総務省 学術雑誌『情報通信政策研究』 第７巻第２号
Journal of Information and Communications Policy Vol.7 No.2

寄稿論文

AIに不適合なアルゴリズム回避論：機械的な人事採用選別と自動化バイアス

平野 晋[1]（中央大学）

要 旨

本稿は先ず、旧ソ連軍スタニスラフ・ペトロフ中佐が、アメリカからの核攻撃警報を鵜呑みにせず、直観的にこれを疑い、無事に核戦争を回避できた事例紹介を通じて、〈自動化バイアス〉に抗う〈意味のあるヒトの監視・決定等〉の重要性について読者の理解を促している。

続く第一章に於いては、広い射程を包含する概念として「AI」という文言を使いながら、AIの予測・決定・推奨等を信頼し過ぎる（over-trustな）ヒトの偏見としての自動化バイアスと、これに類する認知上の誤謬である〈確証バイアス〉等々を解説。その上で、これ等の誤謬に対する対策案を、総務省「AI利活用原則の各論点に対する詳説」、EU『AI規則案』、及びホワイト・ハウス『AI権利章典の青写真』から例示・紹介している。

そして第二章では、自動化バイアスとは真逆の、AIの予測の方がヒトより正しくても、ヒトはヒトによる意思決定を好みがちであるという、AIに対する信頼不足（under-trust）な偏見が、社会的損失を生むと主張する〈アルゴリズム回避〉を紹介。加えて、これに類する幾つかの、ヒトの意思決定尊重主義に対する批判的指摘の代表例――筆者はこれ等をまとめて「アルゴリズム回避論」と呼んでいる――に言及した上で、それ等アルゴリズム回避論に対する筆者の反論を展開している。

第三章に於いては、たとえヒトには〈アルゴリズム回避〉という偏見がはたらく場合があるとしても、それでも人事採用選別や被用者の評価等に於けるAI利用には慎重であるべき理由として、「AIによる予測がヒトよりも正しい」という前提/停止条件が満たされていない場合があること、AIの特性的にAI利用が妥当ではない場合があること、及び前提/停止条件が満たされてもAI利用には謙抑的であるべき倫理上の理由があること、を説明している。

その上で、「おわりに」に於いて筆者は、AI利用はその適した業務に限り、ヒトが適する業務との協働が望ましいと示唆している。

キーワード：採用 AI、アルゴリズム回避（algorithm aversion）、自動化バイアス（automation bias）、スタニスラフ・ペトロフ中佐、Human in/on the Loop、意味のあるヒトの監視・決定等（meaningful human oversight）、〈ルール〉のコモン・ロー対〈スタンダード〉の衡平法、〈相関関係〉対〈因果関係〉、人間の尊厳

[1] 中央大学 国際情報学部 教授・学部長、博士(総合政策)(中央大学)、ニューヨーク州弁護士．なお筆者は、本稿が言及する日本及び国際機関のAIの諸規範を創った内閣府「人間中心のAI社会原則[検討]会議」構成員、並びに総務省「AIネットワーク社会推進会議」副議長及び同会議「AIガバナンス検討会」座長を務めていて、経済協力開発機構「AI専門家会合」（AIGO エイゴー又はエイ・アイ・ゴー）の日本共同代表であったけれども、本稿の内容は私見である。

はじめに

〔[その事件が起こったのは、大韓航空機をソ連が撃墜した直後の、東西冷戦の緊張が高まった頃であった……。突如サイレンが鳴り響き、]ミニットマン大陸間弾道ミサイル５発が[、次々とアメリカから]ソ連に向けて発射されたことを、[衛星からの情報を選り分けた早期警戒システムのコンピュータ]画面は示していた。／報復攻撃を命じることを、赤軍の規則は定めていた。しかし 44 歳の中佐であるペトロフは、警報を無視して、これは誤報であるという自分の本能に従った。〕

サイレンが鳴ったのですが、私は、「**発射**：**Launch**」と明るく表示された大きな赤い画面を見つめたまま、ただそこに座っていたのです。……私は動けませんでした。私はまるで熱いフライパンの上に座っているように感じました。

〔別部門に所属する[地上]レーダー操作員達は、ミサイルが映っていないと言った。しかし彼等は単なる補助業務員であった――規則上は明確に、担当士官のペトロフ中佐が、コンピュータ情報の読み取りに従って決定を下すことになっていた。[更に、その判断は、僅(わず)か数分内に下されなければならなかった……。] それにも拘わらずペトロフ中佐は、ソ連陸軍本部の担当士官に架電して、システムが故障であると報告した。〕

23 分後、何も起こらなかったことを知りました。もし本当に[アメリカからの核]攻撃だったならば、もう結果は分かっていたでしょう。[私が感じたのは]そんな安堵でした。

〔誤報は、太陽光の雲への反射をミサイル発射である、と衛星が誤認した結果だったことが、後に判明した。〕

私達はコンピュータよりも賢いのです。コンピュータは私達が創ったのです[から]。

ソ連軍スタニスラフ・ペトロフ中佐の言葉（拙訳）[2].

---

[2] Marc Bennetts, *Soviet Officer Who Averted Cold War Nuclear Disaster Dies Aged 77*, GURDIAN, Sep. 18, 2017, https://www.theguardian.com/world/2017/sep/18/soviet-officer-who-averted-cold-war-nuclear-disaster-dies-aged-77 (last visited Feb. 11, 2024); Pavel Aksenov, *Stanislav Petrov: The Man Who May Have Saved the World*, BBC NEWS, Sep. 26, 2013, https://www.bbc.com/news/world-europe-24280831 (last visited Feb. 11, 2024). *See* Doug Irving, *How Artificial Intelligence Could Increase the Risk of Nuclear War*, RAND Corp., Apr. 24, 2018, https://www.rand.org/pubs/articles/2018/how-artificial-intelligence-could-increase-the-risk.html (last visited Feb. 23, 2024); Glen Pederson, *Stanislav Petrov World Hero*, 71 FELLOWSHIP 7, 9, Jul./Aug. 2005, https://www.proquest.com/openview/99d8a34911f1c57fa5ebc13ff5286c88/1?pq-origsite=gscholar&cbl=2041863 (last visited Feb. 23, 2024)(同事件を紹介). 拙考「ロボット法と学際法学：〈物語〉が伝達する不都合なメッセージ」『情報通信学会誌』35 巻 4 号 109, 109～10 頁(2018 年) [以下、拙考「ロボット法と学際法学」という] (映画「ターミネーター3」（ワーナーブラザーズ・ピクチャーズ, 2003 年）に於いて、核ミサイル発射の決定権をヒトからネットワーク型 AI の

通常、ヒトは自動システムの決定を疑うよりも、寧ろこれを信頼し過ぎる偏見がはたらき、これを〈自動化バイアス〉（automation bias）という[3]。上の引用文は、そのような自動化バイアスに抗[あらが]ったペトロフ中佐が、核ミサイル発射の警報を信頼せずに[4]、直観に従ってこれを誤報であると上官に伝えたことにより、人類の滅亡を回避できた英雄的エピソードである。中佐が警報を疑った理由の一つは、飛来する核ミサイルの発射数の少なさにあっ

〈Skynet〉に委譲した経緯と自動化—Human-out-of-the Loop—の危険性を紹介）；及び拙書『ロボット法（増補版）：AIとヒトの共生にむけて』49頁及びFig. 2-9（弘文堂, 2019年）[以下、拙書『ロボット法（増補版）』という](ソ連が報復攻撃を自動化した為に人類滅亡に至る恐怖を描いたスタンリー・キューブリック監督の名作「博士の異常な愛情」（コロンビア・ピクチャーズ, 1964年）を紹介)も参照.

[3] Antonio Coco, *Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crime*, 21 J. INT'L CRIM. JUSTICE 1, 2 (2023)； Rebecca Crootof et al., *Humans in the Loop*, 76 VAND. L. REV. 429, 469, 500 (2023)； Kevin Jon Heller, *The Concept of "the Human" in the Critique of Autonomous Weapons*, 15 HARV. NAT'L SEC. J. 1, 51 (2023) [*hereinafter referred to as* K. Heller]; Uuri Benoliel & Shmuel I. Becher, *Termination without Explanation*, 2022 U. ILL. L. REV. 1059, 1075-76； Ben Green, *The Flaw of Policies Requiring Human Oversight of Government Algorithms*, 45 COMP. L. & SEC. REV. 1, 7 (2022)； S Mo Jones-Jang & Yong Jin Park, *How Do People React to AI Failure? Automation Bias, Algorithm Aversion, and Perceived Controllability*, 28 J. COMPUTER-MEDIATED COMM. 1, 2 (2022), https://academic.oup.com/jcmc/article/28/1/zmac029/6827859 (last visited Feb. 11, 2024)； Thomas F. McInerney, *The Emergence of Intelligent Treaty Systems and the Future of International Law*, 2022 U. ILL. J.L. TECH. & POL'Y 259, 317； Andrew Keane Woods, *Robophobia*, 93 U. COLO. L. REV. 51, 98 (2022)； Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring Systems*, 34 HARV. J.L. & TECH. 621, 636 (2021) [*hereinafter referred to as* Ajunwa, *Automated Hiring System*]； Emily L. Drake, Note, *Evaluating Autonomous Weapons Systems: A Dichotomic Lens of Military Value and Accountability*, 53 COLUM. HUM. RTS. L. REV. 297, 325 (2021)； Daniel E. Ho et al., *Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains*, 98 DENV. L. REV. 753, 767 (2021)； David S. Rubenstein, *Acquiring Ethical AI*, 73 FLA. L. REV. 747, 775 (2021)； Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 681 (2020)； Caroline Kemper, *Kafkaesque AI? Legal Decision-Making in the Era of Machine Learning*, 24 INTELL. PROP. & TECH. L.J. 251, 289 (2020)； Johnathan D. Falcone, *Machine Learning Systems in Nuclear Command, Control, and Communications Architecture: Opportunities, Limitations, and Recommendations for Strategic Commanders,* Defense Technological Information Center, at 18, June 4, 2019, https://apps.dtic.mil/sti/citations/AD1079957 (last visited Feb. 11, 2024)； Michael C. Horowitz et al., *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence*, arXiv:1912.05291 [cs.CY], at 4, 7-8, Dec. 2019, https://arxiv.org/abs/1912.05291 (last visited Feb. 11, 2024)； Shin-Shin Hua, *Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control*, 51 GEO. J. INT'L L. 117, 141 (2019)； INTERNATIONAL COMMITTEE OF RED CROSS, ETHICS AND AUTONOMOUS WEAPON SYSTEMS: AN ETHICAL BASIS FOR HUMAN CONTROL?, 13, Apr. 3, 2018, https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control (last visited Feb. 25, 2024)； Thompson Chengeta, *Defining the Emerging Notion of "Meaningful Human Control" in Weapon Systems*, 49 N.Y.U. J. INT'L L. & POL. 833, 853 (2017)； Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271-72 (2008)； 平野晋「（資料1）AIの判断に対するヒトの最終決定権の限界：Human-in-the Loopの問題」30～37頁 *in* 総務省「情報通信法学研究会 令和5年 第1回」, https://www.soumu.go.jp/main_content/000899843.pdf (last visited Feb. 4, 2024) [以下「拙講演資料」という].

[4] Crootof et al., 76 VAND. L. REV. at 500 ("some degree of undertrust is useful: famously, undertrust may have averted what might have otherwise become World War III."(emphasis added)と記述して、システムをある程度信頼しないことが大惨事回避に貢献したと評価).

た――アメリカが核攻撃を仕掛ける場合は、ソ連の基地の全滅を試みてもっと多数のミサイルを一斉発射するはずであると聞かされていたから、僅か５発という警報に疑念を持ったのである[5]。警報を疑ったもう一つの理由は、地平線から上昇して来るはずのミサイルを地上のレーダーが把握していないことだった[6]。そこで彼は誤報であると判断したのである。（なお後述するように[7]、そのように広い文脈を理解して、事案毎に異なる諸要素も考慮に入れて臨機応変な決定を行う能力は、AI[8]には欠けていると云われている。）

ところで重要な決定に関してはAIに全てを委ねずに、ヒトが最終決定を下したりヒトが監視する機能を実装する場合があるけれども、ヒトの決定・監視等は、自動化バイアスによる偏見の為[ため]に機能しないおそれも認識され始めたので、その偏見除去（debias）手続の実装が推奨されたり義務化されつつある[9]。

ところが近年、自動化バイアスによる偏見とは真逆の主張として、AIによる予測がヒトよりも正しいにも拘わらず、これに委ねることを嫌悪する〈アルゴリズム回避〉（algorithm aversion）と呼ばれるヒトの偏見が指摘され、これが不当にAIへの決定機能等の委譲を阻害することは、社会的損失を生むと批判され始めた[10]。

本稿では以上に関して、主に人事採用選別や被用者の評価等に於けるAI利用の問題として[11]、先ず第一章に於いて、自動化バイアスと、対策としての〈意味のあるヒトの監視・決定等〉（meaningful human oversight）について説明する。続く第二章に於いては、〈アルゴリズム回避〉の偏見ゆえにAIの普及が妨げられているという批判を紹介した上で、これに類似する幾つかの、ヒトによる意思決定尊重主義批判に対する筆者の反論を述べておく。最後の第三章では、たとえヒトには〈アルゴリズム回避〉という偏見がはたらく場合があるとしても、それでもAIの利用には慎重であるべき理由として、A.「AIによる予測がヒトよりも正しい」という前提/停止条件が満たされていない場合、B．AIの特性的・

---

[5] *E.g.*, David Hoffman, "*I Had a Funny Feeling in My Gut*," WASH. POST, Feb. 10, 1999, https://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/shatter021099b.htm (last visited Feb. 11, 2024).
[6] *Id.* 更に望遠鏡を使った監視でも、アメリカからの核ミサイルを補足できなかった。Irving, *supra* note 2.
[7] 後掲第三章B節２項参照.
[8]「AI」という文言について、内閣府「人間中心のAI社会原則」が「AI」を「高度に複雑な情報システム一般」と広く定義していることに倣[なら]って、本稿でもロボット、機械、アルゴリズム、自動意思決定（ADM: automated decision-making）、機械学習、ピープルアナリティックス（people analytics）、及び人工知能等の広い射程に及ぶように用いている。*See also* Woods, *supra* note 3, at 53 n.1 (ヒトによる決定に代わって、ヒトでは無いものによる自動決定の社会的受容性を扱う論文なので、「ロボット」「機械」「アルゴリズム」「AI」等を分け隔てなく扱うと説明している). なお本稿が関心を寄せる人事採用選別や被用者の評価等に於けるAI利用に於いて主に利用されるAIの機能は、連邦雇用機会均等委員等が執筆した論文によれば、機械学習と自然言語処理（文章や発話をヒトのように理解する能力）の２つであると指摘されている。Keith E. Sonderling et al., *The Promises and the Peril: Artificial Intelligence and Employment Discrimination*, 77 U. MIAMI L. REV. 1, 14 (2022). また同論文によれば、AIとは、データとコンピュータ技術を用いて決定を下すか、又はヒトが下す決定を補助するシステムである、と定義している。*Id.*
[9] 後掲第一章B節及びD節参照.
[10] 後掲第二章参照.
[11] 主に人事採用選別に於けるAI利用の問題については、拙講演資料・前掲注（3）等参照.

限界的に AI 利用が妥当ではない場合、及び C．前提/停止条件が満たされても倫理的に AI 利用には謙抑的であるべき場合について説明する。

第一章〈自動化バイアス〉と〈意味のあるヒトの監視・決定等〉

A．ヒトの判断の介在：Human in/on/off the Loop

AI に様々な業務（tasks）を任せる際に、その最終決定までも AI に委ねることが許されるか、又は AI に委ねないでヒトが最終的に決定すべきか、という論点が存在する。ロボット兵器の文脈等に於いては、その論点を〈human in the loop〉、〈── on the loop〉、又は〈── out of the loop〉等と表現する[12]。文脈によってその用語の使い方に差異やゆらぎがない訳ではないようであるけれども、主に〈── in the loop〉は決定を常にヒトが行う場合であり、〈── on the loop〉はいわゆる〈キル・スイッチ〉のように AI に最終決定を原則として任せつつもヒトが中止を決定出来るようなヒトの介入を許す考え方であり、〈── out of the loop〉は完全に AI に最終決定も委ねてしまう場合を表す[13]。

ちなみに総務省 AI ネットワーク社会推進会議がとりまとめた日本に於ける AI 利活用の代表的規範である、「AI 利活用原則」は、その各論点に対する詳説に於いて[14]、「人間の判断の介在」が必要な場合の考慮要素や具体例等を示している。特に「⑧公平性の原則」に関しては、「意思決定（判断）に対して納得ある理由を必要とする場合」、すなわち「例えば、人事評価に当たっては、社員に対し評価の理由を説明できることが期待される」と説明している[15]。

欧州に於いても、欧州連合（EU）のいわゆる『AI 規則(*AI Act*)案』が、その附属文書 III（ANNEX III）の 4 に於いて雇用や人事採用選考等に於ける AI システムの利用を「ハイリスク」に分類した上で[16]、同第 14 条 1 項ではその利用に於けるヒトによる監視を義務付けている[17]。また、EU の『一般データ保護規則』（GDPR）の第 22 条は、自動処理のみで最

---

12 拙書『ロボット法（増補版）』・前掲注（2）84 頁.

13 同上； 拙講演資料・前掲注（3）4 頁； Coco, *supra* note 3, at 3-4. *See also* Crootof et al., *supra* note 3, at 441 n.37； Horowitz et al., *supra* note 3, at 9. なお<human out of the loop>は<――off the loop>と表現される場合もある。Crootof et al., 76 VAND. L. REV. at 441 n.37, 444.

14 総務省 AI ネットワーク社会推進会議「報告書 2019 別紙 1 （附属資料）AI 利活用原則の各論点に対する詳説」令和元年 8 月 9 日 5, 31, 34, 52, 53 頁, https://www.soumu.go.jp/main_content/000637098.pdf (last visited Feb. 4, 2024).

15 同上 34 頁. なお評価システム（scoring system）が「公正さというヒトの価値観」――“human values of fairness”――を維持する為には、ヒトによるレビューが必要であると指摘する文献として、see Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 81 WASH. L. REV. 1, 6 (2014).

16 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, at Annex III 3 & 4, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (last visited Feb. 20, 2024) [*hereinafter referred to as AI Act*]. なお本稿は 2024 年以降の *AI Act* 案のあり得べき修正は反映していない。

17 *Id.* at Art. 14 (1). なお本文の次のセンテンスが言及している GDPR に於ける規制については、山本龍彦先生が本誌第 3 巻 1 号 I-25 頁に「『完全自動意思決定』のガバナンス―行為規制型規律からガバナンス統制型規律へ？―」https://www.soumu.go.jp/main_content/000656380.pdf (last visited Feb. 25, 2024)を寄稿されているので参照して欲しい。更に GDPR については、小向太郎＝石井夏生利『概説 GDPR』（NTT 出版, 2019 年）も参照.

終 決定を下すことを禁じており、ヒトの介在を要求している。

アメリカに於いても、大統領府の『AI 権利章典の青写真』（BLUEPRINT FOR AN AI BILL OF RIGHTS）は、労働や教育等の〈機微な分野〉（“sensitive domains”）に於いては一層の権利保障が必要であるとして、ヒトによる考慮を組み込むべきであり、ヒトによる広範囲な監視に服すべきでもあり、ヒトが介入して AI の決定を撤回させるような治癒策も必要である等と指摘している[18]。加えて、適切な場合には、AI による決定を拒否して代わりにヒトによる決定を選択（opt out）する権利を賦与すべきとさえ指摘し、しかもヒトの決定を選択した場合に手続を遅延させられる等の不利益を被らせるべきではないとしている[19]。

なお opt out のような権利を、ニューヨーク市条例もハードローとして採用している事実は[20]特筆に値しよう。

Ｂ.〈意味のあるヒトの監視・決定等〉とは

以上のようなヒトによる監視・決定等は、単に形式的に組み込まれるだけでは足りず、実質的に機能すること迄も要求される傾向が日・米・欧の三極に於いて見受けられる。学説に於いても、Human-in/on the Loop（以下「HITL」と略称する）な形式を採用しても、本稿の〈はじめに〉に於いて触れた〈自動化バイアス〉の影響によってヒトの監視・決定は機能せず、単に見せかけの、「いい加減な押印」（rubber stamping）的役割を担うに過ぎないし、HITL の外見は却ってヒトが監視・決定しているから安心であるという弛緩して誤ったメッセージを与える悪影響さえ指摘されている[21]。すなわち真に「意味のある」（meaningful な）ヒトの監視・決定等が必要なのである。

そこで日本に於いては、例えば前述した「AI 利活用原則の各論点に対する詳説」が、以

---

[18] WHITE HOUSE OFF. OF SCI. AND TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE 6, 7, 47, 49 (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf (last visited Feb. 4, 2024)[*hereinafter referred to as* WHITE HOUSE, AI BILL OF RIGHTS].

[19] *E.g.*, *id.* at 7, 47.

[20] NEW YORK CITY LOCAL LAW 144 (2021); Malika Dargan, Comment, *Model Act for Algorithmic Models: A Regulatory Solution for AI Used in Hiring Decisions*, 13 HOUSTON L. REV. ONLINE 50, 55-56 (2023); Lindsey Fuchs, *Hired by Machine: Can a New York City Law Enforce Algorithmic Fairness in Hiring Practices?*, 28 FORDHAM J. CORP. & FIN. L. 185, 212 (2023); *No.2 NYC Employers Using AI for Screening Beware*, Nov. 21, 2022, 20221121P NYCBAR 1 [City Bar Center for Continuing Legal Education New York City Bar], *available at* 2022WL 18027898; Paul J. Sweeney, *NYC Will Be Watching: Is Your Hiring Program Compliant?*, 29 No. 6 N.Y. EMP. L. LETTER 1, June 2022.

[21] David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn about Machine Learning*, 51 U.C. DAVIS L. REV. 653, 716 (2017). *See also* Green, *supra* note 3, at 1, 7, 9 (実証研究に基づいて、ヒトによる監視が機能していないと指摘し、かつ実際は機能しないヒトの監視の形式が AI 利用を正当化していると批判); Crootof et al., *supra* note 3, at 507 (意味のないヒトの監視は外見だけ倫理性を装っていると批判). なお“meaningful human oversight”に類似する概念を、主にロボット兵器の文脈に於いては“meaningful human control [MHC]”と呼ぶ例が多い。*See, e.g.*, K. Heller, *supra* note 3 (完全自律型兵器への反対運動の文脈から meaningful human control の概念を説明); Chengeta, *supra* note 3, at 836-37; Hua, *supra* note 3, at 117, 119, 131; 新保史生「自律型致死兵器システム（LAWS）に関するロボット法的視点からの考察」13 IEICE FUNDAMENTALS REVIEW [ファンダム・レビュー] 217, 222 頁(電子情報通信学会基礎・境界ソサイアティ, 2020 年 1 月)(完全自律型兵器の使用の文脈で MHC の概念が出て来た背景を説明).

下のように指摘して、AIと異なる判断をヒトが行えるような配慮をすべき、としている。

> AIによりなされた判断について人間が最終判断をすることが適当とされている場合に、人間がAIと異なる判断をすることが期待できなくなることも想定されることから、説明可能性を有するAIから得られる説明を前提として、人間が判断すべき項目を事前に明確化しておくこと等により、人間の判断の実効性を確保することが期待される[22]。

EUに於いても例えば、『AI規則案』第14条が以下のように規定することにより[23]、ヒトによる監視が形式ではなく実質的に機能するよう要求している。

> ……ヒトによる監視を任された個人が、状況に応じて適切に、以下を実施することが出来るようにしなければならない。
>
> (a) ハイリスクなAIシステムの能力と限界を完全に理解し、かつその運用を適切に監視することにより、異常、誤作動、及び予期せぬ履行の兆候を出来るだけ早く発見し、かつ対応出来るようにすること。
>
> (b) 特に自然人が下す決定に対しての情報や奨励を提供する為に用いられるハイリスクAIシステムに関して、ハイリスクAIシステムが提供するアウトプットに自動的に依拠するか又は過剰に依拠する傾向（「自動化バイアス」）があり得ることを認識し続けること。
>
> (c) 特にシステムの特徴と解釈の道具と利用可能な方法を考慮に入れて、ハイリスクなAIシステムを正しく解釈できること。
>
> (d) 如何(いか)なる具体的状況に於いても、ハイリスクAIシステムのアウトプットを使用せず、又は不使用以外にもそのアウトプットを無視したり、これに取って代えたり、あるいはこれを覆す決定が出来ること。及び、
>
> (e) 「ストップ」ボタン又は同様な手続を通じて、ハイリスクAIシステムの運用に介入し、又はシステムを中段することが出来ること。

アメリカに於いては例えば『AI権利章典の青写真』が、ヒトの介入が効果的に機能するように、ヒトへのトレーニングや評価を実施すべきと指摘して[24]、具体的には以下のように記述している[25]。

---

[22] AIネットワーク社会推進会議「詳説」・前掲注（14）5頁（強調付加）.
[23] *AI Act*, *supra* note 16, at Art. 14 (4) (a)-(e)(emphasis added)(筆者拙訳). *See also* Crootof et al., *supra* note 3, at 447-48 (EUのAI規則案が、意味のある監視を義務付けている事実を指摘).
[24] WHITE HOUSE, AI BILL OF RIGHTS, *supra* note 18, at 48 (筆者拙訳)(強調付加).
[25] *Id.* at 50 (拙訳)(強調付加).

トレーニングと評価. 自動システムのアウトプットを管理し、それと相互作用し、又は解釈する者は誰でも、システムの意図された目的に照らしたシステムのアウトプットの適切な解釈方法も含むシステムのトレーニング、及び自動化バイアスの効果を緩和する方法のトレーニングを受講すべきである。トレーニングは、システムに従って更新されることを確かなものとし、かつシステムが適切に使用されることを確かなものとすべく、定期的に繰り返し実施されるべきである。

監視. ヒトに基づく[監視]システムには、自動化バイアスを含む偏見の可能性、及びそのシステムの有効性を限定するかもしれないその他の懸念がある。そのようなヒトに基づく[監視]システムの有効性の評価結果と偏見の可能性の評価結果は、そのような悪影響を緩和できるように、ヒトに基づく[監視]システムの運用を更新させ得るガバナンスの構造によって監督されるべきである。

C．そもそも〈自動化バイアス〉とは

以上のようにEUもアメリカも、〈自動化バイアス〉による偏見のおそれと、その具体的な対策を指摘していた。日本の「詳説」も、「人間がAIと異なる判断をすることが期待できなくなることも想定されることから…しておくこと等により、人間の判断の実効性を確保することが期待される」と指摘して、自動化バイアスのおそれに対する対策を示唆している。更に内閣府の「人間中心のAI社会原則」も、自動化バイアスの危険性を以下のように示唆している。

（１）人間中心の原則

……。AIが活用される社会において、人々がAIに過度に依存したり、AIを悪用して人の意思決定を操作したりすることのないよう、我々は、リテラシー教育や適正な利用の促進などのための適切な仕組みを導入することが望ましい。

そもそも〈自動化バイアス〉とは、本来ヒトが精査すべきところを放棄して、「簡便な判断方法」（heuristics）――以下「ヒューリスティック」という――として自動化に依拠し過ぎるヒトの態度や傾向のことをいう[26]。言い換えれば、AIの予測・決定・推奨等々の正しさや公正さ等をヒトが精査せずに、AIの言いなりになってヒトがこれを追認しがちなヒトの態度や傾向のことを、自動化バイアスという。謂わばAIを「信頼し過ぎ」（over-trust）なヒトの態度や傾向が、自動化バイアスである。

なお自動化バイアスを定義・説明する際にしばしば用いられる「ヒューリスティック」の文言が示唆するように、自動化バイアスの概念は、少なからず不合理な意思決定や行動をするヒトの傾向等を研究する〈行動経済学〉、〈行動科学〉、及び〈認知科学〉等に由来する概

[26] Citron, *supra* note 3, at 1271-72; Cary Coglianese & Alicia Lai, *Algorithm v. Algorithm*, 71 DUKE L.J. 1281, 1293 (2022)(利用可能性ヒューリスティック等を説明); 拙書『アメリカ不法行為法：主要概念と学際法理』359～61頁(中央大学出版部, 2006年) [以下、拙書『アメリカ不法行為法』という] (ヒューリスティックを説明).

念である[27]。

従って自動化バイアスは、行動経済学等で指摘されている〈投錨と調整〉（anchoring and adjustment）に似ているという指摘も見受けられる[28]。投錨と調整とは、一度見せられた数値が投錨のように記憶に残ってしまう為に、被験者がその後に無関係な数字（例えば国連加盟国の数）を言い当てるテストを受けると、先に見せられた数値に近似する数字を無意識に言ってしまうような傾向をいい、記憶を払拭する――すなわち〈調整〉する――ことが難しいヒトの偏見のことである[29]。この投錨と調整の偏見ゆえに、AIによる自動的な決定を最初に示唆されたヒトは、無意識に自らの決定をAIの決定と一致させてしまう偏見がはたらくと指摘されているのである[30]。

更に行動経済学等に於ける〈確証バイアス〉（confirmation bias）も、自動化バイアスに近い概念としてヒトの偏見に繋がると指摘されている[31]。確証バイアスとは、自分の信じるところを支持する情報ばかりを探して（“cherry-pick evidence”）、相反する情報は無視しがちになる偏見である[32]。従って例えば、捜査官が〈予測警備〉（predictive policing）[33]の予測した結果を肯定・補強する情報ばかりを求めつつ、相反する証拠は求めない捜査をしがちである偏見が懸念されている[34]。自動化バイアスや確証バイアス故に、ヒトはシステムを信頼し過ぎであると云われている[35]。

確証バイアスに加えて、「自己満足(不適切な監視)」（complacency）もAIの予測・決定・推奨等を信頼し過ぎるヒトの偏見に繋がると指摘されている。自己満足(不適切な監視)とは、自動化が進むにつれて、ヒトの監視が弛緩しがちな傾向をいう[36]。航空機の自動化が進む

---

[27] 例えば、拙書『アメリカ不法行為法』・前掲注（26）359～60頁等参照．*See also* April Lea Pope, *To Behave or Not to Behave: How Behavioral Science Can Inform Policy and the Law*, 59 -Apr ADVOCATE (IDAHO) 41, 41 (Mar./Apr. 2016)(心理学、神経科学、行動経済学、及び認知科学は全て行動科学の分野であると指摘)． なお本文中の「不合理な意思決定や行動をするヒトの傾向等」は、法と経済学等に於いて「限定合理性」（bounded rationality）と呼ばれる。拙書『アメリカ不法行為法』352～53頁及び脚注10等参照．

[28] Coglianese & Lai, *supra* note 26, at 1295； Kemper *supra* note 3, at 289, 292; Benoliel & Becher, *supra* note 3, at 1075-76. *See also* K. Heller, *supra* note 3, at 38-39.

[29] 拙書『アメリカ不法行為法』・前掲注（26）367～70頁参照． *See also* Shiri Krebs, *Drone Visuals and Compliance with IHL*, 116 AM. SOC'Y INT'L L. PROC. 132, 135 (2022)(投錨を説明)； Jeffrey J. Rachlinski, Essay, *The "New" Law and Psychology: A Reply to Critics, Skeptics, and Cautious Supporters*, 85 CORNELL L. REV. 739, 751 n.60 (2000)(投錨を説明)； Amos Tversky & Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1128-30 (1974)(投錨と調整を解説)． なお「あと知恵の偏見」（hindsight bias）も投錨と調整の一種と捉えることが出来よう。拙書『アメリカ不法行為法』370～74頁参照．

[30] *See, e.g.*, Kemper *supra* note 3, at 276-77； Coglianese & Lai, *supra* note 26, at 1295.

[31] *See, e.g.*, Drake, *supra* note 3, at 325, 327； Chengeta, *supra* note 3, at 853; McInerney, *supra* note 3, at 328.

[32] 拙書『アメリカ不法行為法』・前掲注（26）376頁；及びCoglianese & Lai, *supra* note 26, at 1294-95等参照．なお自動化バイアスや確証バイアスは、ヒトが見たいものを見るという〈同化バイアス〉(assimilation bias)に似ているという指摘もある。Chengeta, *supra* note 3, at 853.

[33] 「予測警備」については、例えば拙書『ロボット法（増補版）』・前掲注（2）94～98頁参照．

[34] *See, e.g.*, Molly Griffard, *A Bias-Free Predictive Policing Tool?: An Evaluation of the NYPD's Patternizr*, 47 FORDHAM URB. L.J. 43, 56 (2019).

[35] *See, e.g.*, Drake, *supra* note 3, at 327, 337.

[36] Joseph Avery, *Fumble! Anti-Human Bias in the Wake of Socio-Technical System Failure*, 53 ARIZ. ST. L.J. 1009, 1025 (2022)； 前東晃礼「自動化システムの使用と信頼の

と、パイロットがこれに依存し過ぎて注意が弛緩し、かつ嘗(かつ)てパイロットが修得していた技能も失って、延(ひ)いては航空機事故に繋がると指摘されていた現象である [37]。AI の利用分野に於いても、自己満足(不適切な監視)によってヒトの決定が弛緩すると指摘されている [38]。

以上のような、AI の決定を信頼し過ぎる偏見が望ましくない結果を生む代表例は、自動化された兵器分野の事故事例に見ることが出来る。例えば、2003 年のイラク戦争当時、アメリカ陸軍の地対空ミサイル〈パトリオット〉が、友軍の英国〈トルネード〉戦闘機を撃墜し乗員２名を即死させた事件が、自動化バイアスの招く惨事として悪名高い [39]。トルネード戦闘機をイラク軍からの対レーダー・ミサイル攻撃であると誤認したパトリオットからの指示を、操作員達が、トレーニング不足等ゆえに疑念を抱くことなく妄信し、裏付け無しにその指示通りにパトリオットを発射して生じた事故であった。逆に、自動化バイアスに抗うヒトの行動が、人類の大惨事を回避できた有名な事例は、冒頭で紹介したペトロフ中佐の例である。（冒頭で紹介した中佐の行動を読み返すと、本章前掲 B. で紹介した EU の『AI 規則案』第 14 条が要求する自動化バイアス除去手続に附合する点が見受けられるので、興味深い。）

Ｄ．自動化バイアス対策の必要性

ヒトの生死に係る自動化された兵器分野に限らずに、本稿が主に例示する人事採用選別や被用者の評価のような分野に於ける AI 利用も、人生の重大事に関わる決定である上に、欧米では「ハイリスク」又は「機微な分野」であると分類されて、より一層慎重な AI 利用が求められていることに鑑みると、そこに於いても、やはり欧米で要求されている対策を日本も採用することが、民主主義と人権保障という共通の価値観の擁護者である日本の地位を保つ上でも重要であろう。

AI の決定を信頼し過ぎる偏見を除去する方策としては、トレーニングを受講する [40]等の、本章前掲 B. にて示した日・欧・米に於ける具体例以外にも、例えば、AI の予測・決定・奨励等を最終的に判断するヒトに、<u>その判断の根拠・理由を十分に説明する責任を課す</u>手法が指摘されている [41]。単に AI(アイ)ちゃんがそう言った（“the machine told me so”）からそれに従った、といった理由は不十分であると取り扱うことにより、AI 以外の情報源にも目を向けさせて [42]——あたかもペトロフ中佐が地上レーダーにミサイルが映っていない事実に目を向けたように——、安易にヒューリスティックに頼らせず、きちんと熟考させるよう

役割」21 COGNITIVE STUDIES 100, 104 (Mar. 2013) https://www.jstage.jst.go.jp/article/jcss/21/1/21_100/_pdf (last visited Feb. 25, 2024)(元々は航空業界で使われていた用語であると指摘).

37 Woods, *supra* note 3, at 103.

38 *See, e.g.*, Jeffrey L. Vagle, *Tightening the OODA Loop: Police Militarization, Race, and Algorithmic Surveillance*, 22 MICH. J. RACE & L. 101, 129 (2016).

39 Coco, *supra* note 3, at 2 & n.2; Horowitz et al, *supra* note 3, at 8.

40 自動化バイアスの偏見除去手法としてトレーニングの重要性を指摘する文献は、本文中で紹介した AI 規則案や AI 権利章典以外にも、多数見受けられる。*See, e.g.*, Citron, *supra* note 3, at 1306; Kemper, *supra* note 3 , at 289, 290.

41 Vagle, *supra* note 38, at 135-36. *See also* Citron, *supra* note 3, at 1307 (行政が AI の予測・決定・奨励等を利用した場合に担当官に説明させることにより自動化バイアスの偏見除去に資すると指摘).

42 *See* Vagle, 22 MICH. J. RACE & L., at 135-36.

仕向ける工夫である。他にも例えば、**そもそもヒトは機械を信頼し過ぎるのだから、逆に信頼しない＝不信（distrust）をシステムに組み込んで、不信がある場合には初期値をAIに頼らない設定にするという対策が重要**という指摘も見受けられる[43]。すなわちヒトがAIに委ねることに不信を抱く業務では、ヒト自身が意思決定を下すことを初期値＝原則とする訳である。確かにAIに不信を感じるヒトの直観は、そもそもAIには不適切な業務を委ねている違和感（perceived mismatch）を示唆する契機（signal）なのだから、その不信、直観、又は違和感を不当であると切り捨てることなく、不信・直観・違和感を契機としてAIに委ねていた業務の妥当性や適合性を見直すべきという指摘も見受けられるので（後掲「おわりに」の脚注で言及する「正確性だけでは十分ではない」論文参照）、AIを信頼しないこと＝ヒトの意思決定等を初期値に設定する主張に合理性があると思われる。

第二章　〈アルゴリズム回避〉の逆襲

以上のように自動化バイアスや確証バイアス等のヒトの偏見の弊害は、AI利用時の意味のあるヒトの監視・決定等の重要性を説明し、AI利用の慎重さをELSIの価値観に基づいて喚起する根拠として主に紹介されていた。しかし近年、これとは真逆に、主にAI利用の萎縮を憂いて、ELSI 的慎重さに批判的な主張の根拠として、〈アルゴリズム回避〉という偏見が指摘されている。

Ａ．〈アルゴリズム回避〉とは

そもそもアルゴリズム回避とは、Dietvorst達の論文「アルゴリズム回避：人々はアルゴリズムの誤謬を見た後に誤ってアルゴリズムを回避する」『実験心理学雑誌』144巻1号114頁（2014年）に於いて命名された偏見である[44]。MBA（経営学修士）課程の入試を模した実証実験結果に基づく同論文によれば、人々は、アルゴリズムが予測を誤るのを目の当たりにすると、アルゴリズムの決定を回避して、正確さの劣るヒトの予測・決定を選択しがちであるという偏見に左右される――これを「アルゴリズム回避」と呼称している[45]――。すなわち人々は、ヒトの犯す誤謬に対しては寛容であるけれども、アルゴリズムの誤謬に対しては非寛容である、と分析している。そして、このように、より予測の正確なアルゴリズムを回避するヒトの傾向は、延いては社会の損失を生むから望ましくない、と結論付けている論文である。

Ｂ．反ELSI的主張：〈アルゴリズム回避論〉

AIに関する偏見は、第一章にて紹介した自動化バイアスのように〈信頼し過ぎる―over-

---

[43] *See* Karen Hao, *We Need to Design Distrust into AI Systems to Make Them Safer*, MIT TECH. REV. (May 13, 2021), https://www.technologyreview.com/2021/05/13/1024874/ai-ayanna-howard-trust-robots/ (last visited Feb. 19. 2024). *Cf.,* Crootof et al., *supra* note 3, at 500-01 & n. 344 (同上Hanoを出典とし、ペトロフ中佐の例を挙げてAIへの不信を初期値とする対策の有用性を示唆しつつも、不信が危険性を生む自動的兵器の例も挙げて、信頼レベルの最適化を推奨).

[44] Berkeley J. Dietvorst et al., *Algorithm Aversion: People Erroneously Avoid Algorithm after Seeing Them Err*, 144 J. EXPERIMENTAL PSYCHO. 114 (2015).

[45] *Id.*

trust な―偏見〉が社会に弊害を生む場合だけではなく、逆に、アルゴリズム回避のように〈信頼しなさ過ぎる―under-trust な―偏見〉が社会に弊害を生む場合もあるという指摘は、一般論としてはそれなりに説得力がない訳ではない。何故ならば行動経済学等に於いては嘗てから、新規な技術をヒトは否定的に捉えがちであるという偏見が指摘されていたので[46]、その延長戦上にアルゴリズム回避が存在すると捉えることが可能だからである。あるいはこれも拙書が指摘・紹介した〈利用可能性ヒューリスティック〉（availability heuristics）や〈代表性ヒューリスティック〉（representative heuristics）の偏見から説明すれば[47]、AI を利用した予測・決定・奨励等が誤謬を犯した事件は、その頻度の多寡や発生蓋然性の高低とは無関係に、目立って報道されかつ引用や再引用が繰り返されて、アマゾン社による AI 採用が女性差別的効果を生じて同社が使用を禁じた事例[48]のようにデジタル・タトゥー化されて目立つから、人々の脳裏に深く広く焼き付く結果、その期待事故費用（expected accident costs）を越えて理不尽な迄に危険であると誤認されるおそれもあろう。

更に、〈法の行動科学〉的学際研究の専門家であり法学と心理学の二つの学位を持つコーネル大学ロースクールの Rachlinski 教授も共著論文に於いて、自動運転車に対しては裁判官が従来型自動車に対してよりも賠償責任を課し易く、前者を否定的に捉えがちであると指摘している[49]。加えて、筆者の経験からも、ロボット・カー（自動運転車）のような新規な製品が重大事故を一件でも起こしたら社会から排除されてしまうところ、通常の従来型自動車は毎日世界中で何百件もの死亡事故を起こしているにも拘わらず禁止されない……、とロボット産業の関係者の感想を耳にしたこともある。従って AI が予測・決定等を下す役務のような〈新規な製品〉に対して、社会は因習的な製品に対する以上に厳しく否定的に捉えがちであるという指摘には、直感的にも説得力がある。

しかしここで筆者が懸念するのは、アルゴリズム回避の問題意識ばかりを声高に主張することにより、自動化バイアスのような AI への過度な依存が生む諸問題や悪影響に対する社会意識を極小化させて AI の普及を促進しようとする姿勢――本稿では以下「アルゴリズム回避**論**」という――である。確かに AI の高度化は、社会問題を解決して経済発展に寄与するエンジンになるという〈正〉の面があり、その役割は重要である。しかし他方、エンジンが高度化し強力になれば、それに伴ってブレーキとハンドルも同時に高度化し強力に制動力・制御能力を高めなければ、クルマは暴走する[50]。例えば AI の予測・決定等に完全に

---

46 拙書『アメリカ不法行為法』・前掲注（26）426～28 頁.
47 同上 361～67 頁。*See also* Coglianese & Lai, *supra* note 26, at 1293-94 (利用可能性ヒューリスティックを説明).
48 *E.g.*, Jeffrey Dastin「焦点：アマゾンが AI 採用打ち切り、『女性差別』の欠陥露呈で」REUTERS, 2018 年 10 月 14 日, https://jp.reuters.com/article/idUSKCN1ML0DM/ (Feb. 13. 2024).
49 Jeffrey J. Rachlinski & Andrew J. Wistrich, *Judging Autonomous Vehicles*, 24 YALE J.L. & TECH. 706 (2022).
50 「エンジン」と「ブレーキ」「ハンドル」の比喩については、総務省「AI ネットワーク化検討会議 第１回 議事概要」５～6 頁, 平成 28 年 2 月 2 日, https://www.soumu.go.jp/main_content/000415482.pdf (last visited Feb. 13, 2024)(理化学研究所・高橋恒一構成員発言)参照．なおアルゴリズム回避が喚起した、ヒトによる意思決定尊重主義に反対する傾向――本稿では「アルゴリズム回避**論**」と呼称している主張――に批判的な論文としては、本稿以外にも、例えば Ethan Lowens, Note, *Accuracy Is Not Enough: The Task Mismatch Explanation of Algorithm Aversion and Its Policy Implications*, 34 HARV. J.L. & TECH. 259, 260 (2020)等参照.

委ねずにヒトに最終決定権を付与するHITLが旧型の〈ドラム・ブレーキ〉とすれば、そのヒトの最終決定が意味のある内容となるように、自動化バイアスに左右されない対策も加えることが、エンジンの強力化に伴って新型の〈ディスク・ブレーキ〉や〈ABS エイ・ビー・エス〉（アンチロック・ブレーキ・システム）を採用するように、必要な改善策なのである。従って、ブレーキやハンドルの高度化（自動化バイアス対策）を軽視しつつ、アルゴリズム回避を理由にエンジンの強化ばかりを偏って主張する姿勢は、バランスを欠くから厳に慎まねばなるまい。しかるに今般、アルゴリズム回避の主張と伴に以下のような主張・アルゴリズム回避**論**が再現されがちである事態を筆者は懸念するので、そのようなアルゴリズム回避論が誤っていることをここで説明しておこう。

１．AIの予測がヒトよりも正確なのに、ヒトを優先させる社会の偏見は是正すべきという主張：

この主張は、アルゴリズム回避論の核心なので、次章でしっかり反論する。

２．ヒトにも偏見があるから、AIの偏見に対して社会は厳し過ぎるという主張[51]：

この主張は、言い換えれば「赤信号、みんなで渡ればこわくない」と主張していることと同義である。すなわちヒトも偏見的な決定等が見逃されてきた／いるのだから、AIの偏見も見逃されるべきという主張であるから、みんなが違法行為を犯しているから自分も許されるべきという考え方に通じる。しかしそのような考え方は、規範意識を重んじる法学者の筆者には受け入れ難い。寧ろ、ヒトの決定が偏見にまみれていた／いるのならばそれも、AIの偏見と同様に、対策を措置して両者共に治癒すべきである、と考えるのが在るべき法学者の姿勢であろう。実際にヒトの決定に於ける偏見に対して対策を措置している近年の例としては、オーケストラで女性の採用率が極めて低い差別的効果（disparate impact）の現状を改善する策として採用された、〈ブラインド・オーディション〉を挙げることが出来よう[52]。その結果実際に女性採用率が上がった事実が示しているように、ヒトの偏見に対してもこれを見過ごさずに措置する方向性・価値観を社会は近年採用してきているのだから、AIだけに対して対策を措置せずに偏見を見逃して良い、という姿勢を採るべきではない。

３．ヒトによる決定等の理由も不透明で説明責任を果たしていないのだから、AIの予測・

---

[51] *See* Woods, *supra* note 3, at 70 (アマゾン社のAIによる採用が女性差別的であったという悪名高い記事を例に挙げて、AIの方がヒトによる採用決定選別よりも劣っていた証拠がないと主張して、企業はPR的視点から「偏見と大きな欠陥にまみれたヒトによる採用手続の継続」 —"to continue rely on biased and deeply flawed human hiring processes"— を奨励されているし、機械の方がヒトよりも向上することが約束されているにも拘わらず、大衆も機械の誤謬に対する寛容さが殆どないようである、と主張して非難); Coglianese & Lai, *supra* note 26, at 1312 (". . . with respect to . . . lack of bias, humans do not necessarily compare with favorably to machine learning"と指摘).

[52] 小宮山純平「(資料5) AIによる意思決定の公平性」38頁 *in* 総務省「AIネットワーク社会推進会議 AIガバナンス検討会（第1回）」平成30年11月29日, https://www.soumu.go.jp/main_content/000587312.pdf (last visited Feb. 13, 2024)(ブラインド・オーディションがオーケストラの女性比率を上昇させたと指摘)参照.

決定等にばかり透明性や説明責任を要求すべきではないという主張[53]：

この主張は、前項２．に対する回答の、「赤信号、みんなで渡ればこわくない」と同様に、ヒトもAIも双方共に不透明で説明責任を果たしていない問題を放置しろと主張していることになるので、法学者である筆者には受け入れ難いし、社会もこれを受容すべきではない。寧ろ、「ヒトによる決定等の理由も AI の決定等の理由も、双方共に透明性をもって説明責任を果たすべき」という考え方の方が、現代の価値観に適合するであろう。実際に社会の価値観も、嘗ての「ヒトによる決定等の理由も不透明で説明責任を果たしていない」状態を許容しない方向に変化している。例えば修士号等の学位の資格審査に於いては、今では単に「合格・不合格」という結論だけを恣意的に決定する姿勢は許容されない。ディプロマ・ポリシーを満たすような幾つかの複数の考慮要素を事前に決めておいて、それ等の各要素をどの程度満たしたのかを点数化して評価した上で、しかも主査単独の決定だけではなく複数の副査の評価点も考慮して合議で決めた上に、最終的には研究科委員会全体の承認を得なければならないように変化してきている。そのように、必要な場合には理由や根拠を示すことが可能な判断が、ヒトの決定等にも求められるように社会の価値観が変化してきているのである。最近のフィギュアスケートの評価点も、似た価値観に基づいているといえるのではないか。従って、AI の決定等も、近年のヒトの決定に求められるような透明性や説明責任を果たせる価値観に従うべきである、という姿勢こそが健全である。

４．SF 等のフィクションは反ロボット・反 AI に偏っているから、〈SF を例示して反対するな！〉〈SF やファンタジーを語るな！〉という主張[54]：

この主張については、既に筆者が他の機会に反論しているので[55]、そちらを読んでいただければ幸いである。一つ指摘しておきたい点は、日本に於ける AI 規範の代表である内閣府

---

[53] *See* Woods, *supra* note 3, at 81 (ヒトも説明なしに決定を下していると批判); Coglianese & Lai, *supra* note 26, at 1312-13 (ヒトも透明性からほど遠いと分析). なお、AI が不透明で説明責任を果たさないから社会的に受容されない根拠の一つとして、理由を示された上で職務質問されれば社会の危険を除去・防止する為に已やむを得ないと感じるところ、理由なしに職質された場合には反感が高まると指摘する文献として、Emily Berman, *A Government of Laws Not of Machines*, 98 B.U. L. REV. 1277, 1325 (2018)参照. なお本文が指摘するような、ヒトも説明責任や透明性が求められるように社会の価値観が変化している点に関連して、例えば人事採用選別に於いては、有名な〈三菱樹脂事件〉（最大判昭 48・12・12 労判 189 号 16 頁）も雇用主側の採用選別に於ける恣意的な広い裁量権を認めている訳ではなく「法律その他による特別な制限がない限り、原則として自由」（強調付加）であると指摘している点を取り上げて、あらかじめ客観的な基準を定めて公正な採用を行う義務 ――「公正採用義務」―― があると分析・指摘する論文として、三井正信「選抜における公正の法的論点」『日本労働研究雑誌』756 号 4, 11～13 頁（2023 年 7 月）参照.

[54] *See* Woods, *supra* note 3, at 60-62 (SF に於けるロボット脅威論を紹介); Rachlinski & Wistrich, *supra* note 49, at 710-11 (AI を利用したロボットへの大衆の否定的イメージについて、「2001 年宇宙の旅」（MGM, 1968 年）や「ターミネーター」シリーズ等のフィクションを例示して指摘).

[55] 拙考「ロボット法と学際法学」・前掲注（2）； 拙考「汎用 AI のソフトローと〈法と文学〉: SF が警告する〈強い AI／AGI〉用規範を巡る記録から」『法學新報』127 巻 5・6 号 561 頁(2021 年 3 月)； 拙書『ロボット法（増補版）』・前掲注（2）5～6 頁. 更に本稿の後掲注（108）でも、説明責任を果たさず一方的に不利益処分を AI が決定することの非人間性について読者の理解を促す為に、フランツ・カフカの『審判』と、映画「ブレードランナー2049」に触れている。*See also* Citron & Pasquale, *supra* note 15, at 12 (評価の理由が不透明な信用評価システムのことを、"Kafkaesque world of credit scoring"と表現することにより、その不条理さを表現).

「人間中心のAI社会原則」の第二原則である「(2)教育・リテラシーの原則」は、人文＋社会科学の必要性を次のように明記しているので、関係者には遵守をお願いしておきたい。

> …、AIの開発者側は、…規範意識を含む社会科学や倫理等、人文科学に関する素養を習得していることが重要になる。

なお折角なのでこの機会に、SFが描く危険性が実現してしまった最新の実例に言及しておきたい。それは、例えば大統領等の実在の人物そっくりの声音で全く別の内容を発話させる〈音声生成AI〉技術が、主権者達を操作して民主主義を危うくさせたり、特殊詐欺で騙される人々を増やす等の、危険性が懸念され始めた実例である。この〈音声生成AI〉技術は、実は有名なSF映画の「ターミネーター」に於いて既に架空の技術として表現されていた[56]。つまりSF等のフィクションを用いて未来の危険性を予測して警笛を鳴らすという、人文+社会科学の学際研究である〈法と文学〉と〈予防法学〉[57]が用いて来た手法は、根拠に基かない手法ではないと示すことが出来る実例がまた一つ追加されたのである。

第三章　ヒトの決定の代わりにAIを利用すべきではない場合

アルゴリズム回避は、AIの予測がヒトよりも正しい場合に、AIよりも劣るヒトの決定等に固執する偏見を捨てて、AIを更に利活用すべきという主張であった。しかし、このような主張は、そもそも「AIの予測がヒトよりも正しい場合」という前提/停止条件が満たされない場合には――すなわち「AIの予測がヒトよりも正しい」旨の立証責任が果たされない場合には――、ヒトが意思決定等を担っている現状（*status quo*）の変更を肯定すべき証拠に欠けるので、AIをもっと利活用すべきという結論が当てはまらない。更にそもそも正確さ云々(うんぬん)を論じる前に、AIに決定させることが妥当ではない場合があることも、複数の学説が指摘している。例えば**正確性を高めること（maximizing the accuracy＝minimization of forecasting error）は、しばしば公正さ（fairness）が減退するというトレードオフな関係――これを「公正さの対価」（a "price of fairness"）という――にあるとい**

[56] The Terminator (Orion Pictures, 1984)(サラ・コナーが架電する相手の母親の声をターミネーターが真似して騙している); Terminator 2: Judgment Day (Tri-Star Pictures, 1991)(ジョン・コナー少年の声音で架電して、里親に化けた受信者のリキッド・メタルのT1000型を騙している).

[57]〈法と文学〉については、例えば拙監訳のリチャード・A. ポズナー著『法と文学（上）（下）』(2011年, 木鐸社) [*hereinafter referred to as* ポズナー著／拙監訳『法と文学』]、及び拙書『ロボット法（増補版）』・前掲注（2）40～50頁参照. 本文が指摘するように、将来の技術発展が社会に及ぼす危険性の警告としてSF作品が有用であると捉える〈法と文学〉研究の実証となる論文としては、see, *e.g.*, Mitchell Travis, *Making Space: Law and Science Fiction*, 23 LAW & LITERATURE 241 (2011)(近年の遺伝子操作技術に関連して、SF小説の古典である19世紀のH.G.ウエルズ著『モロー博士の島』(1896年)が、既に、ヒトと獣のハイブリットな生物を創ってしまう科学者の危険性を予測・伝達していたという具体例を例示). 〈予防法学〉に於ける将来の新興技術等の危険性の予測については、拙書『ロボット法（増補版）』・前掲注（2）6頁参照.

う指摘に鑑みれば[58]、公正さが重要な業務に於いては[59]AI の方が正確だからといって短絡的にヒトの決定を AI に委ねれば良いという結論になるべきではない。以下では、主に人事採用選別や被用者の評価等に於ける AI の利用の文脈において、AI の正確さという前提/停止条件が満たされない場合と、そもそも AI の利用が妥当ではない場合について迄も社会が誤って AI を濫用しないように、注意すべき考慮要素を例示しておきたい。

Ａ.「AI の予測がヒトよりも正しい場合」という前提/停止条件を満たさない場合：立証責任を果たさない場合

日本政府が公言しているように、政策の決定は「evidence based＝証拠に基づく」ことが重要である[60]。更に、国際的に共有される価値観として、AI は「人間中心＝human centered」で「信頼に値する＝trustworthy」ことも重要である[61]。従って、「AI の予測がヒトよりも正しい」という前提/停止条件も、根拠に基かない単なる allegation だけでは、「証拠に基」

---

58 Coglianese & Lai, *supra* note 26, at 1325; Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 818-19, 832 (2020)("The reason is that any fairness constraint imposed by a modeler will prevent the algorithm from simply maximizing accuracy based on all the features that would otherwise be available to the algorithm"とか、"Measured by overall accuracy of predictions, an algorithm instructed only to identify good employees will beat an algorithm instructed to identify good employees subject to a fairness constraint, because of the necessary trade-off with accuracy.(footnote omitted) Arguably, a less accurate predictor . . . may not serve the employer's need to identify good employees as well as the unconstrained predictor"であると指摘・分析).

59 例えばキュウリの大きさや形を出荷前に仕分けする業務には公正さはほぼ無関係であって正確性と効率性・低コストが重要であるけれども、人事採用選別や入試等に於いては公正さの重要性が当然増すであろう。三井・前掲注（53）(人事採用選別等に於いては公正採用義務等が課されると分析・指摘); Karen Hao,「『くたばれ、アルゴリズム』成績予測システムのひいきに英国学生が怒りの猛抗議」, MIT TECH. REV., Aug. 31, 2020, https://www.technologyreview.jp/s/217376/the-uk-exam-debacle-reminds-us-that-algorithms-cant-fix-broken-systems/ (last visited Feb., 29, 2024); Louise Amoore, *Why 'Ditch the Algorithm' Is the Future of Political Protest*, GUARDIAN (Aug. 19, 2020), https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics (last visited Feb., 29, 2024)(成績予測システムの不公正さに対して"Fuck the algorithm"等のプラカードを掲げた若者達がデモを行ったと紹介); Citron & Pasquale, *supra* note 15, at 18, 19 (労働や大学入試や借金等々の人生の機会に対して予測アルゴリズムが重要な影響を与えることに鑑みると、手続的な規制が不可欠であるし、公正さの要求に服すべきであると指摘). *See also* Doaa Abu-Elyounes, *Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness*, 2020 U. ILL. J.L. TECH. & POL'Y 1, 54 (〈公正さ〉対〈正確性＋効率性〉という文脈に於いて、次のように指摘しているので非常に参考になる: "[C]ertain notions that might be dismissed quickly by computer scientists due to some tradeoffs in the accuracy could be suitable for some legal and policy domains where we acknowledge that efficiency and accuracy alone are not the end of the game."(emphasis added)).

60 「内閣府における EBPM [**E**vidence-**B**ased-**P**olicy **M**aking] への取組」最終更新日令和５年６月, https://www.cao.go.jp/others/kichou/ebpm/ebpm.html (last visited Feb. 10, 2024)(強調付加).

61 OECD, OECD AI Principles Overview, https://oecd.ai/en/ai-principles (last visited Feb. 13, 2024)(筆者も OECD「AI 専門家会合−AIGO−」の日本共同代表として起案に参画した『OECD AI 原則』ウエブ表紙ページの見出し下の一行目に、"The OECD AI Principles promote use of AI that is innovative and trustworthy and that respects human rights and democratic values." (emphasis added)と記述されている); OECD, OECD AI Principles, Human-centered Values and Fairness (Principle 1.2)(emphasis added), https://oecd.ai/en/dashboards/ai-principles/P6 (last visited Feb. 13, 2024)(第２原則の表題のウエブページの見出しに於いて「Human-centered Values」と明記している).

きつつ、「人間中心」すなわちヒトが納得するように「信頼に値」するという諸要素を一つも満たしていないから、全く不十分である。つまり「AIの予測がヒトよりも正しい場合という前提/停止条件を満た」す為には、証拠に基づいて、人間が、信頼に値すると評価できる立証が必要である[62]。特に本稿が焦点を当てている人事採用選別や被用者の評価等に於けるAI利用については、そもそも「人事において将来の予測は難しく、AI活用による予測精度向上は小さい」「個人ごとの予測は誤差が大きい。」と専門家が指摘しているので[63]、そのように正確性に問題を抱える分野に於いてヒトよりも正確であると主張する場合には、抱える問題を覆すだけの立証責任を、AI利用の前に果さねばなるまい[64]。

なお立証責任について厳密に言えば、立証責任を果たす為には単に〈証拠提出責任＝burden of production of evidence〉を果たすだけでは不十分である。〈説得責任＝burden of persuasion〉迄も果たして初めて立証責任を果たしたことになる[65]。従って、ヒトには偏見があると非難しつつ、その人々に対する説得責任――すなわちAIの方が正しいからヒトの決定に換えてAIを採用することに人々に同意させる責任――を果たさないままにAIの利用を推進することは許されまい。しかし現状に於いて社会の人々に対する説得責任が果たされているかについては、検証が必要であろう。

次に、如何なる行動をすれば、「証拠に基づいて、人間が、信頼に値すると評価できる立証」責任を果たしたことになるのであろうか[66]。その立証責任は、特に本稿が対象としてい

---

62 *See* Green, *supra* note, 3, at 15 (ヒトによる監視が機能しないことが実証されているので、それでもアルゴリズムを使用したいのならば、機能する旨の立証責任を果たすべきと指摘).

63 大湾秀雄「(資料１）人事データ活用への関心とガイドライン作成に向けての議論」3頁 *in* 総務省「AIネットワーク社会推進会議・AIガバナンス検討会（第2回)」平成30年12月10日, https://www.soumu.go.jp/main_content/000589116.pdf (last visited Feb. 13, 2024).

64 同上9頁(「バイアス、因果関係、解釈や各種リスクに留意した使い方が出来ない限りは、個人に紐づけした利用は避けるべき。」と指摘しているので、AIを利用する前にはこれ等の条件を満たしている旨の立証責任を果たさねばならないであろう). *See also* Robert Sprague, *Welcome to the Machine: Privacy and Workplace Implications of Predictive Analytics*, 21 RICH. J.L. & TECH. 13, 46 (2015)(事実に基づく伝統的な手法であれば、異議を申し立てて正確性を検証させられるけれども、予測分析を用いる場合には、相関関係を示すだけでは足りず、因果関係の立証責任を雇用者が負う[べき]と指摘). 更に、AIの予測・決定・推奨等の影響によって不利な取り扱いを受けた就活生や被用者には、そもそも透明性に欠けて説明責任も果たされていないAIの予測等の正しさについて知る由もないので、〈情報の非対称性〉からも、AIを利用する雇用主やAI役務を提供したベンダ側に立証責任があるといえるのかもしれない。*See* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1727-28 (2020)(雇用差別禁止法上のdisparate impact testの文脈に関して [中立的な採用基準が差別的効果を有していた事実が原告によって立証された場合に]、被告の雇用主側にbusiness necessityの反証義務、すなわち説明責任が課される根拠として、情報の非対称性を挙げている).

65 拙書『アメリカ不法行為法』・前掲注（26）86頁参照.

66 Kemperは、自動意思決定――ADM――への信頼を高める為には、アルゴリズムの決定が**透明かつ公正**でなければならない。そして、特に偏見と差別に関するADMの品質が**「検証されかつ検証され得なければならない」**――"must be ascertained and ascertainable"――、と指摘しているので、参考になろう。Kemper, *supra* note 3, at 293. *See also* Janine S. Heller, *Fairness in the Eyes of the Beholder: AI; Fairness; and Alternative Credit Scoring*, 123 W. VA. L. REV. 907, 917 (2021)(「データに基づく決定」―"data driven decisions"―は人々の基本的な「生活の質」―"qualities of life"―に影響を与えるから、適正手続の保障を広く適用すべきと研究者達が主張していると指摘しつつ、Citron & Pasquale, *supra* note 15, at 20を出典表記しながら、評価システムが深淵なる影響を人々に与えることに鑑みて、適正手続の根拠となる透明性、正確性、説明責任、当

る「ハイリスク」又は「機微な領域」に属する人事採用選別や被用者の評価等のような AI 利用の場合と、それ以外の、例えばキュウリの出荷前の大きさや形の品質選別作業のような単純、反復的、かつ数値化して評価・判断等が容易な場合とは当然異なる。すなわち、人事採用選別や被用者の評価等は「ハイリスク」又は「機微な領域」であり、人生の大きな部分を左右する決定に繋がるから、合理的に考えれば当然、キュウリの場合よりも相当程度厳しい立証責任が課されるべきである。特に人事採用選別等々に於ける AI 利用に関しては、**因果関係が明らかではないにも拘わらずデータ・マイニングと相関関係のみに基づいて安易に予測を行う慣行が厳しく批判され[67]、かつその内容が不透明で検証も不可能である事実も考慮すると[68]**、AI の方が正しいという前提/停止条件は満たされていないと捉えた上で、

事者参加の原則、及び公正さという諸価値を原動力としてこれを監視すべきと主張).

[67] 例えば、大湾・前掲注（63）4, 9 頁（「因果関係を特定できない。相関関係のみに基づき、ブラックボックスのまま、予測に使うのは極めて危険。」「予測は、因果関係を検証し、構造変化の有無を確認しながら行うもの。」「個人に紐づけした予測を目的とする AI の活用(プロファイリング)は、法原理、倫理、経済効率のいずれの面から見ても問題点が多い。」と指摘）. *See, e.g.*, Ajunwa, *Automated Hiring System, supra* note 3, at 637 (動画の顔の表情等を採用決定の考慮に入れることは、因果関係が科学的に立証されていないので疑問であると指摘); Lori Andrews & Hannah Bucher, *Automatic Discrimination: AI Hiring Practices and Gender Inequity*, 44 CARDOZO L. REV. 145, 154, 155 (2022)(プログラムはトップ被用者の仕事能力とは無関係な特性に焦点を当ててしまい、例えば日本の漫画サイトをビジットした者のポテンシャルが高い等とベンダが依頼企業にアドバイスしたり、上手く仕事する者は「Jared ジャレッド」という名で高校時代にラクロス部だった者であると指摘する悪名高い相関関係の例を紹介); Janneke Gerards & Fredrik Zuiderveen Borgesius, *Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence*, 20 COLO. TECH. L.J. 1, 8, 14-15 (2022)(例えば特定のアルファベットと番号の組み合わせの住所―「4A」「20C」等―の住人は事故率が高いという相関関係が示されると、損害保険会社が因果関係の有無と無関係にその住人達の保険料を高くするから、対象住人達は自身の帰責とは無関係な相関関係によって「罰せられる―punished―」から「不当である―unfair―」と感じるし、例えば差別的な結果が含まれるデータを学ばせるとアルゴリズムが差別的な結果を示してしまう、等の相関関係に基づく短絡的の諸問題を指摘); Sonderling et al., *supra* note 8, at 24 & n.125 (アルゴリズムは候補者の特性と将来の仕事上の成功との間の相関関係を示すけれども、それでは因果関係が欠けているし、前者は偶然の産物であると批判); Sandra Wachter, *The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law*, 97 TUL. L. REV. 149, 220 (2022)("Unfortunately, data science and AI are often not concerned with causation or establishing a causal or empirical link between characteristics (e.g., algorithmic groups) and outputs; mere correlation is sufficient to drive decision making. In other words, decision makers often have little incentive to empirically prove a causal relationship between decision criteria (*e.g.*, friends on Facebook) and predicted variables (e.g., ability to repay a loan)."(emphasis added)と指摘); Pauline T. Kim, *Data-Driving Discrimination at Work*, 58 WM. & MARY L. REV. 857, 879-80, 881 & n.95 (2017)(そもそもデータ・マイニングは因果関係を示す訳ではなく相関関係しか示さない等と批判); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1053, 1548 [*hereinafter referred to as* Zarsky, *Transparent Predictions*.] (単なる相関関係で決め付けられずに因果関係の存在が尊厳から要求されると指摘). *See also* Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 401, 409 (2015)(犯罪捜査に於ける予測分析―predictive analytics―は、特定の関連性に基づいて容疑を示すけれども、その容疑はしばしば根拠に欠けると分析しつつ、次のように指摘: "Correlations do not prove causation.").

[68] *See* Tal Z. Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCIENCE, TECHNOLOGY, & HUMAN VALUES 118, 129 (2016) [*hereinafter referred to as* Zarsky, *The Trouble with Algorithmic Decisions*] (AI による自動的決定は、ヒトによる説明がないばかりか、不透明であるから恣意的―arbitrary―であると受け止められるので、判断に影響を与える諸要素や予測の誤謬率等を説明すれば、自動決定に対する懸念を和ら

次に挙げるような厳しい立証責任が課されるべきであろう。具体的には、定期的内部監査[69]、独立して中立的な第三者による監査[70]、監査結果の公表[71]、AI による決定等の理論[72]や決定の根拠になったデータについての平易な事前及び/又は事後的説明[73]、及び独立して中立

げることが出来ると示唆). なお、AI は不透明・ブラックボックスであると嘗てから指摘されている問題に加えて、営業秘密を理由に AI の下した予測・推奨・決定等の理由解明や説明責任の要請にベンダが応じない為に、検証も出来ないので不利益を被った者の尊厳を侵す問題を解決する必要がある。Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1308 (2020)(影響を受けた者達が決定を理解しかつその正しさの検証を要求する権利よりも、ベンダからの営業秘密への要望を優越させることは不適切である、という点に於いて研究者達は一致していると指摘). *See also* Houston Federation of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017)(アルゴリズムによる低評価ゆえに解雇された教員が、そのアルゴリズムの低評価の説明を求めたところ、ベンダが営業秘密を理由に開示しなかったこと等が争点になった事件); Citron & Pasquale, *supra* note 15, at 5, 6, 7, 21, 27 (アルゴリズムが営業秘密によって強力に保護されているから、その予測がしばしば恣意的かつ差別的に人生の機会を奪っていても、評価システムの過程や結果に誰も異議を唱えられないし、意味のある検査も出来ないけれども、秘密が前提・初期値であるという立場を止めて人々が如何にランキングされたのかを知ることが出来るような転換が必要であり、保秘よりも人間の尊厳への脅威の懸念の方が上回るし、予測的評価が個人から重要な機会を奪った場合には少なくとも異議を唱えることが出来る意味のある内容の告知と機会を賦与されるべきと指摘). なお、営業秘密と、被評価者への公正な取扱いとのバランスをはかった手続として、独立した第三者である「信頼された中立的専門家」――"trusted neutral experts"――を用いたり、「保護命令」――"protected order"――を用いる提案としては、Citron & Pasquale, 81 WASH. L. REV., at 28.

[69] Sonderling et al., *supra* note 8, at 80 (継続的な監査の必要性を指摘); Citron & Pasquale, *supra* note 15, at 21 (評価システムの機微性に鑑みれば、労働等に使用される場合には、監査要件に服すべきと指摘).

[70] Ajunwa, *Automated Hiring System, supra* note 3, at 660, 796 n.7 (独立した監査人による監査の必要性を指摘。加えて、自動的雇用決定ツールを使用する場合には、独立した監査人による偏見監査―bias audit―の実施とその結果の開示等を命じている NY 市条例も紹介). なお NY 市条例の文献例については、前掲注（20）. *See also* Citron & Pasquale, *supra* note 15, at 28 (保秘を維持しながら公正さを実現する手続として、independent third parties に評価させる案を提示).

[71] Sonderling et al., *supra* note 8, at 47-48 (独立した監査人による偏見監査やその公開を要求する NY 市条例を紹介). 更に雇用主とベンダによる、もっと徹底的な透明性の実施と独立した監査人による監査が重要であると指摘する文献としては例えば、see, *e.g.*, Aaron Rieke & Miranda Bogen, *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*, UPTURN (Dec. 10, 2018), https://www.upturn.org/work/help-wanted/ [https://perma.cc/UR7U-NPWR] (last visited Feb. 13, 2024).

[72] *See* Citron & Pasquale, *supra* note 15, at 7 (評価システムがどのように予測したのかを理解しないままにその予測を受け入れるべきではないと指摘). なお、尊厳が保たれる為には、ヒトが AI に決定される前に、単なる相関関係の存在だけではなく因果関係の存在が要求されると指摘されている。Zarsky, *Transparent Predictions*, *supra* note 67, at 1548. *See also* Kim, *supra* note 67, at 865-66 ("if employers rely on these models [using data mining techniques], they may deny employees opportunities based on unexplained correlations and make decisions that turn on factors with no clear causal connection to effective job performance" (emphasis added)と指摘して、仕事上の成功との因果関係[の説明]を欠いた相関関係に基づく人事活動を批判).

[73] *See* Citron & Pasquale, *supra* note 15, at 23 (自身に関する全てのデータにアクセスす[ることが許される]べきであるし、理想的には予測的評価システムの論理も検証される為に公開されるべきと指摘). なお説明責任が必要である倫理的根拠については、see, *e.g.,* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines,* 87 FORDHAM L. REV. 1085, 1118-19 (2018) (説明なしの決定に対する恐れは、フランツ・カフカの『審判』が最もよく例示できると指摘しつつ、説明することそれ自体が善である理由として、「自律性、尊厳、及び人間性の尊重」――"a respect for autonomy, dignity, and personhood" ―― が含まれると指摘し、これ等は法制度に不可欠な手続的正義に最も反映されていて、その手続的正義は対象者の参加を尊重し、そこには透明性等のみならず「礼

的な第三者である研究者達による裏付け検証可能性の確保と[74]、裏付け検証によってAIの方がヒトよりも正しいことが検証されること[75]、等々を満たすことが望ましい。何故ならば、人事採用選別や被用者の評価等に於ける上記諸要素の欠如が既に問題であると指摘されているからである[76]。

Ｂ．AIの特性的・限界的にAI利用がそもそも妥当ではない場合

AIの予測の方が正確云々という前に、そもそもAIの決定に委ねるべきではない場合も存在することを理解しておくことは、AIの濫用を避ける為に非常に重要である。そこで以下に於いて、AIの特性的・限界的にAI利用が妥当ではない例を掲げておこう。

１．〈ルール〉対〈スタンダード〉―― 衡平法的な考慮要素の衡量や裁量が必要な場合：

アメリカ法学では、規範を〈ルール〉と〈スタンダード〉に区別する考え方がある[77]。一方の〈ルール〉は、例えば「55ｍｐｈ（時速88ｋｍ）を超えてはならない」という規範のように、事前に客観的に違法な事実を定めて、担当者が事後的に裁量で変更することを原則として許さない[78]。従ってルールは、同じ規範を様々な事件に統一的に当てはめて法的安定

儀正しさ・敬意」―“politeness”―の要求さえも包含される、と分析)； Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessment*, 35 HARV. J.L. & TECH. 117, 133-34 (2021)[*hereinafter referred to as* Selbst, *Algorithmic Impact Assessment*](尊厳の為には説明が必要と指摘).

[74] Kim, *supra* note 67, at 879-80 (そもそもデータ・マイニングは「理論に基づかない」―“atheoretical” ―ばかりか、その正しさが不透明性ゆえに第三者によって検証が出来ない問題を指摘)； Citron & Pasquale, *supra* note 15, at 8, 21, 26 (評価システムとそれが生み出す恣意的で不正確な結果は研究者達や専門家達による検討に服さねばならないし、計算も公表されて全ての過程が検証可能になることが理想的であると指摘).

[75] 検証(verification)を阻害している不透明性への批判としては、see, *e.g.,* Calli Schroeder et al., *We Can Work It out: The False Conflict between Data Protection and Innovation*, 20 COLO. TECH. L.J. 251, 271-72 (2022)(採用面接の動画や音声を分析するアルゴリズムによって候補者の特性が判明するとベンダが主張するけれども、その理由・根拠を開示せず、そのようなAIを採用する企業も理由・根拠を知らないおそれがあり、その不透明性ゆえに検証が困難であると批判).

[76] 拙講演資料・前掲注（3）参照. *See e.g.*, Sonderling et al., *supra* note 8 (企業が透明性と説明責任を果たすべきであり、AI使用を監視・監査すべきであると指摘しつつ 、定期的監査の必要性を含む、正しさを確保する為に必要な施策にも言及)； Citron & Pasquale, *supra* note 15, at 22 (企業は効率性のみを追求するから、公正さを検証する為に企業の自主規制に依存出来ないと指摘). なお本文中では正確さ等を実現させる為の *ex post* 的に必要な手続を主に例示列挙したけれども、本来はそれだけでは足りず、*ex ante* 的なアルゴリズムの設計段階やビッグデータの収集等々から始まる〈ライフサイクル〉の全てにわたって様々な手続を履行することが重要である。*See*, *e.g.*, OECD, *Advancing Accountability in AI: Governing and Managing Risks throughout the Lifecycle for Trustworthy AI*, OECD Digital Economy Papers, No. 349, Feb. 2003, https://www.oecd-ilibrary.org/docserver/2448f04b-en.pdf?expires=1708139441&id=id&accname=ocid56001934&checksum=C20B691C3E4CE4BABFAA88605E9BCF08 (last visited Feb. 17, 2024) (信頼に足るAIである為には「AIシステムのライフサイクルの全てにわたって危険性を統治しかつ管理せねばならない」と指摘)； Lehr & Ohm, *supra* note 21, at 657-58 (デプロイ前のデータを扱う“playing with the data”段階からの対応も重要であり、[human in the] playing-with-the-data loop が必要であると主張).

[77] *See, e.g.*, Cass R. Sunstein, *Problems with Rules*, 83 CAL. L. REV. 953, 956-57 (1995).

[78] Citron, *supra* note 3, at 1301.

性を維持する役割に資する[79]。他方、〈スタンダード〉は、例えば「状況下で理にかなった速度で運転しなければならない」という規範のように[80]、事後的に当該状況に於ける様々な諸要素を考慮した上で、裁量権に基づいてヒトが決定する[81]。従ってスタンダードは、事件毎に異なる諸要素や状況の変化に応じられるように、具体的に妥当な結論を導き出すことに資する[82]。以上のように法は、〈ルール〉と〈スタンダード〉が適切に用いられるように工夫されるのである。

ところで英米法の流れをくむアメリカでは、イギリスに由来する〈コモン・ロー〉と〈衡平法〉という法理の歴史的区別が存在し[83]、いささかダイコトミー的に分類すれば前者のコモン・ローはルールに相当し、後者の衡平法はスタンダードに相当する[84]。柔軟性に欠けるコモン・ローの原則だけでは妥当な救済を与えられなかったことから、より複雑化した事案に適合して公正な救済を付与できるように衡平法が発展していったのである[85]。

そして AI を用いる決定は、単純で明確な規範の機械的当てはめが可能なルールやコモン・ロー的決定に適しているけれども、単純ではなく曖昧な規範を事案毎に異なる諸般の考慮要素に鑑みた上で事案の文脈に合うように当てはめて適切な決定を導き出さねばならないスタンダードや衡平法的な決定には不向きである[86]。

ところで人事採用選別という業務に、ルールとスタンダードの違いを当てはめてみると、例えば定職に半年以上就いていない履歴書を自動的に AI によって足切りすることは[87]、確かに半年以上無職な者を振り落とすという単純・明快な〈ルール〉を統一的に当てはめる機

---

79 *Id.*; Coglianese & Lai, *supra* note 26, at 1309 & n.148.

80 スタンダード的規範の例としては、本文中の例のように法律用語で多用される「reasonable」以外にも、例えば『連邦証拠規則』403 条—FEDERAL RULES OF EVIDENCE 403—が規定する、展示証拠等の提出認容基準である、証明上の価値が不当な偏見のおそれを凌駕しなければならないという場合の衡量を挙げる例も見受けられる。Eugen Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1145 (2019). なお SF を用いて AI やロボットの危険性を論じることの是非を、証拠法のスタンダードを用いて検討すべきと主張する筆者の論文については、拙考「ロボット法と学際法学」・前掲注（2）112 頁参照.

81 Citron, *supra* note 3, at 1301.

82 *Id.* at 1302.

83 *Id.* at 1301.

84 コモン・ローと衡平法の違いについては、ポズナー著／拙監訳『法と文学』・前掲注（57）206～07 頁も参照.

85 *See* Sean Kanuck, Essay, *Humor, Ethics, and Dignity: Being Human in the Age of Artificial Intelligence*, 33 ETHICS & INT'L AFFAIRS 3, 6 (2019). 拙書『体系アメリカ契約法：英文契約の理論と法務』154～56 頁, 194～209 頁（2009 年, 中央大学出版部）も参照.

86 *See* Citron, *supra* note 3, at 1303 (状況に応じた臨機応変な対応が不要な場合にはルール的規範が適切であるけれど、ヒトの裁量が必要な場合は不適切と指摘). *See also* Green, *supra* note 3, at 13 (事前にルールが明確化されている場合にアルゴリズムを使用することは適切であるけれども、他の諸要素にも鑑みた衡量的な予測も係る決定には不適切であると指摘).

87 履歴書の文脈を読めない AI の欠点例については、Schroeder et al., *supra* note 75, at 267-70 ("Applicant Tracking Software [ATS]"と呼ばれる履歴書振り分けシステムの使用と欠点について説明). なお履歴書を振り分ける業務に AI を使用している企業は、何とフォーブス 500 社の内の 98%にも及んでいる。*Id.* at 268. ところで実際に半年以上の離職期間がある候補者を、企業の半数は自動的に足切りしていると云われている。Sonderling et al., *supra* note 8, at 6 n.19. また、本文中の仮想事例のような振り分けの結果、女性の候補者が不釣り合いな程に採用されない場合には、アメリカの雇用差別禁止法違反のおそれが指摘されている。*Id.* at 6.

能としては優れているかもしれない。他方、仮に半年以上無職な理由が結婚・妊娠・出産・育児等であって、今では就労できる状況に至って意欲も能力も優れた候補者を、スタンダード的な考慮を欠く AI では、振り落とさずに面接段階に進めてその意欲や能力等々を評価する機会を賦与することが出来ない。AI はヒトよりも公平であるという幻影が AI の利点としてしばしば主張されがちであるけれども、他方、具体的妥当性に欠けて不当な決定が下されるという欠点が見逃されがちなのである。しかし人事採用選別という業務は、そもそも個人の能力・意欲や個性等を尊重しつつ諸般の事情を考慮すべき業務であるべきなので、そのミッションを果たせない AI に決定させること自体が不適切ではあるまいか。

しかし残念なことに、**高利便性と低コスト故に、本来は衡平法的でスタンダード的な衡量が必要な決定業務に於いてさえも、具体的妥当性を欠いた柔軟性のないルール的な AI 利用が広がっている**と指摘されている[88]。

２．文脈やニュアンスを読まねばならない場合：

この場合は、論理的に上述した〈スタンダード〉的決定が必要な場合に AI 利用が妥当ではないという指摘と同類である。すなわち半年以上無職である履歴書を振り分けるような単純で機械的な決定に AI は適しているけれども、結婚・妊娠・出産・育児等という背景事情や高い意欲と能力が現在備わっている状況等を理解した上で足切り対象から外すようなスタンダード的で衡平法的な決定に AI が妥当しないことを言い換えれば、必要な文脈やニュアンスを汲み取る為には AI 利用が不適切でもある。

文脈・ニュアンスを読み取らねばならない場合に AI が不適切である欠点は、ロボット兵器・完全自律型兵器の場合にも指摘されている[89]。例えば[90]、アフガニスタンに於いてドローンが捉えた動画が、羊の群れの中に AK-47 突撃銃を持つ老人の姿を映し出していたと仮定してみよう。銃を持っているからテロリストであると短絡的に AI が決定して攻撃すると、戦闘員ではない民間人に対する誤爆のおそれがある。何故ならこの地域では羊飼いが自衛の為に銃を携帯するという背景事情・文脈・ニュアンスを AI が考慮できなかったからである。他方、道を歩く男性が同じく AK-47 突撃銃を携帯しているばかりか、近くに羊の群れが居おらず、代わりに近郊にはテロリストのアジトが在り、かつその男性が衛星電話を使っている様子が映ればテロリストの蓋然性が高まる。そのような文脈やニュアンスを読み取って総合判断できるのは、ヒトの判断であると指摘されている。

そもそも AI は、提供されたデータだけしか決定の考慮に出来ず、当該状況に於ける全体を考慮することが出来ない[91]。他方、ヒトは、新しく予想外の情報さえも組み込んで決定が

---

88 Citron, *supra* note 3, at 1303. *See also* Huq, *supra* note 3, at 687 ("machine decision"が正確性と中立性を向上させるという前提で使用されてきたものの、政府が利用する機械的決定は「欠陥だらけ」――"highly flawed"――であると指摘). なお、本文中の「AI はヒトよりも公平であるという幻影が AI の利点としてしばしば主張されがちであるけれども」という記述に関連して、see Citron & Pasquale, *supra* note 15, at 4, 33 (評価システムが評価課程に於けるヒトの欠点を排除して、全ての個人を公平に扱うという評価システム推進派の主張は誤導的であり、評価[点]の単純さから正確で信頼に足るという幻影を与えると指摘・分析).

89 Charles P. Trumbull IV, *Autonomous Weapons: How Existing Law Can Regulate Future Weapons*, 34 EMORY INT'L L. REV. 533, 543, 548, 559, 576 (2020).

90 本文中の仮想事例については、see *id.* at 559.

91 Berman, *supra* note 53, at 1351.

出来る[92]。

以上のように、文脈的な理由（contextual reasoning）やヒトの直観に依存しなければならない業務には自動化が馴染まず、他方、利用者側の臨機応変で柔軟な決定は不要で、反復繰り返し的単純業務の場合には自動化が適している[93]。後者の例は、郵便番号を読み取って送付先住所に郵便物を仕分けするような業務であり[94]、そのような反復的で単純な決定は、疲れて誤謬が増えてしまうヒトよりも、疲れを知らず誤りが少ないAIの方が長けているのである。更に例えば、プロ・テニス試合に於けるボールのイン／アウトを機械的に決定する「Hawk Eye」は、ヒトの審判員よりも正しい決定を下す好例といえるであろう[95]。

３．数値化できない業務：

数値に基づくAIの決定は、理論的に当然、数値化できない業務に不適切である。しかし現実社会は現在のところ、数値化できない要素が沢山存在し、かつそのような諸要素が適切な決定の為には重要な場合も在[あ]る。従って、数値化できない業務や目的を明確化できない業務には、AIが適さないと指摘されている[96]。特に人事採用選別、及び被用者の評価等について、「AIは記録に残らないものは扱えない。人間の目で見た評価情報を代替するものではない。」と指摘されているので[97]、AI利用が妥当ではないおそれを十分配慮すべきであろう。

４．先例のない事態を予測する場合：

AIは過去の出来事のビッグデータに基づいて将来の予測等に使われるので、先例のない事態を予測することへの使用には不適切である。例えば、未曽有なパンデミックによる景気後退のような、先例のない事態の予測には、AIは不適切であると指摘されている[98]。

５．十分な資金繰りや資力を持たない場合：

先ず各業務に適合した正しいアウトプットをAIに出させる為のカスタマイズには、労働集約的な準備と、専門性の高い専門家のスキルと、ビッグデータを貯蔵し分析したりその劣化を防げるようなシステム、ハードウエア、ソフトウエア、及びネットワーク等々の資源が不可欠である[99]。従ってその為の財力や資金繰りが出来ないベンダ等がAIの役務を他の企

---

92 *Id.* at 1323.
93 Vagle, *supra* note 38, at 136. *See also* Coglianese & Lai, *supra* note 26, at 1312 ("Machine learning tends to perform best with tasks involving pattern recognition and high levels of repetition. This means that . . . , digital algorithms still hold great promise for reducing much of the drudgery [単調で決まりきった] work in government.(footnote omitted)"と分析).
94 Coglianese & Lai, *supra* note 26, at 1331. 更に、道路に急に飛び出してきた障害物を回避するような操作はヒトの方がアルゴリズムよりも素早く出来るけれども、大量の書類を即座に読むような業務は後者が適していると Coglianese & Lai は指摘している。*Id.* at 1309-10.
95 Lowens, *supra* note, 50, at 273-74 (例えばイン／アウトを判定する機械の Hawk Eye と、ヒトの審判員とを比べて、前者はイン／アウトの決定に特化した狭い業務を担って、後者は機械が不得手とする、それよりも広い業務範囲を担うように協働することを推奨).
96 Coglianese & Lai, *supra* note 26, at 1324.
97 大湾・前掲注（63）3頁.
98 Coglianese & Lai, *supra* note 26, at 1328.
99 *Id.* at 1323.

業に提供しようとしても、「AI の方がヒトよりも正しい場合という前提/停止条件を満たすこと」が容易ではなくなるであろう。

６．正確な予測に必要なビッグデータが利用不可能な場合[100]：

必要なビッグデータが利用可能ではない場合には、AI による正しい決定の提供が不可能である。更に、或(あ)るデータセットと他のデータセットが異なる様式で貯蔵されていて相互運用性が欠けている場合等にも、利用が不可能になる。

加えて、AI の決定等を人事採用選別や被用者の評価等に利用したい場合に、対象となる企業の被用者達のデータがあればビッグデータとして十分であると捉えてはいけない。「大企業の人事データでもビッグデータではない。精度を上げるのに必要なデータ数が確保できない。企業がデータをシェアすれば話は変わるが、それでも企業ごとの異質性を上手く処理できないだろう。」と専門家に指摘されている[101]。

７．社会の価値観の変化を反映できる新しいデータを利用できない場合：

社会の価値観は日々変化しているので、AI もその変化から時代遅れにならないようにする為には、トレーニングをし直せるように、新しいデータの流れへのアクセスを確保する必要がある[102]。それが出来ない場合には、AI 使用が妥当ではなくなる場合もある。

Ｃ．公正さ、尊厳、又は人道等ゆえに、ヒトの決定をそもそも AI に委譲すべきではない場合

この場合は、本章の冒頭部にて触れた、AI の正確性（accuracy）と公正さ（fairness）がトレードオフな関係にあるという指摘に関連している。すなわち、**たとえ AI の決定等が正しくても――すなわち AI の予測が正しい蓋然性がヒトよりも高いとしても――、公正さの観点からはヒトによる決定を AI に短絡的に委譲してはならない場合**もある[103]。

そのような場合の倫理的根拠としては、「**個人がどの分類に属するのか次第で**――すなわち、個人と他人との間に如何なる関係性が示されるのか次第で――**[その個人についての]評価を自動的に下すことは、その個人を個人として取り扱うことを怠ってしまう**」ことになる[104]。その為に、例えば人事採用選別等に於いて、候補者が採用企業に適しているか否かを AI が決定する場合に問題視されている欠点の一つは、個々の候補者が或る分類に属すると予測されてレッテル貼りをされてしまうと、たとえその候補者個人がその**分類の特性とかレッテルとは異なる〈個性〉（individuality）故にその分類特性・レッテルよりも望ましい行動をする可能性があっても十把一絡げ(じっぱひとからげ)に振り分けられて、是非もなく問答無用に足切りされてしまう**点にある。そのように、或る分類に属することだけで自動的に個

100 *Id.* at 1326.
101 大湾・前掲注（63）3 頁.
102 Coglianese & Lai, *supra* note 26, at 1327.
103 *See, e.g.*, Zarsky, *Trouble with Algorithmic Decisions, supra* note 68, at 129 ("Algorithmic decision-making processes raise a crucial additional set of fairness-based concerns. These are autonomy-related concerns that also involve harms to individual dignity" (emphasis added)と指摘).
104 Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1528, 1542 (2019)(拙訳)(強調付加).

人を判断することは、個人を個人として扱うことを怠ることに問題がある[105]。

以上のように属性・関連性の予測だけで個人を決めつけて決定を下すことの不公正さは、航空機に乗機する前に、テロリストの名前に似ていると AI に決定されてしまうと乗機できない不利益を被るばかりか、その理由の説明や誤りを正させる機会の賦与を政府が懈怠していることから、同様な災難が複数回繰り返される不利益までも被らされる例が象徴的である[106]。本当はテロリストではないのに機械的に AI がそうであると決めつけるこのようなシステムは、正に人間を人間として扱う尊厳や人間性に反しているといえよう。

ところで AI に決定を委ねることが、その正しさとは無関係に本来的に許されるべきではないとする倫理的理由としてもう一つ挙げられている根拠は、そもそも人間は自由な個人として尊敬されるべきであるから、AI の決定に服させることによって個人を「物扱いする」こと（objectify）は個人の人格を危うくする[107]、すなわち「人間に関する決定を機械に許すことは、人々の人間性に対する本来的な敬意を欠き、本来的に人間を物として取り扱うことである」、と指摘する[108]。

---

[105] *See id* ("If algorithmic decision-making does not allow individuals to proclaim their individuality ('I may look like these other people, but I am not in fact like them'), then it violates their dignity and objectifies individuals as their traits, rather than treating an individual as a whole person(footnote omitted)."と指摘). *See also* Wachter, *supra* note 67, at 200-01 (統計的証拠に依存し過ぎることは、人々を個人として取り扱うことを怠ると指摘); Kim, *supra* note 67, at 866 (労働者を振るい分けし又は点数化する為にデータ・アルゴリズムのような分類化する仕組み――classification scheme――に依拠する故に生じる差別的効果を「classification bias:分類バイアス」と呼称して批判).

[106] テロリストによる乗機を阻止する為に連邦国土安全保障省が展開している〈No-Fly List〉と呼ばれるマッチング・プログラムは粗末なアルゴリズムな為に、似た名前を区別できず、何千人という無辜(むこ)な市民が空港でスムーズに乗機できず、有名なエドワード・ケネディ上院議員も乗機を複数回阻止されているけれども、政府は誰がリストに載っているのかを開示せず、乗機禁止決定についての説明もしないので批判されている。*See, e.g.*, Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 801 (2021); Jennifer C. Daskal, *Pre-Crime Restraints: The Explosion of Targeted, Noncustodial Prevention*, 99 CORNELL L. REV. 327, 345-46 (2014); Justin Florence, Note, *Making the No Fly List Fly: A Due Process Model for Terrorist Watchlists*, 115 YALE L.J. 2148, 2150 (2006).

[107] Kaminski, *supra* note 104, at 1541.

[108] *Id.* at 1542 (拙訳)(強調付加). *See* Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1993 (2021)(ヒトを「手段」―"means"―として扱ってはならず「目的」―"end"―として扱わねばならないと説いたイマニュエル・カントの哲学に基づくと指摘). *See also* Alysia Blackham, *Setting the Framework for Accountability for Algorithmic Discrimination at Work*, 47 MELBOURNE U. L. REV. 63, 69 (2023)(雇用主が職場で被用者に automated processes と相互作用するように要求することは、「ヒトの尊厳を徐々に損ない」―"undermine human dignity"―、かつ「ヒトである労働者を非人間化する」―"dehumanize human workers"―と指摘). なお、不透明なアルゴリズムによる融資拒絶の文脈に於いて、ヒトの尊厳と自律性ゆえに説明責任があると指摘し、フランツ・カフカの『審判』（1914-15 年）が「データに基づく説明なしの決定の非人間性」――dehumanization of unexplained decision based on data――を最もよく捉えていると指摘する論文として、Selbst, *Algorithmic Impact Assessment*, *supra* note 73, at 133-34; Selbst & Barocas, *supra* note 73, at 1118 参照. 筆者もこの主張に賛成である。拙講演資料・前掲注（3）12 頁. 更に付言すれば、映画「ブレードランナー2049」(コロンビア・ピクチャーズ, 2017 年)に於いて主人公のレプリカントの「Officer "K"」――『審判』の主人公と同じ名前(!)――が受診を強いられる The Post-Traumatic Baseline/Stress Test にも、機械がヒト[レプリカント]の運命を決定することの非人間性が表されていると筆者は指摘してきた。拙講演資料・前掲注（3）12 頁. 更に付言すれば、Baseline Test は一定の言葉やフレーズに対する K の声や表情への影響を計測した結果、彼が baseline を満たさなければ「退職させられる: retired」＝殺処分されるという仕組みであるところ、現実世界でも採用候補者の面接動画の表情や声を AI に分析させて評価す

以上のように倫理的観点から AI による決定への委譲を問題視する見解からは、AI によって不利益な決定を被ったヒトに対して、その理由を知らせて異議を申し立てる機会も付与されるべきであるという主張が見受けられる [109]。例えば AI の決定によって解雇された教員には、その理由や異議を申し立てたい欲求が生じるが、それは人間の尊厳や自律性に由来するという指摘がある [110]。またこの考え方が規範化した実例としては、日本もその理事会勧告化に大きく貢献し推進した〈OECD AI 原則〉の第３原則である〈1.3 透明性及び説明可能性〉を挙げることが出来る [111]。更に日本の専門家も「AI による自動意思決定によって、採用、異動、研修の機会を否定された時に、納得できる理由が明かされなければ、個人の尊厳が否定されたと見るべき」と指摘している [112]。

おわりに

以上の分析は、AI のこれ迄の能力とその限界・欠点に基づいていたので、将来、AI が、その能力の限界を超える進化を遂げて欠点が治癒・克服されれば、結論が変わり得よう [113]。しかしそのような、予測が難しい将来は別にして、そもそも AI の利用には物理的・能力的限界があってその使用が妥当ではない場合があることを理解しておくことが、重要であろう。特に人事採用選別や被用者の評価等に於ける AI 利用は、欧米に於いては既に「ハイリスク」又は「機微な分野」に於ける AI の利用であると位置づけられた上で、慎重な運用が求められているばかりか、AI にヒトの意思決定を委譲することが倫理的に妥当ではない、という研究者達からの強い指摘も見受けられる。従って、AI の正し

る慣行が、前掲注（75）で指摘したように厳しく批判されている。当然の批判であろう。*See also* Schroeder et al., *supra* note 75, at 272 & n.79;　Will Knight, *Job Screening Service Halts Facial Analysis of Applicants*, WIRED (Apr. 12, 2021), https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/ (last visited Mar. 12, 2024)(面接動画上の就職応募者の表情をアルゴリズムが分析して認知能力、心理特性、感情知能、及び適性を決定できると主張していた HireVue 社の役務が、物議を呼び、遂には人権団体が連邦取引委員会―FTC―へ提訴する事態に至り、同社がその役務提供を中止した事例を紹介).

109 Kaminski, *supra* note 104, at 1531 n.2, 1541;　Bloch-Wehba, *supra* note 68, at 1308-09 (アルゴリズムの決定により影響を被った者が如何に評価され、かつ点数化されたかについて理解するのに十分なだけの情報開示によってはじめて個人の尊厳と自律性と適正手続上の権利を全う出来ると分析).　*See also* Citron & Pasquale, *supra* note 15, at 24, 31 (「透明性、意味のある監視、及び悪影響を与える決定によって相関関係ゆえに不当に分類された個人を救済する手続」がアルゴリズムによる決定には必要である、と連邦雇用機会均等委員会の Ramirez 委員長が力強く主張していると紹介しつつ、不当な扱いを受けたと感じる者達が不透明な評価方法ゆえに自らの訴えを主張することを困難にしている問題を指摘).

110 Kaminski, *supra* note 104, at 1531, 1541;　*See* Houston Fed'n of Tchrs, 251 F. Supp. 3d 1168.

111 OECD, OECD AI Principles, Transparency and Explainability (Principle 1.3), https://oecd.ai/en/dashboards/ai-principles/P7 (last visited Feb. 13, 2024).　*See also* Kaminski & Urban, *supra* note 108, at 1982 (筆者と同じく OECD AI 原則 1.3 条を、異議申立権の規程の例として挙げつつ、OECD 理事会勧告は世界中の立法者に影響を与え得ると指摘).

112 大湾・前掲注（63）7 頁.

113 例えば本稿が焦点を当てている人事採用選別や被用者の評価等に於ける主な AI 利用は、前掲注（8）に於いて前述したように機械学習と自然言語処理であるとされるところ、後者については近年、大規模言語モデル（LLM: large language models）を使ったいわゆる ChatGPT に代表される生成 AI が飛躍的な進化を遂げていることから、現在の AI の欠点の治癒を期待したい。

さをヒトによる意思決定の参考にしても良い場合であっても、その正しさを確保できるような前提/停止条件（独立/中立的監査人による監査＋公表＋独立/中立的第三者による検証等々）を満たした上で、HITL 的にヒトの決定/監視も加え、かつその決定/監視が意味のある内容になるように自動化バイアスへの対策も講じた上で、AI とヒトが各々適する業務に於いて役割を果たすようなハイブリッドな運用を目指すべきであろう[114]。日本も本稿が紹介した欧米に於けるバランスのある指摘に学んで、その使用が妥当ではない場合の AI の利用は慎むべきであろう。

[114] そのような AI と、ヒトの意思決定等とのハイブリッドで適切な運用の望ましさについては、学生 Note 論文ではあるけれども、〈ナッジ〉理論や〈法と行動経済学〉で有名な Cass Sunstein 教授の指導の下で書かれた、『ハーヴァード・法と技術雑誌』34 巻 259 頁（2020 年）掲載の「正確さだけでは十分ではない」という論文が、AI の特性とヒトの特性を組み合わせた業務分担が望ましいという、説得力のある指摘をしているので参考になろう。Lowens, *supra* note 50, at 959.