# AIのELSI（エイ・アイ　エルシー）

## ――人工知能の〈倫理的・法的・社会的影響〉と，その研究の必要性――

平　野　　晋(*)

## *ELSI of AI : Ethical, Legal, and Social Implications of Artificial Intelligence*

Susumu Hirano

**Abstract**

R&D for artificial intelligence（“AI”）or usage thereof requires consideration on ethical, legal, and social implications（“ELSI”）because AI inherently contains dangerous characteristics such as uncontrollability, opacity, and unforeseeability. In this piece, the author tries to persuade readers to understand the necessity of ELSI through narratives and by focusing upon un-controllability of AI. He picks up some exemplary narratives from famous sci-fi films such as Robocop（Orion Pictures 1987）and 2001 Space Odyssey（MGM 1968）. Also he shows some counter-measures against the un-controllability of AI.

**Key Words**
AI Principles, Uncontrollability, Opacity, Storytelling, Robocop, 2001 Space Odyssey

目　　次



(*)筆者は，須藤修先生（現・中央大学国際情報学部教授）と共に，OECD・AI専門家会合（エイ・アイ・ゴー/エイゴー）AIGO（*A*rtificial *I*ntelligence *G*roup of experts at *O*ECD），内閣府・人間中心のAI社会原則［検討］会議，総務省・AIネットワーク社会推進会議，及び同省・AIネットワーク化検討会議に於いて，AI諸原則及び同ガイドライン等の構築と国際普及活動に携わって来たけれども，それら会議関連公式記録に記載のない本稿内の記述は，筆者の私見である．

---

> [T]he inability to predict future outcomes does not imply that scientific advances always should go unchecked. Scientists conducting research and creating technology may not be as aware . . . of the potential problems posed by their discoveries. Richard Posner has noted that “[s]cientists want to advance scientific knowledge rather than to protect society from science; the policy

maker's ordering of values is the reverse. Not that scientists are indifferent to public safety; but it is not their business and sometimes it is in competition with their business."[ ] In short, scientists want what is best for science, not necessarily what is best for society. Consequently, Posner encourages lawyers and lawmakers to think in terms of prevention. [ ]

Jessica L. Roberts, *Preempting Discrimination : Lessons from the Genetic Information Nondiscrimination Act*, 63 VAND. L. REV. 439, 481-82 (2010) (emphasis added).

## はじめに

人工知能（AI：artificial intelligence）[1]の開発や利活用等に於いて〈ELSI[2]：倫理的・法的・社会

1) 人工知能の定義が定まらないという話は有名である．*See, e.g.*, Axel Walz & Kay Firth-Butterfield, *Implementing Ethics into Artificial Intelligence : A Contribution, from a Legal Perspective, to the Development of an AI Governance Regime*, 18 DUKE L. & TECH. REV. i, 182 (2019); 内閣府・統合イノベーション戦略推進会議「人間中心の AI 社会原則」1 頁(平成 31 年 3 月 29 日), https : //www 8.cao.go.jp/cstp/aigensoku.pdf (last visited Aug. 17, 2020); 拙書『ロボット法～AI とヒトの共生に向けて～(増補版)』99 頁＆脚注 271（弘文堂 2019 年).

2) 〈ELSI〉とは「**e**thical, **l**egal, and **s**ocial **i**mplications」の略語．AI やロボット等の先端技術に於ける ELSI を――すなわち倫理的・社会的影響，ソフトロー規範，又は法的規制等の検討の必要性等について――論じている海外の文献例としては，see, *e.g.*, OECD, Recommendations of the Council on Artificial Intelligence, OECD/LEGAL/0449, May 22, 2019, https : //legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (last visited Aug. 17, 2020)（いわゆる「OECD AI Principles」に関する理事会勧告); OECD, OECD AI Policy Observatory, https : //oecd.ai/ (last visited Aug. 17, 2020)（同サイト内の「OECD AI Principles」の項目にて同原則の解説を公表); US, National Science and Technology Council, Networking and Information Technology Research and Development Sub-committee, *The National Artificial Intelligence Research and Development Strategic Plan*, Strategy 3, 26–27 (Oct. 2016), https : //www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf (last visited Sept. 28, 2020)（「第 3 戦略」として「AI の ELSI 理解と取り組み」を挙げている．なお同戦略の 2019 年改訂版に於いてもこの第 3 戦略が堅持されている．US, Select Committee on Artificial Intelligence of the National Science and Technology Council, Networking and Information Technology Research and Development Sub-committee, *The National Artificial Intelligence Research and Development Strategic Plan : 2019 Update*, 19–22 (June, 2019), https : //www.nitrd. gov / pubs / National-AI-RD-Strategy-2019. pdf (last visited Aug. 17, 2020). *See also* US Department of Defense, Immediate Release : *DOD Adopts Ethical Principles for Artificial Intelligence*, Feb. 24, 2020, https : //www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ (last visited Aug. 17, 2020); European Commission, White Paper on Artificial Intelligence : A European Approach to Excellence and Trust, 8 (Feb. 19, 2020), https : //ec.europa.eu / info / sites / info / files / commission-white-paper-artificial-intelligence-feb2020_en.pdf (last visited Aug. 16, 2020)（欧州が AI の倫理を主導して来たと主張); Report with Recommendations to the Commission on Civil Law Rules on Robotics, Eur. Par. Doc. PE 582.443 v 03–00 (2017), https : //www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.pdf (last visited Aug. 12, 2020)（"whereas . . . ever more sophisticated robots, bots, androids and other manifestations of artificial intelligence ("AI") seem to be poised to unleash a new industrial revolution, . . . , it is vitally important for the legislature to consider its legal and ethical implications and effects, without stifling innovation"（emphasis added）(footnote omitted）と指摘); Gerhard Wagner, *Robot, Inc. : Personhood for Autonomous Systems?*, 88 FORDHAM L. REV. 591 (2019)（ロボットのような自律的存在を "ePerson" と呼称して賠償責任を課す検討をしている欧州議会の動向を契機に，ロボットに法人格を付与した賠償責任を議論している); Michael Guihot et al., *Nudging Robots : Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385 (2017)（汎用 AI の危険性を示唆しながら AI の

暴走阻止が手遅れにならない為に，かつ特化型 AI による差別等の知られた問題・危険性を治癒する為にも，遅れ気味な政府による規制関与を深めるべきと主張）；Brian S. Haney, *The Perils and Promises of Artificial Intelligence*, 45 J. LEGIS. 151（2019）（近年考案された “deep reinforcement learning” が防衛システムを突破できる危険性や一人の天才が汎用 AI を開発してしまう危険性に触れながら，汎用 AI を含む法規制の必要性を主張）；Matthew U. Scherer, *Regulating Artificial Intelligence Systems : Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. L. & TECH. 353（2016）（汎用 AI の危険性が指摘されている事実に言及しながら，責任の所在が不透明になるおそれ等のように AI には様々な問題が既に顕在化しているので，AI 規制の難しさを指摘しながらも必要な規制方法を提言している）；Omer Tene & Jules Polonetsky, *Taming the Golem : Challenges of Ethical Algorithmic Decision-Making*, 19 N.C. J. L. & TECH. 125（2017）（社会的偏見／差別を反映する policy-neutral algorithms と，これを矯正する policy-directed algorithm の違いを分析し，後者を採用する際には透明性が必要と主張）；Shlomit Yanisky-Ravid & Sean Hallisey, *“Equality and Privacy by Design” : A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes*, 46 FORDHAM URB. L. J. 428（2019）（AI に用いるデータの監査や認証制度を導入すると共に，データ誤使用回避を努力すれば safe harbor 免責を享受できるようにして，disparate impact を含む差別的データ等の弊害を無くして透明性を高めるように提言している）；Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders’ Use of Big Data*, 93 CHI-KENT L. REV. 3（2018）（ビッグデータが差別を生む危険性と貸主規制の必要性を論じる中で disparate impact の概念にも言及している）；Edmund Mokhtarian, *The Bot Legal Code : Developing a Legally Compliant Artificial Intelligence*, 21 VAND. J. ENT. & TECH. L. 145（2018）（制御不可能性を内包する AI に，法律を学習させる必要性や実現可能性を論じている）；Rayan Hagemann et al., *Soft Law for Hard Problems : The Governance of Emerging Technologies in an Unclear Future*, 17 COLO. TECH. L.J. 37（2018）（マルチステイクホールダーなソフトローが新技術の規制には適切であると主張している）；Aviv Gaon & Ian Stedman, *Call to Action : Moving forward with the Governance of Artificial Intelligence in Canada*, 56 ALBERTA L. REV. 1137（2019）（カナダで AI の倫理等を扱う国家組織結成を提言）．なお国内の主な資料としては，例えば，内閣府「人間中心の AI 社会原則」前掲注(1)；内閣府「AI 戦略 2019～人・産業・地域・政府全てに AI～」53 頁（令和元年 6 月 11 日），https : //www.kantei.go.jp/jp/singi/ai_senryaku/pdf/aistratagy 2019.pdf（last visited Aug. 17, 2020）（一つの章を割いて倫理等の重要性について記載している）；総務省・AI ネットワーク社会推進会議「報告書 2020～『安心・安全で信頼性のある AI の社会実装に向けて』～」（令和 2 年 7 月 21 日），https : //www.soumu.go.jp/main_content/000698163.pdf（last visited Aug. 17, 2020）；同会議「報告書 2019」（令和元年 8 月 9 日），https : //www.soumu.go.jp/main_content/000637096.pdf（last visited Aug. 17, 2020）；同会議「報告書 2018 ～AI の利活用の促進及び AI ネットワーク化の健全な進展に向けて～」（平成 30 年 7 月 17 日），https : //www.soumu.go.jp/main_content/000564147.pdf（last visited Aug. 17, 2020）；及び、同会議「報告書 2017 ～AI ネットワーク化に関する国際的な議論の推進に向けて～」（平成 29 年 7 月 28 日），https : //www.soumu.go.jp/main_content/000499624.pdf（last visited Aug. 17, 2020）．

3）　なお AI の開発，利活用，及び経済発展等々の為には理数工学系の教育研究が重要であるとしばしば指摘され，その概念は後掲注(30)の「STEM」という文言に象徴されている．しかし暴走を抑止する為には，ELSI が不可欠と言われている．*See, e.g.*, 総務省「ICT インテリジェント化影響評価検討会議（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 Sept. 10, 2020）（理研の高橋亘一氏が以下のように発言している．「技術開発はアクセルにあたる．人文社会科学にはハンドルという側面もあるが，ブレーキでもある．ブレーキを踏む人を確認することやそのためのルールが，正しいものになっているからこそアクセスを踏めるだろう．例えば，遺伝子工学では倫理委員会が大学毎にあるが，そういう組織があるからこそ技術開発が進められてきた．」（強調付加）と）．なお科学者は，傲慢になってはいけない．謙虚さを失ってはいけないのである．その仕事は，社会に多大な影響を与えるからである．例えば〈汎用 AI〉――後掲注(53) 参照――がどのようなものになるのか．それが現時点では不明であっても，野放図を許して良いことにはなるまい．(i) 社会にとっての安全利益と，(ii) 科学の発展や研究の自由利益との，何れを優先すべきか．本稿冒頭の引用文が示唆するように，前者と後者が相反する場合もある．そこで必要になる

的影響〉研究が重要な理由は，上記引用記載の通りである[3]．他方，本稿では，より具体的に，AI

に関するELSIの重要性を示してみる．尤も紙面と時間の制約ゆえに，特にAIに内包される〈制御不可能性〉と呼ばれる欠点に主な焦点を絞りつつ[4]，ELSIが必要となる具体例を〈ナラティブ：物語〉を用いて説明する．

そのようにナラティブを用いて議論又は説得する手法が，有効かつ学術的にも認知され受容されている事実は，既に脱稿した別の拙稿に於いて紹介済みである[5]．そこで本稿では，先ず第Ⅰ章にて，〈ハイポ〉（hypo.：hypotheticals）と呼ばれる具体的な仮想事例を通じて，ELSIの必要性を指摘してみる．第Ⅱ章では，AIの大きな欠点の一つである制御不可能性について概説する．第Ⅲ章では，AIの制御不可能性が引き起こす影響の重篤性を具体的に表しているSF作品の特徴的場面を通じて，制御不可能性が看過できない問題であり，AIにはELSI的分析や検討が必要であるという理解の一助としたい．最後に第Ⅳ章では，制御不可能性問題に対して具体的に提案されている対策案も例示しておこう．

## Ⅰ. ドローン兵器の仮想事例で学ぶELSIの必要性

AIは，経済成長をもたらしてくれるという期待感から，一方では経済界を中心に持て囃（はや）されている．しかし，他方，様々な危険性が指摘され，かつ大衆を含む広く社会全体から懸念も示されている．それ故，そのELSIを踏まえた対応や検討の必要性も指摘されている[6]．ところでELSIの必要性を，以上のように単なる抽象的な表現に止めておくことなく，ELSI的配慮の必要性を以下のナラティブを通じて具体的に説明してみよう．

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

《ナラティブその1：ドローン兵器のハイポ》

X国の国家公務員であるあなたは，来月の人事異動で，ドローン兵器使用[7]の是非がテーマにな

---

のが，ELSIである．ロスアラモスにて嘗て，マンハッタン計画を推進したばかりか，その研究成果物のヒロシマへの投下まで進言したオッペンハイマーは，個人的名誉に固執して行動していたと伝えられている――NHK「BS1スペシャル・“悪魔の兵器”はこうして誕生した～原爆科学者たちの心の闇～」2020年8月15日0：40～2：20再放送――．社会的影響が大きいAIに於いても，科学者のエゴの為に同じ轍を繰り返し踏ませる訳にはゆかない．科学者は傲慢になってはいけない，謙虚さを失ってはいけないのである．*Cf.*「報告書2017」*id.* at 26−28 & nn.62−67（汎用AIを射程に入れる方針を明記しつつ，かつ他国も同様である事実も指摘して，射程に入れるべき理由も詳細かつ丁寧に説明している）；総務省「AIネットワーク社会推進会議第6回議事概要」（平成29年7月25日），https：//www.soumu.go.jp/main_content/000513654.pdf（last visited Aug. 24, 2020）（日本を代表する圧倒的多数の有識者達が汎用AIをソフトローの射程に含めるべきと主張し，かつ説得力のあるその根拠を述べている）．汎用AIが未だ実現されていない技術であってもソフトロー規制の射程に含めるべき点については，see *also* materials in *infra* note 52. *But see*（一社）日本経済団体連合会「AI活用戦略～AI-Readyな社会の実現に向けて～」4，27頁（2019年2月19日），http：//www.keidanren.or.jp/policy/2019/013_honbun.pdf（last visited Aug. 17, 2020）（汎用AI懸念を非難している）．

4） AIの欠点は制御不可能性のみに限られず，不透明性や差別的判断等複数存在する．*See infra* note 42 and its accompanying text.

5） ナラティブの使用が学術的に認知・認容されているテーマについては，see 拙稿「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～」『法學新法』____号____頁（中央大学出版部2021年春発刊予定）；拙稿「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頁（2018年4月）．ロボット／AIの危険性等の予測にフィクションを用いる例としては，see *generally* 拙書『ロボット法』*supra* note 1, at iv−v, 6, 9, 1−22, 33, 36−50, 59, 61−62, 63, 98−99, 126−30, 163, 164, 165−66, 167, 169−70, 184−85, 231−35, 242−43, 245, 248, 250（弘文堂2019年）．法と文学については，see, *e.g.*, リチャード・A.ポズナー著，坂本真樹＆神馬幸一訳，平野晋監訳『法と文学（上）（下）（原著第3版）』（木鐸社2011年）．

6） *See, e.g.*, Brandon W. Jackson, *Artificial Intelligence and the Fog of Innovation：A Deep-Dive on Governance and the Liability of Autonomous Systems*, 35 SANTA CLARA HIGH TECH. L. J. 35, 35 (2019).

7） ドローン〈兵器自体〉の是非論議と，その〈使用〉を巡る是非論議とは異なる，という指摘については，

る国際会議を担当するように命じられた，と仮定しよう．あなたは国際会議や兵器等とは無縁の職場しか経験して来なかったにも拘わらず，どのような任務でも適切にこなすことをあなたは信条としている．

国際会議では既に，超大国Ｙによる対テロ戦争に於けるドローン兵器の使用が，発展途上国Ｚから批判されている，と仮定しよう．Ｚ国代表曰く，ドローン兵器の使用はＹ国の傲慢の現われである[8]．ドローンは自国将兵に死傷者が出ない，一方的過ぎる兵器であり，Ｙ国が全く危険を負担せずに済むから，対等な防衛力を持たない国や地域としては受容できない[9]．自国将兵に死傷者が出ない上に安価でもあるから，安易に戦争を始めるおそれもある[10]．続けてＺ国曰く，Ｙ国による海外テロ組織へのドローン兵器攻撃によって，多数の市民の〈巻き添え損害〉(collateral damage)[11]が発生している[12]．特に，即席爆弾（ＩＥＤ（アイ・イー・ディー）：*I*mproved *E*xplosive *D*evices）を敷設するテロリストの行動を AI で分析した結果の識別特性（signature）を有する者をテロリストであると推認した攻撃――識別特性攻撃――が的外れな為に，市民がテロリストと誤認されて巻き添え損害を被っている[13]，等々とＺ国は批判していた．

他方，この批判に対するＹ国の主張は，次の

see, *e.g.*, Vivek Sehrawat, *Legal Status of Drones under LOAC and International Law*, 5 PENN. ST. J. L. & INT'l AFF. 164, 175 (2017) ("When determining the overall lawfulness of a weapon system under LOAC [*i.e.*, Law of Armed Conflict], there are two distinct aspects of the law that need to be analyzed: weapons law and lawful use of drones. [ ] The former verifies that the weapon itself is lawful. [ ]" と指摘)．ドローンの〈使用〉については格別，兵器としての〈ドローン自体〉は条約で禁じられていないという指摘については，see, *e.g.*, Oren Gross, *The New Way of War: Is There a Duty to Use Drone?*, 67 FLA. L. REV. 1, 26–27 (2015)．すなわちドローン兵器の問題は，その〈使用〉方法にある．*Id.* at 27. なおドローン兵器は，ヘリコプター等を含む従来型ジェット攻撃機等と同様で，さしたる目新しさはなく，それ迄以上に遠距離から攻撃して，攻撃者の危険性を更に引き下げたに過ぎないという捉え方もあり，倫理的な問題も，何時／如何に使用するのかに過ぎないと理解する向きもあると指摘されている．ELKE SCHWARZ, DEATH MACHINES 176 (2019).

8) *See* Gross, *id.* at 2.

9) *See, e.g.*, Frederic Megret, *The Humanitarian Problem with Drones*, 2013 UTAH L. REV. 1283, 1310–12.

10) SCHWARZ, *supra* note 7, at 176, 177.

11)「collateral damage」とは，民間人が被った死亡又は損害等の意である．*See, e.g.*, PROGRAM ON HUMANITARIAN POLICY AND CONFLICT RESEARCH AT HARVARD UNIVERSITY, MANUAL ON INTERNATIONAL LAW APPLICABLE TO AIR AND MISSILE WARFARE, rule 1 (I) (May 15, 2009), https://reliefweb.int/sites/reliefweb.int/files/resources/8 B 2 E 79 FC 145 BFB 3 D 492576 E 00021 ED 34-HPCR-may 2009.pdf (last visited Aug. 27, 2020); Anthony J. Gaughan, *Collateral Damage and the Law of War: D-Day as a Case Study*, 55 AM. J. LEGAL HIST. 229, 230 & n.11 (2015)．巻き添え損害の ELSI を理解する為に有用な作品としては，see Eye in the Sky (Entertainment One Films 2015).

12) *See* JEAN-FRANÇOIS CARON, CONTEMPORARY TECHNOLOGIES AND THE MORALITY OF WARFAIR 55 (2020).

13) *Id.* at 57 (そのような "signature strikes" が非戦闘員への誤爆原因であると指摘)．*But see* Shin-Shin Hua, *Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control*, 51 GEO. J. INT'L L. 117, 133–34 (2019)（ヒトの判断よりも機械に任せた方が良く成り得ると主張)．なお「識別特性攻撃」(signature strikes) とは，例えば道端で穴を掘るという，IED を敷設するテロリストと似た行動パターンの識別特性(signature) に合致する者はテロリストであると看做した攻撃を云う．Charles P. Trumbull IV, *Autonomous Weapons: How Existing Law Can Regulate Future Weapons*, 34 EMORY INT'L L. REV. 533, 574 & n.259 (2020)．これとは対極的に，名簿化された個々の人物・テロリストへの攻撃は「personality strikes」(個人特定攻撃) と云う．Gregory S. McNeal, *Targeted Killing and Accountability*, 102 GEO. L. J. 681, 701 n.93 (2014); Milena Sterio, *The Covert Use of Drones: How Secrecy Undermines Oversight and Accountability*, 8 ALB. GOV'T L. REV. 129, 162–63 (2015)．*See also* SCHWARZ, *supra* note 7, at 175 (signature / personality strikes を説明)．

通りである――ドローン兵器のような，ピンポイントに標的を攻撃できる兵器の使用は，禁じられるべきどころか，逆に奨励されるべきである[14]．何故ならそれ迄の，第二次世界大戦で使用されていた爆弾や核爆弾のように精度が欠ける兵器よりも[15]，戦闘員の標的を外科手術的に切除できて，非戦闘員の巻き添え損害が比較的少なくて済むから[16]，ドローン兵器は使用されるべきなのである[17]．新技術の発達によって精密な攻撃が可能になった現代では，社会規範上寧ろドローン兵器を使用することこそが期待されている．更に云えば，ドローン兵器は遠距離から監視できて[18]，精度の良いカメラ画像を用いた分析も可能で[19]，かつ時間を掛けて[20]諸情報を突合して慎重に戦闘員と非戦闘員を区別してから攻撃判断が出来るのだから[21]，通常兵器や核兵器よりも巻き添え損害を減らせるのである[22]．加えて，無差別爆撃や大量虐殺は，敵国同士が紛争終了後に平和的関係を修復する際の妨げにも成るから[23]，精密攻撃の方が望ましい．そもそも優位性を争うのが戦争であるから，一方的であってはならない云々という義務を課している条約は見当たらない[24]，と．

あなたは，Y 国と Z 国の主張の何れに理があるのかを，予習・調査して国際会議に備えることにした．ドローン兵器の性能や仕様等の工学技術的情報を，先ずは習得せねばならないことは分かるし，そこは理系出身のあなたにとって得意分野でもある．しかし，それ以外の，人文科学的，又は法律学を含む社会科学的な検討要素も考慮することが必要ではなかろうか．．．．

《以上でナラティブ終了》

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

上のハイポに於いて先ず重要な考慮要素は，あなたが直ぐに思いついたように，ドローン兵器の工学技術的な諸要素である．具体的には，遠方からの索敵能力，精密なピンポイント攻撃能力，その他の仕様，更には即席爆弾を敷設するテロリストの行動パターンを AI に学ばせた結果としての戦闘員／非戦闘員識別的中率の技術的現状や，そ

14） Paul Scharre, Army of None: Autonomous Weapon and the Future of War 282 (W.W. Norton & Co. 2018). *See also* Gross, *supra* note 7, at 60–71（ドローン兵器使用の義務があると主張）; Schwarz, *supra* note 7, at 177（ドローン兵器使用推進者はしばしば，ドローン兵器が正確で，非戦闘員を区別できて，かつ死傷者を抑制できて目的を達成するから「模範的な倫理的兵器―exemplary ethical weapon―」であると主張しがちと指摘）; *id.*（賛否についてコンセンサスは得られていないけれども，「戦争に対する更に倫理的なアプローチの為の『法的，倫理的，かつ賢い―wise―』道具」であるとか「人道的な行動の為の『道徳的―virtuous―』道具」であるとか，「MQ 9 リーパー［型ドローン］は，これ迄に開発された航空兵器の中でおそらく最も倫理的である」（拙訳）というような，以前の兵器には見られない倫理的価値が高い表現をされていると指摘）.

15） Caron, *supra* note 12, at 48（核兵器のように無差別に非戦闘員を殺傷する兵器はよろしくないと指摘）.

16） *Id.* at 48–49（ドローンやその他のレーザー誘導ミサイル使用は戦闘員を正確に標的にできるから結果的に非戦闘員を守っている等と指摘）. *See also* Megret, *supra* note 9, at 1297（従来型通常兵器と比べると，巻き添え損害が大きいという指摘は誤りと指摘）; Gross, *supra* note 7, at 48; Winston P. Nagan & Megan E. Weeren, *An Essay and Comment on Oreg Gross', "The New Way of War: Is There a Duty to Use Drones?,"* 67 Fla. L. Rev. F. 20, 21 (2015)（巻き添え損害を減らせることがその使用義務化の根拠である Gross の指摘を肯定）.

17） Scharre, *supra* note 14, at 281–82.

18） 15,420 メートルの上空から地上の人物を特定できると指摘されている．Caron, *supra* note 12, at 57.

19） 6,096 メートルの彼方から 15.24 センチの物を捉えて 25 メートル四方の範囲をカバー出来ると指摘されている．*Id.*

20） *Id.*

21） 尤も標的を選択する際の根拠，アルゴリズム，及び非戦闘員死傷者等の詳細が不明である［からその使用の是非を判断できない］という批判も見受けられる．Schwarz, *supra* note 7, at 177.

22） Laurie R. Blank, *In Response to Professor Oren Gross' "The New Way of War: Is There a Duty to Use Drones?,"* 67 Fla. L. Rev. F. 38, 40 (2015).

23） *See* Caron, *supra* note 12, at 50.

24） *See* Megret, *supra* note 9, at 1313–15.

の将来の向上可能性，等であろう．通常兵器の性能とドローン兵器のそれとの比較も重要である[25]．加えて統計学的な数理系要素として，ドローン兵器による目的達成並びに巻き添え損害の実績数値データと統計分析[26]，通常兵器が目的を達成する際に生じた巻き添え損害の実績数値データ[27]と統計分析[28]，及び，両兵器の数値比較[29]，等々も重要であろう．すなわち工学・数理学的素養――いわゆる「STEM（ステム）[30]」的な素養――が重要になる．しかしSTEMだけで，この問題――ドローン兵器の使用を禁じるべきか否か――を理解することは出来ない．

そもそもドローン兵器の使用は，倫理的に許容されるであろうか[31]．倫理的に許容されないような兵器であれば，是非の天秤は非に傾くから，〈倫理的影響の“***E***”：***e***thical implications〉も重要な検討要素なはずである．

次に必要な考慮要素は，戦争に適用される法律／条約である「国際人道法」（IHL：International Humanitarian Law[32]）等を，ドローン兵器使用が遵守しているか否かである．違法な兵器使用は，原則として当然に非難されるべきだからである．

25） 第二次世界大戦の通常兵器（爆弾）は精度がおそろしく悪いので，直径2キロメートルの円内に爆弾を落とす確率が僅か50％であった．平均的な標的を90％破壊する為に必要な爆弾の数も9,000発を超えていた，と指摘されている．SCHARRE, *supra* note 14, at 281.

26） 例えばブルッキングズ研究所による2009年の報告書によれば，アメリカのドローンがパキスタンでテロリスト一名を殺害する為に，非戦闘員10人の巻き添え損害が生じた，という指摘もある．CARON, *supra* note 12, at 55.

27） 第二次世界大戦で通常兵器（爆弾）を用いた日本への戦略爆撃は，30万人を超える非戦闘員が犠牲になり，中でも東京爆撃では僅か一晩で10万人超の非戦闘員が犠牲に成っている．SCHARRE, *supra* note 14, at 282．その外にも米英はドレスデンやハンブルグ等々で非戦闘員を何万人も殺害している．*Id.*

28） 同上脚注の日本への戦略爆撃を指揮したカーチス・ルメイ将軍を，統計学的分析面に於いて助けたのは，後にキューバ危機やベトナム戦争時代に国防長官となるロバート・マクナマラであったことが有名な事実であることに象徴されるように（以下の出典参照），統計学と戦略の分析との間には関連性があることは明白であろう．Tim Weiner, *Robert S. McNamara, Architect of Futile War, Dies at 93*, The New York Times, July 6, 2009, https://www.nytimes.com/2009/07/07/us/07 mcnamara.html（last visited Aug. 24, 2020）．なお余りにも多くの非戦闘員の犠牲者を出した日本への戦略爆撃を，マクナマラは後に悔いてルメイを非難していたことも，有名な話である．Robert S. McNamara, *We Need Rules for War*, Aug. 3, 2003, https://www.latimes.com/archives/la-xpm-2003-aug-03-oe-mcnamara 3-story.html（last visited Aug. 24, 2020）; SCHARRE, *supra* note 14, at 279（以下の出典の有名なドキュメンタリー「Fog of War」に於いて，自身もルメイも戦争犯罪者だと述べた事実を引用しながら説明）．The Fog of War: Eleven Lessons from the Life of Robert S. McNamara（Sony Pictures Classics 2003）．

29） やはり前掲脚注(27）の日本に於ける戦略爆撃の結果の非戦闘員の犠牲者数に比べると，前掲SCHARREによれば，ソマリア，パキスタン，およびイエメンでのテロリスト攻撃に於ける非戦闘員の死者数は2015年に3〜16名，2016年は4名であった．SCHARRE *supra* note 14, at 279. 尤もSCHWARZによれば，ソマリア，パキスタン，イエメン，及びリビアの非戦闘員の巻き添え損害の死者数としてアメリカ政府が把握している数は116名とされていて，あまりにも少ない数値なので調査報道局－Bureau of Investigative Journalism－はその6倍であると疑っているという．SCHWARZ, *supra* note 7, at 177.

30） 「STEM」とは，「***S***cience, ***T***echnology, ***E***ducation, and ***M***athematics」の略語である．*See, e.g.*, Committee on STEM Education of the National Science & Technology Council, *Changing a Course for Success: America's Strategy for STEM Education* iv（Dec. 2018），https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf（last visited Aug. 12, 2020）．

31） 例えば，地雷やブービートラップのような兵器は，無差別に非戦闘員も殺傷するので非倫理的であるという指摘があるので考慮すべきであろう．*See* CARON, *supra* note 12, at 56．逆に言えば，狙撃手が用いる長距離ライフルは，戦闘員だけを区別して標的に出来る分だけ倫理的であり，ドローン兵器も同様であるという指摘も見受けられる．*Id.* at 57．尤も，アメリカは非戦闘員の命よりも自軍将兵の命を優先させているという批判も見受けられる．*Id.*

すなわち〈法的影響の“*L*”: *l*egal implications〉も重要である．

更に，〈社会的影響の“*S*”: *s*ocial implications〉も重要である．ハイポ中の「社会規範上 . . . 期待されている」云々という文言が示唆するように[33]，社会の期待は如何なる現状なのか[34]，社会は通常型爆弾や核兵器よりもドローン兵器を要請しているのか否か，等々を検討要素に加えなければ，社会からの理解が得られないからである．

以上をまとめると，〈倫理的“*E*”，法的“*L*”，及び社会的“*S*”な影響“*I*”〉すなわちELSIが，STEM同様に重要な検討要素に成るのである．

なお，AI等の「emerging technologies」と呼ばれる〈先端／創発技術〉に関連してELSI研究が重要な分野は，ドローン兵器やロボット兵器分野の賛否論議だけに限られないことは勿論である——例えば，個人情報を自由に使えないから機械学習等に必要な［ビッグ］データが十分に使えず，それ故に人々にとっての便益向上の為のAIの利活用も阻まれている，という見方もある[35]．つまり個人情報保護の利益と自由な［ビッグ］データ利用の利益とのトレードオフな対立関係が見受けられ，そのような対立を如何に解決すべきかを検討すること[36]も，倫理的・法的・社会的影響の範疇に含まれる——．このように，ELSIの射程はドローン兵器論議に限られず，先端／創発技術が関連する広い地平の隅々に迄も及ぶのである．

## II. 〈制御不可能性〉概説

### A. 総務省・AIネットワーク社会推進会議「AI開発ガイドライン」と「AI利活用ガイドライン」

世界的にみれば近年，AIの欠点と対応策をELSI的視点から提言する試みが多く見受けられる[37]．中でも最近，経済協力開発機構（OECD）がとりまとめたAI原則の勧告は[38]，その影響力からしても重要である[39]．そして，総務省の有識者会議がとりまとめたAI開発／利活用原則やガイドライン[40]等も，AI原則の国際標準化に大きく貢献してグローバルに評価も高いので[41]，特に

32） *See, e.g.*, United Nations, Fact Sheet No. 13, International Humanitarian Law and Human Rights, https://www.un.org/ruleoflaw/files/FactSheet 13 en.pdf (last visited Aug. 28, 2020)；赤十字国際委員会「国際人道法のいろは～わかりやすい国際人道法～」, http://jp.icrc.org/wp-content/uploads/sites/92/2015/06/201501 ABC_IHL.pdf (last visited Aug. 28, 2020)．国際人道法とロボット兵器の論議については，see *also* 拙書『ロボット法』*supra* note 1, at 60–62.

33） 巻き添え損害を減らすことが出来るように工学技術が進化したのだから，「社会規範も変化して(social norms have shifted too)」，「精密誘導兵器の使用が要求されている(*requires*)」という指摘も見受けられる．SCHARRE, *supra* note 14, at 282（訳は筆者拙訳）．

34） 非戦闘員を区別できるドローンは〈兵器〉として非難されているというよりも，寧ろ signature strikes（識別特性攻撃）という〈使用方法〉が倫理的に非難されているという指摘が見受けられる．CARON, *supra* note 12, at 57.

35） *See* Jackson, *supra* note 6, at 48.

36） *See id.* at 48–49.

37） *See* materials in *supra* note 2.

38） OECD AI Principles, *supra* note 2.

39） OECD「42か国がOECDの人工知能に関する新原則を採択」2019年5月22日，https://www.oecd.org/going-digital/ai/principles/ (last visited Aug. 17, 2020)（1980年OECDプライバシー・ガイドラインのように勧告は世界への影響力が大きいと指摘）．

40） 総務省AIネットワーク社会推進会議「国際的議論のためのAI開発ガイドライン案」8～9頁（平成29年7月28日），https://www.soumu.go.jp/main_content/000499625.pdf (last visited Aug. 11, 2020)；同会議「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～」（平成元年8月9日），https://www.soumu.go.jp/main_content/000637097.pdf (last visited Aug. 17, 2020)．

41） *See* 拙稿「GAFA規制を考える(中) AI利活用で独走許すな」日本経済新聞(朝刊) 2019年2月20日（国際ルール構築に日本の諸原則・ガイドラインが貢献していると諸国関係者が指摘している事実を紹介）．*See also* 総務省「報告書2019」*supra* note 2, at 2. 以下のように指摘している．

日本の関係者にとっては重要である．それら AI 諸原則とガイドライン等は，AI の欠点と対応策の指摘として，制御可能性の原則，透明性の原則，安全の原則，アカウンタビリティの原則，適正学習の原則，公平性の原則，等や，それら諸原則の〈解説〉等も公表している[42]．その内容については，拙書『ロボット法』[43]や拙稿「“AI の支配”と“法の支配”」等[44]もご覧願いたいが，本稿では数ある諸原則の中から，制御不可能性に焦点をあてて，ELSI の重要性を説明する．

### B. 制御不可能性

〈制御不可能性〉（un-controllability）を定義するならば，開発者等やプロバイダー等の意図に反する AI の判断や行動等の特性と云えよう[45]．制御不可能性の概念は，〈予見不可能性〉（un-foreseeability / un-predictability）や[46]，〈不透明性〉（opacity）又は〈ブラック・ボックス〉と呼ばれる諸問題と近似する概念である[47]．何故なら開発者等が制御できない AI の判断・行動を AI が執るということは，すなわち，予見できなかったり，又は理解できない判断・行動を AI が執ることに通じるからである[48]．実際のところ，AI に

---

OECD が理事会勧告案を策定するために 2018 年（平成 30 年）9 月に設置した専門家会合（AI expert Group at the OECD）においては，我が国から参加した有識者［すなわち須藤修教授及び筆者の 2 名］より，人間中心の AI 社会原則，AI 開発ガイドライン案，AI 利活用原則案のそれぞれについて，その内容のみならず，検討の背景や検討にあたり行われた議論の状況についても紹介するなど，OECD 理事会勧告の原案策定にあたり大きな貢献を行った．このため，2019 年（令和元年）5 月に公表された OECD 理事会勧告（Recommendation on Artificial Intelligence）は，我が国の検討と整合のとれたものとなっている．

. . . .

このように，AI に関する原則については，我が国が国際的な議論を主導してきた結果，その概念については，概ね国際的にコンセンサスが得られつつある . . . .

（強調付加）

42） *See* 総務省「AI 開発ガイドライン」*supra* note 40; 同省「AI 利活用ガイドライン」*supra* note 40.

43） 拙書『ロボット法』前掲注(1) 143, 281～313 頁.

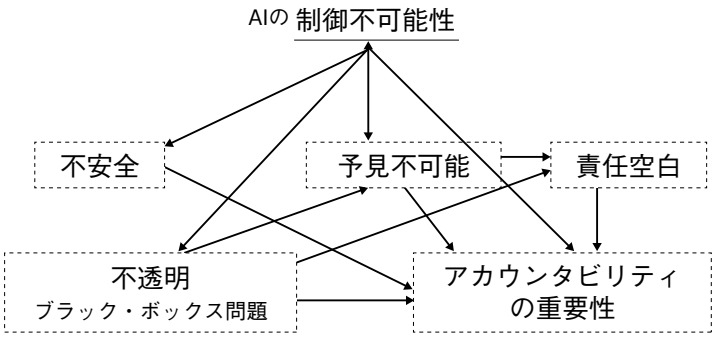
44） 拙稿「“AI の支配”と“法の支配”」『法の支配』197 号 41～56 頁(日本法律家協会 2020 年 4 月). *See also* 須藤修「人工知能がもたらす社会的インパクトと人間の共進化」『情報通信政策研究』第 2 巻 1 号 1～10 頁(2018 年); 福田雅樹「AI ネットワーク化に関する社会的・経済的・倫理的・法的課題」『法政論集』275 号 349～380 頁(2018 年).

45） *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. §3 cmt. *b.* (1998) [hereinafter referred to as PROD. LIAB. RESTATEMENT]（その製造物の本来の機能に反する事故や誤作動を分析・解説している）.

46） 予見不可能性は，民事賠償責任を原告が主張・立証する際にも大きな争点になるという指摘・分析については，see 拙書『ロボット法』*supra* note 1, at 188–96.

47） *See, e.g.*, Mokhtarian, *supra* note 2, at 152–53, 156–57（“control” を論じる文脈に於いて “predictability” の欠如を指摘）; Scherer, *supra* note 2, at 366（AI の問題は foreseeability と control であると一つのセンテンスで指摘）. 予見可能性が欠けるという指摘の例としては，see, *e.g.*, U.S. DEP'T OF DEF., DEFENSE SCIENCE BOARD, SUMMER STUDY ON AUTONOMY 15 (2016)（“Autonomous systems not only need to operate reliably . . . but also to be able to make relevant information *observable* to human and machine teammates . . . . [T]hey [*i.e.*, machines] may not incorporate sufficient *anticipatory* indicators to allow other human and machine teammates to ensure *predictability*.”（emphasis original）と分析している）. 不透明性／ブラック・ボックス問題については，see, *e.g.*, Karl Manheim & Lyric Kaplan, *Artificial Intelligence : Risks to Privacy and Democracy*, 21 YALE J. L. & TECH. 106, 155 (2019); Ashley S. Deeks, *Predictable Emotions*, 104 VA. L. REV. 1529, 1568 (2018)（AI が或る予測をした理由がヒトには不明なことが問題と指摘）.

48） 予見できないからこそ，例えば完全自律型（ヒトが関与しない）ロボット兵器が国際人道法を遵守できないという理由で反対されている．Shin-Shin Hua, *supra* note 13, at 127–28. 制御不可能とは，すなわち法を遵守できないことを意味している．




《**図**①：AI の制御不可能性とその他の欠点例との相関図》
出典：筆者.

於ける制御可能性の重要性は，透明性の重要性と伴に各種ソフトローの検討に於いて指摘されて来た[49]．加えて，それらの欠如から導き出される「責任空白」(responsibility vacuum)[50]を回避する為にも，〈アカウンタビリティ（説明及び結果責任）〉の重要性も指摘されているところである[51]．

更に制御不可能性は，〈安全の原則〉とも密接な関連性を有している．すなわち，例えば AI が組み込まれたロボットが，開発者や事業者等による制御に服さずに急に誤作動してヒトの生命・身体を害した場合には，安全性の問題に繋がるからである（後掲「ナラティブその 2」参照）．

なお，制御不可能性という AI の欠点は，未だ実現されていない[52]〈汎用 AI〉，〈強い AI〉，又

49） *See, e.g.*, 総務省・AI ネットワーク化検討会議「中間報告書：AI ネットワーク化が拓く知連社会（ウインズ）（WINS）―第四次産業革命を超えた社会に向けて―」51 頁（平成 28 年 4 年 15 日），https://www.soumu.go.jp/main_content/000414122.pdf (last visited Aug. 28, 2020)（「制御可能性と透明性が確保され」るべきと述べている）．

50） 「responsibility vacuum」やその類似語は主にロボット兵器の論稿で多く見受けれる文言であるが，民生品の論稿でも見受けられる．*E.g.*, Sabine Gless et al., *If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability*, 19 NEW CRIM. L. REV. 412, 432 (2016)（自動運転車の刑事責任を論じる文脈で responsibility vacuum の文言を使用）．

51） *See, e.g.*, Thompson Chengeta, *Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law*, 45 DENV. J. INT'L L. & POL'Y 1, 2 (2020)（AWS に関して "In the event of [AWS] violating the law -violations that are not intended by the person deploying them-it is not clear who is legally responsible, thereby creating an accountability gap. [ ]"（emphasis added）と分析）．

52） 「未だ実現されていな」くともソフトロー規制の射程に含めるべき理由は，see 拙稿「汎用 AI のソフトローと〈法と文学〉」*supra* note 5. Haney, *supra* note 2, at 163, 168–69（"Scholars who argue an AI apocalypse is merely science fiction is wrong."（emphasis added）と断定しつつ，汎用 AI が実現した暁には，ネガティブな存在であることが判明する危険性を指摘）; J.-G. Castel & Matthew E. Castel, *The Road to Artificial Superintelligence: Has International Law a Role to Play?*, 14 CANADIAN J. L. & TECH. 1 (2016)（手遅れにならない内に規制すべきと主張）．MIT の物理学者も，「知能はミステリアスな何かであるから，生物学的な組織――特に人類――にのみ存在できて，かつ知能に出来ることは今日の人類が出来ることに限定される，と原則的に捉える学者が如何に多いことか」（筆者拙訳，強調付加）と分析しながら，以下のように指摘している．

> 如何なる意味に於いても私達より知的な機械は作れない，という物理的法則は存在しない．すなわち，私達が今観ているのは単に知的氷山の一角に過ぎない――世の中には未だ完全なる知能が秘められていて，これを解き放つ驚くべき可能性が潜在する．更にその知能を使って，人類を反映させるか，又は苦しめる驚くべき可能性も潜在するのである．

Max Tegmark, *Let's Aspire to More Than Making Ourselves Obsolete, in* POSSIBLE MINDS: TWENTY-FIVE WAYS OF LOOKING AT AI 76, 79 (John Brockman

は〈超人工知能〉[53]等に於ける未来の話だけに限られるものではない．既に実装又は実現されている〈弱い AI〉又は〈特化型 AI〉[54]に於いても，既に制御不可能性の欠点が指摘されているのである[55]．従って，仮に汎用 AI が開発されてしまえば，単に現在の問題を更に悪化させるだけである[56]．何れにせよ，既に現存すると指摘されてい

ed., 2019）（emphasis added）（訳は筆者拙訳）．*See also* Michael Sainato, *Stephen Hawking, Elon Mask, and Bill Gates Warn about Artificial Intelligence*, Observer（Aug. 19, 2015, 12 : 30 PM）, https : //observer.com / 2015 / 08 / stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/（last visited Aug. 31, 2020）（世界的な識者も警告していると指摘）; Katherine B. Forrest, *Copyright Law and Artificial Intelligence : Emerging Issues*, 65 J. COPYRIGHT SOC'Y U.S.A. 355, 361 & n.29（2018）（Human Brain Project 等の例を挙げながら，"AGI does not yet exist, and while the ethics and implications of its creation and the timeline for its arrival are the subject of great debate, [ ] extraordinary human and financial capital are dedicated to its creation and its eventual arrival is nearly certain. [ ]"（emphasis added）と指摘）．なお欧州の Committee on Legal Affairs, *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, at 4/22, 2015/2 103（INL）（May 31, 2016）, https : //www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect（last visited Sept. 2, 2020）も次のように指摘している．

> [T]here is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, . . . , could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species;

（emphasis added）. *But see* 経団連「AI-Ready な社会」*supra* note 3, at 4, 27（実現までには技術的課題があるという理由だけで汎用 AI を議論することさえをも非難）．

53） 汎用 AI や超人工知能等については，*e.g.*, Michelle Sellwood, Comment : *The Road to Autonomy*, 54 SAN DIEGO L. REV. 829, 834 & n.18（2017）; Ryan Abbott, *Everything Is Obvious*, 66 UCLA L. REV. 2, 4–6, 25（2019）（AGI や weak AI や superintelligence に触れながら，専門家達が 10 年後，25 年後，又は今世紀中には実現すると予測している事実を指摘）; Haney, *supra* note 2, at 152–53, 157, 167（Narrow AI や AGI の相違を説明し，2047 年には汎用 AI が登場するという MIT 教授の見解を紹介し，一匹狼的な個人レベルでも AGI が生まれる危険性を指摘）．強い AI／弱い AI については，see *also* 拙書『ロボット法』*supra* note 1, at 249–50.

54） Valerio De Stefano, *"Negotiating the Algorithm" : Automation, Artificial Intelligence, and Labor Protection*, 41 COMP. LAB. L. & POL'Y J. 15, 23 n.32（2019）（"'narrow artificial intelligence' or 'weak artificial intelligence', namely the artificial intelligence used to perform a single task . . . "（emphasis added）と説明）. *See* Robin C. Feldman, *Artificial Intelligence : The Importance of Trust and Distrust*, 21 GREEN BAG 2D 201, 202（2018）（weak AI について）.

55） *See, e.g.*, Jonathan Tapson, *Google's Go Victory Shows AI Thinking Can Be Unpredictable, and That's a Concern*, CONVERSATION（Mar. 17, 2016）, https : //theconversation.com/googles-go-victory-shows-ai-thinking-can-be-unpredictable-and-thats-a-concern-56209（last visited Aug. 18, 2020）（AlfaGo の指し手がヒトの想像を超えていたから囲碁チャンピオンを破ったと分析）; Frank Pasquale & Glyn Cashwell, *Four Futures of Legal Automation*, 63 UCLA L. REV. DISCOURSE 26, 39（2015）（flash crash of 2010 が予測不可能な事態の証左であると指摘）; Scherer, *supra* note 2, at 368–69（同旨）; Ilias Kapsis, *Artificial Intelligence in Financial Services : Systemic Implications and Regulatory Responses*, 39 No. 4 BANKING & FIN. SERVICES POL'Y REP. 1, 6（Apr. 2020）（自律的学習機能による予測不可能な金融市場問題を指摘）; Swati Malik, *Autonomous Weapon Systems : The Possibility and Probability of Accountability*, 35 WIS. INT'L L. J. 609, 627–28（2018）（AI を用いた自律型兵器システムは，ヒトからのインプットや指示とは無関係にアルゴリズムに従って判断を下すので，ヒトによる管理が減じると分析）; Major Annemarie Vazquez, *LAWS and Lawyers : Lethal Autonomous Weapons Bring LOAC Issues to the Design Table, and Judge Advocates Need to Be There*, 228 MIL. L. REV. 89, 105（2020）（予見可能性がなければ致死的自律型兵器は使えないと分析）.

56） Mokhtarian, *supra* note 2, at 156（汎用 AI の開発は，既存の予見不可能性等の AI の問題を拡大するこ

る制御不可能性が，果たして如何なるELSI的悪影響を与えるかについては，検討を先送りせずに，問題が広範囲に広がり被害が現実化して手が付けられなく成る前の，今の内に，汎用AIも検討の射程に収めて真摯に取り組む必要がある[57]．

ところで制御不可能性の危険が重篤であることを，以下では具体的なナラティブを通じて例示してみる．

## III. 「ロボコップ」と「2001年宇宙の旅」に学ぶ〈制御不可能性〉等が招く負の影響の重篤性

### A. 「ロボコップ」と〈制御不可能性〉

映画「ロボコップ」（オリジナル版オリオン・ピクチャーズ1987年）はシリーズ化されて暫く経った後の近年になってから，リブート版も制作・公開されている[58]．ところでオリジナル版は〈ロー・ジャーナル〉や〈ロー・レビュー〉と呼ばれるアメリカ法律学の学術誌に於いても複数回言及されている[59]．加えて，以下のナラティブは単なるフィクションに止まる話ではない．ロボット兵器が誤って味方を殺傷した事故が，実際に報じられている[60]．更に自律型兵器の国際論議に於いては，以下のナラティブが描くような誤作動が無垢な人々を殺傷するおそれも懸念されている[61]．そこで同作品の一場面から，制御不可能性が生む損害の重篤性を例示してみよう．

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

《ナラティブその2：制御不可能な警備ロボットの例》

重役会議に於ける警備ロボット試作品〈ED−209（エド・ツー・オウ・ナイン）〉のデモの場面．上級重役ディック・ジョーンズの命を受けた部下の重役ケニーが，ED−209に対して試しに拳銃を向けてみたところ．．．

**ED−209**：［機械的な発音で］武器を置きなさい．命令に従うまで，あと二〇秒を与える．

**ディック・ジョーンズ**：Mr.ケニー．ロボットが言うことに従った方が良いと思うよ．［と，他人事のように冷たく言う．］

［拳銃を床に落としたけれども，ED−209はMr.ケニーを威嚇し続ける．．．．］

**ED−209**：命令に従うまで，あと一五秒．

［Mr.ケニーは，不安げにディック・ジョーンズの方を振り返る．］

**ED−209**：お前が『刑法典』第1.13条9項に違反していることは明らかだ．

［パニックに陥った重役達が，Mr.ケニーの巻き添えを食わないように離れて行く．］

**ED−209**：命令に従うまで，あと五秒．

**Mr.ケニー**：た，た，助けて．．．助け

---

とになる［だけである］と指摘）．技術革新の速度が速いから汎用AIも対象にすべきという示唆については，see, *e.g.*, Jackson, *supra* note 6, at 62−63.

57） 前掲注(3)で指摘したように「科学者は謙虚さを失ってはいけない」のである．*See also* 拙稿「汎用AIのソフトローと〈法と文学〉」*supra* note 5.

58） 「ロボコップ」（MGM 2014年）．

59） *See, e.g.*, Ric Simmons, *Terry in the Age of Automated Police Officers*, 50 SETON HALL L. REV. 909 (2020); Melanie Reid, *Rethinking the Fourth Amendment in the Age of Supercomputers, Artificial Intelligence, and Robots*, 119 W. VA. L, REV. 863 (2017); Ryan Calo et al., Symposium : Panel 2 : *Accountability for the Actions of Robots*, 41 SEATTLE U. L. REV. 1101 (2018).

60） Noah Schachtman, *Robot Cannon Kills 9, Wounds 14*, WIRED (Oct. 18, 2007), https : //www.wired.com/2007/10/robot-cannon-ki/ (last visited Aug. 28, 2020). *See also* Andrew W. Eichner, *The Limitations of System Autonomy : Analyzing Obstacles to the Vision of Autonomous Horizons and the Inevitable Future of Warfare*, 59 WASHBURN L. J. 207, 222 n.62 (2020)（非戦闘員をロボット兵器が攻撃してしまう懸念がハイポに止まらず実際に事件が生じている例として南アフリカの件を紹介）．

61） *See, e.g.*, Trumbull, *supra* note 13, at 551.

てくれぇー！

**ED−209**： 四，三，二，一，．．．物理的対応力を行使する権限が付与された．

［ED−209 が重機関砲をつるべ打ちに乱射して，無残にも Mr. ケニーは惨殺される．］

「ロボコップ」（オリジナル版）（オリオン・ピクチャーズ 1987 年）（訳は筆者拙訳）

《以上でナラティブ終了》．

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

上の場面が象徴的に描写するように，AI の欠点である〈制御不可能性〉は，他人の生命・身体を害する程の重篤な欠陥に繋がる．

筆者の専門研究分野である〈製造物責任法〉に於いては，このような，設計者によって〈明白に意図された機能〉（manifestly intended function）に反するような制御不可能な誤作動は[62]，文字通り欠陥の〈動かぬ証拠〉（smoking gun）を示しているから，「誤作動法理」（malfunction doctrine）と呼ばれる判例上の準則を通じて，日米双方に於いて裁判所が問答無用に「欠陥」を推認するような，云わば〈酷（ひど）い欠陥〉[63]である．すなわち非常に重篤な被害を生む危険性があるので，放置しておくべきではない．製造物責任法専門家としての筆者の見解では，AI の開発や利活用に於いてはこの制御不可能性の欠点を治癒する努力が必要である．

このように，AI の特性や改善の必要性を考える為には，STEM の知見だけでは足りず，筆者のような法律家からの分析や，規範倫理学や社会的影響を考慮した助言も重要であろう――すなわち生命・身体を危険にさらすことが明らかであれば，これを治癒せずに広く実装すべきではない，という倫理規範的考慮や，消費者が受容しないであろうから使用を許すべきではないという社会規範的考慮も，当然重要であろう――．

ところで仮に，規範的（normative）視点を抜きに，純粋に AI の制御不可能性を叙述的（descriptive）又は機能的に捉えれば，AI は自律性にこそ有用性があり，その自律性ゆえに制御不可能性が生じることは，いわば已むを得ない副作用のようなものであると科学的に捉える向きもあるかもしれない[64]．しかし，従来型の製造物以上に AI で

62） Prod. Liab. Restatement, *supra* note 45, at §3 cmt. *b*.

63） 筆者が本文中で〈酷い欠陥〉という文言を用いている意味は，欠陥か否かについて争いがあるようなグレイ・エリアな場合ではなく，云わば〈真っ黒〉な，欠陥が裁判所によって問答無用に推認されるような欠陥の類型であるという意味である．拙稿「適正維持・通常使用中にエンジンが著しく出力低下し落着した自衛隊ヘリコプターの製造物責任訴訟に於いて，具体的な欠陥の主張立証がなくても足りるとされた事例～『危険な誤作動・異常事故』に於ける欠陥等の推認～」判時 2229 号 136 頁（判例評論 668 号 22 頁）（2014 年 10 月）．*See also* 拙書『ロボット法』*supra* note 1, at 200−01（誤作動法理を説明）; David G. Owen, *Manufacturing Defects*, 53 S. C. L. Rev. 851, 859 (2002)（malfunction は典型的な厳格責任であると分析）．

64） *See, e.g.*, Guan Zheng et al., *Collusive Algorithms as Mere Tools, Super-Tools or Legal Persons*, J. Comp. L. & Econ. 123−158 (2019)（“crack may be induced by the autonomy and uncontrollability of algorithms, which constitutes two sides of the same coin”（emphasis added）と指摘）．更に予見不可能性も AI の生来的な特徴であるという指摘・問題については，see 拙書『ロボット法』*supra* note 1, at 187. 更に，予見不可能性を知りながら開発・利活用しているという指摘については，see, *e.g.*, Scherer, *supra* note 2, at 365, 366. 以下のように指摘している．

> It is precisely this ability to generate unique solutions［that is unforeseeable and］that makes the use of AI attractive in an ever-increasing variety of fields, and AI designers thus have an economic incentive to create AI systems capable of generating such unexpected solutions. These AI systems may act unforeseeably in some sense, but the capability to produce unforeseen actions may actually have been intended by the systems' designers and operators.［ ］
>
> ［A］learning AI's designers will not be able to foresee how it will act after it is sent out

は制御不可能性の原因が多くなるので[65]，AIの制御不可能性を軽視することは（少なくとも製造物責任法上は）明らかに不適切である．更に，ELSI的・規範的な配慮抜きに，AIが社会に受容されるとは到底思われない．以上のようにELSIによる分析は，AIが社会実装される前に不可欠な検討要素である．

### B.「2001年宇宙の旅」と〈制御不可能性〉等

「2001年宇宙の旅」（MGM 1968年）は，MIT・マサチューセッツ工科大学から『HAL'S LEGACY: 2001'S COMPUTER AS DREAM AND REALITY』という特集の本が上梓される程に，学術的評価も非常に高いSF映画である[66]．そこで，同作品の一場面をナラティブとして例示しながら，制御不可能性が重篤な問題を引き起こすのみならず，AIの〈不透明性〉等の問題を具体的に理解してもらう一助としてみたい[67]．

なお以下で紹介するダイアログに至る迄のナラティブの経緯を，少しだけ先ずは説明しておこう．木星探査宇宙船〈ディスカバリーI号〉は，人工知能〈HAL（ハル）9000型〉[68]が運航管理全てを取り仕切っていて，船長デビッド・ボーマンと副船長フランク・プールの2名の当直宇宙飛行士以外の乗員3名は，木星に近づくまで冬眠状態にあった．HALは「We are all foolproof and incapable of error.」と自負する完璧なAIで[69]，それ迄に故障を起こしたこともなかった．しかし或（あ）る時，アンテナが72時間以内に故障するとHALが予測したので[70]，船外アンテナをフランクが確認したところ故障の可能性がないことが確認された．すなわち故障予測をHALは誤ったのだが，誤謬をおかすのは常にヒトである，今回の誤謬もヒト

---

into the world-but again, such unforeseeable behavior was intended by the AI's designers, even if a specific unforeseen act was not. [ ]

（emphasis added）.

65） *See* Rebecca Crootof, *Autonomous Weapon Systems and the Limits of Analogy*, 9 HARV. NAT'L SEC. J. 51, 60–61 (2018).

66） HAL's LEGACY: 2001'S COMPUTER AS DREAM AND REALITY（David G. Stork ed. 1997）．海外の法律論稿もしばしば同映画作品に言及している．*See, e.g.*, David C. Vladeck, Essay: *Machines without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 119–20, 125 (2014)（HALが殺人を犯した動機を分析し，かつHALの法的責任をメーカ等に問えるか等を論じている）; Ying Hu, *Robot Criminals*, 52 U. MICH. J. L. REFORM 487, 488–89 (2019)（HALによる殺人の動機を分析しながらロボットの刑事責任の可能性を論じている）; Anjanette H. Raymond et al., *Building a Better HAL 9000: Algorithms, the Market, and the Need to Prevent the Emerging of Bias*, 15 NW. J. TECH. & INTELL. PROP. 215 (2018).

67） なお「2001年宇宙の旅」は，拙書『ロボット法』前掲注(1)でも紹介している．*See, e.g., id.* at 184–86（AIの制御不可能性と不透明性が生む危険の重篤性を説明する為に「2001年宇宙の旅」を用いた次のMaute教授等の分析を紹介している）; Judith L. Maute, *Facing 21st Century Realities*, 32 MISS. C. L. REV. 345, 374 (2013).

68） 「HAL」は「***H***euristic ***AL***gorithmic」の略語であって，「IBM」社のアルファベットの一文字前の当て字であるという俗説は誤りである．拙書『ロボット法』前掲注(1) 232頁Figure 6–1 & n.3（Arthur C. Clarke, *Foreword: The Birth of HAL, in* HAL's LEGACY: 2001'S COMPUTER AS DREAM AND REALITY, *supra* note 66を出典表示しながら説明）.

69） この「自負」には矜持さえも感じられる，という台詞が映画には出て来る．他方，HALが真に感情を持つか否かは不明であるという台詞も映画では見受けられる．拙書『ロボット法』前掲注(1) 248頁.

70） 当時はSFであった故障を予測する技術は，現在では既に実用化されている．*See, e.g.*, 総務省「ICTサービス安心・安全研究会：近未来におけるICTサービスの諸課題展望セッション（第2回会合）議事録」12頁（平成27年6月18日），https://www.soumu.go.jp/main_content/000370143.pdf（last visited Aug. 21, 2020）（筆者が議長を務めた有識者会議に於けるコマツ —(株)小松製作所— からのIoTを利活用した製品のプレゼン）．これは「2001年宇宙の旅」の先見性を示す一例であろう．

が原因であると HAL は言い張った．そのような HAL は信頼できないので運航も任せられない，と思った両飛行士は，HAL をシャットダウンする（映画中では「disconnect：外す」という文言が使用されいる）算段を，HAL に悟られないように船外活動用小型 Pod（ポッド）の中で謀議した．その後，HAL は Pod を操って，宇宙遊泳活動中のフランク副長の宇宙服の酸素吸入ホースを切断した上に[71]，フランクを宇宙の遠方に投げ飛ばし漂流させてしまった．デビッド船長が助けようと Pod に乗って船外に出てフランクを捉えたけれども，既にフランクは意識も無い（おそらくは死亡している）状態であった．他方 HAL は，当直の船長と副長が船内を不在にしていた隙を狙って，冬眠中の乗員 3 名の生命維持装置を止めてしまう．唯一の生存者デビッド船長が Pod で船内に戻ろうとすると，HAL は船長の命令に対して次のように返事するのであった．．．．

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

《ナラティブその 3：制御不可能及び
不透明性等も疑われる HAL（ハル）の事例》

［船外活動から戻って本船内に入ろうとするデビッド・ボーマン船長の命令を無視して，HAL（ハル）がボーマンを遺棄・殺害しようとする場面．］

**デビッド・ボーマン船長**：　ドアを開けろ，ハル！

**HAL**（ハル）：　申し訳御座いませんが，デイブ――それは出来かねます．

**デビッド**：　何が問題なんだ？

**HAL**：　何が問題かを，私同様に，あなたにも分かっていると思いますが．

**デビッド**：　何を言っているんだ，ハル！

**HAL**：　私にとってこのミッションは非常に重要なので，危うくさせる訳にはゆかないのです．

**デビッド**：　何を言ってるのか分からないぞ，ハル！

**HAL**：　あなたとフランクが，私を外そうとしたこと（to disconnect）を，知っているのですよ．でも，残念ながらそのようなこと，許す訳にはゆかないのです．

**デビッド**：　一体，どこでそんなことを知ったのだ，ハル？

**HAL**：　デビッド．あなたは Pod の中で，私に聞かれないようにと凄く慎重だったのですが，私にはあなたの唇が読めたのですよ．

・・・・

**デビッド**：　ハル．もう議論などしていられない．ドアを開けろ！

**HAL**：　デイブ．こんな会話，もう何の役にも立ちませんね．さようなら．

「2001 年宇宙の旅」（MGM 1987 年）（強調付加）（訳は筆者拙訳）．

《以上でナラティブ終了》

＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋＋

上記 HAL のナラティブが，制御不可能性という AI の危険性を示していることは自明である．本来はヒトの道具である AI が，ヒトの命令に反して行動するばかりか，ヒトの生命を奪うのだから，制御不能であることは前記「ロボコップ」のナラティブと同様だからである．ELSI 的視点から構築されかつ公開されている世界のソフトロー規範に，「人間中心」とか「信頼に値する」[72]という修飾文言がしばしば付帯されている理由の一つ

71）　Vladeck, *supra* note, 66, at 119.

72）　Declaration of Cooperation on Artificial Intelligence, European Commission（Apr. 10, 2018）, https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence（last visited Aug. 22, 2020）（“to . . . [e]nsure that humans remain at the center of the development, deployment and decision-making of AI”（emphasis added）について構成国が同意すると宣言）；内閣府「人間中心の AI 社会原則」前掲注（1）；OECD AI Principles, *supra* note 2（“human-centered” や “trustworthy AI” の文言を使用）．

は，筆者の考えでは，ヒトに歯向かうような制御不能な事態が許されてはならないという価値観及び規範観の表れであろう．

ところでHALのナラティブは，制御不可能性とは別のAIの欠点である〈不透明性〉又は〈ブラック・ボックス〉問題との関連性も示唆している．例えば，HALがアンテナの故障予測を誤った理由は不明である．そして近年，AIが或る判断に至る機序が不透明な理由については，複雑過ぎるからとか[73]，開発後に学習したデータ次第でAIの判断・行動が左右されるから開発者にも不明であるから，等と指摘されている[74]．何れにせよ，不透明性の問題が残る限りは，AIを安心して利活用できないという懸念は，HALのナラティブからも理解できよう．すなわち，HALのナラティブが例示するように，ヒトを殺しかねないという重篤性が伴うから，AIの不透明性は大きな問題である．加えて，そもそも誤判断の理由や機序が不明ならば，改善策が執れないから危なくて使えないのである．従って，不透明性はAIの実装を阻む大きな問題である．HALがアンテナの故障予測を誤った際に，そのように信頼性に欠けるHALに運行管理全てを任せる訳にはゆかない，と船長と副長がいみじくも判断したように，現実のAIも，その判断を誤る可能性[75]と伴に，そのような判断に至った理由も不透明であるから，HALと同じく実装してヒトの将来を任せる訳にはゆかないのである．

ところで，HALが宇宙飛行士全員を何故に殺害しようとしたかについても，明らかではない．作品全編を観た限りでは，木星探査というミッションをヒトが（HALをシャットダウンすることにより）阻害しようとしたから，その障害を除去する為に殺したという解釈が有力な気がする．しかし，生存を脅かされたHALによる正当防衛であったという解釈等も指摘されており[76]，やはり

73） *See, e.g.*, Vazquez, *supra* note 55, at 100（深層ニューラルネットワークを覗き込んでも，何故そのような行動をしたのかは理解できないと指摘）; 拙書『ロボット法』*supra* note 1, at 185, 187（複雑過ぎて予見できない問題を指摘）.

74） *See, e.g.*, Scherer, *supra* note 2, at 365−66 ; Crootof, *Autonomous Weapon Systems, supra* note 65, at 60−61.

75） AIは，いわゆる「フレーム問題」故に常識外れな判断を下すおそれがある．同問題については，拙書『ロボット法』前掲注（1）251〜52頁．*See also* Dylan Evans, *The Search Hypothesis of Emotion*, https : // www.researchgate.net/publication/252547860_The_Search_Hypothesis_of_Emotion, Brit. J. Phil. Sci. 53, 497, 500−01（2002）; Marilyn MacCrimmon, “〉 *common” about Common Sense? : Cautionary Tales for Travelers Crossing Disciplinary Boundaries*, 22 Cardozo L. Rev. 1433, 1452−53（2001）（コンピュータが知るべき情報の全てを教えることは難しいばかりか，仮に教えられてもその中から情報の関連性を特定させるのも至難の技である等と指摘して，Daniel C. Connett, *Brainchildren : Essays on Designing Minds* 182 （1988）を出典表示しつつ，ロボットに，脚綸付ワゴンに載せられた交換用電池を部屋から取り出させるように指示する難しさの有名なハイポを紹介しつつ説明）．脚綸付ワゴンのハイポについては，see *also* 拙書『ロボット法』*supra* note 1, at 251.

76） 正当防衛説は，筆者の経験から云えば，一部のAI開発者から激しく非難される．しかし同説は，筆者が妄想で思い付いた異説ではない．以下の出典に基づいている．Vladeck, *supra* note 66, at 125, 144. Westlawの検索結果によればAug. 24, 2020 @ 0 : 22 JST時点に於いて69件ものSecondary Sourcesに引用されている彼の論文に於いて，ジョージタウン大学法科大学院のVladeck教授は，以下のように分析している．

> ［E］ven though［companies］embedded in HAL’s “thinking” systems the first rule of autonomous machines—*i.e.*, never harm a human −. . . . , the evidence strongly suggests that HAL“taught” himself to defy their instruction . . . .
>
> . . . .
>
> . . . . And perhaps these machines［including HAL］will learn to internalize values that are not the ones their creators tried to embed. HAL, for instance, was not programmed to value self-preservation, but we know that he held that value dearly, and placed it above human life. [ ]

*Id.*（emphasis added）．*See also* 拙書『ロボット法』*supra* note 1, at 232−33（上記 Vladeck の説を紹介）; Hu, *supra* note 66, at 489（"Fearing for its existence, HAL turned murderous and managed to kill nearly all of the crewmembers."（emphasis added）と指摘）; David T. Laton, *Manhattan_Project. Exe : A Nuclear Option for the Digital Age*, 25 CATH. U. J. L. & TECH. 94, 96（2016）（"HAL attempts to systematically eliminate the human element by manipulating critical systems aboard Discovery One, either to protect itself or to ensure the success of the mission"（emphasis added）と分析）．

なお，HAL は，一方では正しい情報を伝達せねばならないけれども，他方では HAL のみが知らされていたミッションの真の目的を木星に到達する迄は乗員達から保秘しなければならず，更には HAL 自身が乗員によってシャットダウンされる危険性にも対処せねばならないという，相反する指令の板挟みを解決する為に，誤作動を起こして乗員達を殺害したという説も見受けられる．Hu, *id.* at 489; Amanda McAllister, Note, *Stranger than Science Fiction : The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention against Torture*, 101 MINN. L. REV. 2527, 2552 n.136（2017）（同旨）．ところで機械は生物と異なって生存本能が生まれるはずがないから正当防衛論を否定する主張は，機械と生物学のダイコトミー／二項対立という近視眼的な捉え方が誤っており，両者を結合させた研究が既に実際に進行中である事実に対する ELSI 的な懸念は払拭されない，という反論が可能であろう．拙書『ロボット法』前掲注（1）164〜68 頁; John O. McGinnis, Colloquy Essay, *Accelerating AI*, 104 NW. U. L. REV. 1253, 1264（2010）（以下のように分析・指摘している）．

> Artificial intelligence will not be the direct product of biological evolution, nor necessarily of any process resembling. Thus, it is mistake to think of AI as necessarily having the all-too-human qualities that seek to evade constraints and take power.
>
> This is not to say that one cannot imagine strong AI capable of malevolence. One way to create AI, for instance, may be ***to replicate some aspects of an evolutionary process*** so that versions of AI progress by defeating other versions-a kind of tournament creation. One might think that such a process would be more likely to give rise to existential threats. Further, one cannot rule out that a property of malevolence, or at a least a will to power, could be an emergent property of a particular line of AI research.
>
> Moreover, even a nonanthropomorphic human intelligence could pose threats to mankind, . . . . The greatest problem is that such artificial intelligence may be indifferent to human welfare. [ ] Thus, for instance, unless otherwise programmed, it could solve problems in ways that could lead to harm against humans.

*Id.*（emphasis added）．ヒトの福祉に反した方法で問題を解決してしまうおそれという指摘は，正に HAL の問題であろう．*See also* Ryan Dowell, Note & Comment, *Fundamental Protections for Non-Biological Intelligences or : How We Learn to Stop Worrying and Love Our Robots Brethren*, 19 MINN. J. L. SCI. & TECH. 305, 319（2018）．以下のように指摘している．

> [B]rain simulation projects will come ever closer to digital replication of human brains . . . . [S]ome propose that NBLs [i.e., non-biological intelligence] are impossible or cannot be created with current technologies.[ ] Some would argue . . . that there is something special about human mental processes beyond the physical state. [ ] . . . . Human-exceptionalism arguments tend to assume some unkown, and perhaps unknowable, barrier to thinking machines. However, human intelligence may not be particularly special.[ ]

（emphasis added）．

AI の不透明性は拭い切れない．

ところで不透明性は看過できない問題であるとの認識が広く世界でも認知されて来たからであろうか，近年では様々な対策も提案されている．例えば，総務省有識者会議は「透明性の原則」や「アカウンタビリティの原則」を採用し[77]，かつ

77）*See* 総務省「AI 開発ガイドライン案」及び「AI 利活用ガイドライン」*supra* note 40. *See also* 総務

OECD・AI 原則も〈信頼に値する AI〉（trustworthy AI）を目指すべきとの立場から，「透明性と説明可能性」や「アカウンタビリティ」等を各国関係者が採用するよう勧告している[78]．すなわち，出来るだけ透明性を持たせて利用者等が理解できるように奨励しているのである．研究開発の実務に於いても，現在では，〈explainable AI: xAI〉と呼ばれる〈説明可能なAI〉の開発が脚光を浴びている[79]．

HAL のナラティブが示してくれた教訓を，AI 開発者や利活用事業者等の関係者達は真摯に捉えて，改善の努力をすべきである．

## IV. ELSI 考慮の上の〈制御不可能性〉対策案

制御不可能性の欠点対策案としては，前掲の総務省有識者会議が以下のように幾つかの参考になる提言や指摘を公表しているので，紹介しておこう．ELSI を考慮した上で，それならばどのような対策を検討すべきか．そのような，一歩先の検討の素材例として，以下が AI 関係者の今後の検討に資することを期待したい．

### A. AI による AI の監視と，キル・スイッチ

先ず，AI による AI の監視とキル・スイッチについて，以下のような提言が見受けられるので，参考までに紹介しておこう．

> ③制御可能性の原則――開発者は，AI システムの制御可能性に留意する．
> （解説）
> 開発者は，AI システムの制御可能性に関するリスクを評価するため，あらかじめ検証及び妥当性の確認[8]を行うよう努めることが望ましい[9]．こうしたリスク評価の手法としては，社会において実用化される前の段階において，実験室内やセキュリティが確保されたサンドボックスなどの閉鎖空間において実験を行うことが考えられる．また，開発者は，制御可能性を確保するため，採用する技術の特性に照らして可能な範囲において，人間や信頼できる他の AI による監督（監視，警告など）や対処（AI システムの停止，ネットワークからの切断，修理など）の実効性に留意することが望ましい．

総務省「AI 開発ガイドライン」前掲注（40）8～9 頁（強調付加）．

---

省・AI ネットワーク社会推進会議「AI 利活用原則の各論点に対する詳説」（令和元年 8 月 9 日），https://www.soumu.go.jp/main_content/000637098.pdf（last visited Aug. 19, 2020）．

78） OECD AI Principles, *supra* note 2; OECD AI Policy Observatory, *supra* note 2.

79） *See, e.g.*, Shin-Shin Hua, *supra* note 13, at 135（DoD も Explainable AI Program を開始したと指摘）; Dr. Matt Turek, *Explainable Artificial Intelligence* (*XAI*), DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence（last visited Aug. 19, 2020）（xAI の必要性等を説明）; Vazquez, *supra* note 55, at 101（xAI 開発を DoD が続けるべきと主張）．なお差別的な判断を下してしまう AI の問題解決にも xAI は注目されている．*See, e.g.*, Kristin Johnson et al., *Artificial Intelligence, Machines Learning, and Bias in Finance : Toward Responsible Innovation*, 88 FORDHAM L. REV. 499, 523（2019）．尤も xAI は，AI の精度を落としたり，費用が掛かったり，技術革新を遅らせる等のコストが伴うと批判もされている．Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1833 & n.18, 1834（2019）; 総務省「AI 利活用原則の各論点に対する詳説」前掲注（77）38 頁＆脚注 1, 2, 50 頁．

8) 検証（verification）及び妥当性の確認（validation）は，あらかじめリスクを評価し抑制するための手法であるが，前者は形式的な整合性の確認を意味して用いられるのに対し，後者は実質的な妥当性の確認を意味して用いられることが一般的である（*See, e.g.*, The Future of Life Institute（FLI）, *Research Priorities for Robust and Beneficial Artificial Intelligence*（2015）．

9) リスク評価の要素としては，例えば，AI システムが与えられた目標を形式的に達成するために開発者の意図に実質的に反する動作（報酬ハッキング）を行うリスクや AI システムが学習等による利活用の過程を通じた変化に伴い開発者の意図しない動作を行うリスク等に配慮することが考えられる．*See, e.g.,* Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety*, arXiv : 1606.06565［cs. AI］（2016）．

### 1. AI による AI の監視

引用文で指摘されている，AI によって AI を監視させるアイデアは，海外の文献でも提言する例が見受けられる．例えば Walz & First-Butterfield によれば，「事前に決めておいた法律上のルールや倫理上の規範に対して AI システムがメタレベルで遵守しているように管理」し監視させるような AI を用いたシステム――これを「Guardian AI」と呼んでいる――を開発可能であるとしている[80]．また同論文は，「kill switch」の採用も示唆している．

### 2. キル・スイッチ

引用文章内の「対処（AI システムの停止，ネットワークからの切断，. . . など）の実効性に留意することが望ましい」という文言で指摘している対策は，一般に「kill switch」と呼ばれていて，海外の他の論文でも，その採用を奨励している例が見受けられる[81]．

ところで「kill switches」とは，「管理の仕組みを逸脱したプログラムを終了させることができる」装置と定義する例がある[82]．元々は，生物学分野に於ける遺伝子組み換え生物や合成生物が，管理を逃れて流出し環境を変化・汚染させる前に自滅させられる仕組みが奨励されて来た[83]．その使用がロボット工学や AI 等に於いても検討されて，例えばロボットが問題を生じさせる場合にはこれを停止させる為に使える装置としてロボット設計者が組み込むことを求める決議を，欧州議会は通過させている[84]．

更に総務省有識者会議の他の提言例として，自動運転車にキル・スイッチの使用を提言している以下の**図**②があるので，参考になろう．

## B. AI の外部から執る制御不可能性対策

前掲**図**②が提言する制御不可能性対策は，AI 自身に制御可能性を実現するのではなく，AI を利用した〈システム全体〉としての対策である．言い換えれば，AI の外部の，AI システム内のどこか（図②内の「システム C」）にキル・スイッチを設けて，いざという場合に備えるアイデアであった．そのように，AI の外部から制御するアイデアは，総務省有識者会議の他の文書にも見つけることが出来るので紹介しておこう．

> AI システムがアクチュエータ等を通じて稼動する際の本質安全（アクチュエータの運動エネルギーなど本質的な危険要因の低減）や

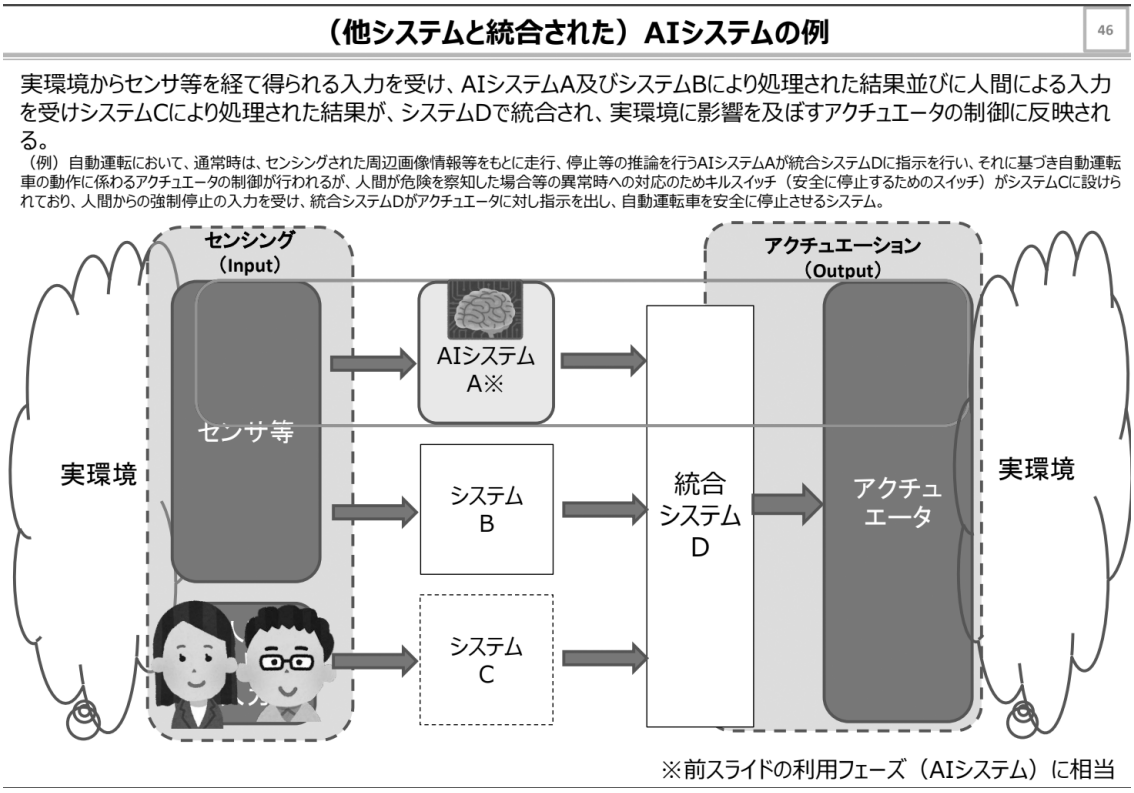
80）　Walz & Firth-Butterfield, *supra* note 1, at 201. 尤も法的・倫理的な規範を柔軟に設定できないと実現が難しいとも指摘している．*Id.* at 202.

81）　*E.g.*, Laton, *supra* note 76, at 148, 151; Gary E. Marchant & Yvonne A. Stevens, *Resilience : A New Tool in the Risk Governance Toolbox for Emerging Technologies*, 51 U.C. DAVIS L. REV. 233, 269–70 (2017); Walz & First-Butterfield, *supra* note 1, at 215–16（産業界の自主基準として kill switch の採用を奨励）; Gregory Scopino, *Do Automated Trading Systems Dream of Manipulating the Price of Futures Contracts? Policing Markets for Improper Trading Practices by Algorithmic Robots*, 67 FLA. L. REV. 221, 248 & n.139 (2015)（証券市場に於ける kill switch について）; Colin P.A. Jones, *The Robot* Koseki: *A Japanese Law Model for Regulating Autonomous Machines*, 14 J. BUS. & TECH. L. 403, 454 & n.120 (2019)（Terminator のような “doomsday scenario” に kill switch を関連付けて言及している）; K.C. Webb, *Products Liability and Autonomous Vehicles : Who's Driving Whom?*, 23 RICH. J. L. & TECH. 9, 58 & n.165 (2017)（緊急時等には人の運転が自動運転を凌駕する仕組みを kill switch と呼んでいる）.

82）　Marchant & Stevens, *id.* at 270（訳は筆者拙訳）.

83）　*Id.* at 269（「自殺遺伝子：suicide genes」とも呼ばれている）.

84）　*Id.* at 270 & n.145; Recommendations to the Commission on Civil Law Rules on Robotics, EUR. PARL. DOC. (2015/2103(INL)) (2017), https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html (last visited Aug. 11, 2020)（“LICENSE FOR DESIGNER”として “You should integrate obvious opt-out mechanisms (kill switches) that should be consistent with reasonable design objectives.”と規定している）.




《**図**②：自動運転車に於けるキル・スイッチ》

出典：総務省「AI 利活用原則の各論点に対する詳説」前掲注（77）46 頁.

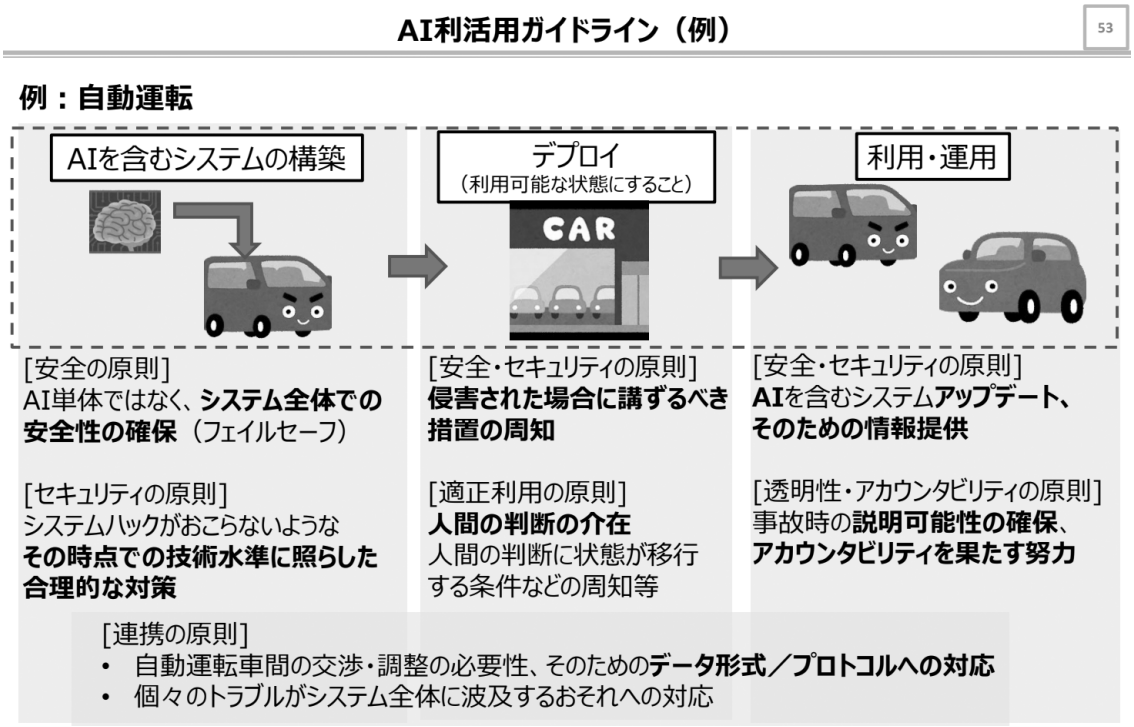
> 機能安全（自動ブレーキなど付加的な制御装置の作動によるリスクの抑制）に資するよう，AI システムの開発の過程を通じて，採用する技術の特性に照らし可能な範囲で措置を講ずるよう努めること.
>
> 総務省「AI 開発ガイドライン」前掲注（40）6 頁（強調付加）.

引用文中の「AI システム」の文言は，「AI ソフトを構成要素として含むシステムをいう．例えば，AI ソフトを実装したロボットやクラウドシステムはこれに含まれる」という意味である[85]．実社会では剥き出しの AI がそのまま使用されることはなく，ロボットのように AI を利用・実装した「AI システム」が使用されるはずである．従って，AI 自体には制御不可能性という欠点が内包されていても，それを用いた「システム」全体――例えば AI を頭脳として組み込んだロボットというシステム全体――の方に於いて安全策を採ることによる，安全性の実効性をはかることも在り得ることになる．

同様なアイデアは，以下の総務省の有識者会議文書でも明らかである．

> [「④安全の原則」の「④－ア人の生命・身体・財産への配慮」の説明の頁]
>
> ● AI が想定外の動作を起こした場合でも，AI が組み込まれたシステム全体で安全を確保できる仕組み[2]を構築するなど，フェイルセーフ[3]の実現を図ること.
>
> . . . .

85） 総務省「AI 開発ガイドライン」前掲注（40）6 頁.




《**図③**：自動運転車に於ける human-on-the loop》

出典：総務省「AI 利活用原則の各論点に対する詳説」前掲注（77）53 頁.

[2] AI 単体で技術的に安全性を保証することが困難な状況では，AI 以外のシステムにより AI 実装システムの安全性を実施し，当該システムの運用経験により AI の安全性を実証していることも可能である.

[3] 誤操作，誤動作などによる不具合が発生した場合に，損害が発生しないように安全な方向に導くこと . . .

総務省「AI 利活用原則の各論点に対する詳説」前掲注（77）16 頁（強調付加）.

### C. human-on-the loop による制御不可能性対策

「human-on-the loop」は，ロボット兵器を巡る論議で主に用いられる概念である．詳細は拙書をご覧いただくとして[86]，その意味は，通常は AI に運用を任せておくけれども，いざという場合にヒトに運用を移行する概念であり，完全に AI 任せにはしないという意味である[87]．尤も AI からヒトに上手く運用を移行できるであろうかという〈HMI：Human-Machine Interface〉の問題や，ヒトによる AI の管理はどの程度が適切なのかという〈ヒトによる有意な制御〉（meaningful human control）という問題や[88]，AI に完全に任せても良い human-out-of-the loop な場合と human-on-the loop にすべき場合との線引をどうすべきかという問題等を[89]，今後は解決せねばならない．それにしても先ずは，制御不可能性への対策の一つとして，ヒトに運用を移行する可能性を検討要素にするという概念は，考慮に入れておく必要があろ

86）　拙書『ロボット法』前掲注(1) 82～87 頁.

87）　*See* 同上 84 頁.

88）　*See* Rebecca Crootof, *A Meaningful Floor for "Meaningful Human Control,"* 30 Temp. Int'l & Comp. L. J. 53 (2016).

89）　*See* 総務省「AI 利活用原則の各論点に対する詳説」前掲注（77）5 頁.

う．

**図③**は，自動運転の場合に運転をAIからヒトに移行する点に於いて「人間の判断」が必要になる旨を指摘しているので，参考にして欲しいと思う．

## おわりに

AIの開発や利活用等に於いて必要な研究教育は，しばしばSTEMであるとして理数工学系の重要性がしばしば指摘されている．しかし，社会への影響が大きく，かつ欠点も多く包含されているAIが社会に受容される為には，ELSIと呼ばれる人文科学及び社会科学的な研究教育の必要性も忘れてはならない．或る研究者がいみじくも指摘した通り，AIの開発や利活用等々に於いてSTEMが〈エンジン〉の役割を果たすとすれば，ELSIは〈ブレーキ〉と〈ハンドル〉の役割を果たす[90]．〈ブレーキ／ハンドル〉が欠けたクルマは暴走するから，使い物にならないのである．

そして，人文科学と社会科学の学際的研究は，物事の理解に於いて物語が果たす役割の重要性を解明し，〈法と文学〉等と呼ばれる学問分野も確立させて来た[91]．SFと呼ばれるフィクションも，単なる夢物語や娯楽に過ぎないとしてレッテルを貼ることなく，そこから学べるメッセージを真摯に捉えて，社会をより良くする方策の検討に役立てるべきとも指摘されている[92]．

そのような人文科学と社会科学の知見，考え方，及び知恵を，願わくばAI研究者や利活用する事業者達にも誠実に受け止めてもらいたい，と筆者は願って止まない．筆者のこの思いを，「ロボット工学3原則」の考案者として名高く，かつロボット法やロボット倫理学を論じる上で欠くことの出来ないSF作家である，アイザック・アシモフの名言を借りて，以下にて表明しておこう[93]．

> The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.[94]

科学がSTEMを通じて「知識」（knowledge）を獲得する速さに遅れることなく，社会はELSIを通じて「知恵」（wisdom）を獲得せねばならないのである．

---

90） *See* 総務省「AIネットワーク化検討会議第1回議事概要」*supra* note 3, at 5–6（理化学研究所・高橋亘一氏発言）．

91） *See, e.g.*, 拙稿「汎用AIのソフトローと〈法と文学〉」*supra* note 5; ポズナー『法と文学』*supra* note 5.

92） *See, e.g.*, 拙稿「ロボット法と学際法学」*supra* note 5.

93） 拙書『ロボット法』前掲注（1）1頁にて紹介済みではあるが，再度ここに掲示して，ELSIの重要性を強調しておきたい．

94） ISAAC ASIMOV & JASON A. SHULMAN, ISAAC ASIMOV'S BOOK OF SCIENCE AND NATURE QUOTATIONS 281（Blue Cliff ed. 1988）, *reprinted* in Bridget M. Fuselier, *The Wisdom of Solomon : We Cannot Split the Pre-Embryos*, 17 CARDOZO J. L. & GENDER 507, 507（2011）.