

刑事司法における「ブラックボックス・アルゴリズム」の 脱却と「責任あるAI」の確立

——ブラックボックスからクリアシステムへの転換——

竹村 典良

目 次

- I プロローグ
- II ブラックボックス・アルゴリズム刑事司法
- III 刑事司法における「機械決定」の問題性
- IV アルゴリズムの透明性～ブラックボックスからクリアシステムへ～
- V 「説明可能なAI」のための法的要請
- VI アルゴリズム規制と民主主義～批判的検討～
- VII 「責任あるAI」への道程
- VIII エピローグ

I プロローグ

第四次産業革命（第5の革新）の黎明期に、人工知能（AI）は日常生活に急速かつ広範に取り入れられ、アルゴリズム社会への移行が加速している。これまで経験したことのない進歩にもかかわらず、AIを基礎とするシステムの使用について主たる障害となっているのは、そのようなシステムに透明性が欠如していることである。これらのシステムの「ブラックボックス」的特徴は、強力な予測を可能とするが、しばしばその結論を導いた経緯について説明することができない。この問題は、説明可能なAI（XAI）をめぐる新たな議論を惹起する。現在、新しい「説明可能なAI」及び「責任あるA

I」に関する研究が急速に発展している。

本稿では、刑事司法の分野においても使用され始めている人工知能（A I）およびアルゴリズムについて、その「ブラックボックス」的特徴に伴う問題点を明らかにし、「透明性」「説明可能性」「解釈可能性」「A I 基本原則」などの諸概念の分析・検討を通じて、「責任ある A I」を基盤とするクリアな刑事司法システムの確立について探求する。

II ブラックボックス・アルゴリズム刑事司法

アルゴリズムの不透明性が刑事司法、とりわけ刑事裁判に及ぼす影響について、Francesca Palmiotto の論考「公判におけるブラックボックス：アルゴリズム不透明性が刑事手続きにおける公正な裁判を受ける権利に与える影響」(Palmiotto) にしたがって検討する。

1. 刑事司法におけるアルゴリズム使用

刑事手続きのすべての段階でアルゴリズムが使用されることが多くなっている。公判前段階では、ソフトウェアは、捜査における証拠獲得の手段として、警察を援助している（例、被疑者の発見に役立つ顔認証ソフトウェアなど）。公判段階では、アルゴリズムが立証を目的として（例、被疑者の G P S 位置情報）、また量刑のために（例、リスク評価アルゴリズム）使用される（Palmiotto: 51）。

アルゴリズムはインプットをアウトプットに変形する一連の計算過程である。アルゴリズムは命令のセットであり、コードに内在する命令に基づいて、インプットが異なるセットのアウトプットに変形される（Palmiotto: 51）。

アルゴリズムは多様に使用されているが、(1) 決定支援アルゴリズム、(2) アルゴリズムに基づく証拠、の 2 つのグループに分けることができる。第一の類型は、公判前、公判、公判後の各段階で、人間の決定に情報を提供するために使われる（例、犯罪データマイニング、保釈審理、量刑、プロベーション、パロールのためのリスク評価ソフトウェア）。第二の類型は、立証目的で使用され、コンピュータから導出された証拠（例、GPS 位置情報）

とコンピュータにより生成された証拠（例、酒気検査器、DNA 型鑑定、コンピュータ・シミュレーション）を含む（Palmiotto: 52）。

いま一つ別の基準により、(1) 単分析アルゴリズム、と (2) 人工知能システム（以下、A I と表記）に分類することができる。第一の類型には、手作業で巧みに作られたインプットとアウトプットの関係が含まれる。これらはしばしばフローチャートのような規則に基づくシステムであり、前もってプログラムされた命令まで、ステップ、方法、結果を追跡できる。第二の類型は、機械学習アプローチを内在させる、より進歩的なアルゴリズム・システムであり、既存の規則や理論で望ましいインプット・アウトプット関係を捕捉できない問題に使用される。これらのシステムは、インプット・アウトプット変換を規定するパラメーターを事前に決定できず、データあるいは複雑なコンピュータ・シミュレーションから導出するため、ノンパラメトリックと称される（The Law Society of England and Wales）。刑事手続において、A I は、ダイナミックに展開するサイバー攻撃との戦い、証拠の拠り所、裁判官や警察官の決定支援として、益々効果を発揮するようになっている（Palmiotto: 52-53）。

2. ミスコードを隠蔽するアルゴリズムの不透明性

通常、アルゴリズムは秘密と不透明性の中に隠されており、それらの機能についての洞察を制限している。しかしながら、どのようにしてアウトプットが生成されたのか、を理解できなければ、その正確性、有効性、正当性を検証することはできない（Palmiotto: 56; Pasquale; Diakopoulos）。

アルゴリズムの不透明性には多様な種類がある。Burrell によれば、不透明性は、企業や国家の意図的な秘密、技術的誤り、解釈可能性の欠如から生じる（Burrell）。他方、Pasquale は、3つのタイプの秘密、法的、現実的、困惑的、を区別する（Pasquale）。不透明性の概念は、アクセスと情報の間の障壁ばかりでなく、そのような情報の理解可能性の問題とも関係する。とりわけ、第二の見方は、A I システムの台頭後に非常に重要になる。事実、伝統的なソフトウェアは、固定された既知のモデルを基盤とするが、A I システムは、システム・デザイナーでさえも、せいぜい部分的にしか知られていない環境で作動する（Burrell）。この場合、どのようにしてアウトプットが

生成されたかを解明することは難題である (Palmiotto: 57)。

このような状況において、異なる学問領域に属し、多様な視点からアルゴリズムの不透明性について研究する研究者が増加している。主要な傾向は、「アルゴリズムの透明性」を、これらのシステムの説明可能性を高め (Kroll *et al.*; Ananny *et al.*; De Laat; Diakopoulos)、その使用の影響を受ける個人に説明し (Burrell)、「ブラックボックス」の社会への有害な影響を制限する (Pasquale; Citron *et al.*; O'Neil, Eubanks) 主たる道具と考えることである。しかしながら、アルゴリズムの透明性もその短所について批判されてきた。とりわけ、Ananny と Crawford は、透明性はシステムを観察することができるだけで、アルゴリズムを理解することと一致しない、と論じる (Ananny *et al.*)。さらに、透明性はシステムの戦略的ゲームを認めるような望ましくない帰結に至ることもあり得る。透明性の限界の観点から、説明と解釈可能性の概念に言及する流れも生まれている (Lepri *et al.*)。しかしながら、現在のところ、これらについて研究者の間に概念的な明確性と統一性は存在しない (Palmiotto: 57)。

3. アルゴリズム不透明性の「公正な裁判を受ける権利」への影響

刑事手続において、アルゴリズムのアウトプットは、立証あるいは決定支援のために使用される「供述」に相当する。しかしながら、どのようにしてアウトプットが生成されたのかを理解できなければ、供述に対して異議申し立てをすることができない (Palmiotto: 58)。

公正な裁判の保障は、公判前段階、量刑手続き、有罪・量刑に対する上訴審、終局判決の執行など、全ての手続きにおいて適用される。立証あるいは決定を支援するためのアルゴリズムの使用は、ヨーロッパ人権条約 (ECHR: European Convention of Human Rights) 第 6 条に定められる「公正な裁判」の保障に適合しなければならない (Palmiotto: 59-60)。

アルゴリズムに基づく証拠について、とりわけ、異議申し立て権が関係する。公正な裁判を保障するために、被告人は自身に不利な証拠と対決する機会が保障されなければならない。特に、ECHR 第 6 条 (3)(d) は、被告人に有罪判決が下される前に、「被告人が出席する公開の法廷で、対審的議論の観点から、被告人に不利益なすべての証拠が提示されなければならない」と規

定する。本規定は、裁判所で採用された「証人」の自律的で広範な解釈の観点から、(1) 証拠が公判及び上訴手続において使用され、(2) 有罪判決の根拠として重要な場合、アルゴリズムを基礎とする証拠に適用される。リーディングケースである *AI-Khawaja and Tahery vs UK* によれば、対審の権利は、証拠の誠実性と信用性に対して異議申立てる地位にあらなければならないばかりでなく、真実性と信頼性を検証することができなければならない。上述したように、アルゴリズムの不透明性はこの検証の障害となる。対審目的を達成するために、被告人はアルゴリズム・システムについての洞察が必要であるが、被告人には情報とその源へのアクセスが欠如するがために、これは容易に達成することができない (Palmiotto: 60)。これに関連し、法手続きにおいて、個人が、弁護人の援助を受けても、証拠の処理に用いられる複雑なアルゴリズム・システムを理解し異議を唱えることができない場合には、適正手続き権に対する重大な脅威となる (The Law Society of England and Wales)、と指摘される。

決定支援アルゴリズムに関して、主たる法的問題はそのような決定の論拠と正当化の欠如に関係する。ECHR 第 6 条 (1) は、法的決定の論証を求める。確立された判例法によれば、裁判所及び審判の判断は根拠理由を適切に述べなければならない。論証は、防御権、上訴権の保障のもとで、被告人の審理が行われたことを証明する道具である。その上、裁判官は論証を客観的な議論に基づいて行うよう義務付けられている。不透明性が判決の根拠理由を隠蔽する場合には、被告人は上訴権を行使できない (Palmiotto: 60-61)。

要約するならば、アルゴリズムの不透明性は、両当事者が相手方の陳述に対して反駁できない場合は対審原則に、当事者間に知的不均斉が生じる場合は武器の対等原則に、アルゴリズムが被告人によって尋問されない場合は対審の権利に、どのようにして特定の決定に至ったのかをアルゴリズムが説明あるいは証明できない場合には論証された法的決定を受ける権利に、それぞれ有害な影響を及ぼす。公正な裁判権はアルゴリズム・システムに関する見識を必要とする。アルゴリズムに基づく証拠は裁判官によって評価されるが、決定支援アルゴリズムは人間の裁量を完全に取り除くことはできない。被告人には、アルゴリズムに基づく証拠と決定にアクセスし、理解し、異議申立てをする機会が保障されなければならない (Palmiotto: 61; Abiteboul *et al.*;

Lyons *et al.*)。これに関連し、論争可能性は、法の支配の中心にあり、手続きの要である。対審手続の重要性はより良い決定に貢献することであり、それがために人工法知能 (Artificial Legal Intelligence) の統合は新しいヘルメノイティックと新しい対審設計過程を必要とする (Hildebrandt 2018)、と指摘される。

4. 小括

刑事司法におけるアルゴリズムの使用は、広範囲に及び、利益と危険のトレードオフ (二律背反の調整) に対する関心を生んでいる (The Law Society of England and Wales)。刑事司法システムの「自動機械システム」への変化は現実には生じている。したがって、公正な裁判と適正手続きという基本原則を保護するために、刑事手続における「アルゴリズム統治」の適切なモデルが必要である (Palmiotto: 67)。

不透明なアルゴリズムに基づく証拠と決定支援アルゴリズムは、防御権に対する重大な挑戦である。不透明性は、アルゴリズム・システムの探求を妨げ、アルゴリズムに基づく証拠と判決との対決を妨げる。アルゴリズムの不透明性から生じる悪影響を取り除き、軽減する効果的な方法を考えることが重要である (Palmiotto: 68)。

Ⅲ 刑事司法における「機械決定」の問題性

刑事司法における機械決定の問題点と解決方法について、Teresa Scantamburlo、Andrew Charlesworth、Nello Cristianini による論考「機械決定と人的帰結」(Scantamburlo *et al.*) にしたがって検討する。

1. 機械学習アルゴリズムと規範原則

機械学習アルゴリズムは刑事司法においても激しい論争を惹起している。刑事司法において、知的ソフトウェアの使用は広範囲の人々に直接の影響を及ぼさないかもしれないが、決定の結果は、刑事司法システムに関わる人々の権利、及び、過程の公平性、公正性に関する一般の人々の知覚にとって、

極めて重要である。ソフトウェア・ツールは、米国およびイギリスの裁判所と矯正機関において、公判前における犯罪者の釈放に関する決定を支援するために（例、拘禁・保釈の決定）、及び、被告人に科される刑罰の種類・期間を決定するために（例、量刑とパロール）、既に使用されている。通常、「リスク・アセスメント・ツール」と称されるそのようなソフトウェアは、それらのリスクと利益が主張されるがために、公衆の注意と議論の焦点となっている（Scantamburlo *et al.*: 50; Council of Bars and Law Societies of Europe）。

民主主義社会において、裁判官、刑事施設高官、ソーシャル・ワーカーのような人間の決定権者が、関係する個人に向けられた現実的な結果を伴う何らかの利益あるいは負荷が与えられる権利に関する決定をする目的で、個人を評価する時、一般に理解されている規範原則に従って行われなければならない、と理解されている。例えば、

- ・正義（例、法の下での平等、適正手続き、公平性、公正性）、
- ・適法性（例、法規コンプライアンス）、
- ・権利保護（例、表現の自由、プライバシー権）。

通常、決定者がこれらの原則に従って行動するばかりでなく、決定権の行使に対して異議申し立てがなされた場合には監督に従う、ことが求められる。これらの原則が作動するシステムは、例えば、透明かつ説明可能で、一般人の参加と関与を容易にするような、一定の規範的義務に従うことが求められる（Scantamburlo *et al.*: 50-51; Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies）。

また、多くの決定は再検討や精査を受けることはない。決定がなされるシステムの正当性は、規範原則を内在化し適用する決定者に対する信頼、及び、批判し異議を唱える能力を付与するシステムの基本的義務の実効的な履行、に基づいている。一般人が信頼を発展させ、あるいは不信を示す能力が著しく減じられている決定システムは、正当性がなく機能的でないと捉えられる（Scantamburlo *et al.*: 51）。

機械決定が、規範原則に従い、規範的義務が組み込まれたシステムで実行されるのを保証することは、重要な課題である。なぜなら、これらの原則の遵守と義務の履行の正確な性質と程度は、文化的・規律的期待によって変化

するからである。現代世代の知的アルゴリズムは、明示的プログラムではなく、事例から学習する規則に基づいて決定を行う。したがって、知的アルゴリズムが、必要的に、関係する規範原則を正確に内在化し、効果的に適用することも、規範原則が埋め込まれたシステムが、特定の規範義務を有意的に履行するのを促進する手段を持つであろうことも、保証の限りではない (Scantamburlo *et al.*: 51)。

2. 機械決定と正当性

機械学習を人間による決定に適用する際に共有される関心は以下のようなものである。どのようにして決定がなされるのか。どのようにして、機械学習は信頼を育みあるいは不信を示す能力に影響するのであろうか。そのような決定の正確性、公平性、透明性について、私たちは何を知っているであろうか。トレーニング事例あるいはそれらの使用環境によって、バイアスが知的アルゴリズムに影響するのであろうか。これらの状況について考えるために、どのような種類のアナロジーが使われるのか。どのような技術的・法的解決方法が開発されなければならないか (Scantamburlo *et al.*: 51)。

ある人間が一定の期間内にある種の行動をするリスク (例、罪を犯す、問題なく卒業する、任務を完遂するなど) の指標として点数が使用される、という機械決定の使用の根底にある原則は、決定の結果がコンテキストにより著しく異なるものの、広範囲に及ぶ使用において変わらない。アルゴリズムは、算出された点数に基づいて、「犯罪者の拘束」あるいは「申し込まれたローンの拒否」のような決定的な応答行動をする。しかしながら、これらのアルゴリズムがどのように作動したのか、及び、それらが人々の生活に及ぼす影響、についての一般の人々の理解の欠如は、アルゴリズムへの依存度の増大が徹底的な検査の必要性を提案したとしても、ソフトウェアと結果に関する情報へのアクセス制限と結びついて、そのような決定過程を不透明にする (Scantamburlo *et al.*: 51-52; Citron *et al.*)。

機械決定において人間が果たす役割が減少しあるいは皆無になりつつある現在、機械決定への依存度の高まりの個人的・社会的結果、及び、そのような重要決定に基づくシステムが、刑事司法過程・行政決定の正当性に対して基本的である信用の有無にとって必要な要素を支えることができるか否かにつ

いて、議論する必要がある。そのようなシステムにおいて、信用を醸成し、不信を表明することの要請は、単純な正確性の保障を超え、正義、合法性、特別な権利の保障のような主要な規範原則に従って決定がなされることの保障まで含まれるまで拡張される。そのような制限は、決定に至る際に規範的に排除される情報（例、新人採用決定における性別、性的指向の使用）、予め決められた公正性概念の尊重、どのように決定に至ったか、及び、計画され承認された条件のもとでシステムが作動していること、を含んでいる（Scantamburlo *et al.*: 52）。

主要な技術的・規範的概念の相互作用に関する批判的な評価は、機械決定を効果的かつ正当にするために必要である。これらの基準は、正確性、透明性／説明可能性、公平／平等な処遇、プライバシー権／表現の自由である。これらは、社会構造への技術の浸透に関する人々の関心についての最も代表的な要素である。ここでの主たる目的は、機械決定の背後にあるテクノロジーを解明し、それを理解可能にし、技術に疎い聴衆がアクセスできるようにし、補足的な技術的・規範的視点が、どのようにして特定の要望と実現方法を蓄積し結びつけるのに役立つかを示すことである（Scantamburlo *et al.*: 52）。

3. 機械決定における「透明性」と「説明責任」

透明性は法的・行政的システムの望ましい資産である。それは、自身に影響する決定を導く手続・データに関する情報への市民のアクセス権として理解される。いくつかの理由から、それは有益である。一方で、それは、組織が遂行した任務について説明する、すなわち特定のアウトプットに至った全過程を追跡する助力となり（説明責任）、そして、他方で、それは、市民が過程と結果を理解し異議を唱える助力となる（参加）（Scantamburlo *et al.*: 71）。

透明性を求める権利は、国連人権宣言やヨーロッパ人権条約のような人権関連文書で明示されていないが、周縁あるいは背景の権利としての透明性なくして、明確に保護されたこれらの権利が尊重されてきた、と断言することは困難であろう。現代のデータ保護法は、個人データ処理において、透明性原則を順守する必要性を明示する（例、GDPR 第12条）（Scantamburlo *et*

al.: 72)。

透明性の考えは、しばしば、潜在的権利濫用あるいは差別に光を当てることができ、必要な場合には相当な制裁を科す監督機関（例、監査委員会、規制機関）を必要とする。例えば、GDPR の下では、透明性の義務に関するコンプライアンスを順守することは組織のデータ保護官の責任の一部であり、国家監督機関の精査に従う（Scantamburlo *et al.*: 72）。

通常、透明性は、何らかの失敗が生じた時（例、記録紛失など）、責任の発見を人に許可する説明責任システムのデザインを支援する。透明性と説明責任は、別々の問題に言及するが（「不可解な証拠」「追跡可能性」）、両者は決定システムによって惹起された潜在的・現実的危害の予測に向かっている。確かに、情報へのアクセスができ、精査に開かれている場合には、いつどこで危害が発生し、だれが責任を負うべきか、を理解することが容易である（Mittelstadt *et al.* 2016; Scantamburlo *et al.*: 72; Diakopoulos）。

機械決定のコンテキストにおいて、情報公開は新たな問題を生むかもしれない。主要な要因が公開されると、最終結果に影響する変数を変更しようとする「システム競争」が生じるかもしれない。取り扱いに慎重を要する情報がインプットとして使用されると、ひとまとまりのデータの公表に関心が集まる。しかしながら、ソースコードの公開は、正当な商業利益に影響するかもしれないが、過程・結果を理解するための透明性に必ずしも影響を及ぼさない。通常、ソースコードは非専門家には理解不可能であり、専門家でもコードを検査することによってアルゴリズムの行動を予測することはできないであろう（Scantamburlo *et al.*: 72）。

さらに、機械決定の透明性について考察する時、コードの読解能力と単純な関係にはないが、データ駆動機械学習方法と人間の解釈の要請に係る、いま一つの問題に直面する（Burrell）。機械学習方式は微妙な相関関係を発見するために設計されたのであり、研究対象の現象の背景にある原因を理解するためではないので、結果としての分類はしばしば一般のユーザーに説明することができない複雑な定式となる。これらの定式は予測的であり、特定の条件下では最適であることが証明できるが、個別決定について説明することはできない。種々の方法（Random forest、boosted combination of trees、linear combination）でさえも、アルゴリズムによって自由が否定された市

民にありのままの説明をすることができない（Scantamburlo *et al.*: 72-73; Magos）。

近年、研究者集団は、機械決定の透明性を高めるいくつかの方法を開発しつつある。

Ⅳ アルゴリズムの透明性～ブラックボックスからクリアシステムへ～

アルゴリズムの透明性とは何か、Cary Coglianese と David Lehr による論考「透明性とアルゴリズム統治」（Coglianese *et al.*）に従って、ブラックボックスをクリアにするための解決方法を探求する。

機械学習アルゴリズムは、医療、交通、ビジネスなど、多数の分野における重要な機能を向上させ、自動化している。公共機関の公務員等も、アルゴリズムの正確性とスピードを注視し、公共部門の業務に利用し始めている。機械学習アルゴリズムは伝統的な分析手段と比べ予測能力が格段に優れているが、アルゴリズムの予測は理解し説明するのが難しい。機械学習の特徴である「ブラックボックス」は、「アルゴリズム統治は政府の透明性に関する法原則と調和するか」という問題を提起する（Coglianese *et al.*: 2-5; Pasquale; Brevini *et al.*; Citron; Yeung; Brauneis *et al.*; Bathaee）。

1. 透明性の類型：「ガラス張りの透明性」と「理由を付した透明性」

「開かれた政府」が何を意味するかは多様であるので、政府の透明性を評価するのは現実には難しい。政府の透明性の概念は、多様な法学者、政治理論家、公務員によって、多様な方法で援用されてきた。アルゴリズム統治の概念を明確にし、分析の基盤を築くために、政府の透明性を、「ガラス張りの透明性」（fishbowl transparency）と「理由を付した透明性」（reasoned transparency）の2つの類型に分けることができる。前者は、政府が何をしているかに関する情報の開示を優先するのに対して、後者は、なぜ政府がそれをしているのかの理由を理解することを目的とする。いずれの透明性も、違った方法で、政府による機会学習の使用と関係する（Coglianese *et al.*: 19）。

機械学習はブラックボックス的な特徴を有するがために、機械学習に関する最も特徴的な質問は、とりわけ、人間による決定に代替して利用される場合、政府がその行動を理由づける能力に対して、機械学習がどのような影響を及ぼすかに集中するであろう (Coglianese *et al.*: 19)。

ガラス張りの透明性は、政府の内部を凝視し、公務員の公務内容に関する情報を手に入れる一般人の能力に関係する。政府が所有する情報と政府の行動に関する情報への一般人のアクセスに焦点が当てられる。これらには、政府の公聴会、ファイル・キャビネットに蓄積された記録、政府のコンピュータで入手可能な資料などが含まれる。ガラス張りの透明性が、政府が行っていることに関する情報への一般人のアクセスを強調するのに対して、理由を付した透明性は、政府が行動理由を開示するか否か、という情報の有用性を強調する。理由を付した透明性は、政府が理由づけにより行動を説明することを重視する (Coglianese *et al.*: 20-21)。

ガラス張りの透明性と理由を付した透明性は良き政府にとって重要であり、両者は、政府が機械学習を利用する時はいつも果さなければならない法的義務を伴う。これら2つのタイプの透明性は、理由を付した透明性は本来的に何らかのレベルのガラス張りの透明性に依存するということで、相互関係にある。政府は、特定の行動をとった理由を公に説明するために、どのような行動をとったのかを公開しなければならない。論証は、政府が収集した事実、及び、政府が行動を正当化するために行った分析、を開示しなければならない。両者の違いは、機械学習が理由を付した透明性を要求するため、アルゴリズム統治において問題となる。機械学習アルゴリズムがブラックボックス性を特徴とするがために、学習アルゴリズムがある予測に至った理由を十分に説明できるかどうか、の問題が提起される (Coglianese *et al.*: 21-22; Committee of Experts on Internet Intermediaries (MSI-NET))。

2. アルゴリズム統治における「理由づけ」

手続的デュープロセスは、政府がアルゴリズムの正確性に関する情報を提供するように求める。なぜなら、手続的デュープロセスは、政府が個人を手続的に公正に扱うばかりでなく、政府の手続きが重大なエラーに傾かないようにすることを目的とするからである (Coglianese *et al.*: 40)。

機械学習において、エラー修正情報の手続きの要請は、政府による3つの情報開示によって処理される。第一に、政府は、アルゴリズム判決に異議申立てをする関係当事者に対して、正しさを証明するために収集されたデータを提供しなければならない。第二に、政府は、テスト・データセットで評価がなされた場合、アルゴリズムの正しさに関する情報を関係当事者に提供しなければならない。アルゴリズムが誤ったインプット情報を使用していない場合でも、容認できないエラー傾向があるように、粗悪なジョブをするアルゴリズムは手続的デュープロセスを侵害するおそれがある。第三に、政府は、判決が正しくかつ問題なく適用されたアルゴリズムの結果であることを証明する、論証手続きから得られた結果を公開しなければならない（Coglianese *et al.*: 41）。

手続的デュープロセスが普遍的な原則である状況において、アルゴリズムによる判決は法的要請を充足することはそれほど困難ではないであろう。手続的デュープロセス分析のための最も重要なフレームワークは、次の三要素のバランスを要求する。①政府の決定に関わる私的利益、②決定過程から生じる「誤った奪取のリスク」、③特定の手続き配置に関係する「財政的・行政的負担」。機械学習は、より正しい自動的な決定を促進し、エラーのリスクを減少し、政府の決定に関わる時間と支出を節約することができる。機械学習を判決手続に取り入れる機関は、アルゴリズムツールが上記の利点を実現できることを実証し、アルゴリズムが正しく適用されたことを保証する情報へのアクセスを個人に提供しなければならない（Coglianese *et al.*: 42; Cerrillo; Martinez）。

3. 小括

データ科学者は、ブラックボックス・アルゴリズムから説明情報を導く以上のことをする方法を発見しつつある。政府はアルゴリズムの使用における説明可能性の要請を充たし、アルゴリズム統治が人々の信頼と政府の正当性を高めるであろう。よくデザインされたアルゴリズムは、政府機関と共働し、決定に関係する人々に迅速かつ公正な結果をもたらすことにより、人々の信頼を増大させるであろう。将来は、いわゆるブラックボックス・アルゴリズムを利用する政府もブラックボックス政府である必要がなくなる。アルゴリ

ズム統治は、法による透明性の要請を充足し、政府の効力、効率、正当性を高めるであろう。

V 「説明可能な AI」のための法的要請

法の適正手続き原則を充足するために、不透明な AI 処理過程を透明にする方法として、「説明可能な AI」の開発が進められている。Ashley Deeks による論考「説明可能な人工知能に対する法的要請」(Deeks)に従って、現状と問題点を明らかにする。

機械学習アルゴリズムに関する現代の関心は、それらが「ブラックボックス」のように作動し、アルゴリズムが特定の決定、勧告、予測に至った方法と理由を明らかにすることが困難であることである。しかしながら、裁判官は、刑事、行政、民事事件において、機械学習アルゴリズムと対面する頻度が増加している。裁判官も、これらのアルゴリズムの結果について説明を求めるべきである。「ブラックボックス」問題と取り組む一つの方法は、アルゴリズムが結論あるいは予測に至った方法を説明するシステムをデザインすることである。裁判官がこれらの説明を求めた場合、そのようなシステムは、「説明可能な AI」(XAI: explainable AI)の特徴と形態を形作る生産的な役割を果たすであろう。裁判所は、コモンローの道具を使って、XAI が多様な法的コンテキストにおいて何を意味するのか、を発展させることができる。裁判所がこのような役割を果たすことには利点がある。ケースバイケースで事実を検討して微妙な陰翳のある結論を導く、ボトムアップ処理による法的論証は、XAI のための規則を開発するプラグマティックな方法である。また、裁判所は、独特な配置と聴衆に応答する多様な形態の XAI の生産を鼓舞する傾向がある。より一般的には、これまで主として民間に委ねられてきた、XAI の考案に公的機関をより多く関わらせるべきである (Deeks: 1829)。

1. 説明可能な AI の開発

機械学習の利用には潜在的利益があるが、とりわけ、システムが人々の自由、安全、プライバシーに関わる予測をする場合、多くの懸念が生じる。

一つの批判の流れは、科学者がアルゴリズムを教育する際に用いるデータの観点から、これらのアルゴリズムが社会的バイアスを反復し悪化する方法に焦点を当てる。いま一つの批判の流れは、多様な機会学習による予測の正確性に疑問を呈し、刑事司法アルゴリズムのような道具による再犯予測は人間の予測と比べより正確でない、と主張する (Deeks: 1833)。

第三の最も重要な懸念は、アルゴリズムがその結論に至った方法に関する情報の欠如（「ブラックボックス」問題）に集中する (Pasquale; Citron)。アルゴリズムによる勧告の背後にある理由を説明できないことは、勧告の関係者を害する可能性がある。不透明なアルゴリズムは、とりわけ政府によって使用された場合に、人々の公正の感覚と信頼を損ない、刑事司法においては、被告人の防御権を侵害する可能性がある (Deeks: 1833; Pasquale; Citron; Kitchen)。

コンピュータ・サイエンティストは、将来有望な「説明可能な AI」(XAI) を開発することにより、アルゴリズムの透明性の欠如と解釈可能性の問題に果敢に取り組んできている (The Royal Society)。XAI は、特定の機械学習モデルが結論に至った方法を説明する、あるいは、人間が解釈するのを援助する、一連の作業を包含する。ここにおける説明概念は、アルゴリズムの内部状態に関する洞察、あるいはアルゴリズムに関する人間が理解可能な接近、を提供することである (Wacher *et al.* 2018; Mittelstadt *et al.* 2019)。XAI は、人間とシステム間の信頼を促し、システムにバイアスがかかりあるいは不公正であるおそれのある事例を明らかにし、世界がどのように作動しているかに関する知識を強化する (Samek *et al.*)。法の分野では、XAI は、決定をサポートするアルゴリズムを頼りにする裁判官、アルゴリズムを使用することが防御になることを裁判官に説得しようとする訴訟当事者、危険予測に異議申し立てをする被告人、にそれぞれ利益をもたらすことができる。しかしながら、XAI は何らかのコストを要する。最も重要なのは、アルゴリズムを説明可能にすることにより、正確性が減じる結果をもたらすことがある。XAI は、イノベーションを抑制し、開発者に取引上の秘密を明らかにすることを強制し、XAI を構築するには多額の費用が掛かるがために高額のコストを課すことがあり得る (Deeks: 1833-1834; Doshi-Velez *et al.* 2017a; Doshi-Velez *et al.* 2017b)。

2. 外的要因アプローチと分解アプローチ

現在、多種多様な XAI が存在し、コンピュータ・サイエンティストも新しい形態の XAI の開発を続けている。一方で、説明可能性を内在させる機械学習モデルは、その結果としてしばしば複雑性が弱く、予測の正確性が弱まる傾向にある。他方で、説明可能性を内在させないモデルが存在するが、これらには 2 つの基本的なアプローチがある。「外的要因アプローチ」(exogenous approach) と称される 1 つの類型は、機械学習アルゴリズムの内部動作 (論証過程) を説明しようとせず、モデルが外因子直交 (変数が統計的に独立した) 方式を使って動作する仕組みに関する関連情報を、アルゴリズムのユーザーあるいは主体に提供しようとする。第二の類型のアプローチは、モデルの論証を説明あるいは反復することを試みるもので、時々「分解アプローチ」(decompositional approach) と称される (Deeks: 1834-1835; Edwards *et al.*)

さらに、外的要因アプローチは、「モデル中心的」(model-centric) か「主体中心的」(subject-centric) のいずれかに分類される。一方で、モデル中心アプローチは、グローバルな解釈可能性と称され、モデル化過程の背景にあるクリエイターの意図、システムが使用するモデルのファミリー、システムのトレーニング前にクリエイターが特定するパラメーター、モデルのトレーニングに使用される入力データの定性、新しいデータに基づくモデルの成果、などの説明が含まれる。換言すれば、モデルを構成する各部分に関する重厚な記述である。また、他の類型のモデル中心のアプローチは、機械学習システムの結果を検査する (Kroll *et al.*)。このアプローチは、システムの結果や勧告について、バイアスやエラーの出現を急速に探し回る。モデル中心アプローチは、特定の事例に関する作業ではなく、モデル全体についての説明を標榜し、決定が手続的に通常の方法で行われたことを保証することに役立つであろう (Deeks: 1835-36; Edwards *et al.*)。

他方、主体中心のアプローチは、ローカルな解釈可能性と称され、類似の決定を受け取った個人の特徴に関する情報とともに、勧告あるいは決定の主体を提供する (Edwards *et al.*)。いま一つ別の主体中心アプローチは、反事実的条件文 (counterfactuals) の使用を伴う。そこでは、アルゴリズムの勧告に最も影響を与えた要因が何か、を理解しようとする人々は、同じアル

ゴリズムでインプット要因を微調整し、所与の要因がもとの勧告でどの程度作用したかを検証する。例えば、犯罪で有罪判決を受けた者が再犯の危険性が高いと考えるアルゴリズムは、対象者が10歳若く、逮捕歴が少ないとしたら、異なる勧告が出されていたかどうか、反事実的条件文によって検証される。外的要因アプローチの長所の一つは、モデルの内部ロジックを理解するのに、データ主体を必要としないことである。主体中心アプローチは、とりわけ、どのような条件のもとでどのように異なる結果が得られるか、を探索する個人に有用である（Deeks: 1836-37; Wachter *et al.* 2018; Edwards *et al.*; Citron *et al.*）

以上の外的要因アプローチに代替するのは、モデルの論証を分解して説明しようとするXAIのカテゴリーである。その最もわかりやすい方法は、機械学習モデルのソースコードを明らかにすることであるが、このアプローチはしばしば不十分であることが証明される。なぜなら、機械学習の作動の仕組みが複雑で、大抵の人々はコードを理解できないからである（Kroll *et al.*）。しかしながら、微妙な差異のある代替方法がある。その一つは、オリジナルな「ブラックボックス」モデルに並んで、「代替モデル」（surrogate model）と呼ばれる第二のモデルを創り出すことである。代替モデルは、モデルの内的加重値にアクセスせずに、特徴的な入出力対を分析することによって作動する。例えば、研究者が、ブラックボックス・モデルの計算法に類似する決定樹（decision tree）を組み立てる。決定樹は、ブラックボックス・モデルがそのリスク評価をする際にどの要因にウェイトを置いたのか、をコンピュータ・サイエンティストが追跡することを可能にする。これらのシステムは、基本モデルによる予測に類似する（Deeks: 1837; Kroll *et al.*; Pruett *et al.*; Martini; Rodu *et al.*）

3. 小括

将来、機械学習アルゴリズムが法廷で使用される多数の方法がある。その結果、機械学習エコシステムにおいて裁判所自体が重要な関係者となる。個々の事例において、多様な問題を検討することが必要になるであろう。誰が説明の聴取者であり、説明はどの程度シンプルあるいは複雑であるべきか。ユーザーが説明を理解するのに要する時間はどのくらいか。XAIはどのよ

うな構造あるいは形態をとるべきか、コードの行数、ヴィジュアルなプレゼンテーション、操作可能なプログラム。説明はどの要因に焦点を当てるべきか。XAI は何時にモデル中心で、何時に主体中心であるべきか。より一般的には、何が「有意的説明」(meaningful explanation)を構成するのか。このエコシステムにおいて、裁判官は、機械学習技術の専門家でないとしても、XAI への実践的なアプローチを開拓するのに良い立場にある (Deeks: 1837; Hildebrandt 2018)。

Ⅵ アルゴリズム規制と民主主義～批判的検討～

1. アルゴリズム権力の法的批判

「ビッグ・データ」、ユビキタス・コンピューティング、クラウド・ストレージ・システムの台頭を含む、ネットワーク化されたデジタル・コミュニケーション・テクノロジーの革新は、アルゴリズム規制として知られる新しい社会秩序化システムを生み出している。アルゴリズム規制は、リスク管理及び行動変容のために活動領域を規制する決定システムを指す。予め特定された目的を達成するためのシステムのオペレーションを確認し、必要な場合には自動的に洗練するために、規制された環境に関係する多数のダイナミックな構成員から直接引き出されるリアルタイムあるいは継続的なデータをシステムティックに収集し、知識が継続的にコンピュータ的に生み出される (Yeung 2017: 1; Beer; Waldman)。

アルゴリズム決定システムは、情報プライバシーや他の基本的人権、手続的・実体的公正性などに対するリスクを含み、憲法的・民主主義的価値と敵対する虞がある。いくつかの法的関心は、影響が及ぶ個人がアルゴリズム決定に対して異議申立てをする機会が欠如していることを強調する。「適正手続き」(due process)や「手続的公正性」(procedural fairness)のような英米ほかの憲法上の権利は、公務員の決定によって影響を受ける者は、その決定に参加する機会が与えられなければならない、そのような決定はバイアスのないようになされなければならない、ということを保証することが意図されている。執行裁量が合法的かつ公正に、個人の適正手続き権に従って行

われなければならないことを求める英米行政法の中心原則は、伝統的に政府あるいは公的な決定機関にのみ適用されてきたが（Citron）、有力な企業が個人に異議申立ての機会を与えることなく一方的な行動をする自由について不安が増大している。決定における隠れたバイアスに対する関心は、人間及び自動形態の決定のいずれに関しても高まっている。自動決定の利点の一つは、人間による決定に必ず影響する意識及び下意識のバイアスを回避しながら、一貫性を貫けることである。他方、アルゴリズム決定システムが、捕捉するのが難しい潜在的・非意図的なバイアスの影響を受ける、また、一発で何百万というユーザーに影響を及ぼす可能性の、それぞれ程度について、多大な注意が向けられている（Yeung 2017: 23; Citron; Citron *et al.*; Danaher *et al.*; Gueydier）。

2. アルゴリズム説明責任と民主主義の社会的基盤

洗練された計算過程、企業秘密としての非公開、著しく重大な影響を伴う決定権のため、不透明で不可解な「ブラックボックス」とされるアルゴリズムの特徴は、アルゴリズム説明責任に対する要求を促している。説明責任は、決定や行為について説明し、証明し、誤りやエラーを修正するように関係者に求めることである。説明責任の要請に応えることは、機械学習アルゴリズムに基づく決定過程にとってとりわけ重要である。なぜなら、それらは、因果関係のあるいは説明的行動理論ではなく、データポイント間のパターンや相関関係に基づき、過去のインプット・アウトプット・データを考慮して、変更するからである。自由民主主義社会では、個人は、本人に不利益となる決定の理由を知る権利が与えられているばかりでなく、その作動と原則がよく知られ、公の理解と精査ができる透明な秩序であることも求められる（Yeung 2017: 26; Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence MSI-AUT 2018; Bovens）。

アルゴリズム転換が透明性と説明責任という集団的価値に与える影響に対する関心は、どのようにアルゴリズム決定システムへの移行が、民主主義と個人の自由が拠り所とする集団的な道徳的・文化的骨組みを侵食し、自由民主主義的政治秩序を傷つけるリスクを負うか、に光を当てる（Yeung 2017:

26)。

法哲学者の Mireille Hildebrandt によれば、民主主義の本質に照らせば、主権者の規則は透明性の二重の形態によって正当化される。第一に、人々が自身で作成した規則の下で生きており (例、民主主義的参加)、第二に、これらの規則の適用に対して、解釈のブラックボックスを開くことができる対審手続により異議を唱えことができる (法の支配)。これらの 2 つの要素によって、現代法システムは、不完全ではあるが最も成功したサイバネティック・システムの一つである本質的民主主義を確立することができる。それは、支配者と被支配者の間の双方向において作動する完璧なフィードバック・ループを制定する一連のチェック・アンド・バランスに頼る統治システムである。したがって、法の支配のもとに生きるすべての者は、単なる統制される客体ではなく、自主管理に参加し、政府および相互に対して自身の行動に責任を負う、主体と考えられなければならない (Hildebrandt 2016)。アルゴリズム規制への移行とそれを支えるデータによって動く機関はこの均衡の脅威となる。著しく微細なレベルで継続的に個人を追跡することにより、アルゴリズム規制は、見下ろす者が下位の者を監視するのを認めるマジックミラー (片方からしか見通せない鏡) として作動する。しかしながら、彼らは、日常生活の規制を増加しつつあるアルゴリズムのブラックボックスを見通し理解するいかなる現実的な視点を持たない (Zuboff 2015: 81-82; Yeung 2017: 27; Zuboff 2019a; Zuboff 2019b)。

3. 小括

アルゴリズム規制の台頭は、多様な視点と分析レンズによって検討できる多数かつ緊急の問題を引き起こしている (Pruett *et al.*)。学問的及び政策的議論において繰り返される重要なテーマは、確かな「アルゴリズム説明責任」である (Busuioc; Mainzer)。確かで有意義なアルゴリズム説明責任には何が必要かを理解するためには、アルゴリズム権力についての理解を深めなければならない (Yeung 2017: 29)。

巨大なデータセットを動力とする機会学習アルゴリズムを拠り所とする複雑な形態のアルゴリズム規制は、とりわけ、将来の行動の予測システムとして設定される場合、まったく新しいものである。これらのコントロールシス

テムは、伝統的な形態の建築上の規制と比べ著しくパワフルである。なぜなら、これらのシステムは、一人のユーザーだけでなく、広範囲に拡散するすべてのユーザーの行動を追跡し介入することができ、予測し、先んじて行動するために、全人口規模のリアルタイムのデータを収集・分析し、個人的ならびに集団的決定や行動の機先を制することができるからである（Yeung 2017: 30）。

アルゴリズム転換期において、著しい法的、社会的、民主主義的影響が個人あるいは社会に及んでいる。自由で民主主義な社会における中心的法原則である「公正で尊厳のある個人の処遇」「透明性と説明責任」「適正手続きと法の支配」が危機に晒されている。

Ⅶ 「責任あるAI」への道程

刑事司法における「ブラックボックス・アルゴリズム」依存を脱却し、「責任あるAI」を基盤とする刑事司法システムの構築に向かう動きについて、INTERPOLとUNICRIによる共同レポート『責任あるAI革新に向かって：法執行のための人工知能に関する国際刑事警察機構・国連地域間犯罪司法研究所第二レポート』（INTERPOL-UNICRI 2020）の考察にしたがって検討する。

1. 緒論

法執行における「責任あるAI」の使用は、人権、民主主義、正義、法の支配を尊重する一般原理を基盤とする。これらの原理を実行するために、法執行機関は、AIの設計と使用が公正性、説明責任、透明性、説明可能性（FATA）の必要条件に従うように、作動しなければならない。これらの必要条件は、AIコミュニティ内において、アルゴリズムが自身に対する信頼および適切なレベルの安全を保証するために何が必要か、についてのコンセンサスから現れた（INTERPOL and UNICRI 2020: 33; European Union Agency for Fundamental Rights; Latonero; Rodrigues; Access Now）。

これらの4つの必要条件に加え、AIシステムの信頼性と攻撃に対するレ

ジリエンス、及び、セキュリティの観点から、安全とエラー強さの概念も挙げられる。これらの付加的必要条件も、責任ある A I のために、法執行機関によって注意深く考慮されなければならない。A I システムが安全でエラー強くあるために、A I システムを開発し、配備する際には、2つの主要な実践が制度化されなければならない。第一に、内部エンジニア、(使用者たる)警察官、および、必要な場合には、信頼のおける外部パートナーによる定期的システム点検とアップデート、第二に、システム操作の容易性、および、他の将来システムとの共働を保証する相互情報交換・利用可能性、である (INTERPOL and UNICRI 2020: 35; Council of Europe Commissioner for Human Rights)。

法執行において、これらの高レベルの原則と必要条件に反して、非倫理的に、さらには違法に A I が使用されるならば、法執行が奉仕し保護する任務を負う市民は、反発し、法執行によるこのテクノロジーの使用に脅威を感じるであろう。これらは法執行における A I アプリや他の最新技術の使用に対する反抗と批判を生む可能性がある。法執行が A I の積極的潜在可能性を利用し続けるためには、市民の信頼を保持しなければならない (INTERPOL and UNICRI 2020: 35; Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence MSI-AUT 2019)。

2. 合法性

A I の使用において、法執行に関する重大な法的難題がある。法執行において、これらの難題を解決できなければ、A I の使用は、プライバシー権、平等権、差別されない権利のような基本的人権を侵害し、無罪推定、自己負罪拒否特権、合理的疑いを超える証拠のような法原則に違反する虞がある。何よりもまず、法執行における A I の使用は法に従って行われなければならないことが保証されなければならない。A I がデータをエネルギーとすることを鑑みるならば、この問題は、データ・プライバシー (データ収集の規制を含む) とデータ保護 (データの保持、保存、処理を含む) に関する法律に関係する。さらに、法執行は、A I システムの使用時に法に従うばかりでなく、その開発においても法に従わなければならない (INTERPOL and

UNICRI 2020: 36)。

法執行ばかりでなく、他のエンドユーザーのコミュニティによるAIの使用に関する議論においていつも登場する2つの重要な法的文書がある。すなわち、「EU一般データ保護規則」(GDPR: EU General Data Protection Regulation)と「EU刑事司法指令」(LED: EU Law Enforcement Directive 2016/680)である。いずれの文書もAIに特化して発展してきたのではないが、いずれもデータに関係するがために、AIの発展と実行に直接関係する文書とされている。その上、GDPRはEU市民の個人情報と関係し、それ故にEU内外で機能している諸機関にも適用可能であることを鑑みるならば、GDPRはグローバルに関係する文書であり、注目に値する (INTERPOL and UNICRI 2020: 36; 宮下; 星)。

GDPRは、個人の私的情報を保護する法律を近代化するように設計され、4年以上の議論と交渉を経て、2016年に欧州議会 (European Parliament) 及び欧州理事会 (European Council) で採択され、2018年5月に施行された。それは、個人情報の収集と処理のための6つの主要原則を含んでいる。すなわち、1) 合法性、公平性、透明性、2) 目的制限、3) データの最小限化、4) 正確性、5) 保存制限、6) 完全性と秘密保持 (セキュリティ)、である (INTERPOL and UNICRI 2020: 36)。

LEDは、刑事司法指令として知られ、GDPRの個人情報の統制に関する規則を法執行活動に適用することを目的として、2016年5月に施行された。EU基本権憲章 (EU Charter of Fundamental Rights) にしたがって、高レベルのデータ保護と共に、自由、安全、正義の領域を確立する際に、その役割が先導された。個人の私的データを保護し、同時に、高レベルの公共の安全を保障することを目的として、LEDは、データ主体に権利を付与し、法執行目的、すなわち、犯罪の予防、捜査、探知、刑事訴追、刑罰の執行のためのデータ処理における義務を公権力に課す (INTERPOL and UNICRI 2020: 36)。

GDPR、LEDのいずれもAIを念頭に置いて採択されたのではないので、いくつかの重要な規定は、警察活動にAIが使用されるコンテキストにおいて検証されている。GDPRとLEDの必要条件に関する多数の可能性が発生し、注目に値する議論が交わされている (INTERPOL and UNICRI 2020:

36-37)。

GDPR と LED の下における自動手続の規制範囲はそのような議論のうちの一つである。自動手続にのみ基づく決定、いわゆる自動決定システム (ADM: automated decision-making system) は、何処で犯罪が発生し、あるいは、誰が犯罪と関わるのか、を予測するためにデータ分析をする予測的警察取締り (predictive policing) に次第に使用されるようになっており、責任および説明責任の観点から実践的な関心が高まっている。裁判所における自動決定システムの許容性についても同様に議論を呼び起こしている (INTERPOL and UNICRI 2020: 37)。

GDPR は明確に A I に言及していないが、自動決定の役割について言及し、22 条のもとで、プロファイリングを含む「自動決定について説明を求める権利」を示唆している。それは、「管理者」がアルゴリズムを透明で予測可能で証明可能な方法で設計・開発・適用しなければならない、ことを意味する。興味深いことに、GDPR は、「情報主体の明確な同意に基づく場合を除いて、情報主体は自動手続だけにに基づく決定に従わない権利を有する」と規定する。このコンテキストにおいて、十分な情報が参加者に提供され、可能な限り包括的あるいは総合的なデータが使われなければならない (INTERPOL and UNICRI 2020: 37)。

LED も 11 条のもとで特別に自動決定システムに適用される。11 条は、管理者がその管轄下にあり、データ主体の権利と自由に対する保護手段、最低限でも管理者が自身の見解を表明し、決定に異議申し立てをするために人間の介入を得る権利、を提供する E U あるいは加盟国の法律で権限が付与されている場合を除いて、データ主体に不利益な影響を生み、あるいは、重大な影響を及ぼす自動決定システムは禁止される、と規定する。しかしながら、LED に応じて、過敏なデータ処理に基づいて自然人に対する差別的な結果となるプロファイリングは、いかなる条件下においても、認められない (INTERPOL and UNICRI 2020: 37)。

したがって、法的保護を効果的にするために、法執行機関は、自動決定システムにおいては証明が困難な背景ロジックではなく、個々の決定についての説明を提供できなければならない (INTERPOL and UNICRI 2020: 37)。

法執行において遭遇する困難を鑑みるならば、A I の開発・利用過程には

法律の専門家が関与すべきである。とりわけ、自動決定システムが使用される場合に、データ主体あるいは他のいかなる個人にも不利益な影響が生じないようにするために必要である。さらに、法執行機関と他の関係機関は、GDPR や LED のような既存の文書に依拠して、どのように合法的に行動すべきかを形作り導くために、A I 使用における特別な規則を作成し、あるいは、既存の法執行規則を改正することを検討することを望むであろう（INTERPOL and UNICRI 2020: 37）。

3. A I 倫理

1) E UにおけるA I 倫理

近年、E Uにおいて、注目に値するいくつかの展開が見られる。欧州委員会（European Commission）は、これらの倫理的、法的、社会的問題と取り組み、A I システムが人間中心であり続け、このテクノロジーの利益を最大にし、同時に、リスクを防ぎ最小にすることを目指すのを保証するように努めてきた（INTERPOL and UNICRI 2020: 40; INTERPOL and UNICRI 2018: 12-14; Commission Européenne pour l'Efficacité de la Justice (CEPEJ)）。

2018 年 4 月、E U の 24 の加盟国は、「人工知能に関する協力宣言」（Declaration on Cooperation on Artificial Intelligence）に署名し、適切な法的・倫理的枠組みを発展させ、この目的のために協力する必要性を確認した。これに続いて、2018 年 6 月、欧州委員会は、学界、産業界、市民社会を代表する著名な 52 人で構成される独立組織の「人工知能に関する高レベル専門化グループ」（AI-HLEG: High-Level Expert Group on AI）を設立した。この組織は、将来的な政策展開及び社会経済的な課題を含む、A I に関係する倫理的・法的・社会的問題に関する勧告を作り上げることを任務とした。この目的のために、2019 年 4 月、AI-HLEG は『信頼される A I のための倫理ガイドライン』（*Ethics Guidelines for Trustworthy AI*）を公表した。欧州委員会はそのようなガイドラインを公表した最初の機関ではないが、この倫理ガイドラインはこの領域における最初の政府主導のものであり、グローバルレベルにおける A I への人間中心のアプローチであるばかりでなく、A I 倫理という考えをめぐる国際的なコンセンサスの構築に向かう重要なステッ

プである (INTERPOL and UNICRI 2020: 40)。

倫理ガイドラインによれば、A I が信頼されるときには、全ライフサイクルを通じて、1) 合法的 (すべての関連法規に従う)、2) 倫理的 (倫理規範への同調)、3) エラー強く (技術的・社会的視点から)、なければならない。信頼される A I システムの開発は、人間の尊厳の尊重、民主主義、正義、法の支配のような確立された基本的価値に基づく、と同時に、平等で差別のないことを確実にするために、個人の自由と市民の権利を保障しなければならない。これに基づいて、倫理ガイドラインは、A I システムの開発・配置・使用の基礎をなす 4 つの最も重要な倫理原則 (人間の自律性の尊重、危害の予防、公平性、説明可能性) を提示する (INTERPOL and UNICRI 2020: 40)。

上述の原則が実践される際に、それらの間に何らかの対立・矛盾が生じるのは自然なことである。例えば、A I が予測的警察取締りに使用されるならば、犯罪を減じることに役立つかもしれないが、それには個人の自由とプライバシーを侵害する監視が含まれ、危害の予防原則と人間の自律性原則を葛藤に陥れ、最も適切なトレードオフ (二律背反) について熟慮が必要となる。これらのトレードオフと解決方法を適切に確認、評価、記録、共有するために熟慮が必要である (INTERPOL and UNICRI 2020: 41)。

A I システムを信頼される方法で開発・配置・使用するための 7 つの必要条件 (人間機関と管理、技術的エラー強さと安全性、プライバシーとデータの統治、透明性、多様性、差別のないことと公平性、社会環境福祉、説明責任) が確認される (INTERPOL and UNICRI 2020: 41)。

倫理ガイドラインは、これらの必要条件を充足するために、技術的ならびに非技術的方法を提示する。技術的方法に関して、信頼される A I アーキテクチャーは、設計、システムの有効性テスト、サービス指標の質によって、倫理と法の支配を履行する。必要条件を充足するための非技術的方法は、規制、行動コード、標準化、検定、説明責任に関する参加と行動、倫理的思考様式を促進するための教育と意識改革、利害関係者の参加、社会的対話、多様で包括的な設計チームの設立、などである。倫理ガイドラインには、これらの必要条件の補完として、評価リストあるいは網羅的でない一連の質問も含まれている。2019 年 6 月、欧州委員会は、利害関係者に対して評価リス

トを試験的に実施し、リストを修正するためのフィードバックを行った（INTERPOL and UNICRI 2020: 41）。

2020年2月、欧州委員会は、EUにおける可能なAI立法の発展過程の開始の合図となる白書、『人工知能について～卓越と信頼へのヨーロッパ・アプローチ』（*On Artificial Intelligence: A European Approach to Excellence and Trust*）を公表した。白書は、AI-HLEGが倫理ガイドラインを開発する際に行った基礎研究に基づいており、AIは人間中心的で、倫理的で、持続可能で、基本的な権利・価値を尊重しなければならない、ことを強調する（Bostrom; Ganascia; Russel; Tegmark）。また、AIは標本データに基づいて訓練され、企業はAI開発の経緯に関する詳細な資料を保存し、市民がAIシステムと相互作用する場合には十分な情報が提供される、のような特定の法的必要条件を提案する（INTERPOL and UNICRI 2020: 42）。

2) 他のAI倫理

AI倫理と責任あるAIの発展に関して、基本権と一致協力する議論は、ヨーロッパに限られない。いくつかの公共機関、研究施設、民間企業は、原則とガイドラインについて声明を発し、AIへのアプローチの仕方に関する政策文書を作成するAI専門家委員会を設置した。Nature誌の最近の研究によれば、AI倫理原則あるいはガイドラインを含む84の資料が確認される。これらの資料は、貴重な文書であり、法執行において責任あるAIの開発と配置を推進するための評価基準となる（INTERPOL and UNICRI 2020: 42; Bundesverband Digitale Wirtschaft (BVDW) e.V.; Demiaux; Kraemer *et al.*）。

おそらく、最も注目に値するものとして、経済協力開発機構（OECD: Organization for Economic Co-operation and Development）は、2018年5月に、社会におけるAI原則を作成するために、専門家グループを設立し、2019年5月に、AIに関する最初の政府間政策ガイドラインである『人工知能に関する原則』（*Principles on Artificial Intelligence*）を採択した。この原則は、42か国によって採択され、革新的で信頼が置け、人権と民主主義的価値を尊重するAIを促進することに焦点を当てる（Organization for Economic Co-operation and Development, OECD）。国家レベルでは、人工知能とデータの倫理的使用に関する諮問会議（シンガポール）（Advisory

Council on the Ethical Use of Artificial Intelligence and Data) のような、A I の倫理面を探究するためのいくつかの委員会や専門家グループが設立された (Interpol and UNICRI 2020: 42)。

民間部門においても、とりわけ、そのビジネスを A I に頼る企業の間で、同様の努力がなされている。Google、IBM、Microsoft のような主要かつ巨大な企業は、A I 配置の探究を続けるための倫理原則を確立した (INTERPOL and UNICRI 2020: 42)。

4. 小括

A I は、法執行においてその業務を一変させる潜在力を認識させ、あらゆる形態の犯罪との闘いにおいて効果を高め、既存の能力を増大させることができる強力な道具である。それはまた、プライバシー権、平等権、差別されない権利のような基本的人権を侵害しないように、また、無罪推定、自己負罪拒否特権、合理的疑いを入れない証拠のような法原則を傷つけないように、注意深く行使しなければならない「諸刃の剣」である (INTERPOL and UNICRI 2020: 5)。

実際に、法執行において A I の使用は増加しており、「法執行における責任ある A I の使用」が最も必要な問題になっている。人権の尊重、民主主義、正義、法の支配、および、関連する必要条件として、公平性、説明責任、透明性、説明可能性は、法執行において厳守されなければならない一般原則である (INTERPOL and UNICRI 2020: 6; Dreyer *et al.*; Wachter *et al.* 2017)。

VIII エピローグ

現代社会における A I 使用の到達点として、アルゴリズムの説明可能性 (説明責任) は不可欠の要素である。説明可能性には多様な概念が存在し、内在的なメカニズムから外在的な解釈可能性にまで展開している。最近では、公正性、透明性、プライバシーなどを含む数々の「A I 基本原則」を義務付ける新しいパラダイム「責任ある A I」(Responsible AI) についても議論されるようになってきている。今後、「責任ある A I」についての認識が高ま

るのは必然であり、将来において「A I 基本原則」について研究することがますます重要になるであろう。現在、「ブラックボックス・アルゴリズム」に依拠する刑事司法を脱却し「責任あるA I」を基盤とするクリアな刑事司法システムを構築することが喫緊の課題となっている。

【注】

- 1) 本稿は、JSPS 科研費基盤研究（C）「南北統合グローバル・グリーン犯罪学の生成と国際環境裁判所の創設に関する調査研究」(課題番号 19K01353) の研究成果の一部である。

【参考文献】

- Abiteboul, S., et G'Sell, F. (2019). Les algorithmes pourraient-ils remplacer les juges? *Le Big Data et le droit*, Dalloz, 2019, Thèmes et Commentaire. Hal-02304016v2.
- Access Now (2018), *Human Rights in the Age of Artificial Intelligence*.
- Adadi, A., and Berardi, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160.
- Ananny, M., and Crawford, K. (2018). Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Society*, 20 (3): 973–989.
- Andrews, L., Benbouzid, B., Brice, J., Bygrave, L. A., Demortain, D., Griffiths, A., Lodge, M., Mennicken, A., and Yeung, K. (2017). *Algorithmic Regulation*. Discussion Paper No.85. London: Centre for Analysis of Risk and Regulation at the London School of Economics and Political Science.
- Arrieta, A. B., Rodriguerz, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcis, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusin*, 58: 82–115.
- Bathae, Y. (2018). The Artificial Intelligence Black Box and the Failure of

- Intent and Causation. *Harvard Journal of Law and Technology*, 31 (2): 889–938.
- Beer, D. (2017). The social power of algorithms. *Information, Communication and Society*, 20 (1): 1–13.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press. (倉骨彰訳『スーパーインテリジェンス 超絶 AI と人類の命運』日本経済新聞出版社、2017 年)
- Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13 (4): 447–468.
- Brauneis, R., and Goodman, E. P. (2018). Algorithmic Transparency for the Smart City. *Yale Journal of Law and Technology*, 20: 103–176.
- Brevini, B., and Pasquale, F. (2020). Revisiting the Black Box Society by rethinking the political economy of big data. *Big Data and Society*, 2020: 1–4.
- Bundesverband Digitale Wirtschaft (BVDW) e.V. (hrsg.) (2019). *Mensch, Moral, Maschine: Digitale Ethik, Algorithmen und künstliche Intelligenz*. Berlin: Bundesverband Digitale Wirtschaft (BVDW) e.V.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data and Society*, 3 (1): 1–12.
- Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, 0 (0): 1–12.
- Cerrillo i Martinez, A. (2019). How can we open the Black Box of Public Administration? Transparency and Accountability in the Use of Algorithms. *Revista Catalana de Dret Públic*, 58: 13–28.
- Citron, D. K. (2008). Technological Due Process. *Washington University Law Review*, 85 (6): 1249–1313.
- Citron, D. K., and Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89 (1): 1–33.
- Coglianese, C., and Lehr, D. (2019). Transparency and Algorithmic Governance. *Administrative Law Review*, 71: 1–56.
- Commission Européenne pour l'Efficacité de la Justice (CEPEJ) (2018). *Charte*

éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement. Adopté lors de la 31e reunion plénière de la CEPEJ (Strasbourg, 3-4 décembre 2018). Conseil de l'Europe.

Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence MSI-AUT (2018). *Addressing the impacts of Algorithms on Human Rights*. Draft Recommendation of the Committee of Ministers to member States on the human rights impacts of algorithmic systems. Council of Europe.

Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence MSI-AUT (2019). *Responsibility and AI*. Council of Europe study DGI (2019) 05. Council of Europe.

Committee of Experts on Internet Intermediaries (MSI-NET) (2017). *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. Council of Europe study DGI (2017) 12. Council of Europe.

Council of Bars and Law Societies of Europe (2020). *CCBE Considerations on the Legal Aspects of Artificial Intelligence*.

Council of Europe Commissioner for Human Rights (2019). *Unboxing Artificial Intelligence: 10 steps to protect Human Rights*. Council of Europe.

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., Paor, A. D., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., and Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data and Society*, 2017: 1-21.

Deeks, A. (2019). The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review*, 119 (7): 1829-1850.

De Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology*, 31 (2): 525-541.

- Demiaux, V. (2017). *How Can Humans Keep the Upper Hand? The ethical matters raised by algorithms and artificial intelligence*. Report on the public Debate led by the French Data Protection Authority (CNIL) as part of the ethical discussion assignment set by the Digital Republic Bill. Commission Nationale Informatique et Liberté.
- Diakopoulos, N. (2014). *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Tow Center for Digital Journalism.
- Doshi-Velez, F., and Kortz, M. (2017a). *Accountability of AI Under the Law: The Role of Explanation*. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. (<http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584>)
- Doshi-Velez, F., and Kim, B. (2017b). Towards a Rigorous Science of Interpretable Machine Learning. (<http://arxiv.org/pdf/1702.08608.pdf>)
- Dreyer, S., and Schulz, W. (2019). *The General Data Protection Regulation and Automated Decision-making: Will it deliver? Potentials and limitations in ensuring the rights and freedoms of individuals, groups and society as a whole*. Gütersloh: Bertelsmann Stiftung.
- Ebers, M., and Gamito, M. C. (eds.) (2021). *Algorithmic Governance and Governance of Algorithms: Legal and Ethical Challenges*. Cham: Springer.
- Edwards, L., and Veale, M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law and Technology Review*, 16 (1): 18–84.
- Eubanks, V. (2017). *Automating Inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.
- European Union Agency for Fundamental Rights (2020). *Getting the Future Right: Artificial Intelligence and Fundamental Rights*. Luxembourg: Publications Office of the European Union.
- Ganascia, J.-G. (2017). *Le Mythe de la Singularité: Faut-il craindre l'intelligence artificielle ?* Paris: Éditions du Seuil. (伊藤直子監訳『そろそろ、人工知能の真実を話そう』早川書房、2017 年)
- Gueydier, P. (2018). *Pouvoir Régalien et Algorithmes, Vers L'Algocratie ?*

OPTIC.

Hildebrandt, M. (2016). Law as Information in the Era of Data-Driven Agency. *Modern Law Review*, 79 (1): 1–30.

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A*. 376: 20170355.

星周一郎「GDPRと刑事司法指令・PNR指令の相関——データの越境移転の規律を中心に」ジュリスト No.1521（2018年）

INTERPOL-UNICRI (2018). *Artificial Intelligence and robotics for Law Enforcement*. Lyon: The International Criminal Police Organization (INTERPOL) and Torino: United Nations Interregional Crime and Justice Research Institute (UNICRI).

INTERPOL-UNICRI (2020). *Towards Responsible AI Innovation: Second INTERPOL-UNICRI Report on Artificial Intelligence for Law Enforcement*. Lyon: The International Criminal Police Organization (INTERPOL) and Torino: United Nations Interregional Crime and Justice Research Institute (UNICRI).

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication and Society*, 20 (1): 14–29.

Kraemer, F., van Overveld, K., and Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, 13: 251–260.

Kroll, J. A., Huey, J., Barocas, S., Felton, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165 (3): 633–705.

Latonero, M. (–). *Governing Artificial Intelligence: Upholding Human Rights and Dignity*. Data and Society.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenge. *Philosophy and Technology*, 31: 611–627.

Lyons, H., Velloso, E., and Miller, T. (2021). Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. arXiv:2103.01774v1

[cs.CY] 23 Feb 2021.

Magos, V. (-). *Résister à L'Algocratie: Rester humain dans nos métiers et dans nos vies*. Bruxelles: yapaka.be.

Mainzer, K. (2019). *Künstliche Intelligenz: Wann übernehmen die Maschinen?* Berlin: Springer.

Martini, M. (2019). *Blackbox Algorithmus: Grundfragen einer Regulierung Künstlicher Intelligenz*. Berlin: Springer.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3 (2): 1-21.

Mittelstadt, B., Russel, C., and Wachter, S. (2019). Explaining Explanations in AI. In: *FAT '19: Conference on Fairness, Accountability, and Transparency (FAT '19), January 29-31, 2019, Atlanta, GA, USA*. ACM, New York, NY, USA: 279-288.

宮下紘『E U一般データ保護規則』（勁草書房、2018 年）

O'Neil, C. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. New York: Crown. (久保尚子訳『あなたを支配し、社会を破壊する、AI・ビッグデータの罠』インターシフト、2018 年)

Organization for Economic Co-operation and Development, OECD (2019), *Artificial Intelligence in Society*. OECD. (経済協力開発機構 (OECD) 編著、齋藤長行訳『OECD 人工知能 (AI) 白書：先端テクノロジーによる経済・社会的影響』明石書店、2021 年)

Palmiotto, F. (2021). The Black Box on Trial: The Impact of Algorithmic Opacity on Fair Trial Rights in Criminal Proceedings. In: Ebers, M., and Gamito, M. C. (eds). *Algorithmic Governance and Governance of Algorithms: Legal and Ethical Challenges*. Cham: Springer.

Pasquale, F. (2015). *The Black Box Society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.

Policy Department for Citizens' Rights and Constitutional Affairs Doctorate-General for Internal Policies (2020). *Artificial Intelligence and Law En-*

- forcement: Impact on Fundamental Rights*. European Parliament. European Union.
- Pruett, W.A., and Hester, R. L. (2016). The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes. *PLos One*. (<http://doi.org/10.1371/journal.pone.0156574>)
- Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4: 100005.
- Rodu, J., and Baiocchi, M. (2020). When black box algorithms are (not) appropriate: a principled prediction-problem ontology. arXiv:2001.07648v2 [stat.OT] 3 Apr2020.
- Russel, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking. (松井信彦訳『AI 新生 人間互換の知能をつくる』みすず書房、2021 年)
- Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries*, Special Issue No.1, 13 Oct. 2017.
- Scantamburlo, T., Charlesworth, A., and Cristianini, N. (2019). Machine Decisions and Human Consequences. In: Yeung, K., and Lodge, M. (eds.). *Algorithmic Regulation*. Oxford: Oxford University Press.
- 竹村典良「「法の支配」から「アルゴリズムの統治」へ～A I による刑事司法の予測化・自動化における最低基本三原則：「公平性」「説明責任」「透明性」～」桐蔭法学 27 巻 1 号（2020 年）43-72 頁。
- Takemura, N. (2021). AI-Algorithm-Big Data, Predictive/Automated Criminal Justice, and Hyper Crime/Social Control: 'Surveillance Capitalism' after 'Singularity' and Prospects of Information Civilization. *Toin University of Yokohama Research Bulletin*, 44: 51-58.
- Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin Books. (水谷淳訳『LIFE 3.0 人工知能時代に人間であるということ』紀伊國屋書店、2020 年)
- The Law Society of England and Wales (2019). *Algorithm use in the criminal justice report*.

- The Royal Society (2019). *Explainable AI: the basics. Policy briefing*.
- Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7 (2): 76–99.
- Wachter, S., Mittelstadt, B., and Russel, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31 (2): 841–887.
- Waldman, A. E. (2019). Power, Process, and Automated Decision-Making. *Fordham Law Review*, 88 (2): 613–632.
- Yeung, K. (2017). Algorithmic Regulation: A Critical Interrogation. *Regulation and Governance*, 12 (6): 505–523. (本文引用はドラフト原稿の頁数)
- Yeung, K., and Lodge, M. (eds.) (2019). *Algorithmic Regulation*. Oxford: Oxford University Press.
- Zuboff, S. (2015). Big Other: Surveillance Capitalism and the Prospects of an Informal Civilization. *Journal of Information Technology*, 30: 75–89.
- Zuboff, S. (2019a). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books. (野中香方子 訳『監視資本主義：人類の未来を賭けた闘い』東京経済新報社、2021 年)
- Zuboff, S. (2019b). Un capitalisme de surveillance. *Le Monde diplomatique*, Janvier 2019: 1, 10–11.

(たけむら・のりよし 桐蔭横浜大学法学部教授)