「倫理的に調和した設計」の論点整理
――異分野・異業種によるワークショップからの示唆――

東京大学政策ビジョン研究センター特任講師
江　間　有　沙
EMA Arisa
科学ライター
長　倉　克　枝
NAGAKURA Katsue

Ⅰ　人工知能と倫理をめぐる議論の背景

人工知能（AI）の社会的，倫理的，法的な影響を考える国内外の議論は，2015年以降，産学官民の垣根をまたいで行われている[1)]。

そのような中，The IEEE Global Initiative（The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems）が公開したIEEE Ethically Aligned Design（「倫理的に調和した設計」）の第一版と第二版（以下EADv1とEADv2）は倫理そのものではなく，設計論や設計思想，それをどのように技術に落とし込めるかといった論点が整理されており，人工知能，法，倫理，哲学など他領域にわたる研究者，政策関係者，企業関係者の意見を集約して作成された。現在，作成に関わっている人は全世界で250人以上に上る。EADは公開性や多様性を求めるため，誰でもが参加でき連携していけるという「プロセス」を重視しており，多様な人々を巻き込み，かつロビーイング活動をしていくことでデファクトスタンダード化を狙っている[2)]。

国際的な影響力も大きな報告書であるが，EADv1もv2も日本語訳は既にIEEE公認版が公開されている[3)]。そこで本稿では，EADv2本体の解説ではなく，EADv2を論点のインデックスとして用いて行われたワークショップにおける論点を概説する。それによって，現在のIEEEに欠けている視点や日本ならではの視点，今後のAIと社会関連の議論に必要な論点を考察したい。

Ⅱ　IEEE「倫理的に調和した設計」

1　IEEE「倫理的に調和した設計」について

2016年12月に公開されたEADv1は，自律的で知的なシステム（Autonomous and Intelligent System: AI/S）に対する恐怖や過度な期待を払拭すること，倫理的に調和や配慮された技術をつくることでイノベーションを促進することを目的としている。

続く2017年12月には第二版であるEADv2が公開され，以下の13項目から構成される[4)]。

1）2018年7月に公開された総務省情報通信政策研究所のAIネットワーク社会推進会議による「報告書2018」第1章などで国内外，国際的な議論の動向がまとめられている。http://www.soumu.go.jp/main_content/000564147.pdf（2018年7月23日閲覧）

2）IEEE Global Initiativeの活動の方法論については，江間有沙，「倫理的に調和した場の設計：責任ある研究・イノベーション実践例として」，人工知能 vol.32, No. 5, 2017, pp.694–700. http://id.nii.ac.jp/1004/00008794/（2018年7月23日 閲覧）を参照。

3）IEEE Ethically Aligned Design version 2, workshop seriesのサイトにて日本語版概要と英語版が入手可能である。https://sites.google.com/view/ethically-aligned-design-ws/eadv2-workshop （2018年7月23日閲覧）

1　倫理的に調和した設計をするための一般原則
2　自律型知的システムに価値観を組み込む
3　倫理的な研究や設計のための方法論やガイド
4　汎用人工知能や人工超知能の安全性や便益
5　個人データとアクセス制御
6　自律型兵器システムの再構築
7　経済／人道的課題
8　法律
9　アフェクティブコンピューティング
10　政策
11　情報通信技術における伝統的倫理観
12　複合現実
13　ウェルビーイング

2　用語集の作成

EADv2 の公開と同時に，「自律的で知的なシステム」に関連する用語集（Glossary）委員会が組織された[5]。分野も業種も違う人々が集まって作成している報告書だからこそ，定義を明確化する必要がある，というニーズから派生した委員会である。委員会では，アカウンタビリティ，データ，人権などの多義的なキーワードに対し，それぞれ⑴一般用語，⑵情報系，⑶工学系，⑷政策や社会科学系，⑸倫理や哲学など，他領域での定義の整理が行われている。

3　標準化プロジェクト

EAD 報告書執筆陣を中心として，標準化を目指すプロジェクトが IEEE-SA（Standard Association：標準規格）に承認されている[6]。2018 年 7 月現在，IEEE-SA の P7000 シリーズとして 14 ワーキング・グループが標準化を目指しドラフトを作成中である。

1．システム設計時に倫理的問題に取り組むモデルプロセス（P7000）
2．自律システムの透明性（P7001）
3．データプライバシーの処理（P7002）
4．アルゴリズム上のバイアスに関する考察（P7003）
5．子供と学生のデータガバナンスに関する標準（P7004）
6．透明性のある雇用者のデータガバナンスに関する標準（P7005）
7．パーソナルデータ人工知能エージェントに関する標準（P7006）
8．ロボット及び自動システムを倫理的に駆動するためのオントロジー標準（P7007）
9．ロボットや知的自動システムを倫理的に「ナッジ（そっと促す）」して駆動するための標準（P7008）
10．自律及び半自律システムのフェイルセーフ設計に関する標準（P7009）
11．倫理的な人工知能と自律システムのウェルビーイング測定基準に関する標準（P7010）
12．ニュースソースの信頼性を特定し評価するプロセス標準（P7011）
13．機械可読のパーソナルプライバシー条項の標準（P7012）
14．自動化された顔分析技術のための包含及び適用基準（P7013）

Ⅲ　イベントの概要

EADv1 を題材としたワークショップを筆者らは 2017 年の 4 月から 5 月にかけて名古屋と京都で 1 回ずつ，東京で 2 日間連続開催し，総勢 44 名の方にご参加いただいた。本ワークショップの反省として，⑴少ない時間ですべての項目を扱うのが難しく表面的な議論になりがちであったこと，⑵日本の具体的な事例などを埋め込むことができなかったこと，が挙げられた[7]。

そのため 2018 年 2―3 月に EADv2 を対象とした第 2 回ワークショップを開催するにあたっては，

4）The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems: Ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems, https://ethicsinaction.ieee.org/（2018 年 7 月 23 日閲覧）

5）A glossary for discussion of ethics of autonomous and intelligent systems, version 1, Glossary Committee, October 2017, https://standards.ieee.org/develop/indconn/ec/eadv2_glossary.pdf（2018 年 7 月 23 日閲覧）

6）個別の SA の概要は IEEE のウェブサイトから確認できる。https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html （2018 年 7 月 23 日閲覧）

EADv2の全13項目を読んだ上で，扱っているテーマが近い項目をオーガナイザ側で6つに再分類し（表1），それぞれの項目について，EADv2を参照して話題提供をいただけるゲスト講師をお招きする形式をとった。

ゲスト講師は，各テーマに対し研究者やジャーナリスト，企業の方で，⑴技術的なアプローチをされている方と，⑵人文社会科学的なアプローチをされている方をそれぞれお招きした。またEADv2を読み込んで解説を行っていただくのではなく，EADv2を1つのインデックスと見なし，それに対して各話題提供者が知りえている情報や研究について具体的な事例を紹介していただくようお願いした。

全6回の順番は，まずは自律型兵器や汎用人工知能（第1回），アフェクティブコンピューティングや複合現実（第2回）など技術的にイメージしやすい項目から開始し，そこから社会や倫理について議論するための方法論（第3回），法（第4回），基盤（第5回），最後に倫理・原則（第6回）といった抽象度の高いテーマへと進めていった。

表1　各回のテーマ，項目とゲスト講師

第1回 リスク	4　汎用人工知能や人工超知能の安全性や便益 6　自律型兵器システムの再構築 嘉幡久敬氏（朝日新聞社科学医療部専門記者） 山川宏氏（ドワンゴ人工知能研究所所長／全脳アーキテクチャ・イニシアティブ代表）
第2回 感情	9　アフェクティブコンピューティング 12　複合現実 木村忠正氏（立教大学社会学部教授） 鳴海拓志氏（東京大学大学院情報理工学系研究科講師
第3回 社会	3　倫理的な研究や設計のための方法論やガイド 7　経済／人道的課題 10　政策 井﨑武士氏（NVIDIAエンタープライズ事業部事業部長／日本ディープラーニング協会理事）
	城山英明氏（東京大学大学院法学政治学研究科教授）
第4回 法	5　個人データとアクセス制御 8　法律 中川裕志氏（東京大学情報基盤センター教授〔当時〕／理化学研究所革新知能統合研究センター） 山本龍彦氏（慶應義塾大学大学院法務研究科教授）
第5回 基盤	11　情報通信技術における伝統的倫理観 13　ウェルビーイング 安藤英由樹氏（大阪大学大学院情報科学研究科准教授） 久木田水生氏（名古屋大学大学院情報科学研究科准教授）
第6回 倫理・原則	1　倫理的に調和したデザインをするための一般原則 2　自律型知的システムに価値観を組み込む 神嶌敏弘氏（産業技術総合研究所人間情報研究部門情報数理研究グループ主任研究員） 神崎宣次氏（南山大学国際教養学部教授）

また，本ワークショップは多様な観点を持つ人に参加してもらうべく，オーガナイザや話題提供者が所属している組織だけではなく，各種関連学会にもパートナーとして後援いただいたほか，メーリングリストでも告知を行った。

各回10名程度の参加者（初回だけ22名）であり，リピーター含め全回を通じて総勢36名の参加があった。

Ⅳ　各回での議論

各回のゲスト講師による話題提供や資料はウェブサイト[8]で公開されているが，本章ではその概要を紹介する。

1　第1回　リスク

⑴　嘉幡氏による話題提供

EADv2「6　自律型兵器システムの再構築（Re-

7）江間有沙，長倉克枝，工藤郁子，「IEEE『倫理的に調和した設計』を用いた議論の場とコミュニティの設計」，2018年度人工知能学会全国大会予稿集，pp. 3H1-OS-25a-05, https://confit.atlas.jp/guide/event-img/jsai2018/3H1-OS-25a-05/public/pdf?type=in（2018年7月23日閲覧）

8）https://sites.google.com/view/ethically-aligned-design-ws/eadv2-workshop （2018年7月23日閲覧）

framing Autonomous Weapons Systems）」に関連して，嘉幡久敬氏（朝日新聞社科学医療部専門記者）は，科学技術取材の中で軍事研究を取材する立場として，最近の自律型兵器システム（Autonomous Weapons Systems: AWS）をめぐる各国の動きを紹介した。

1990年代より米軍主導で兵器の無人化が進んでおり，2000年代には米軍はアフガニスタンなどで無人機による空爆を行っている。一方で，米軍はこうした自律兵器は「Human in the loop」，つまり人間を除外したシステムではないと説明している。

また，現在国際的な議論となっている自律型致死兵器システム（Lethal Autonomous Weapons Systems: LAWS）は，自律性に加えて攻撃性，致死性を持つ兵器システムとされるが，明確な定義はない。自律性を持つ兵器システムとして，イスラエル軍がガザ地区付近に設置しているミサイル防衛システム「Iron Dome」があるが，これもLAWSに含まれるのかはグレーであるとされている[9)]。LAWSに関しては，誰がやったかわからない匿名性と，要素技術は容易にテロリストの手に渡るという拡散性の問題を嘉幡氏は指摘している。

現在，国連において2017年11月から特定通常兵器使用禁止制限条約（CCW）の枠組みで，LAWSについての議論が始まっている。国連での議論の下地となる「適用可能な国際法規則」とは国際人権法と国際人道法であり，それぞれがLAWSにどのように適用されるかが議論の1つの焦点である。

最後に，あまり議論されていない点として「サイバー攻撃」があることを嘉幡氏は指摘した。最近のサイバー攻撃はほぼ自律的に攻撃／防御を行っているとされ，特に社会インフラが攻撃になる場合は「致死的」になりかねないと指摘した。

⑵　山川氏による話題提供

続いてEADv2「4　汎用人工知能や人工超知能の安全性や便益」に関連して，山川宏氏（ドワンゴ人工知能研究所所長／全脳アーキテクチャ・イニシアティブ代表）が，汎用人工知能（Artificial General Intelligence: AGI）研究の取り組みと，EADv2でのAGI記述を紹介した。

AGIは人間のような汎用性をAIに持たせることを目的としている。こうしたAGIの研究をする団体やプロジェクトは世界で45ほどあり，「Safety（安全性）」に関しての研究が進められている。またチェコのスタートアップであるGood AIなどは特定の組織だけがAGIを作って独占し，さらに技術を発展させて人工超知能（Artificial Super Intelligence: ASI）に到達しうることへの懸念から，それを防ぐための方法を考えるワークショップやグランドチャレンジを行っていることを紹介した[10)]。

次に山川氏は「汎用性」と「自律性」概念の整理を行い，汎用性と自律性のループを回していくことが重要であると指摘した。「汎用性」とは様々な知識の組み合わせから未知の状態を推定できることを意味する。「自律性」の定義は議論があるものの，「自動化」「副目標生成」「知識獲得」「生存志向」の4点を挙げた。このうち山川氏はある上位の目標からそれを実現しうる手段として副目標を設定する「副目標生成」が重要であると指摘した。現存する強化学習技術を用いて自動的な副目標設定を実現することは既に可能である。AIが自己改善的に知能を高度化した場合，制御可能になるのではないかという懸念に対し，EADv2では，明確な対策はないとしつつもテスト環境の用意，透明性を高めるセーフ・バイ・デザインなどが提案されていることを紹介した。

EADv2ではAGIに関して「技術」の面から考えるべきことと，「一般原則」に分かれて課題が挙げられている。技術的には前述のような副目標設定におけるセーフ・バイ・デザインが挙げられるが，一般原則においては，予期せぬ事態を考慮

9）人工知能学会2018年年次大会の企画セッション「AIに関わる安全保障技術を巡る世界の潮流」において，外務省の南氏からLAWS規制に関する国連の会議が紹介されたが，2018年4月に行われた最近の特定通常兵器使用禁止制限条約の会議においても，規制対象に攻撃用兵器だけではなく防御用システムも含めるべきかなどは論点としてあがっている。人工知能 Vol. 33, No. 6, In Press.

10）Good AI主催によるAI Race Challengeの結果は2017年7月に公開され，59の応募のうち6つに賞を授与していることがブログ（https://medium.com/goodai-news/solving-the-ai-race-finalists-15-000-of-prizes-5f57d1f6a45f）（2018年7月23日閲覧）に公開されている。

するための審査委員会の設置などが提案されている。

最後に山川氏はAGI研究においては「クローブン（クローズドとオープンを合わせた造語）」が重要であると指摘した。AGIはオープンに開発を進めることが重要であるが，悪用を防ぐためにも「何をどこまでオープンにするか」を考えていくことが重要であると締めくくった。

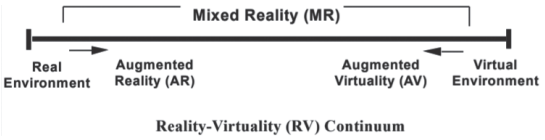
2　第2回　感情

(1)　鳴海氏による話題提供

バーチャルリアリティ（VR），拡張現実感，人間拡張工学を専門に研究をしている鳴海拓志氏（東京大学大学院情報理工学系研究科講師）が，EADv2「9　アフェクティブコンピューティング（Affective computing）」「12　複合現実（Mixed Reality: MR）」は結びついて研究されているとしたうえで，MRの定義を紹介した。

現実環境（リアル）に情報を足してリッチにするのが拡張現実感（Augmented Reality: AR），リアルを計測してVR世界に持っていくことであたかもリアルと同じ空間を再現するものをAugmented Virtuallyとし，その先にあるものがVRというスペクトラムとなる。このうちリアルを除く，「情報が付与されたもの」をまとめてMRと呼ぶため，ほとんどのVRはMRに含まれる（図1）[11]。

図1　リアルとヴァーチャルの連続性



鳴海氏はアフェクティブコンピューティングには，(1)感情を計測してAIなどで分析する，(2)感情を活用して人に働きかけたり人の感情を喚起したりする，の2つがあると説明した。人の感情に働きかけることで無意識のうちに行動を変えさせるのはナッジ（nudge）とも呼ばれる。

人の感情や行動を誘発することは議論が必要であり，鳴海氏はいくつか具体的な研究を紹介した。例えば「Incendiary reflection／煽情的な鏡」では，カメラ付きのディスプレイを鏡に見立て，カメラに映った表情を，笑顔にしたり悲しい顔にしたりと変化させて表示をする。鏡の利用者は，その映った表情の変化によって，自身の感情が変化するという。この研究には心理学的に「悲しいから泣く」のではなく「泣くから悲しい」との知見が反映されている[12]。

また，身体に合わせて態度や思考が変わることは「プロテウス効果」と呼ばれるが，VR空間で白人が黒人のアバターを体験することで，人種差別意識が変化したとする研究もある[13]。

一方で，VRや感情を扱う技術の副次的な効果は長期的に考えていく必要があり，ガイドラインを作るのであればトップダウンではなくユーザを巻き込んで常に更新していくようなものが必要であるとまとめた。

(2)　木村氏による話題提供

続いて木村忠正氏（立教大学社会学部教授）が情動と倫理（モラル）を考えるにあたり，道徳心理学者のジョナサン・ハイトが提唱した道徳基盤理論（Moral Foundations Theory: MFT）を紹介した[14]。MFTでは，倫理的判断が論理に基づくのではなく，感情が論理に先立って大きな役割を担っており，論理はそれを正当化するため使われるという立場を取る。

11）Augmented reality: A class of displays on the reality-virtuality continuum, P Milgram, H Takemura, A Utsumi, F Kishino, Telemanipulator and Telepresence Technologies (1994) – SPIE Vol. 2351, 282–292

12）Shigeo Yoshida, Sho Sakurai, Takuji Narumi, Tomohiro Tanikawa and Michitaka Hirose: Incendiary reflection: evoking emotion through deformed facial feedback, In SIGGRAPH 2013 Emerging Technologies, Jul, 2013. YouTubeの動画はhttps://youtu.be/1ORSFameZZI（2018年7月23日閲覧）から閲覧できる。

またこの表情を変化させる鏡を試着室の鏡に使った実験も紹介された。2種類のマフラーの試着のうち，一方のマフラーを試着した際に，表情を変化させて笑顔にすると，そちらを選ぶ確率が有意に高くなったという。そのほか，ビデオチャットで相手の顔を画像処理で笑顔にさせた状態でブレストをしたところ，アイディアが多く出てきたという。

13）Domna Banakou et al., Virtual Embodiment of White People in a Black Virtual Body Leads to a Sustained Reduction in their Implicit racial bias, Frontiers in Human Neuroscience, vol. 10, Article 601, 2016.

欧米社会では倫理（モラル）は，個人の自律性を尊重する，ケア（危害），権利，正義（公正）といった個人主義的価値に基づくと概念化されてきた。他方，「集団への帰属」「伝統や権威の尊重」「神聖なものの崇拝」といった価値は，集団主義的，全体主義的なものとして倫理（モラル）の議論からは遠ざけられてきた。しかし，文化心理学では，人間社会において，個人の自律性と並んで，集団性，神聖性に基づくモラル（判断基準）も極めて重要な役割を果たしていると議論されており，MFT は，集団主義的価値観もまた倫理（モラル）の基盤として積極的に導入する。

木村氏はハイトが道徳基盤として区別した〈ケア／危害〉〈公正／欺瞞〉〈内集団（忠誠）／背信〉〈権威／転覆〉〈聖／不浄〉〈自由／抑圧〉の 6 つの情動ベクトルと，日本における政治的体動や志向性との相関を調査し，実際の投票行動やネットでの荒し行為や書き込み行為とも関連があることを明らかにした[15]。ネット世論では保守パターンを示す人が発言をすることが多いため，因果応報的な公平さ，頑張った人は報われるというロジックが強くなる。さらにリベラルが書き込まないことでより増幅されるという。

最後に木村氏は心理人類学者のアラン・フィスクの Relational Model Theory（RMT）に基づいた社会的関係性を理解するための 4 つの基本モデル「Communal Sharing（CS）」「Authority Ranking（AR）」「Equality Matching（EM）」「Market Pricing（MP）」を紹介した[16]。CS は皆が平等で皆で分かち合おうというもの，AR は集団の中で絶対的な順位付けをするというもの，EM は「クリスマスでいくらもらったから今度いくらかえさないといけない」といったギブアンドテイク，MP は市場での値付けである。人類の進化から見ると，現在の価値観は MP へ向かっていると木村氏は指摘した。

MFT も RMT も人間の強い情動が倫理的判断の基盤として働いていると考えている。これらの議論が IEEE の議論にどのように組み入れられるかと問題提起をして締めくくった。

3　第 3 回　社会

(1)　井﨑氏による話題提供

企業の立場から井﨑武士氏（NVIDIA エンタープライズ事業部事業部長／日本ディープラーニング協会理事）は，「3　倫理的な研究や設計のための方法論やガイド」について，「自分では倫理的に問題がないと考えているにもかかわらず，自分の気づかぬところで実は問題となるということがあるとすると，それを排除するのは非常に難しい。ただ明示的に間違っていることや，法を犯していることは簡単に排除できる」と指摘した。

井﨑氏は前者の例として，六本木ヒルズの回転ドアに子供がぶつかって死亡した事故を紹介した。回転ドアはもともと人がぶつかっても安全なようにアルミニウムで作られた軽い設計であった。しかし六本木ヒルズはビル風が強いため，利用者の安全のため頑丈なステンレス製に変更され，ドアが重くなった[17]。風で動かない安全なドアが，結果的に危険なドアとなるリスクを，技術者は意識していなかっただろうと指摘する。

一方，後者の例として，明示的な倫理違反は大学や企業での教育により排除が可能であると指摘し，前職である「倫理教育が徹底している企業（most ethical company）」として表彰されたテキサス・インスツルメンツの倫理教育を紹介した。テキサス・インスツルメンツには ethics office があり，報告されたハラスメントや倫理違反などを，事業部門や人事部門と独立した状態で調査を行い，場合によっては懲戒免職などの処分を行う仕組みがあるという。

問題点は必ずしも既定されたものだけとは限らない。ある時点で考えうる範囲内の不確定要素を洗い出し，問題がないかをテストする仕組みが企

14）　ジョナサン・ハイト，『社会はなぜ左と右にわかれるのか――対立を超えるための道徳心理学』，紀伊国屋書店，2014 など。

15）　木村忠正，「『ネット世論』で保守に叩かれる理由――事象的調査データから」，中央公論，132 (1), pp. 134–141, 2017.

16）　Alan Fiske, “The four elementary forms of sociality; framework for a unified theory of social relations,” Psychological review, 99, pp. 689–723, 1992.

17）　失敗知識データベース「六本木回転ドア事故」。http://www.shippai.org/fkd/cf/CZ0200718.html （2018 年 7 月 23 日閲覧）などを参照。

業内部に必要である。しかし一方で，過剰な規制が商品開発や技術の進歩を止めてしまうことにも井﨑氏は懸念を示した。例えば，ディープラーニングで医療画像の診断支援をする技術を開発する際に，個人情報を含む医療画像を施設から持ち出すことが課題となり，技術開発が進まないというケースが現在起きている。様々な課題に配慮しつつも，バランスの取れた倫理の基準を様々なステイクホルダーと共に作っていくべきであると最後に問題提起した。

⑵ 城山氏による話題提供

続いて政治学が専門で政策研究に携わる立場から，城山英明氏（東京大学大学院法学政治学研究科教授）がEADv2は「倫理」を幅広く捉えていると指摘した。例えば，「3　倫理的な研究や設計のための方法論やガイド」には研究機関や企業は倫理的課題を考慮した研究開発や設計をどのように進めていくべきか，が書かれており，「10　政策」ではそのための政策が書かれている。前者は価値に根差したアプローチ（Value based approach）を，後者は権利に根差したアプローチ（Right based approach）を取ると書かれている。そして，後者は権利をベースとして割り切ることが求められるが，前者の実際の設計においては，ある価値を絶対に守れとは必ずしも言っておらず，バリュー・トレード・オフ，つまり特定の価値を守ると他の価値を守れなくなる中での判断が必要となると指摘した。

トレードオフが必要になるとき，どの価値を重要と誰が決めるのかが次に問題となる。EADv2では「チーフ・バリュー・オフィサー（CVO）」という価値の観点から全体を総括する人や倫理委員会の必要性が提案されている。しかし，総括者や委員会の委員に誰がなるのか，実質的に機能する仕掛けをどのように作るのかが今後のステップとして必要であると指摘した。

EADv2では，そのために学際的な議論が強調されるが，分野だけではなく学術と社会の当事者両方を巻き込むトランスディシプリナリティーといった観点が重要であると指摘した。

4　第4回　法

⑴ 中川氏による話題提供

EADv2の「5　個人データとアクセス制御」に関連して中川裕志氏（東京大学情報基盤センター教授〔当時〕／理化学研究所革新知能統合研究センター）が，経済社会において「AIのブラックボックス化」が進んでいることを指摘した。具体例として2010年の株価暴落に言及し，AGIへの懸念よりは，日常的に既に機能しているAIの危険性があり，そのためにも法整備が必要になると指摘した。特にネットワークで接続された複数のAIがそれぞれ振舞うことで制御不能になる危険性を指摘した。

さらに，EADv2にも書かれている「AIの悪用」の問題は大きいと言及し，その発見と対策，救済が必要であるが，そのためにはAIの透明性と説明責任（アカウンタビリティ）が必要になるという。しかし，ブラックボックス化が進んでいる中で説明が困難になりつつあるとし，EADv2には書かれていないが「trust（信頼）」概念を入れるべきであると提案した。

続いて，中川氏は個人情報保護をめぐる世界の潮流として，「忘れられる権利」や「Do not track」を紹介した。「忘れられる権利」に関しては，情報の公共性とプライバシー保護のバランスが重要である。例えば，政治家の汚職と個人の犯罪歴などの扱いなどは，問題ごとに対応が異なり複雑であるが，複雑だからこそAIによるスクリーニングなどが必要となるだろうと指摘した。

「Do not track」では個人がプロファイリングされることを拒否する権利と追跡拒否権がある。しかし，これが実行されると企業のビジネスモデルが成り立たないため実効性がほぼない状態であると指摘した。

そこで欧州では現在，個人のデータは個人が管理し，同意に基づいた利用ができる仕組みに移行していると説明した。2018年5月から施行されているEU一般データ保護規則（GDPR）においてもデータポータビリティが20条に記載されている。また，追跡拒否権については，中国で国民のレーティングとして使われている芝麻信用を紹介した上で，中国のプライバシーの考え方は，EUの個人の人権をベースとした考え方と根本的に違

うことを指摘し，今後の動向について注視していく必要性を指摘した。

最後に，AIを用いたアートや知的財産などの創作物の扱いについて，AIに権利を付与することもEADv2で書かれていることを紹介し，今後の動向について考える必要があると締めくくった。

⑵　山本氏による話題提供

続いて，山本龍彦氏（慶應義塾大学大学院法務研究科教授）がEADv2「8　法律」に関連して，憲法学の視点から解説した。憲法は，「人間がどうあるべきか，社会はどうあるべきか」について一定のコンセンサスを獲得した価値観をfix（固定）している。そのため現在行われている「AI社会はどうあるべきか」の議論も，ゼロベースで議論するのではなく，まずは憲法が掲げる価値観や原理を前提に議論する必要があると説明した。

山本氏は新卒採用にAIを用いたプロファイリングシステムを活用するシナリオを紹介し，この場合「個人の尊重」と「個人の自律」（憲法13条）が危ぶまれるのではないかと指摘した。AIによる採用評価アルゴリズムは共通の属性をもつ「集団」（セグメント）として個人を評価するため，集団的属性から短絡的に個人を評価するのではなく，その個人の具体的事情を斟酌すべきとする「個人の尊重」原理や，身分や集団的属性によって生き方を事前に規定されることなく，個人が自分の人生を自らデザインすべきという「個人の自律」原理に反し，前近代的な社会に舞い戻る可能性があると指摘した。

またAI活用社会では，大量のデータ収集が必要になるが，これは「忘れられる権利」と矛盾する。山本氏は1994年の「ノンフィクション『逆転』判決」を引き合いに出し，「忘れられる」ことによって人生を再構築・再創造できたり，自律的な生き方ができたりする可能性を指摘した。

さらに山本氏は個人の遺伝情報を保険リスクの評価に活用することの問題も指摘した。具体例として，遺伝情報は個人自らが選択・修正できない属性であるから，遺伝情報によって不利益を課すことは「個人の尊重」と矛盾し得るとの考えを，2013年の民法の非嫡出子相続分規定を違憲とした最高裁決定を引用しながら提示した。

3つ目の例として，実際にAIプロファイリングが再犯リスク評価のため刑事裁判に導入されているアメリカの事例を紹介した。このシステムを使うことは2006年ウィスコンシン州の最高裁判決で「合憲」とされた。しかしその条件として，裁判官はAIプロファイリング結果に「依拠」してはいけないとしている。

山本氏は最後にAIによる評価の予測精度を高めるためには個人のプライバシーや個人の尊重など，憲法的価値と矛盾する課題も出てくると指摘する。そうした社会を望むのかどうかを主権者国民の憲法的選択に委ねるためにも，「価値の『天井』としての憲法」の意義や理念について，具体事例と共に考えていく必要があると締めくくった。

5　第5回　基盤

⑴　安藤氏による話題提供

安藤英由樹氏（大阪大学大学院情報科学研究科准教授）は「13　ウェルビーイング」に関連して自身の研究を紹介した。安藤氏がウェルビーイングについて考えるきっかけになったのは「心臓ピクニック」というワークショップであった。マイク内臓の聴診器のようなものを心臓にあてると，手の上で自分の鼓動を感じることができる装置を作ったところ，「生きていることを再確認した」などのコメントがあった。そこから，人の心をポジティブにする情報技術の設計論を考えるに至ったという。

類似の研究を行っている本に『ポジティブ・コンピューティング』[18]がある。この本は「情報技術は世の中を便利にしているが，人を幸せにしているか」を問う。さらに機械にウェルビーイングを実装する方法論を考えることを提案しているが，実際の設計方法については書かれてはいないという。また，安藤氏は欧米の「ポジティブ・コンピューティング」という考え方をそのまま日本に持ち込んでよいか問題提起をした。

ウェルビーイングの議論は国際会議でも行われ

18）「ポジティブ・コンピューティング」に関する本は『ウェルビーイングの設計論――人がよりよく生きるための情報技術』として翻訳が安藤氏の研究メンバーによって出版されている。

ているが，「良い状態」になるように指標など数値的に示して気づきや行動を促すシステムが多いという。しかし答えを教えてくれるのではなく，共感させて腑に落ちる体験をするような仕組みを考えるのが日本的なウェルビーイングではないかと考えている。

また，ウェルビーイングとは多様であるという考えのもと，「自分にとってのウェルビーイングとは何か」，「そのウェルビーイングを最大化する仕組みを考える」ワークショップを開催しており，現在数百人分のデータが集まっているという。

さらに，ウェルビーイングは「満足」の本質であるが，それが砂糖的な満足なのか（摂取した瞬間は嬉しいが，その後減衰する），脂肪的な満足か（摂取したらもっと欲しくなる），うまみ的満足か（摂取すると少しの違いにも気づき敏感にもなる）などの様々な満足の感覚があるという予防医学の石川善樹氏の考えを紹介した。

安藤氏は，現在，ウェルビーイングをめぐっては，心理学的研究，指標などの理論を作る研究，現象を集めていく研究などがあると指摘した。最後に，ウェルビーイングを考えるにあたって，人の無意識に干渉してしまうことの難しさについても問題提起して締めくくった。

⑵　久木田氏による話題提供

続いて久木田水生氏（名古屋大学大学院情報学研究科准教授）が，「11　情報通信技術における伝統的倫理観（Classical Ethics in AI）」に関連して，哲学の専門家の立場から解説した。Ethics（倫理）には，⑴我々が従うべき規範や道義と，⑵学問的探究，倫理学という２つの側面があると紹介し，EADv2 全体としては規範や道義を指しているが，この章では特に学問としての倫理と AI/S の関係を考えていると解説した。

古代ギリシャに起源をもつ西洋倫理学は論理的，論証的，弁証法的な方法や分析的，解釈学的なアプローチなどの「科学的方法」に基づくという特徴を持つ。EADv2 においては，AI/S が人間の自律性にどう影響するのか，義務論的な倫理学理論をどのように機械に実装できるかなどが議論されており，久木田氏はその可能性と有用性に関心があるという。この研究分野としては例えばアメリカで AI に道徳的な判断をさせる研究が 2006 年に発表された[19]。また，2014 年には善悪とその帰結を理解できる自律ロボットを研究するチームも出始めた[20]。

これらの研究に対し久木田氏は，道徳をアルゴリズムに落とし込んでいるだけで「道徳的」だろうかと問題提起した。道徳をアルゴリズムに落とし込むのは昔の記号的 AI パラダイムと類似しており，その理由として西洋の伝統的な倫理学では，論理が倫理において重要であると考えているからではないか，と説明した。

一方近年，心理学・神経生理学などの分野の研究成果からは意思決定において合理性だけではなく感情が重要だと指摘されている[21]。久木田氏は，感情は外からその人の内面と行動を推測する手掛かりであると指摘する。また人間は感情をコントロールできないがゆえに，感情は社会的関係において重要な役割を果たしてきたという。ロボットは感情と行動を切り離すようプログラムもできるため，感情を表現できることによって人を欺くことにも使え，信頼ができないのではないか，と指摘する。

これに対し，道徳的判断や行為に感情は不要とする意見もある。戦争においては怒りや恐怖などにかられることなく行動するロボットのほうが倫理的になるとする考え方もあるという[22]。

最後に，久木田氏は現在の機械倫理の議論は，心理学の知見などを合わせることによって，道徳や倫理をさらに良く理解できるのではないかと提案した。

19）Michael Anderson, Susan Leigh Anderson(eds), Machine Ethics, Cambridge University Press, 2011.

20）http://www.kurzweilai.net/can-robots-be-trusted-to-know-right-from-wrong（2018 年 7 月 23 日閲覧）

21）脳損傷を受けた人が意思決定できなくなった理由を調査すると，推論能力や記憶には問題がなく，感情がなくなっていたという古典的な研究がある（アントニオ・ダマシオ，『無意識の脳・自己意識の脳』，講談社，2003）。また，ハイトの道徳基盤理論（MFT）においても感情の重要性が研究されているほか，ジョシュア・グリーンという哲学者かつ心理学者も道徳的意思決定において論理だけではなく感情も重要であると考えている。

22）例えば Ronald Arkin, Governing Lethal Behavior in Autonomous Robots, Chapman and Hall/CRC, 2009.

6　第 6 回　倫理・原則

⑴　神嶌氏による話題提供

神嶌敏弘氏（産業技術総合研究所主任研究員）は EADv2「2　自律型知的システムに価値観を組み込む」に関連して，自身が行っている「公平性配慮型データマイニング（Fairness-Aware Data Mining）」技術を説明した[23]。機械学習などを用いて採用などの人生に影響の大きな選択を行う時，社会的な公平性，ジェンダーや人種などに影響をされないように決定することが求められている。

例えば，検索キーワードに応じて広告が表示されるが，アフリカ系の人の名前を検索すると「その人が逮捕されましたか」のようなネガティブな印象の広告が表示されていたことがあった。その理由を調査した結果，アルゴリズム自体に作為的な偏りはないことが判明した。広告はクリック率を最大化することが目的であり，アフリカ系の人の検索結果では悪意を持って，「その人が逮捕されましたか」という広告をクリックする人が多く，それを学習してしまったことが原因であったという。

またデータと客観的な根拠に基づいて決定を下すものの例として再犯リスクスコアを予測する「COMPAS」というソフトウェアに関するデータジャーナリズム NPO の ProPublica による分析事例を紹介した。実際には将来の再犯はなかったが，過去に「リスク有り」と判定されてしまっていた割合にアフリカ系とヨーロッパ系の差があるという問題が指摘された。

機械学習はデータから予測をするが，キーワードクリックのようにそのデータ事態に偏りがあってはいけない。また COMPAS の例のように予測結果に偏りがあってもいけない。それを補正するのが公平性配慮型データマイニング技術である，と神嶌氏は説明した。

また，不公平でない決定をするとき「レッドライニング・エフェクト」があると神嶌氏は説明する。アメリカの公民権法で，「借金できるかを人種に基づいてはいけない」と定められた時，銀行は地図上に線を引き，特定の地域の人に貸さない方針を取った。その線で囲まれた地域にはアフリカ系の人たちが多く住んでいたという。ここから代替情報を使って結果として差別が生じる，つまり単純に人種や性別の情報を取り除いても公平にならないということがあると神嶌氏は指摘した。

⑵　神崎氏による話題提供

続いて，神崎宣次氏（南山大学国際教養学部教授）が「1　倫理的に調和した設計をするための一般原則」の解説を行った。まず，3 つ挙げられている一般原則が満たすべき条件の第 1 原則“Embody the highest ideals of human beneficence as a superset of Human Rights”の内実が不明瞭だという指摘がなされた。

用いられている用語の定義が未整備であるために，解釈を確定できないことがその一因と考えられる。例えば，善行（beneficence）は「利己的ではないような行い」，つまり自分や身近な者の利害を特別視しないことという意味で使われているのかもしれない。このような理解の例としては，同じ 10 ドルでも自分の子供におもちゃを買ってあげるより，アフリカで蚊帳などを買ってあげるほうが世界全体としてより良い結果をもたらすなどの「効果的な利他主義」を説くピーター・シンガー[24] のような功利主義的な解釈がある。

仮に beneficence を功利主義的に解釈した時に，義務論的な用語である Human rights とどのように関係づけられるのかの詳細も不明だという。また EADv2 における Human rights の扱いは，人権宣言に代表される様々な国際条約等への参照によっている。しかし，諸権利の対立や侵害が起きた時，それを調停するための国内における裁判所に該当する機関はどこになるのか，という実質的な問題も提起された。

さらに神崎氏は「調停やバランスを考える」という主張自体が，既に偏った利益に基づいている場合があることを指摘し，日本の公害対策基本法をめぐる議論を事例として挙げた。日本では公害対策に対し産業界から反動があり，「公害対策と

23）公平性配慮型データマイニングの資料は神嶌氏のウェブサイト（http://www.kamishima.net/fadm/）（2018 年 7 月 23 日閲覧）で公開されている。

24）ピーター・シンガー，『あなたが世界のためにできるたったひとつのこと――〈効果的な利他主義〉のすすめ』，NHK 出版，2015。

経済発展は両立可能で調和する」という条項が取り入れられた。しかし，公害対策と経済発展のバランスを取るという言い方が既に経済発展に傾いてしまっているという批判もあり，この条項は後に削除された[25]。

同様の議論は個人情報の保護と利活用による利益のバランスと調停にも言える。このバランスが主張される時点で，既にある程度，経済発展に価値の天秤が傾けられているのではないか，と神崎氏は指摘する。

さらに社会にとって必要なのは総体としてのイノベーションであって，特定の技術のイノベーションが何らかの理由で規制や予防されても，イノベーション自体がスローダウンしつつも完全に止まるということはない，とする立場を神崎氏は取るという。その時に，各社会がどのような価値を重視するかを打ち出すことが重要であり，その観点からするとGDPRはある種の見識を示しているだろうと指摘した。

最後に，第1原則で書かれているhighest ideals（至高の理念）を特定するにあたって，みんなの意見を聞くアンケートのようなアプローチは，適切な方法論ではないかもしれない。highest idealsを誰がどのようにして特定するのかも考える必要がある，と問題提起して締めくくった。

V　考　　察

各回で出された論点は，EADv2各章固有の問題もあれば，横断的に議論された課題もあった。本章では，各回の内容も参照しながら，本ワークショップから得られた論点の考察を行う。

1　EADv2に対する日本からの示唆

本ワークショップの目的の1つは，日本ならではの事例や視点を，EADv2を辞書的に使うことによってあぶりだすことであった。

いくつかの回では，論理だけではなく情動をどのように考えるかが議論にあがった。第5回で安藤氏はウェルビーイングを西洋は指標や数値に落とし込む論点が多いと指摘し，また久木田氏も西洋ではロジックが倫理の基盤に置かれると指摘している。これに対し，安藤氏は答えを教えてくれるのではなく，考えさせる仕組みを作ることを，久木田氏はハイトの道徳基盤理論を参照し，感情を機械に落とし込むときに心理学的な知見が使えるかを議論した。

一方で感情に働きかけることで，第2回の鳴海氏や第5回の安藤氏は，技術が「無意識」に干渉してしまうことの怖さについても言及している。

また，第2回の木村氏は，現代の価値観は市場の価値づけに向かっていると指摘したが，これは第6回で神崎氏が指摘した「経済発展と権利などの保護の天秤」を議論すること自体がある程度経済発展に傾いた議論になっていることを自覚すべきである，という論点とも関連しているように思われる。AIの議論は「リスクとベネフィットのトレードオフ」として語られることが多いが，問題設定そのものにも議論の目を向けることが重要だろう。

2　論点整理（What）から方法論（How）へ

EADv2は，全世界的に共有できる原則や論点を13の章に整理している。そしてその論点が特定のステイクホルダーの視点に偏らないためにも，ステイクホルダーの多様性が重要であるとされる。

一方で，どの章においても「どのように」その倫理的な観点を技術や制度に実効性のあるものとして落とし込んでいけばよいのかに関して具体的には議論されていないと指摘された。多くは倫理を「アルゴリズム」に落とし込むことや，指標を作ること，技術を評価する倫理委員会設置などのシステム的な提案にとどまっている。

第4回で山本氏が指摘するように，各国の憲法等では「人間がどうあるべきか，社会はどうあるべきか」についてコンセンサスが得られた価値観が定められているが，EADv2で参照されている国際的に合意ができる権利体系は，国際条約など

25）昭和50年版の「環境白書」第1節 2「公害行政の体系化（昭和40年代前期）」には，公害対策基本法の改正と「経済発展との調和条項」の削除の経緯が書かれている。https://www.env.go.jp/policy/hakusyo/s50/1736.html（2018年7月23日閲覧）

しかない。さらに，それすらも加盟していない国が存在する。第 1 回で嘉幡氏が紹介した LAWS に関する議論も国連で行われているが，コンセンサスが得られるかどうかは 2018 年 7 月現在，予断を許さない。そのような中，価値の対立が起きた時，誰が判断を下し調停するのかといった HOW の議論をしなければならないステージに現在移っていると言えよう。

その中において，AI という茫漠とした概念ではなく，第 4 回で中川氏が紹介したように「データ」に絞ってその保護の在り方を打ち出した GDPR は興味深く，今後の実効力の在り方含め，日本においても注目していく必要がある。

また第 3 回で井﨑氏や城山氏が指摘したように，異分野，異業種の人たちによる議論の場作りや，第 2 回で鳴海氏が指摘したように，トップダウンではなくユーザや関係者を巻き込みながら動的にガイドラインを作っていく方法についても模索する必要があるだろう。

3　学問の探求

今回のワークショップからは，AI という技術に対し，まったくゼロベースで議論が始まるのではなく，各学問分野における研究体系が基盤とされていること，また AI という概念を用いることで各学問分野そのものにも新たな研究テーマを提示している様子がうかがえた。

技術的には，第 1 回の山川氏が指摘するように AGI における「自律性」や「汎用性」をめぐる安全性確保，第 2 回の鳴海氏の VR 研究，第 6 回の神嶌氏の機械学習における公平性配慮型データマイニングなどに関する研究が進められている。ただしその際，例えば「安全」や「公平性」とは何か，あるいはどのような現象がそこで新たに起こりうるかの検証に関しては他分野との接点が重要となることがいずれも指摘された。

AI という分野がそもそも分野や領域横断的である。第三次ブームの現在，その議論は AI 技術だけではなく，他の学問領域においても新たな研究課題を生み出していると言えよう。

Ⅵ　結果と今後の展開について

今後，IEEE は約 1 年間のフィードバック期間を経て，2019 年に最終版となる第三版を公開する予定であり，本ワークショップからの知見もフィードバックする予定である。

2017 年と 2018 年の春に行われた日本でのワークショップシリーズであるが，話題提供者や毎回の参加者のネットワークが形成されていることも 1 つの成果であるといえる。実際，本ワークショップ以外にも様々な団体が AI と社会について考え議論する機会を提供している[26)]。

今後，日本がどのようなガイドやシステム設計，技術設計の方法論を考えていくのか，本稿がそれを考える上での一助となれば幸いである。

※　本ワークショップを開催するにあたって，工藤郁子氏（マカイラ KK）からの助言とサポートをいただいた。また本稿を執筆するにあたっては，ワークショップ話題提供者の方々にご確認とアドバイスをいただいた。

本研究は，国立研究開発法人科学技術振興機構　戦略的創造研究推進事業（社会技術開発）「人と情報のエコシステム（HITE）」研究開発領域による研究成果の一部である。

26)　例えば「AI 社会論研究会」（http://aisocietymeeting.wixsite.com/ethics-of-ai〔2018 年 7 月 23 日閲覧〕）は毎月研究会を開催しており，様々な観点からの議論が行われている。