特集

羽鳥 剛史 愛媛大学社会共創学部（環境デザイン学科）教授

AIアカウンタビリティの意義と課題

1. はじめに

Society5.0（超スマート社会）やスーパーシティ構想において、人工知能（AI）やビックデータが社会変革を導く中心的な技術として位置付けられている。アセットマネジメント分野においても、デジタルトランスフォーメーション（DX）の推進と共に、膨大なインフラデータをAI技術により分析・解析し、異常の検知、損傷の発見、劣化予測、修繕計画の立案など、維持管理の効率化・高度化に活かそうとする取り組みが本格化しつつある。AIやビックデータ等の先端技術は、財政の逼迫や技術者の不足といった厳しい条件の下でも、アセットマネジメントの高度化を図り、インフラ施設の適切な維持管理を遂行する上で不可欠な役割を期待されている。

スマートシティに関わる政策提言の多くは、AIやビックデータの積極的な活用と同時に、その開発・利用にあたり、人間中心的な価値に基づいて、市民に対する説明責任（アカウンタビリティ）の確保を求めている[1)]。特にアセットマネジメントは、市民の負託によって成り立っており、インフラ管理者は市民に対してその適切な管理・運営に関するアカウンタビリティを果たすことが求められる。しかし、インフラマネジメントに関わる意思決定プロセスにAI技術を導入することにより、市民に対するアカウンタビリティがより困難となる可能性が懸念される。第1に、AIアルゴリズムの判断は、その判断に至った理由や根拠を外部から理解することが難しいという「ブラックボックス問題」を抱えている。特に、機械学習や深層学習を用いたAIでは、開発者にとってさえもAIアルゴリズムがなぜその判断を下したのかを十分に説明することが困難となる。近年では、AI判断の根拠を明確化する「説明能力のあるAI（explainable AI）」の技術開発も進められているが、未だ発展段階にあり、AI判断の根拠を一義的に理由付けすることは難しいのが実情である。

以上の課題に加えて、第2に、AIを用いた意思決定は、そこに多くの関係主体が介在するという「多くの手の問題（problem of many hands）」を抱えている[2)]。AIアルゴリズムは、通常、数多くのアルゴリズムから構成されるシステムに組み込まれており、システム全体として機能を発揮する。こうしたシステムを簡略的に「AIシステム」と呼べば、AIシステムの設計、開発、照査、修正には、アルゴリズム設計者、データ保持者、政策立案者、サービス提供者、サービス利用者をはじめ、多くの関係者が関与する。さらに、AIシステムは、設計者等の設計・運用環境、利用者の使用環境、ソフト・ハードウェアの技術環境、法制度や社会規範を含む、異質な人的・物的ネットワークが錯綜する複雑な社会技術的な文脈に埋め込まれている。こうした状況においてシステムが下した意思決定に対して、その責任を同定することは容易ではない。

AIシステムのアカウンタビリティ問題は、近年の科学技術論やAI倫理学の中心テーマの一つであるが、その要件や枠組みについて統一的な見解に至っているわけではない。本稿では、AIシステムのアカウンタビリティ概念を概観すると共に、特にAIの開発・利用をめぐる「責任のギャップ」という観点から、AIシステムに関するアカウンタビリティのあり方や要件について検討し、AI技術を活用したアセットマネジメントへの示唆を導くことを試みる。

2. AIシステムのアカウンタビリティ概念

アカウンタビリティ概念に関して、会計学、社会学、政治学、経済学、心理学等、多くの学問分野において研究が蓄積されている[3)]。その定義や概念は多義的であるが、伝統的には、委託者と受託者との間の2者関係を前提として発達してきた。すなわち、アカウンタビリティの委託－受託関係は、「受託者Aが委託者Bに対して自己の行為Cを正当化する義務を果たし、もし委託者Bが受託者Aの正当化が不十分であることを発見した場合、一定の制裁を受ける可能性がある時、受託者Aは委託者Bに対して行為Cに関して説明可能である」という形式により定義される[4)]。近年の研究では、このアカウンタビリティ概念をアルゴリズムによる意思決定の文脈において捉え直し、アルゴリズムに関わるアカウンタビリティにおける意思決定者とその対象者の関係について次のように記述している[5)]。すなわち、「意思決定者は、アルゴリズムによる意思決定システムの設計と運用に関する理由及び説明を意思決定の対象者に提供しなければならない。意思決定の対象者は、この正当化が適切であるかどうかを判断することができ、適切でない場合には、意思決定者は何らかの制裁を受けたり、特定の意思決定の撤回や修正を余儀なくされる可能性がある」とされる。

現実のアルゴリズムによる意思決定は、特定の委託者と受託者の間で実施されることは稀であり、不特定多数の市民（委託者）から構成される社会の中で、複数の意思決定者（受託者）間の連携の中で実施される場合が一般的である。この点を踏まえて、アルゴリズムによる意思決定プロセスが社会の規範や手続きに則っているか否かという観点から、アルゴリズムに関わるアカウンタビリティを広義に捉える見方も提示されている。例えば、OECDのAI原則に拠れば、アカウンタビリティは「組織や個人が、自らの役割や適用される規制の枠組みに従って、設

計、開発、運用、または導入したAIシステムが、そのライフサイクルを通じて適切に機能することを確保し、自らの行動や意思決定プロセスを通じてこれを実証すること」とされている。また、ガバナンスと関連付けて、「法的・倫理的な義務、方針、手順、仕組みにコミットし、社内外のステークホルダーに倫理的な実行を説明・実証し、適切な行動が取れなかった場合には是正することを含む、ガバナンス構造に集約される一連のメカニズム、実践、属性」と定義するものもある[6]。

3. 責任のギャップ問題

AIシステムのアカウンタビリティ問題は、AIを用いた判断や意思決定に関して、人間の責任をいかにして担保できるかという問題と不可分に結び付いている。AIの判断や意思決定に対してどの程度まで人間が責任を負うことが出来るか、もしくは責任を負うべきかについては、様々な議論がある。その中でも、AIを用いた判断や意思決定が「責任のギャップ（responsibility gap）」を生み出す可能性が指摘されている[7]。責任のギャップとは、厳密に言えば、AIの判断が危害を及ぼし、その責任を問うことが適切であるにも関わらず、誰もその責任を問われない状況を表している。技術哲学者のデ・シオらは、責任概念の多義性を認めた上で、責任のギャップを1）過失責任のギャップ（culpability gap）、2）道徳的説明責任のギャップ（moral accountability gap）、3）公的説明責任のギャップ（public accountability gap）、4）能動的責任のギャップ（active responsibility gap）に大別している[7]。第1に、過失責任は、故意や過失による違法行為や他者への損害に対する責任を表すが、AIの判断が介在することにより、そのシステムの挙動に対する予測や統制が困難となり、過失行為に対する正統な理由が生まれたり、言い訳が許容されたりする可能性が考えられる。第2に、道徳的説明責任は、自らの行動の理由を他者に説明する義務を表しているが、AIシステムの不透明性や複雑性に起因して、その責任の所在が曖昧になる可能性が考えられる。冒頭で述べた通り、機械学習による判断はその設計者でさえもその判断の理由を説明することが困難となる場合が少なくない。さらに、AIシステムにおいて、設計者、行政、市民、データ管理者等、多様な関係者が介在し、関係者間の情報・データの複雑な交換が行われる。道徳的説明責任のギャップは、こうした状況において、AIシステムの全体的な挙動や個々の判断について、関係者の説明能力や理解能力が低下する状況を表している。第3に、公的説明責任は、公的機関が社会的な判断や意思決定の理由について一般市民に説明する義務を表している。しかし、AIの導入により、その判断や意思決定の裁量や権限（の一部）が技術者やデータ管理者に委譲された結果、市民に対するアカウンタビリティが低下する可能性が懸念される。第4に、能動的責任は、上記3つの受動的な責任と異なり、社会的に共有された目標や価値を追求する前向きの義務を表す。AIの導入に伴う能動的責任のギャップは、設計者や技術者がその技術的な利点のみに着目し、その潜在的なデメリットを防ぐという自らの責任を十分に認識していない、あるいはこの義務を果たす能力や動機を持たない状況に相当する。

アセットマネジメントにAI技術を導入したとしても、それが過失責任、道徳的説明責任、公的説明責任のギャップを許容するものであってはならないことは言うまでもないことであろう。加えて、インフラ管理者には、こうした受動的な責任だけでなく、AIの活用がどのような社会的な目標や価値の達成につながるかという能動的責任を果たすことも求められるだろう。特に、AIの導入に際しては、その利点のみを強調するのではなく、その潜在的な負の影響や限界についても広く社会に共有する義務があると言える。

4. 意味のある人間によるコントロール要件

一方で、上記の責任ギャップ問題のうち、道徳的説明責任のギャップは、AIシステム特有の課題（説明困難性やシステムの複雑性）に起因するものであり、この意味ではより根深い問題と言えるかもしれない。この問題に対して、それを解決不可能と見なす立場（「運命論者」）、問題でもないと過小評価する立場（「デフレ論者」）、新しい技術的・法的手段によって解決できると考える立場（「解決論者」）等、様々な見解があるが[7]、残念ながらどれも本質的・建設的な解決を導くものとは言い難い。AI開発／利用をめぐる責任問題に対処することは容易ではないが、一つの有力なアプローチとして「意味のある人間によるコントロール（meaningful human control; MHC）」という考え方を取り上げてみたい[8]。MHCは、AIシステムの反応や挙動が関連する関係主体の理由と能力と整合し

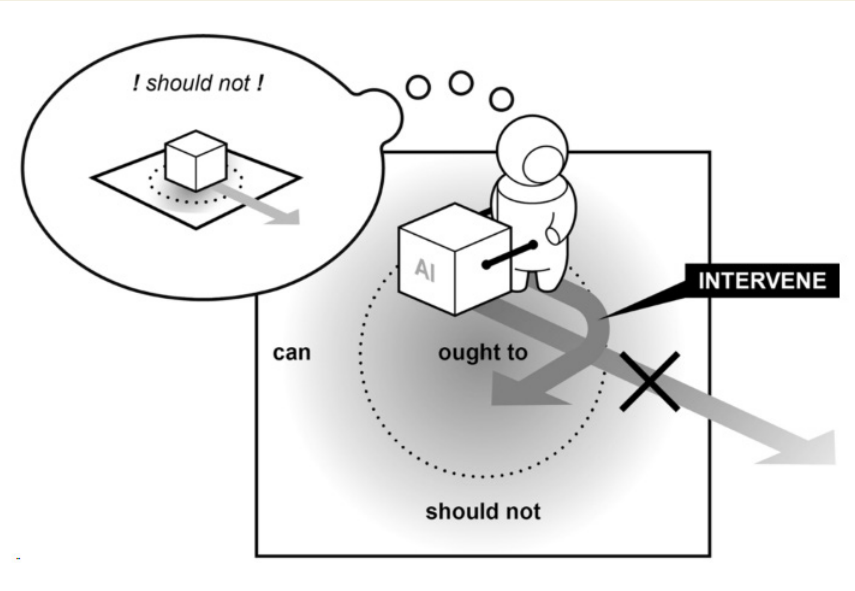

図1　意味のある人間によるコントロール[9]

ていることを表し、AI技術を導入しながらも、そこに意味のある形で人間の判断と入力を維持・介在させることを推奨する（図1）。デ・シオによれば、MHCは、1）人間の理由に関するトラッキング条件と2）人間の能力に関するトレーシング条件という2つの追跡条件によって評価できる[7]。

第1に、トラッキング条件は、AIシステムが、関係主体の関連する理由に対して明白に応答するように設計されていることを要請する。トラッキング条件の評価方法として、自動運転システムを事例として、関係主体の様々な意図や理由をAIアルゴリズムと直接に関連するもの（近位意図、例えば、「歩行者が通れば停車する」など）からシステム全体の遠因となるもの（遠位意図、例えば、「街なかの交通渋滞を緩和する」など）までの連続的なスケールに沿って配置し、それぞれがシステムの振る舞いにどのように反映されるかを分析・評価するマッピング手法が提案されている[10]。アセットマネジメントにおいても、例えば、ロジックモデルを活用することにより、AIを活用したマネジメント業務がその直接的なアウトプットだけでなく、広く社会ニーズに関わるアウトカムとどのように結び付いているかを明確化することにより、トラッキング条件を評価することが出来る。

第2に、トレーシング条件は、AIシステムが関係主体の道徳的・技術的能力と関連付いていることを要請する。より具体的には、システムの設計、開発、使用の各段階において、(i) システムの能力と限界に関する十分な知識と、(ii) システムの振る舞いに対する責任者としての自己の役割に関する十分な道徳的意識とその役割を遵守する能力の両方を備えた、少なくとも1人の関係主体を特定できるような設計であることが求められる。AIシステムをコントロールする上で、必ずしもすべての関係者がその能力や機能について十全に理解する必要はないであろうが、その挙動に対する道義的責任と能力を有する特定の技術者がおり、彼らに対する社会的な認証が担保されていることが必要であろう。

MHCは、責任ギャップ問題の短絡的な解決ではなく、より包括的な観点から、AIシステムと社会のニーズや人間の限定的な能力との関わり合いを反省的に吟味する、より粘り強いアプローチであると言えよう。このアプローチを進める上では、市民や技術者をはじめ、関係者が協働しながら、AI技術の開発から実装までの一連のプロセスを通じて、どのような責任のギャップが生じ、この問題に対していかにして意味のある人間によるコントロールを実現できるかを検討する必要がある。例えば、市役所窓口の混雑予測にAIを導入した事例[11]では、単にAI技術を開発するだけでなく、技術開発者、市民、市役所職員が開発段階から継続的に対話を行い、社会実装に向けて種々の検討（実験サイトの公開、自治体職員が扱えるサイト構築、関係者会議、アクセス数の分析、現場確認、市民アンケート調査、ワークショップの実施など）を重ねて、当該技術の実用性を評価している。こうした地道な協働のプロセスを経ることにより、AI導入をめぐる責任ギャップを克服し、市民に対するアカウンタビリティを果たすことが期待できるものと言える。

5. おわりに

AIアカウンタビリティの問題は、人々がAIとどのように関わり合うべきかという本質的な問いを投げかけている。さらにそれは、我々がAIという新しい技術を用いてどのような社会を目指したいかという問題とも無縁ではない。本稿で紹介したMHCの考え方は、AIシステムに関する専門的・技術的な議論だけでなく、その批判的な検証を含め、我々の社会ビジョンに照らして「何が意味のある人間の関わり方なのか」を追求するアプローチである。AI技術のアセットマネジメントへの活用にあたっては、こうしたアプローチに則って、人間とAIとのつきあい方や目指すべき社会ビジョンに関する市民的議論を蓄積することが重要であろう。

【参考文献】

1）内閣府：人間中心のAI社会原則，2018.

2）クーケルバーク，M.（直江清隆他訳）:AIの倫理学，丸善出版，2020.

3）越水一雄，羽鳥剛史，小林潔司：アカウンタビリティの構造と機能：研究展望，土木学会論文集D，Vol.62，pp.304-323，2006.

4）Bovens, M., Goodin, R. E., & Schillemans, T.: *The Oxford Handbook of Public Accountability*, Oxford: OUP Oxford, 2014.

5）Binns, R.: Algorithmic accountability and public reason, *Philosophy & Technology*, Vol. 31, No. 4, pp.543–556, 2018.

6）Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R.: *A Governance Framework for Algorithmic Accountability and Transparency*, Brussels: European Parliamentary Research Service, 2019.

7）de Sio, F., & Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos. Technol*, Vol. 34, pp.1057–1084, 2021.

8）de Sio, F., & van den Hoven, J.: Meaningful human control over autonomous systems: A philosophical account, *Front. Robot. AI*, Vol.5, 2018.

9）Cavalcante Siebert, L., Lupetti, M.L., Aizenberg, E. et al.: Meaningful human control: Actionable properties for AI system development, *AI Ethics*, Vol. 3, pp.241–255, 2023.

10）Mecacci, G., & de Sio, F.: Meaningful human control as reason-responsiveness: The case of dual-mode vehicles, *Ethics Inf. Technol*, Vol. 22, No. 2, pp.103–115, 2020.

11）浦田真由，谷口友隆，堀涼，遠藤守，安田孝美：市役所窓口における混雑度のAI予測と予測カレンダーの公開―岐阜県高山市における市民課窓口の事例から，実践政策学，Vol.10，No.2，pp.155-162，2024.

はとり　つよし／2006年に京都大学卒業後、東京工業大学助教を経て、現在、愛媛大学社会共創学部（環境デザイン学科）教授。博士（工学）。専門は土木計画学、合意形成論。新居浜市上下水道事業運営審議会会長、今治市中心市街地創生デザイン会議会長等を務める。