究に従事．

© 2021 Information Processing Society of Japan