総務省 学術雑誌『情報通信政策研究』 第３巻第１号
Journal of Information and Communications Policy Vol.3 No.1

特別寄稿

「完全自動意思決定」のガバナンス
―行為統制型規律からガバナンス統制型規律へ？

山本　龍彦[1]（慶應義塾大学）

要　旨

EU の GDPR（一般データ保護規則）は、データ保護に関連した様々な権利を保障している。そして、こうした権利の侵害があった場合には、高い制裁金が科されることでも知られる。そうすると、一見、GDPR は、厳格なペナルティをもって権利侵害行為ないし違反行為を直接統制する法令であるように感じられる。ところが、実際上、権利の具体的な内実や範囲はいまだ確定的ではなく、また権利侵害行為ないし違反行為があっても、これを外部から発見することは非常に困難であるという問題を抱えている。GDPR は、かかる法的不確定性と執行困難性の問題を前提に、事業者自らが行動規範等の策定を通じて不確定性の隙間を埋めたり、データ保護影響評価（DPIA）やアルゴリズム監査といった内部統制システムを整備したりして、想定される違反行為等を未然に防ぐガバナンス体制を構築することを、かかる体制構築の努力と制裁金の免除・軽減とを結び付けることで（明示的なインセンティブ設計）実効的に促しているように思われる。

本稿は、プロファイリングに関連する GDPR の諸権利、とりわけ、重要事項についてプロファイリング等の結果のみに依拠して決定されない権利（22 条）、「説明を受ける権利」（15 条）をめぐる解釈論に照準して、上述のような GDPR の傾向、すなわち、行為ベースの規律（行為統制型規律）からガナバンス・ベースの規律（構造統制型規律）への焦点変動について若干の分析を加えるものである。本稿は、先行して同様の焦点変動が起きた（雇用に関する）反差別法の実践などにも視点を向け、法的不確定性と執行困難性を抱える法領域では「構造統制型規律」が一定程度有効であること、したがって、これらの問題が前景化するであろう AI ネットワーク社会において、かかる規律モデルが中心的な法的アプローチとなる可能性についても言及を加える。

キーワード：GDPR、プロファイリング、完全自動意思決定、説明を受ける権利、データ保護影響評価、ガバナンス

１．はじめに

日本でも、様々な個人情報から人工知能（Artificial Intelligence, AI）を用いて個人の趣味嗜好、健康状態、社会的信用力、職業適性、内定辞退予測率などを自動的に予測・分析する「プロファイリング（profiling）」[2]が一般化しつつある。最近は、いわゆるターゲティン

[1] 慶應義塾大学法科大学院教授

[2] GDPR は、「プロファイリング」を以下のように定義している。「自然人の特定の個人的側面を評価するための、特に、当該自然人の職務遂行能力、経済的状況、健康、個人的選好、

グ広告の配信を目的とした利用を超えて、与信審査や企業の採用活動など、個人の人生を左右しかねない重要場面で、こうした手法が用いられるようになっている。筆者は既に、少なくとも要配慮個人情報のプロファイリングについては、要配慮個人情報の「取得」に本人の事前同意を求める個人情報保護法 17 条 2 項との平仄から、透明性と本人の事前同意が必要であること[3]、与信・採用など重要場面におけるプロファイリングの実施については、後述する EU の一般データ保護規則（General Data Protection Regulation, GDPR）[4]の完全自動意思決定の原則禁止（22 条）の考え方を参考に、本人関与の仕組み等を整備する必要があることを説いてきた[5]。こうした問題関心は、日本社会でも一定程度共有されるに至ったが[6]、筆者は、複雑かつ重要な問いは「その先」にあると考えている。例えば、完全自動意思決定に関して本人が有する権利（またはデータ管理者の義務）とは具体的に何を意味するのか（「説明を受ける権利（right to explanation）」の範囲など[7]）、いかにして権利侵害ない

興味関心、信頼性、振舞い、所在および移動に関する側面を分析または予測するための、個人データの利用によって構成される、あらゆる形式の個人データの自動的処理」（2 条 4 項）。プロファイリングに関する問題の所在については、石井夏生利「プロファイリング規制」ジュリスト 1521 号（2018 年）32 頁以下、山本龍彦『プライバシーの権利を考える』（信山社、2017 年）257 頁以下参照。

[3] 例えば、山本龍彦「個人情報保護の今日的重要性」都市問題 110 巻 2 号（2019 年）46 頁以下参照。周知のとおり、自動意思決定およびプロファイリングに関する GDPR ガイドラインは、以下のような考え方を示している。「データ管理者は、9 条 2 項に規定された条件のうちの 1 つを満たした場合にのみ、特別な類型の個人データを処理することができる。これは、プロファイリングから派生し、または推論された特別な類型データも含む」（強調筆者）。ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, 17/EN, WP 251rev.01(Feb. 6, 2018), at 15[hereinafter GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING]（以下、日本語では「GDPR ガイドライン」とも呼ぶ）． 同ガイドラインはさらに以下のように述べる。「センシティブな選好や特性がプロファイリングから推論される場合、データ管理者は、①当該処理が、もともとの目的と矛盾しないこと、②特別な類型データの処理のための合法的根拠を確認していること、③当該処理についてデータ主体に伝えている（inform）ことを確かめなければならない」。*Id.*

[4] GDPR に関する最近の解説書として、小向太郎＝石井夏生利『概説 GDPR』（NTT 出版、2019 年）、宮下紘『EU 一般データ保護規則』（勁草書房、2018 年）がある。

[5] 例えば、山本龍彦「信用スコアの課題と今後」月刊経団連 67 巻 10 号（2019 年）26 頁以下参照。

[6] 例えば、最近公表された AI ネットワーク社会推進会議（総務省）『AI 利活用ガイドライン～ＡＩ利活用のためのプラクティカルリファレンス～ 』（2019 年 8 月）も、「AI を利用したプロファイリングを行う場合における不利益への配慮」（22 頁）、「人間の判断の介在」（に関する検討）（14 頁、26 頁）、「説明可能性の確保」（24-25 頁）などが期待されるとしている。
http://www.soumu.go.jp/main_content/000637097.pdf

[7] 「説明を受ける権利」は、GDPR 上、そのような権利が存在するか自体が争われている。こうした論争を手際よくまとめたものとして、Bryan Casey, Ashkon Farhangi and Roland Vogl, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L. J.

し違法行為を外部から発見・同定し、救済を与えることができるのかといった問いは、容易に答えの出るものではないからである。

実際、GDPR も、完全自動意思決定の禁止を、「プロファイリングを含む自動的処理のみに依拠した決定に服しない権利（*right* not to be subject to a decision based solely on automated processing, including profiling）」（強調筆者）として規定しているが、どの程度人間が関与すれば「自動的処理のみに依拠した決定」ではなくなり、権利侵害性が否定されることになるのか、解釈上明確になっているわけではない。また、仮に人間の実質的関与がなく、上記「権利」の侵害があったとしても、これを外部から発見・同定することは困難である。さらに GDPR は、完全自動意思決定を例外的に実施する場合、事業者はデータ主体に対して「決定に含まれているロジックに関する意味のある情報（meaningful information about the logic involved）」等を提供しなければならないとしているが（13～15 条）[8]、深層学習のような複雑高度な学習方法を採用すると、その「ロジック」が「ブラックボックス化」するため、何をどこまで説明すれば当該義務を果たしたことになるのか（「説明を受ける権利」を実現したことになるのか）が不確定にならざるを得ない。後に詳述するように、GDPR も、完全自動意思決定に関する基本的なスタンス（基本的な権利概念）は明らかにしているものの、解釈の不確定性（indeterminacy）や、個々の権利侵害ないしは違法行為の同定困難性（執行困難性）の問題を多分に残しているのである。しかし、GDPR は、これらの問題を無責任に放置しているわけではない。解釈の不確定性や違法行為の同定困難性の問題が生じることを前提に、あるいはそれを見越して、「ガバナンス」による規律を要求しているように思われるからである。

本稿は、まず AI を用いた完全自動意思決定に関連する GDPR の諸規定を概観し、その内容の不確定性と違法行為等の同定困難性の問題を指摘したうえで、これらの問題に対する GDPR の対応をみる。そこでは、GDPR が、総合的なガバナンス体制の構築を管理者に要求することで、上記問題を乗り越えようと試みていることが明らかになるだろう[9]。次に本稿は、ガバナンス体制構築の必要性を強調するため、同種の問題を抱える他の法領域でも、法的規律の焦点が個別の行為からガバナンスへと移ってきている可能性を指摘する。最後に本稿は、「ガバナンスとしての法」という観念は、完全自動意思決定の文脈を超えて、アルゴリズムを用いた意思決定が一般化する AI 社会の中心的な法的アプローチになる可能性

143(2019).

[8] GDPR 13 条・14 条は、こうした説明をデータ管理者の義務として規定している。他方で 15 条は、かかる説明に対するデータ主体の「アクセス権（right of access by the data subject）」を規定している。こうみると、13 条・14 条は、単に管理者の義務規定のように読めるが、GDPR ガイドラインはこれらの条文の見出しに「知らされる権利（right to be informed）」という用語を充て、「権利規定」として解釈している。GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 16, 24.

[9] GDPR の主たる目的は「協働的なガバナンスレジーム」の構築にあると指摘するものに、Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L. J. 189, 195(2019).

について言及する[10]。

２．完全自動意思決定に関する GDPR 上の課題と対策

２．１．課題

(1)「人間の関与」とは何か？

先述のとおり、GDPR は、情報処理の一段階として「プロファイリング」を定義し、これを前景化した（2 条 4 項）[11]。そして、特定の目的（公共の利益やデータ管理者によって追求される正当な利益〔legitimate interests〕、ダイレクトマーケティング等）のために行われるプロファイリングに対して「異議を述べる権利（right to object）」をデータ主体に認めた（21 条）。この権利が行使された場合、データ管理者は、「やむを得ない正当な根拠（compelling legitimate interests）」があることを証明しない限り、プロファイリングを中止しなければならない（21 条 1 項）。もっとも、「GDPR は、何がやむを得ない正当な根拠としてみなされ得るかについていかなる説明も行って」おらず[12]、データ主体の異議を受け入れてプロファイリングを中止するか否かを判断するに当たっては、常にデータ主体の側の利益と管理者の側の利益との比較衡量を要する。完全自動意思決定とは異なる一般的なプロファイリングの文脈であるが、ここにも解釈の不確定性を看取できる（なお、ダイレクトマーケティング目的のプロファイリングに異議が述べられた場合は、管理者は絶対にその処理をやめなければならない。21 条 2 項・3 項。その意味で、ダイレクトマーケティング目的のプロファイリングに異議を述べる権利は「無条件の権利（unconditional right）」である[13]）。

GDPR は、さらに、「プロファイリングを含む自動的処理のみに依拠した決定に服しない権利」をデータ主体に認めた（22 条 1 項）。この権利により、データ管理者は、①データ主体に（契約解除、社会保障の給付・撤回、入国許可の取消しといった）法的な効果（legal effects）を与える決定、または、②それと同様に重大な影響をデータ主体に及ぼす（similarly significantly affects）決定（融資、採用、医療機会、教育機会に関する決定など）を、AI 等を用いたプロファイリング「のみ」に依拠して行ってはならないものとされた[14]。本稿が照準する完全自動意思決定の原則禁止である。

同規定の解釈上まず問題になるのは、「のみに依拠した（solely based on）」の意味である。かかる文言が、いわゆる重要決定を行うに際して人間の関与がない、ということを意味することは明らかである。しかし、解釈上、どこまでの人間の関与があれば「のみ」でない、ということになり、同規定の適用を回避できるのかは明確ではない。GDPR のガイドラインは、

---

10 デジタル時代において、より一般的に「ガバナンス」の可能性を探るものとして、西山圭太「データをめぐる国際情勢（DFFT／G20）」月刊経団連 67 巻 10 号（2019 年）22 頁以下参照。

11 前掲注(2)参照。

12 GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 18.ただしガイドラインは、伝染病の拡散を予測するためのプロファイリングが「やむを得ない正当な根拠」に当たる可能性を示唆している。*Id.*

13 GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 19.

14 *Id.* at 21-22.

ある者がプロファイリング結果を自動的（automatically）に個人に適用するなど、「人間の関与をでっち上げること（fabricating）により 22 条を回避することはできない」とし、GDPR が要求しているのは人間の実質的関与であるとしている[15]。またガイドラインは、この関与は「決定を変更する権限と能力をもった者によってなされなければなら」ず（こうした権限・能力をもっていない者の関与は実質的に無意味である）、彼らは「関連する全てのデータを考慮すべきである」とも述べ、「自動的処理のみに依拠した決定に服しない権利」の意味を明確化しようと努めている[16]。しかし、かかる権限保持者が具体的にどこまで決定プロセスに関与すれば権利侵害との謗りを免れるのかについてはなお不確定性が残ると言えよう。AI の確率的評価と人間の判断との関係はいかにあるべきかという問いは、AI 社会における最重要論点であり、この不確定性は、今後の議論に向けた合理的な余白とも考えられる。

加えて、仮に両者のあるべき関係性が規範的に確定されたとしても、データ管理者がこれを実際に遵守しているかどうかは外部から発見しづらい。したがって、ここには権利侵害行為の同定困難性（執行・救済困難性）の問題も認められる。

(2)「決定に含まれるロジックに関して意味のある情報」とは何か？

これも周知のとおり、GDPR 22 条 2 項は、(a)「データ主体とデータ管理者の間の契約の締結またはその履行のために必要となる場合」、(b)「データ主体の権利および自由ならびに正当な利益を保護するための適切な措置も定め、かつ管理者が服する EU 法または加盟国の国内法によって認められる場合」、または（c)「データ主体の明示的な同意に基づく場合」には、重要事項について自動的処理のみに依拠して決定すること（完全自動意思決定）を許容している。しかし、同条 3 項は、この場合でも（上記 2 項（b）の場合を除く）、データ管理者は、「データ主体の権利および自由ならびに正当な利益……の保護を確実にするための適切な措置（*suitable measures*）を実装しなければならない」（強調筆者）とし、「少なくとも、……人間の関与を得る権利、データ主体の見解を表明する権利およびその決定を争う権利」を保障するために適切な措置を講ずることを要求している[17]。

そして、13 条および 14 条は、「公正かつ透明性のある情報処理を確保するために」、完全自動意思決定を行う場合にデータ管理者に課される情報提供義務を（ただしガイドラインはデータ主体の「知らされる権利（*right* to be informed)」と言い換えている)、15 条は、かかる情報に対するデータ主体の「アクセス権（right of access)」を規定している。13～15 条は、ここでデータ主体がアクセスできる（管理者がデータ主体に提供しなければならない）情報として、①「完全自動意思決定が存在すること」（管理者が完全自動決定を行っていること)、②「少なくとも当該決定が存在する場合（at least in those cases)」には（i)「当該決定に含まれるロジックに関して意味のある情報」、(ii)「当該処理のデータ主体にとって

---

[15] *Id.* at 21.　後掲注(16)も参照。

[16] *Id.*

[17] ガイドラインによれば、ここで「鍵」となるのは人間の関与である。「あらゆる審査は、適切な権限と決定を変更する能力をもった者によって実施されなければなら」ず、「審査する者は、データ主体から提供された追加的情報を含む、全ての関連データを徹底的に評価しなければならない」。*Id.* at 27.

の重要性および想定される帰結に関して意味のある情報」を挙げている（13 条 2 項（f)、14 条 2 項（g)、15 条 1 項（h))。なお、ガイドラインは、22 条 1 項の「完全自動意思決定」に当たらない場合（すなわち人間が一定程度重要決定に関与する場合）でも、上記②（i)（ii）の情報を提供することが「ベスト・プラクティス」であるとしている[18]。

ここで解釈上問題になるのは、上記②において管理者がデータ主体に提供しなければならない情報の範囲、とりわけ、②（i)「完全自動意思決定に含まれるロジックに関して意味のある情報」とは何か、である[19]。ガイドラインは、「機械学習の成長と複雑性は、自動意思決定のプロセスまたはプロファイリングがいかに機能しているのかの理解を難しくしている」との認識を示したうえで、「管理者は、背景にある理論的根拠（rationale）や、決定にたどり着くうえで依拠した基準についてデータ主体に伝えるシンプルな方法を探すべきである」[20]と述べる。そして、GDPR はあくまで「意味のある情報」の提供を求めているのであり、一方において「使用されるアルゴリズムに関する複雑な説明や、完全なアルゴリズムの開示までは必要とされない」が[21]、他方において「データ主体が当該決定の諸理由を理解するために十分な広がりをもったものでなければならない」としている[22]。

ガイドラインは、管理者が個人のローン申請を査定・拒否するために信用スコアを利用する事案を例にとり、さらに具体的な解説を加えている。それによれば、信用スコアが信用情報機関等から提供されている場合でも、管理者がそのスコアに頼っている限りは、管理者は、当該スコアについて、そしてその理論的根拠についてデータ主体に説明できなければならないとされる。換言すれば、「管理者は、当該決定に到達するうえで考慮された主要な特徴（main characteristics）の詳細、当該情報の源泉（source）および関連性（relevance）を提供するもの」とされているのである[23]（なお、「管理者は、データ主体が 22 条 3 項に沿うかたちで拒否決定の再検討を要求するのに必要な問い合わせ先も提示する」こととされる[24])。

ガイドラインによるこのような解釈明確化に向けた努力にもかかわらず、管理者が完全自動意思決定を行う際に提供すべき情報の範囲――データ主体の「アクセス権」の範囲――は、なお不確定性が残っている。それは、セルプストとバロカス（Andrew D. Selbst & Solon Barocas）がいう、機械学習の「不可解性（inscrutability）」と「非直観性（nonintuitiveness）」によるところが大きい。セルプストとバロカスによれば、機械学習の強みは、「コンピュータのために明示的な指示を作成する面倒な仕事からプログラマーを解放する能力だけではなく、人間が見落とすか、認識できないデータ上のわずかな関係性を学習する能力」に由来するが、この強みこそが、「機械学習によって発展したモデルを、極めて複雑なものにし、

---

18 *Id.* at 25.
19 例えば、「この要求の意味や適切な解釈はいまだ不明確である」と指摘するものに、Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1100(2018).
20 *Id.*
21 しかし、「複雑さはデータ主体に情報を提供しないことの免罪符としてはいない」とも述べる。*Id.*
22 *Id.*
23 *Id.* at 25-26.
24 *Id.* at 26.

人間が分析することを不可能にする」のだという[25]。彼らは、こうして「意思決定を規律するルールが極めて複雑で、多数で、相互依存的なもの」となり、機械学習は、人間による「実質的な点検を受けつけず、理解を拒む」「不可解性」をもつことになるというわけである[26]。機械学習を用いた意思決定は、しばしば「秘密性」、「曖昧性（opacity）」、「不透明性」、「ブラックボックス」などと呼ばれるが、それらはいずれも機械学習に固有の問題ではなく[27]、「モデルのとてつもない洗練性と範囲」によって生ずるこの「不可解性」こそが、機械学習の強みであるとともに、問題の本質であると述べる[28]。

セルプストとバロカスは、さらに、意思決定の根拠となる統計的関係が発見されたとしても、その関係は、「特定の基準の適切さに関する直観的期待を寄せ付けない」という「非直観性」をもつという[29]。例えば、「靴の好み」と「朝フルーツをよく食べる」ということが統計的に関係するとしても、人間はこの関係を直観的に理解できない。「学生時代、授業中によく居眠りをした経験をもつこと」と「高い職務能力」とが仮に統計的に関係しているとしても、直観的な理解はやはり困難となろう（直観的には、居眠りの経験はむしろ低い職務能力と関係するように感じるからである）。統計的関係が明らかになるとしても、機械学習においては、なぜそうした関係が存在するのかは神秘的（mystifying）なものになり得る。セルプストとバロカスは、それにもかかわらず人間は、意思決定について、我々の直観的理解と一致した理由付けを求める傾向があるという。それは、我々人間には、「妥当性の問題として、規範性の問題として、ある意思決定の根拠が健全なものかどうかを査定する方法が存在することを確実にしたい」という欲求、ある意思決定について「それを評価（evaluate）できる存在でありたい」という欲求があるからである[30]。そうであるがゆえに、人間は、相関関係ではなく、「我々にとってよく知られた、許容可能なパターンと一致した」「因果関係」に基づく説明を好む傾向があるのだが、しかし、そもそもアルゴリズムや機械学習は、人間の直観を寄せ付けない関係を発見するためにデザインされたものであり、非直観的な性格を有するというのである[31]。

セルプストとバロカスによれば、機械学習は「不可解性」と「非直観性」という、人間に対する「説明」に本質的になじまない性格を有しているために、「説明を受ける権利」の解釈は必然的に難しいものになるという。人間がこれまで一般的に要求してきたような「説明」を機械学習に関して求めることは、「不可解性」と「非直観性」によって生み出される機械学習の本質的な強みを削ぐことにもなるからである（例えば、直観的説明が要求されるとすれば、結果に対する直観的関係をもった変数しか利用できないことになる）。かくして、機械学習、とりわけ深層学習を用いた完全自動意思決定について、いかなる目的のために、何をどこまで説明すべきなのかは、GDPR をめぐる――AI 社会の在り方をめぐる――最もホットな論点を形成しつつあるのである。

---

[25] Selbst & Barocas, *supra* note 18, at 1094.
[26] *Id.*
[27] *Id.* at 1091-1094.
[28] *Id.* at 1094.
[29] *Id.* at 1097.
[30] *Id.* at 1098.
[31] *Id.* at 1097-1098.

セルプストとバロカスも、「ロジックに関して意味のある情報」の「意味や適切な解釈はいまだ不明確である」[32]としたうえ、様々な解釈上の課題を抽出している。例えば、データ管理者が機械学習を採用する場合に求められる「説明」には、その主たる「目的」に応じて大要２種類のものがあると指摘している。

１つは、説明することでその[ママ]データ主体に事後の行動指針を与え、自律的で自己決定的な生き方を可能にするという目的（自己決定的目的）に出た「結果ベースの説明（outcome-based explanation）」[33]ないしは「主体中心（subject-centric）の説明（SCEs）」[34]である。この「説明」とは、「意思決定のルールそれ自体に関する記述ではなく、ある決定との関連性が示された諸事実に関する記述」を意味する[35]。したがって、結果ベースの説明では、「ある[ママ]特定の[ママ]自動意思決定の根拠、理由、個別的事情」（強調筆者）に焦点が当たる[36]。Berkman Klein Center の作業部会は、より具体的に、①ある決定における主要な要因（main factors）、②ある決定の結果を変えるために必要な最小限の変化、③異なる結果を伴う同様の事案に関する説明または同様の結果を伴う異なる事案の説明を提供すべきとする[37]。また、ワッチャーら（Wachter et al.）も、そこでは、「異なる答えをもたらし得る最も小さな変化」といった「反事実的（counterfactual）」な説明が要求されると指摘している[38]。上述のように、こうした「結果ベースの説明」の主たる目的は、個人が、よりよい結果を得るために次になすべき行動を自ら決定・選択することを可能にする点（自己決定的目的）にある。

もう１つは、説明させることで意思決定の妥当性や正当性（例えば不適切な情報を使っていないこと、差別的でないこと）を担保するという目的（正当化目的）に出た「ロジックベースの説明（logic-based explanation）」[39]または「モデル中心（model-centric）の説明」[40]である。この「説明」とは、「ある決定の背景にあるリーズニングの記述であって、単に、当該決定に関連するインプットの記述ではない」ということになる[41]。したがって、ロジックベースの説明では、「自動意思決定システムのロジック、重大性、想定される帰結、一般

---

[32] *Id.* 1100.
[33] *Id.* at 1099. もともとのアイデアは、Sandra Wachter, Brent Mittlelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Dose Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 78(2017). この論文では、説明には「システムの機能性（system functionality）」に関するものと、「具体的な決定（specific decision）」に関するものがあるとされた。
[34] Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 55-56(2017).
[35] Selbst & Barocas, *supra* note 18, at 1100.
[36] Wachter et al., *supra* note 32, at 78.
[37] Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* 2-3(Harvard Univ. Berkman Klein Ctr. Working Grp. On Explanation & the Law, Working Paper No. 18-07, 2018), https://dash.harvard.edu/handle/1/34372584.
[38] Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH 841, 845(2018).
[39] Selbst & Barocas, *supra* note 18, at 1099.
[40] Edwards & Veale, *supra* note 33, at 55-56.
[41] Selbst & Barocas, *supra* note 18, at 1110.

的な機能性」に焦点が当たる[42]。上述のように、このロジックベースの説明の主たる目的は、意思決定の技術的および規範的妥当性を担保することにあるが、当然「ロジック」の説明は個別の具体的決定とも関連しており、この点で自己決定的目的に資する側面がないわけではない。

セルプストとバロカスによれば、GDPR 13～15 条が「ロジック」という文言を使用している以上、これらの条文はロジックベースの説明を要求しているように見えるが、それも完全に確定的なものではないという[43]。例えば、ワッチャーらの考えのように、ロジックベースの説明はデータ主体本人にとっては抽象的かつ一般的で、そこから具体的な行動指針を得ることが難しい。他方、結果ベースの説明は「そのデータ主体に特定的（*specific*）な何かを提供できるため、効果的な対応方法を伝えるのにより有益である」[44]（強調筆者）とする見解がある。そうなると、データ主体にとって「意味のある情報」という文言を、結果ベースに重点を置くようなかたちで解釈することもあり得るというわけである[45]。もっとも、「ある決定について〔個人が〕争うためには、具体的な決定の反事実的な説明を単に得るだけでなく、意思決定のロジックを理解する必要がある」[46]との指摘にも注意する必要がある。そうなると、やはりロジックベースの説明も必要ということになるが、かかる「ロジック」について何をどこまで説明すべきは明確に定義付けられるものではないだろう。

近年、こうした解釈上の不確定性の解決を、一定程度技術の進展に委ねようとする方向性も出始めている。AI 自身がその判断のロジックを説明する「説明可能な AI（Explainable AI, XAI）」の開発などはその典型である。ただ、こうした技術が解釈の不確定性を全て解決できるわけではない。

例えば、解釈可能なモデルを意識的に構築するというアプローチがあり得る。端的に、変数のセットを小さくし、分析を制約すれば、学習過程において発見される関係性の総数も限定され、人間にも理解可能なものとなるし（5 つの特徴をもつモデルは、500 の特徴をもつモデルよりも解釈可能性は高い）[47]、深層学習よりも決定木型学習（decision tree learning）や線形モデルの方が解釈可能性は高い。しかし、解釈可能性と正確性・パフォーマンスとは一般的にトレードオフの関係に立つ。解釈可能性を高めようとすればパフォーマンスはある程度犠牲になるし、パフォーマンスを高めようとすれば解釈可能性はある程度犠牲になる。GDPR が両者のトレードオフの在り方を一義的に確定しているわけではないとすれば、理解可能な説明のために正確性やパフォーマンスをどの程度犠牲にすべきかは、なお開かれた問題であると言えよう。したがって、GDPR が求める「程よい」説明とは何かという問題は残り続ける。

事後的（post hoc）アプローチと呼ばれる技術的方法論もある。「モデル構築後に、より

---

[42] Wachter et al., *supra* note 31, at 78
[43] Selbst & Barocas, *supra* note 18, at 1100.
[44] *Id.* at 1121
[45] 前掲注(21)のように、ガイドラインは、「当該決定に到達するうえで考慮された主要な特徴の詳細」などの説明を求めている。この点を踏まえれば、GDPR 13～15 条が要求する説明は、「結果ベースの説明」を排除するものではないだろう。
[46] Selbst & Barocas, *supra* note 18, at 1122.
[47] *Id.* at 1110-1111.

簡単に理解できる形式（in a more readily intelligible form）でモデルに近似（approximate）するような、あるいは、具体的な決定にとって最も顕著な特徴を同定するような特別な技術を適用する」という考え方である[48]。しかし、前者、すなわち、複雑な方法で学習されたモデルに、よりシンプルで、より解釈可能な方法で近付こうという「追体験型」アプローチについては、その実際上の困難性が指摘される。例えば、深層学習により開発されたモデルに接近するために決定ツリーを用いる場合、あまりに多くの枝葉が必要となり、実際上理解可能なものとはならない。そうなると、理解可能性の確保には、最初のモデルのパフォーマンスを犠牲にせざるを得ず、結局、先述したようなトレードオフの問題が出てくることになる。

また、後者、すなわち、特定の結果における種々の特徴の相対的重要性のみを説明させるアプローチ――これは、先述した「結果ベースの説明」になじむ――についても、モデルが余りに多くの特徴セットを考慮している場合には限界があると指摘される。例えばセルプストとバロカスは、このような場合、「重要性があるとみなされる特徴の数があまりに多くなるために、このアプローチが克服しようとした不可解性の問題を単純に再生産してしまう」[49]と述べる。さらに、理解可能となるよう、リスト化する特徴を絞り込むという考え方もあり得るが、その場合には、かかる絞り込み基準をどのように決定するかが問題となる。こう考えると、このアプローチを採用したとしても、GDPR 上、データ管理者に提供が求められる情報の範囲が確定されるわけではない。

以上みてきたように、完全自動意思決定を行う際にデータ管理者に課される情報提供義務ないしデータ主体のアクセス権の範囲は――「不可解性」と「非直観性」という機械学習の本質的な特徴ゆえに――明確になっているとは言い難く、当面はこうした不確定的な状況が続くように思われる。

（3）「特別な種類の個人データに基づいてはならない」の意味

GDPR 22 条 4 項は、「9 条 2 項（a）〔明確な同意がある場合〕または（g）〔法律上の根拠の下、重要な公共の利益のために必要な場合〕が適用され、かつ、データ主体の権利および自由ならびに正当な利益の保護を確保するための適切な措置が設けられている場合を除き、〔22 条〕2 項に規定する〔完全自動意思〕決定は、9 条 1 項に規定する特別な種類の個人データ（special categories of personal data）を基礎としてはならない」と規定している。9 条 1 項が規定する「特別な種類の個人データ」とは、人種的・民族的出自、政治的見解、宗教的・哲学的信条、遺伝的データ等のいわゆるセンシティブ情報を意味する。したがって、22 条 4 項は、完全自動意思決定にセンシティブ情報を使用することを原則的に禁止する規定であると、まずは解釈することができる。しかし、不確定的な部分も少なくない。

例えば、この規定は、本人の明確な同意と、「データ主体の権利および自由ならびに正当な利益の保護を確保するための適切な措置」が設けられている場合に、「特別な種類の個人データ」を完全自動意思決定に使用することを例外的に許容しているが、ここで言う「適切な措置」とは具体的にどのような措置を言うのかが明確ではない。「特別な種類のデータ」が差別や偏見と関係していることから、差別・偏見を防止するための手続を含むように解されるが[50]、それが具体的に何を意味するのかはなお不確定的である。

---

[48] *Id.* at 1110.
[49] *Id.* at 1115.
[50] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 28.

より重要なのは、「特別な種類のデータを基礎としてはならない」の意味である。例えば、プロファイリングを用いた与信の自動審査に「特別な種類のデータ」を意図的に使用してはならないという意味を含むことについては議論の余地がない。しかし、このことを超えて、「特別な種類のデータ」と密接に関連し、実質的にそれに代替し得るデータ――「代理変数（proxy）」となり得るデータ――の使用禁止も含むのだろうか。こうしたデータの使用を無制限に許せば、結果的に、「特別な種類のデータ」それ自体の使用を認めることと同様の差別的効果を発生させるだろう。しかし、どのようなデータが「特別な種類のデータ」の代理変数として機能するかは、予測モデルの事後的で経時的な検証を経ないと判明しないこともあり、管理者の責任の範囲を過度に広げることにもなりかねない。

GDPR そのものの検討からはやや逸脱するが、ここで、アメリカにおける「差別的インパクト（disparate impact）」論について簡単に触れておきたい。差別的インパクト論とは、事業者が差別を意図せず、一見中立的な方策を採用したとしても、結果的に特定の集団に不均衡な効果[51]をもたらすならば、違法な差別となり得るという考え方で（日本では「間接差別」とも呼ばれる）、1970 年代以降、人種、性別等による差別を禁止する公民権法タイトルⅦ（Title Ⅶ of the Civil Rights Act of 1964）の解釈として裁判所で採用されるようになったものである（現在では公民権法タイトルⅦ703 条（k）（1）として条文化されている）。

もちろん、こうした考え方の下でも、特定集団に対する差別的影響の全てが違法とされるわけではない。例えば、事業者が、問題となる方策が「業務上の必要性（business necessity）」と関連することを立証すれば、差別としての認定を免れることなる（もっとも、原告が差別的影響を発生させずに、業務上の必要性を実現する代替的手段の存在を立証すれば、やはり差別的インパクトに基づく法違反が成立する）。しかし、事業者が完全自動意思決定を導入した場合、GDPR 22 条 4 項の場合と同様、「差別的インパクト」の解釈――さらに言えば「差別」それ自体の解釈――を確定することが難しくなる[52]。

先述のように、事業者に差別の意図はなく、あるいは差別防止のためにセンシティブ情報を積極的に排除していたとしても、何らかのデータが代理変数として機能し、結果的に差別的インパクトを生じさせることはあり得る。この場合、その実際の影響は、事後的で経時的な影響評価や検証等を経てはじめて把握されることになろう[53]。しかも、先述した深層学習

---

51 「差別的インパクト」の存在を認定するルールとして、「5 分の 4 ルール」がある（EEOC ガイドライン）。これは、保護されたクラス（マイノリティ）の選択率・採用率が、保護されていないクラス（マジョリティ）の選択率・採用率の 5 分の 4（80%）未満である場合に、異なるインパクトの存在を認めると考えるものである。29 C. F. R. § 1607.4.

52 問題の所在として、例えば、Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671(2016); Stephanie Bornstein, Antidiscriminatory Algorithms, 70 ALA. L. REV. 519(2018).

53 差別的インパクト論のリーディングケースとされる Griggs v. Duke Power Co., 401 U.S. 424(1971)では、企業が採用条件として高校卒業資格を課したことが問題とされた。高校卒業資格は一見したところ人種的に中立であるが、実際には黒人を締め出す効果をもった。差別的インパクト論の典型的な事案では、差別的インパクトを引き起こす具体的行為の同定が比較的容易であったが、機械学習が差別的インパクトを引き起こす場合、その原因となる「行為」の同定が困難となろう。

の「不可解性」を踏まえれば、差別的インパクトの要因を突き止めることは容易ではなく、その修正・調整が困難になることも予想される（アルゴリズム上の調整をかけたとしても、それが差別的影響の除去にどの程度有効に作用しているか、「逆差別」のような副作用的な影響を生じさせていないかを判定するのにさらに一定の時間を要する）。

また、「業務上の必要性」との関連性をいかに認定するかも難しくなる。例えば、セルプストとバロカスは、ある予測モデルが、求職者の「音楽の好み」（例えばヒップホップ音楽を好んで聴く）が職務能力と、特定の人種（例えば黒人）の双方と相関することを学んだ、というシナリオを検討している[54]。このシナリオの場合、当該モデルの運用によって特定人種に対して差別的なインパクトが生じ得る。したがって事業者は、「音楽の好み」と職務能力等との関連性を立証する必要があるが、これまでの伝統的思考枠組みの下では、両者の関係について人間にとって直観的に理解可能な因果的な物語を描くことは困難となるため、その立証は成功しない可能性が高いという。セルプストとバロカスは、しかし、もともと深層学習の強みは「人間の直観を超えたパターン」[55]ないしは隠れた奇妙な相関関係を発見することにあるために、もしかするとこの予測モデルは、人種とは一応区別される「隠れた変数」を学習している可能性があるとする[56]。この「隠れた変数」が――従来型の因果関係論では把握できない――「音楽の好み」と職務能力等との正当な関係性を説明する可能性があるというわけである。かくして、深層学習による差別的インパクトの場合、直観的な因果関係を要求する「業務上の必要性」の抗弁も変容を迫られる可能性があると指摘される。そうでないと、深層学習による差別的インパクトには抗弁が一切成立せず、かような影響が生じたものは全て違法な「差別」と認定されることになってしまうからである（このことは深層学習の利用を過度に制約することになり得る）。

以上のようにみると、完全自動意思決定の導入場面では、GDPR 22 条 4 項の解釈だけでなく、より一般的に、「差別」行為それ自体の解釈――何が「差別」か――が不確定的になるように思われる。

２．２．GDPR におけるガバナンス・アプローチ

以上、プロファイリングを用いた完全自動意思決定が行われる場面で、GDPR 上、どのような行為が権利侵害または違法行為に当たるのかが確定的でないうえに、仮に一定程度確定されたとしても、かかる行為の発見・同定が困難にならざるを得ないこと（執行・救済困難性）をみてきた。本章では、GDPR が、こうした状況にどのように対応しようとしているのかについて考察を加える。結論を先取りすれば、GDPR は、個別の違法行為の解釈不確定性・同定困難性を、「ガバナンス」の観点から乗り越えようとしているように思われる。

以下、完全自動意思決定に関する GDPR のガバナンス・アプローチの詳細を確認しておこう。

（1）倫理委員会の設置等

先述のように、GDPR 22 条 3 項は、例外的に完全自動意思決定を行う場合に、「少なく

[54] Selbst & Barocas, *supra* note 18, at 1124-1125.
[55] *Id.* at 1129.
[56] *Id.* at 1125.

とも、人間の関与を得る権利、データ主体の見解を表明する権利、その決定を争う権利」を保障することを求めている。しかし、「少なくとも」という文言が示すように、これらはあくまでも例示であり、同規定のポイントは、「データ主体の権利および自由ならびに正当な利益を保護するための適切な措置を実装する」ことにある。

ガイドラインは、収集・共有されたデータに含まれるエラーやバイアス、自動意思決定過程に含まれるエラーやバイアスが、不正確な分類をもたらし、それが個人に対してネガティブな影響を与えうるとし、管理者は、「適切な措置」として、「バイアスをチェックするために、自らの保持するデータセットについて定期的なアセスメントを実施し、偏見的要素に対処するための方法を発展させなければならない」としている[57]。そして、「アルゴリズム監査の仕組みや、プロファイリングを含む自動意思決定の正確性や適切性の定期的審査（regular reviews）が有益な措置である」とも述べている[58]。ベスト・プラクティスについて記載したガイドラインの添付文書（Annex）は、より具体的に、「機械学習にかかわる監査プロセスのための行動規範」の策定や、「プロファイリングのための特定の適用が社会に及ぼす潜在的な損害および恩恵を評価（assess）する倫理審査委員会」の設置を推奨している[59]。GDPR は、例えば、先述した「説明」内容の不確定性による自動意思決定に対する個別的挑戦の困難性（決定を争う権利の実現困難性）や、「特別な種類の個人データに基づいてはならない」とする 22 条 4 項の意味の不確定性――さらには「差別」的行為それ自体の不確定性――と、それに基づく同項の執行困難性などを踏まえて、管理者自らがアルゴリズム監査や倫理審査委員会を含む制度的・組織的な措置を講じて、具体的な権利侵害や差別的インパクトの発生を実効的に防止していくことを構想しているように思われる。

(2) データ保護影響評価（Data protection impact assessments, DPIA）

GDPR は、アカウンタビリティの中心的なツールとして、データ保護影響評価（DPIA）を規定している（35 条）。DPIA は、「自然人の権利および自由に高いリスクを発生させるおそれのある」情報処理に関して一般的に求められているが（35 条 1 項）、重要決定についてプロファイリングを含む自動的処理が使われる場合には「とくに必要とされる（in particular be required）」（同条 2 項）。いまここで「完全自動意思決定」という言葉を使わなかったのは、DPIA に関する 35 条が、完全自動意思決定について規定する 22 条 1 項に挿入されていた「のみに依拠して（solely based on）」という言葉を含んでいないからである。したがって、融資・採用にかかわるような重要決定においてプロファイリングを用いる場合には、その決定をプロファイリング結果「のみに依拠して」行うかどうかにかかわらず、つまりは人間を一定程度関与させたとしても、DPIA が「とくに」要求されることになる[60]。これは、先述のように、22 条のいう「のみに依拠して」の意味が不確定で、22 条の権利が狭く解釈される可能性があるために、DPIA については、重要決定にプロファイリングを用いているのであれば、人間がどの程度関与しているかにかかわらず広く実施させようという趣旨に出たものと解される。その意味で、DPIA は「のみに依拠して」という文言の不確定性とそれによるリスクをカバーするものと考えることができる（なお、ガイドラインは、

[57] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 28.
[58] *Id.*
[59] *Id.* at 32.
[60] *Id.* at 29.

DPIA の一部として、管理者は意思決定プロセスにおける人間関与の程度やタイミングを調査し、記録すべきとしている[61]。この点、DPIA によって 22 条が求める人間の関与が自らチェックされると考えることもできる)。

GDPR 35 条 7 項は、DPIA は「少なくとも」以下の事項を含めなければならないとしている。

(a)「予定されている情報処理業務および当該処理の目的の体系的な記述。該当する場合には、管理者によって追求される正当な利益を含む」。

(b)「当該処理業務が、当該目的との関係で必要で、かつ比例的なものになっているか、に関するに評価」

(c)「データ主体の権利および自由に対するリスクの評価」

(d)「個人データの保護を確実にし、かつ、本規則の遵守を立証するためのセーフガード、安全管理措置およびメカニズムを含む、リスクに対処するために予定されている措置」

文言上「少なくとも」とあるように、上記（a）～（d）はあくまで例示であり、重要決定についてプロファイリングを用いる場合、さらに、「目標変数」、「新たなデータが収集されたかどうか、それがどのように収集されたかどうか」、「どのような特徴が考慮されたか」、「考慮はされたが選択されなかったオプションとその理由」(この理由には、「開発者が直面した実際上の制約や、決定を方向付けた諸価値に関する議論も含まれる」)、「選択された手法、選択されなかった手法それぞれの予想される影響」、「ありうる影響軽減手段の効果」などを DPIA に含めることが検討されている[62]。

先述のように、13 条～15 条がデータ管理者に提供を義務付ける情報――データ主体がアクセスし得る情報――の範囲は確定的ではなく、これを「結果ベースの説明」に重点を置くものと解釈することも不可能ではない。しかし、そうなると、13 条～15 条の「説明」では、予測モデル全体の機能や影響が十分に伝えられないことになる。そのことは、データ主体による予測モデル全体の理解を妨げ、個別の決定について実質的に争うことも難しくしよう。DPIA は、こうした「結果ベースの説明」の限界を補完し、予測モデル全体の機能・影響を伝えることで、データ主体が個別の決定について争いやすくすることにも寄与し得る[63]。

また、これも先述のように、機械学習の時代においては、違法とされる差別的行為、とりわけ「差別的インパクト」の意味が不明確化するうえ、データ主体が個別的決定の文脈でかかるインパクトを捕捉することも困難となる。その意味で、とくに上記 (c) の影響評価は、特定集団に対するネガティブな影響を管理者自身がチェックし、同集団のメンバーに対する実質的不利益を未然に防ぐために重要な手段ということになろう。

GDPR は、文言上、DPIA の「公表」まで義務付けていないが、ガイドラインは、その全てか一部かは別として、その公表を強く推奨している[64]。さらに、35 条 9 項は、「適切な場

---

61 at 21.

62 Selbst & Barocas, *supra* note 18, at 1134-1135.

63 GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 2, at 30.

64 ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON DATA PROTECTION IMPACT ASSESSMENT(DPIA) AND DETERMINING WHETHER PROCCESSING IS "LIKELY TO RESULT IN A HIGH RISK" FOR THE PURPOSE OF REGULATION 2016/679, at 18, WP 248(Apr. 4, 2017).

合に、管理者は、予定されている処理に関して、データ主体またはその代理人から意見を求めなければならない」と規定している。これらの要求は、プロファイリングを含む情報処理のアカウンタビリティと透明性を高め、ガバナンスの実効性を担保するために重要であろう。

なお、GDPR 36 条は、「35 条に基づく DPIA が、リスクを軽減させるために管理者によって講じられる措置が存在しない状況下で、その情報処理が自然人の権利および自由に対して高いリスクをもたらすおそれがあることを示している場合、管理者は、その処理を開始する前に監督機関と協議しなければならない」と規定している（事前協議。1 項）。

（3）行動規範（Codes of conduct）

GDPR 40 条は、「データ管理者または処理者を代表する団体およびその他の組織」が、「本規則の適用を具体化するために（for the purpose of specifying the application of this Regulation）」、行動規範を策定することを推奨している（1 項および 2 項）。この行動規範（案）は、所轄監督機関が「十分に適切な保護措置を提供するものであると判断するとき」、当該機関により承認される（5 項）。

この規定は、権利規定の不確定性にデータ管理者自身が――その代表団体等を通じて――正面から応対し、具体的な規範形成に積極的に関与することを求める規定であるように思われる。

（4）認証制度

GDPR 25 条および 42 条は、ガバナンス体制の構築と関連した認証制度（certification mechanism）の創設を奨励している。まず、「データ保護バイデザイン（Data protection by design）」を定める 25 条 1 項は、「技術水準、実装費用、情報処理の性質・範囲・文脈および目的ならびに当該処理によって引きこされる自然人の権利および自由に対する様々な蓋然性と深刻度のリスクを考慮に入れたうえで」、データ管理者は、「本規則の要求を満たし、かつ、データ主体の権利を保護するため」、「データ保護の基本原則を効果的な方法で実現し、情報処理のなかに必要なセーフガードを統合する……適切な技術的および組織的措置を実装しなければならない」（強調筆者）と規定している。そして、同条 3 項は、同条 1 項がいう「適切な技術的および組織的措置の実装」を行っていることを「証明するために」、GDPR 42 条の定める認証制度を利用できるとしているのである（認証は、43 条に規定される認証機関または所轄監督機関から発行される）。

フィンク（Michèle Finck）は、完全自動意思決定に関する諸権利の不確定性と、違法行為の発見困難性から、司法による事後的な救済よりも、データ保護バイデザインの思考と結び付いた認証制度が、権利保護のために有効である旨を説いている[65]。

（5）インセンティブの作出

以上みてきたように、GDPR は、完全自動意思決定に関する諸権利の不確定性と、違法行為等の同定困難性を、総合的なガバナンス体制の構築によって埋め合わせ、克服しようと試みているように思われる。しかし、ここで問題になるのは、データ管理者がこのようなガバナンス体制を構築するインセンティブをどのように作出するか、であろう。ガバナンス体

[65] Michèle Finck, *Smart Contracts as a Form of Solely Automated Processing under the GDPR*, 9 INT'L DATA PRIVACY L. 78, 92(2019).

制の構築・強化には当然コストがかかる。したがって、一定のコストをかけてでも行動規範の策定に参画し、倫理審査委員会のような組織体を創設し、積極的にアルゴリズム監査および DPIA を行い、認証制度にコミットするようなインセンティブをどのように作り出すかが問題になるのである。

結論から言えば、GDPR は、ガバナンス体制の構築・運用と制裁金制度とを連動させることで、こうしたインセンティブを作出しているように思われる。仮に権利侵害とみなされるような事態が生じたとしても、GDPR が要求・推奨するガバナンス体制を構築し、誠実に運用していた場合には制裁金を免除または軽減するものと予め規定しておくことで、いま述べたようなインセンティブを作り出しているように考えられるのである。

以下、そのインセンティブ構造を具体的にみておこう。まず、本稿が検討の対象としている完全自動意思決定に関する諸権利の侵害があった場合、一般には、「2000 万ユーロ以下の制裁金」または「事業の場合、直前の会計年度における世界全体における売上総額の 4%以下の金額、もしくは、いずれか高額の方の制裁金」に服するものとされる（83 条 5 項）。しかし、GDPR 83 条 2 項は、「個々の案件において、制裁金を科すか否かを判断する場合、および、制裁金の額を判断する場合、以下の事項を適正に考慮に入れなければならない」とし、例えば以下のような考慮事項を挙げている。

（d）「25 条〔データ保護バイデザイン〕および 32 条〔安全性確保の措置〕により管理者または処理者によって実装された技術上及び組織上の措置を考慮に入れたうえで、管理者または処理者の責任の程度」（強調筆者）

（f）「違反を解消するための、および、違反の潜在的な悪影響を低減させるための監督機関との協力の程度」（強調筆者）

（j）「40 条による承認された行動規範の遵守、または 42 条による承認された認証方法の遵守」

これらの事項は、いずれも先に検討したガバナンス体制の構築・運用と関連している（上記（f）は、35 条の DPIA の実施と関連している。36 条は、DPIA によって権利・自由に高いリスクをもたらすおそれがあることが示された場合、監督機関と事前協議を行うことを管理者に要求している。（f）はこの協議を誠実に行った場合、制裁金が免除・軽減される可能性を示唆している[66]。そうなると、ガバナンス体制を構築し、誠実に運用していた場合、完全自動意思決定に関する諸権利を万が一侵害したとしても、ともすると科されていたかもしれない莫大な制裁金を免除または軽減される可能性があるわけである。このことは、完全自動意思決定を導入しようとする事業者にとっては大きなメリットであり、コストをかけてでもガバナンス体制を整えておく強いインセンティブになり得るように思われる[67]。

GDPR は、完全自動意思決定の関連規定を含め、その規定の曖昧さや執行困難性などが厳しく批判されることがあるが、実際には、ガバナンス体制の構築・運用と制裁金との連動を法文上明記し、同体制を構築・運用するインセンティブを明示的に作出することで、権利侵害を未然に防ぐことをまずは企図しているように解される。この点で、GDPR は、行為

66 さらに 39 条 1 項（c）も参照。
67 この点は、個人情報保護法関連の検討会の席上で、森亮二弁護士から示唆を得た。記して感謝申し上げる。

ベースの規律よりもガバナンス・ベースの規律に重点を置くものと考えることができるのである[68]。

３．ガバナンス型統制の実例

もちろん、前章で検討したガバナンス・ベースの規律は真新しいものではない。法解釈の不確定性と執行困難性が問題となるような領域で、これまでも広く取り入れられてきた。本章ではその例をいくつか概観しておきたい。

３．１．労働法領域――雇用差別を中心に

アメリカでは、2000 年代以降、雇用差別に対する新たな法的アプローチとして、「構造的アプローチ（structural approach）」[69]が有力に説かれるようになった。それは、職場における「差別」が、意図的で個別的なものから、無意識的で組織的・構造的なものへと変容を遂げてきたことによる。かような「差別」形態には、明確に同定可能な個別の差別的行為を前提とした従来型の法規制では対応が困難であると考えられたわけである。「従来型の法的アプローチ（rule enforcement approach：ルール強制アプローチ）では、今日の複雑な構造をもつ差別が適法か違法かを判断することが難しいだけでなく、使用者のルールを潜り抜けようとする表面的な対応（責任回避的な行動）を促すことにもつながり、企業の組織や文化にも関わる複雑な問題を根本から解決していくができない」[70]と認識されたのである。確かに、「差別」が意図的でないとすると、何が禁止される（非難可能な）「差別」なのかが確定されないし、仮にそれが一定程度確定されたとしても、企業の組織内部で行われる限り外部に顕在化せず、その執行が困難となる（これは AI 社会における「プライバシー侵害」や「差別」の不確定性および発見困難性と同系の問題である）。そこで、有力な労働法研究者であるスタン（Susan Sturm）やエストランド（Cynthia Estlund）らが、個別的行為に着目した規律から、組織の構造・ガバナンスに着目した規律への転回を主張し、企業自身が――業界団体や政府組織と協働しながら――差別的行為を未然に防ぐ構造ないしガバナンス体制を構築することを促していくような法設計を考案したのである。以下、やや長いがスタン自身の説明を引用しておこう[71]。

「『適法性（legality）』は、情報収集、問題発見・認識、改善・矯正、そして評価という相

---

68 もちろん、このことは、個別的行為への視点が重要でないということを意味しない。個別の違法行為を燻り出すにも、まずはガバナンスによる監視の「目」が必要ということである。カティヤルは、この点と関連して、公益通報者保護制度の重要性を指摘している。Sonia K. Katyal, Private Accountability in the Age of Artificial Intelligence, 66 UCLA L. REV. 54, 126(2019). 日本でも、違法行為を外部から同定することが困難となる個人情報の取扱いについて、公益通報者保護制度を積極的に利用することが広く検討されてよい（もちろん、既に公益通報者保護法は「通報対象事実」として個人情報保護法違反に関する一定の事実を挙げている。2 条 3 項）。

69 *See e.g.*, Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458(2001); Cynthia Estlund, *Rebuilding the Law of the Workplace in an Era of Self-Regulation*, 105 COLUM L. REV. 319(2005); Michael C. Dorf, *Legal Indeterminacy and Institutional Design*, 78 N.Y.U. REV. 875(2003).

70 水町勇一郎『集団の再生』（有斐閣、2005 年）167 頁。

71 Sturm, *supra* note 68, at 463. 訳は水町・前掲注（70）168 頁による。

互作用的なプロセスから生まれ出てくる。この規制は、観察・発見された問題に対して、既存の概念的、職業的、組織的な境界線を越えてダイナミックに相互に作用しあうことを促すものである。このアプローチは、ある明確な法規制システムの一部として、情報の収集、組織のデザイン、インセンティヴをもたせる構造、実効性を向上させる措置、そして説明責任を制度化する方法に関する実験を行っていくことを奨励する。職場や、職場慣行に影響を与える非政府組織は、この規制体制のなかでは、単に国家や市場の規制の対象としてでなく、法を作り出す主体として取り扱われるのである」。

このような構造的アプローチを推進するうえでとくに重視されるのが、事業者が――コストをかけてでも――ガバナンス体制を構築し、これを誠実に運用していくためのインセンティブをいかに作出するか、である。この点で、エストランドは、ガバナンス体制の構築と司法審査との関係、すなわちガバナンス体制の構築・運用が裁判所においていかに評価されるのか、に注目した。そこで取り上げられるのは、職場での環境型セクシャル・ハラスメントに対する使用者責任を争った連邦最高裁の Burlington Industries, Inc. v. Ellerth 事件判決[72]（1998 年）と、Faragher v. City of Boca Raton 事件判決[73]（1998 年）である。これらの事件で、最高裁は、問題とされた上司の言動が敵対的環境をもたらすセクハラに該当するとしたうえ、原告は使用者に対して責任を追及し得るとした。しかし最高裁は、使用者に責任を免れるための積極的抗弁を認め、①使用者がセクハラ行為の予防および迅速な是正のために合理的な配慮を行っており、②セクハラの被害者たる被用者が、合理的な理由なく使用者の提供した防止・是正手段を利用しなかったことを証明すれば、使用者は責任を免れることができると述べた[74]。使用者が自ら誠実にガバナンス体制を構築・運用していたことと法的責任の免除とを関連付けたこれらの判決は、使用者が当該体制を構築・運用するインセンティブを作出したものとして、「構造的アプローチ」の提唱者によって積極的に評価されるのである[75]。

さらに、エストランドは、雇用差別からは離れるが、職業安全法（Occupational Safety and Health Act, OSHA）のアプローチにも注目している[76]。OSHA は「自発的な保護プログラム（Voluntary Protection Program, VPP）」を導入し、安全衛生基準の遵守と安全衛生状況改善を行う内部的なガバナンス体制を構築した使用者については一般的な（職業安全衛生局による）臨検を免除ないし一部簡易化するという仕組みを採用している。ＶＰＰには、ガバナンス体制の強度等に応じて「Star」、「Merit」、「Demonstration」という 3 つの地位があり、「Star」を得ると、臨検は 3 年ごとの地位更新時のみ受ければよいことになる。これも、ガバナンス体制の構築・運用のインセンティブを作り出す一つの法設計であると考えられる。

---

72 524 U.S. 742(1998).
73 524 U.S. 775(1998).
74 水町勇一郎編『個人か集団か？　変わる労働と法』（勁草書房、2006 年）186 頁（長谷川珠子執筆）、竹内（奥野）寿「アメリカにおける新たな労働者参加の試みとその法理論的基礎づけ」RIETI Discussion Paper Series 13-J-026（2013 年）8 頁。
75 *See also*, Kolstad v. American Dental Association, 527 U.S. 526(1999).
76 Estlund, *supra* note 68, at 343-344.　竹内（奥野）・前掲注（74）6‐7 頁参照。

３．２．情報プライバシー法領域

「構造的アプローチ」は日本の労働法学において既に好意的に紹介されているが[77]、かかるアプローチは、権利侵害ないし違法行為の解釈不確定性や同定・発見困難性を同様に抱える他の法領域においても有用であろう。例えば、情報プライバシー法の「構造論的転回（Structural turn）」が語られることがある。リチャーズ（Neil M. Richards）は、情報プライバシーの「権利」としての内容不確定性や、情報ネットワークシステム内部における違法行為等の同定困難性（例えば、利用目的に反するプロファイリングの発見困難性）などから、同法領域においては、システムの構造やガバナンスの規律に着目するアプローチが有力になっていると指摘している[78]。こうした見解を受け、日本においても、いくつかの裁判例を根拠に「構造論的転回」の可能性・必要性を主張する見解がある[79]。

かような見解において第１に挙げられるのは、最高裁による平成 20（2008）年の住基ネット事件判決[80]であろう。最高裁は、住基ネットの合憲性が争われたこの事件で、憲法 13 条は「個人に関する情報をみだりに第三者に開示又は公表されない自由」を保障していると述べたうえで、住基ネットの「構造」まで審査し、①漏えい等を防ぐ堅牢な（セキュリティ）システムが構築されていること、②目的外利用や漏えいが罰則等をもって禁止されていること、③監視機関等「情報の適切な取扱いを担保するための制度的措置」が講じられていることから、「住基ネットにシステム技術上又は法制度上の不備があり、そのために本人確認情報が法令等の根拠に基づかずに又は正当な行政目的の範囲を逸脱して第三者に開示又は公表される具体的な危険が生じているということもできない」とし、よって「上記自由を侵害するものではない」と結論付けた。ここでは、一般に「構造審査」と呼ばれる司法審査手法を通じて、「制度的措置」を含むガバナンス構造と権利侵害の認定とが関連付けられている[81]。

第２に、N システム（自動車ナンバー自動読取システム）の合憲性を認めた平成 19（2007）年の東京地裁判決[82]が挙げられる。この事件で東京地裁は、最終的には N システムの合憲性を認めるのだが、その審査場面において、「公権力による国民の私生活に関する情報の収集・管理が〔憲法 13〕条の趣旨に反し，国賠法上の違法性を有するか否かは，①公権力によって取得，保有，利用される情報……の性質はどのようなものか，②公権力がその情報を取得，保有，利用する目的が正当なものであるか，③公権力によるその情報の取得，保有，利用の方法が正当なものであるか，④公権力によるその情報の管理方法の厳格さはどの程度か，などを総合して判断すべきである」と述べた。とくに④は、まさにガバナンスの在り

---

77 前掲注（70）・（74）参照。構造的アプローチの限界を指摘するものに、Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1(2006).

78 Neil M. Richards, The Information Privacy Law Project, 94 GEO. L. J. 1087, 1092(2006); DANIEL J. SOLOVE, THE DIGITAL PERSON 97(2004).

79 山本・前掲注(2)3 頁以下参照。

80 最判平成 20 年 3 月 6 日民集 62 巻 3 号 665 頁。

81 山本龍彦「住基ネットの合憲性」長谷部恭男ほか編『憲法判例百選Ⅰ〔第 6 版〕』（有斐閣、2013 年）46 - 47 頁。

82 東京地判平成 19 年 12 月 16 日訴月 55 巻 12 号 3430 頁。

方（安全管理措置の適切さ）を審査しているように思われる。そうなると、下級審の判決ではあるが、ここでも、構造・ガバナンスと権利侵害の認定とが関連付けられていると考えられる[83]。

第３に、民事事件ではあるが、平成 18（2006）年の Yahoo!BB 顧客情報流出事件判決[84]がある。本件では、個人情報の適切な管理のために必要な措置を講じなかったために情報が不正に取得された場合、管理主体は不法行為責任を負うか、という点が争われた。大阪地裁は、この事件で、管理主体は、リモートアクセス実施に当たり「外部からの不正アクセスを防止するための相当な措置を講ずべき注意義務を怠った過失があり，同過失により本件不正取得を防ぐことができず，X らの個人情報が第三者により不正に取得されるに至った」とし、管理主体は，「本件不正取得により原告らの被った損害を賠償すべき不法行為責任がある」と結論付けた。不正アクセスが起こった場合に最も責められるべきは、不正アクセスの行為者であるが、本判決は、ガイドラインや個人情報保護法の規定を参照しながら、管理主体は、不正アクセスの防止等、個人情報の適切な管理のために必要な措置を講ずべき注意義務を負っているとし、この注意義務を怠った場合には管理主体が不法行為責任を負う場合もあり得るとしたのである。ここでも、管理主体の具体的行為ではなく、ガバナンス体制の不備が、不法行為責任の肯定と関連付けられていた。

以上のようにみると、ガバナンス・アプローチは、権利内容の不確定性や違法行為の同定困難性が問題となるような領域において既にみられてきたもので、完全自動意思決定に関する GDPR の同アプローチは決して真新しいものではない。GDPR の対応に新奇性が認められるとすれば、それは、ガバナンス・アプローチを明文化し、ガバナンス体制構築のインセンティブを可視化したところにあろう。

４．終わりに

以上、本稿は、完全自動意思決定に関する GDPR 上の諸権利の不確定性および権利侵害ないし違法行為の同定困難性（執行困難性）と、これを補完・克服するためのガバナンス・アプローチについて概観してきた。本論で述べたように、GDPR は、個別の権利侵害ないし違法行為を確定的に定義することの困難性、かかる個別行為を外部から同定・発見することの困難性から、事業者自身が行動規範等の策定に関与し、データ保護影響評価（DPIA）を行い、倫理審査委員会のような組織体を創設し、認証を受けることなどを通じて、個別の違法行為等の未然防止に積極的に取り組むためのインセンティブを作出しているように思われる。そして本稿では、個別的な行為に着目した行為ベースの規律というより、ガバナンスの在り方や構造（アーキテクチャ）に着目したガバナンス・ベースの規律を重視するこうした方向性は、他のいくつかの法領域で既に実践されており、法内容の不確定性や違法行為等の同定困難性を抱える完全自動意思決定の実施場面において基本的に肯定されるのではないかと主張した。

今後、完全自動意思決定のあるべき方法――AI の評価と人間の判断とのあるべき関係性

---

83 山本・前掲注（2）85 - 86 頁。また、コンビニの防犯カメラに関する東京地判平成 22 年 9 月 27 日判タ 1343 号 153 頁も、安全管理措置を問題にしている。
84 大阪地判平成 18 年 5 月 19 日判時 1984 号 122 頁。

――に関する議論が進むと、法内容の不確定性をめぐる問題は次第に解決されていくかもしれない。しかし、深層学習のような複雑高度な学習方法が広く用いられるようになると、何が権利侵害になるのか、何が違法行為になるのかが理論的に確定されたとしても、かかる「行為」を同定し、個別の救済を与えていくことは困難を極めよう。例えば、AI 社会における違法行為は、個人情報の持ち出しや漏えいといった従前の典型的な違法行為とは性質が異なり得るのである。個別の行為は、データ・エコシステムの一部として全体のなかに溶け込み、不可視化する。そうなると、本稿が完全自動意思決定との関係でひとまず検討したガバナンス・アプローチ[85]は、過渡的で暫定的なアプローチというのではなく、AI 社会における中心的な規律メカニズムとして長く定着していくことになるのかもしれない。

[85] ガバナンス・アプローチといわゆる共同規制モデルとの異同についてはまた別の機会に検討することにしたい。共同規制モデルについては、生貝直人『情報社会と共同規制』（勁草書房、2011 年）参照。