人々のための AI（AI4People）

良い AI 社会に向けた倫理的枠組み：機会、リスク、原則そして勧告

L. フロリディ他

出典：

Luciano Floridi[1]., Josh Cowls., Monica Beltrametti., Raja Chatila., Patrice Chazerand., Virginia Dignum., Christoph Luetge., Robert Madelin., Ugo Pagallo., Francesca Rossi., Burkhard Schafer., Peggy Vaicke., Effy Vayena., 'AI4people ― An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendation', *Minds and Machines,* 28, 2018, pp. 689-707.

キーワード：

・人工知能（AI）
・人々のための AI（AI4people）
・データガバナンス
・情報倫理学（digital ethics）
・ガバナンス
・AI 倫理学（ethics of AI）

凡例：

原文中の“　”は「　」で、イタリックなどの強調は傍点で、著作名は『』で表記している。また、＜　＞は紹介者による読みやすさのための区切り、（　）は原語表記などのために用いられている。

要旨（全訳）

この報告書は、人々のための AI についての研究成果、つまり「良き AI 社会」のための基盤を提示するために立案された EISMD 発議という、ある画期的研究を報告するものである。我々は社会のための AI が持つ機会やリスクを紹介し、AI を発達させ応用することをより強力に推し勧める 5 つの倫理的原則の全体を提示し、そして、よき AI 社会を評価し発展させ

[1] 本論文は、Luciano Floridi が座長を務める、12 人の専門家から構成された「人々のための AI 科学委員会」との共同で執筆されたものである。本論文の第一著者を務める、オックスフォード大学教授 Luciano Floridi は、情報哲学、技術哲学、情報倫理学における世界的権威である。

奨励し、そしてそれを支えるための、20 箇条の具体的な勧告を提示する。これらの勧告は、ある場合には、国家的あるいは超国家的な政策機関によって採択されるかもしれないし、他方、他の場合には、他のステークホルダーによって先導されるかもしれない。このような勧告は、もし採択されるならば、良き AI 社会を作り上げることに資する堅固な基礎としての役割を果すだろう。

1. 導入

AI とは、我々の生活や相互交流や環境を既に形作っている、強力な力を持った、新たな形のスマート・エージェンシー（a new form of smart agency）である。そしてこの「人々のための AI（AI4People）」とは、この強力な力を社会や人々や環境が持つ善に向けて導くために、作り上げられたものである。

この論文は以下の三つの部分を統合している。①人間の尊厳をはぐくみ人間の開花繁栄（flourishing）を促進するために AI 技術が提示する、機会とそれに伴うリスク（*opportunities and associated risks*）②AI の応用を強力に推し進める原則（*principles*）③そして 20 箇条の勧告（*recommendations*）であり、これらの勧告が採用された場合には、あらゆるステークホルダーは、機会を獲得することが可能になり、リスクを避けあるいは少なくともそれを最小化したりバランスをとったりすることが可能になり、原則を尊重することも可能になり、かくして、良い AI 社会を発展させることができるようになる。

本論文は以下の四つの章に分かれる。第 2 章では、AI によって提出された、人間の尊厳と開花繁栄を促進することに資する中心的機会（core opportunities）と、それに伴うリスクを述べる。第 3 章では、組織が AI の開発と使用に対して倫理的アプローチを取ることの利益を概観する。第 4 章では、既存の分析に基づきながら、社会において AI が倫理的に用いられることを強力に推し進める、AI のための 5 つの倫理的原則を定式化する。最後に、第 5 章では、ヨーロッパにおいて良き AI 社会を開拓するという目的にむけた、20 箇条の勧告を提示する。

2. ＜社会のための AI＞による機会とリスク

AI が社会的に大きな力を持っているということはもはや疑いがない。従って、ここではこの問いを省いて、AI が持つ肯定的あるいは否定的な力は、誰によって、どのように、どこで、そしていつ、感じられるのか、という正当な問いについて考えよう。

以上の問いをより明文化された仕方で実践的に考えるために、ここで我々は、AI が提示する＜社会のための 4 つの中心的な機会＞とは何であると思われるのか、提示することにしよう。これらの機会が 4 つであるというのは、人間の尊厳と開花繁栄を考える際に、これらの機会は、以下の 4 つの基礎的な事柄を取り扱うことになるからだ。つまり、①我々はど

の様な人になりうるのか（自律的な自己実現）②我々は何をなしうるのか（人間の行為者性）③我々は何を達成しうるのか（個人的及び社会的能力）④我々はどのようにして各人及び世界と相互交流することができるのか（社会的結合）という 4 点である。それぞれの場合において、AI は使用される（*used*）ことで、人間の本性を育み、かくして機会を創出することができるが、他方、使用されない（*underused*）ことで、機会コスト（opportunity cost）を作り上げることもできる。つまり、AI の使用とは、この技術を良く開発し、そしてそれを肯定的に応用することなのであるが、恐れや、無知や、誤った懸念や、過度な反応によって、AI 技術の完全な可能性の発揮の手前側で、社会はそれを使用しないということもありうるのである。これは大きな機会コストを生み出すだろう。このように、AI 技術が社会によって十分に実現化されないという危険は、大概、意図しない結果（unintended consequences）に起因するのだし、典型的には、結果として失敗に終わった良き意図（good intentions gone awry）に結びつくのだが、しかしながら我々は、AI 技術の、不注意な過度の使用（inadvertent *overuse*）あるいは故意の誤用（wilful *misuse*）に結びついたリスクをも考えなければならない。E メールからサイバー戦争に至るまでの全ては、技術の悪意ある使用によって、加速化されあるいは過激化されうるのである。さらに、過度の使用あるいは誤用を恐れるがゆえに、AI が使用されないかもしれない、という大きなリスクも考慮に入れなければならない。我々は、以上のようなリスクを下記の図 1 のようにまとめ、以下の節でより詳細な説明を与えることにしよう。

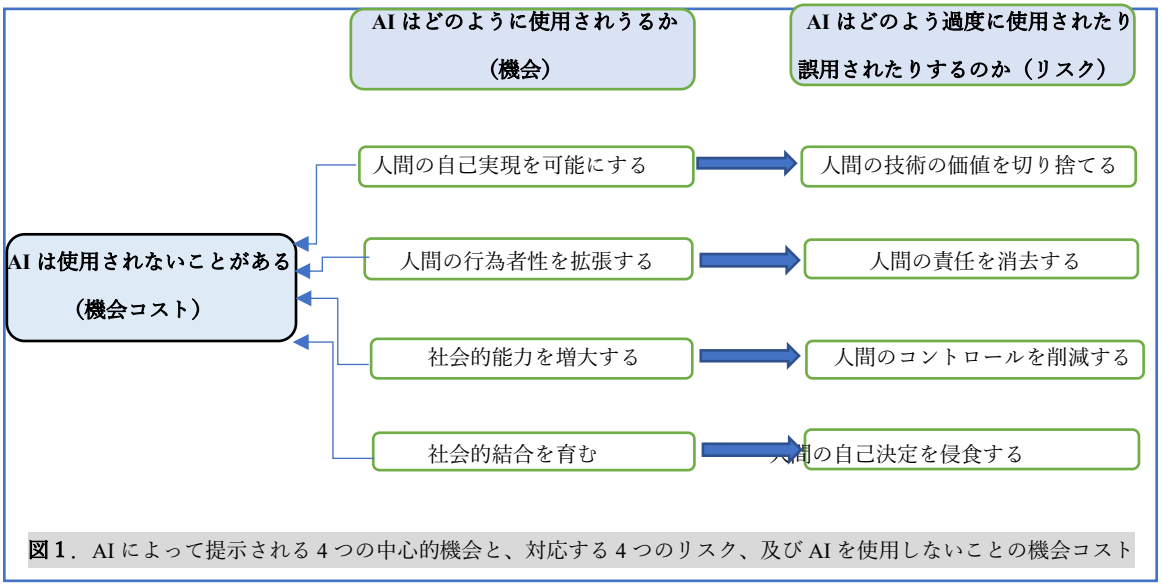

図 1．AI によって提示される 4 つの中心的機会と、対応する 4 つのリスク、及び AI を使用しないことの機会コスト

2.1. 我々はどの様な人になりうるのか：人間の活動の価値を切り捨てることなく、人間の自己実現を可能にする

AI は自己実現（self-realisation）を可能にするかもしれない。この自己実現という言葉に

よって意味されていることとは、性格特性や関心、可能的能力あるいは技巧、切望や人生設計にという点において、人々が開花繁栄する能力である。食洗器などの多くの技術革新が人間を家事労働から解放してきたように、人生における他の日常を「コンピューターによって（smart）」自動化することによって、より多くの時間が解放され、文化的、知的そして社会的な目的に差し向けられるかもしれない。より多くの AI によって、より知的な事柄に費やされた、より人間らしい人生が簡単にもたらされるかもしれないのだ。この場合のリスクとは、一部の古い技巧が廃れてしまうということや新たな技巧の出現そのものではなく、このような技術革新が起きる速度や、その結果引き起こされるコストとベネフィットの不平等な分配である。古い技巧の価値がとても速く失われていくことで、労働市場は短期的に混乱するだろうし、雇用は、個人的なレベルでも社会的なレベルでも、本質的に変化することになる。個人のレベルで言うと、職業はしばしば、個人的アイデンティティー、自尊心、社会的役割といった要素に緊密に結びついている。さらにその上、社会のレベルで言うと、健康診断や航空産業といった繊細で高い技術を要求する領域において、技術革新による単純作業化がおこることによって、AI が機能不全に陥ったり敵意のある攻撃を受けたりする際に問題となる、危険な脆弱性（vulnerability）が生み出されるかもしれない。AI が与える古い能力や技巧への影響を予測し緩和しながら、新たな能力や技巧を支持し、AI の発展を促進するということは、ある種の「普遍的ベーシック・インカム（universal basic income）」の提案において見られるように、綿密な研究と潜在的に急進的な概念の両者を要求するのである。つまるところ、我々は、現在と未来の間にある分裂が、誰にとってもできる限り公平なものとなることを保証するために、＜不利益を被る現在の人＞と＜利益を得る未来の人＞の間にある世代を超えた連帯（solidarity）を必要としているのである。

2.2. 我々は何をなしうるのか：人間の責任を消去することなく、人間の行為者性を拡張する

AI は、「スマート・エージェンシー（smart agency）」をますます蓄えていく、貯蔵庫（reservoir）の役割を提供する。この資源は人間の行為者性を大きく拡張することができる。つまり、AI によって提供された支援のおかげで、我々はより多くのことをより良い仕方でより早くなすことができ、AI によってもたらされた機会やこのような貯蔵庫に由来する利益を享受する人の数が増えれば増えるほど、我々の社会はより良いものとなるだろう。従って、どのような種類の AI を我々は発達させるのか、どのように我々は AI を使用するのか、そして、我々は AI に由来する利点と利益を全員と共有することができるのかという観点から、責任について問うということが不可欠となる。つまり、この人間の行為者性の拡張に対応するリスクとして、責任が不在のものになってしまうのではないか、という問題を明らかに考えなければならないのである。この責任の不在という問いは、我々が誤った社会的・政治的枠組みを持っているという場合だけでなく、意志決定（decision-making）を行う AI システムが我々の理解やコントロールを超えているという場合にも問題となる。そして、このような責

任に対する懸念は、自動運転車による死亡事故といった、目立った事例に対してだけでなく、誓言（parole）や信頼性（creditworthiness）の自動的な決定といった、よりありふれていながらも依然として重要な AI の使用に対しても、また抱かれるのである。

しかしながら、人々が享受する行為者性の程度とその質のあいだにある関係や、我々はどれほどの行為者性を自律的システム（autonomous system）に委譲するのかという問題は、実用的に言っても倫理的に言っても、全か無かという問題なのではない。実際、AI が思慮深く発達させられるのならば、AI は、人間の行為者性に対する可能性を改善させたり増殖させたりする機会を提供する。AI システムに委譲された一連の機能の中に、道徳的に良い帰結を生みだす見込みが高くなるようにデザインされた、「促進的枠組み」を埋め込むことによって、人間の行為者性は、完全に支えられ、洗練され、拡張させられることになるかもしれない。AI システムは、効果的にデザインされた場合には、共有された道徳システムを豊かにし強固にするかもしれないのだ。

2.3. 我々は何を達成することができるのか：人間のコントロールを削減することなく、社会的能力を増大する

AI は、個人や社会一般が持つ能力を、改善したり増大させたりする無数の機会を提示する。病気の予防や治療においても、あるいは、輸送やロジステッィクスの最適化においても、AI 技術の使用は、人間が集団的に可能であることを根本的に増進させることによって、社会の再投資のための無数の可能性を提示するのである。より多くの AI によって、より良い状況、より野心的な目標が支持されるかもしれないのだ。しかし、まさしく、このような技術が、とても力強くまた混乱をもたらす可能性を持っているがゆえに、技術は、これに相当するリスクも合わせ持つのである。課題を AI に委譲することができる場合には、我々は、その課題遂行の過程の一部、あるいは少なくともその過程のコントロールにしか、関わる必要がなくなる。しかし、我々の能力を高めるために誤った仕方で AI 技術を使用する場合には、我々は重要な課題や決定を、少なくともその一部は我々の監視のもとに置かれているべき、自律的システムに委譲してしまうことになるだろう。これは、かえって、これらのシステムのパフォーマンスを監督する我々の能力を削減してしまうことになるだろう。従って、一方で、AI によって提示された、人間の生と人間の達成しうることを改善するという、野心にあふれた機会を追求することと、他方で、我々は AI の主要な発達やその効果をコントロールし続けていると、保証することとの間で、バランスを取ることが必要不可欠なのである。

2.4. 我々はどのように相互交流することができるのか：人間の自己決定を侵食することなく、社会的結合を育む

気候変動といったグローバルな問題は、ますます複雑性の程度を増している。このことは、全てのステークホルダーが、その解決策を共に描き、共有し、協力することによってのみ、

その問題は解決するということを意味しているだろう。AI はデータ集約的（data intensive）であり、解決策はアルゴリズムに基づいて導かれる（algorithmic-driven solutions）ために、AI は、より一層の社会的結合や連携を支えつつ、このような複雑性に対処することの役に立ちうるのだ。例えば、気候変動についていうならば、我々は、気候を直接に変えることと、有害物質の放出を劇的に削減することを促す社会的枠組みをデザインすることとの間で、すぐさま決定を下さなければならない、といった場合がある。そして、この後者の選択肢は、社会的結合を育むアルゴリズムに基づいたシステムによって、より強力なものとなるかもしれないのだ。さらに、このようなシステムは、外側から押し付けらえたものではなく、自分自身に与えた決定（a self-imposed decision）である。この意味で、これは社会的に望ましい仕方で振舞うために「自分自身の肩をおす（self-nudging）」ものであり、最良のナッジの形式であり、自律を保った唯一のナッジであると言うことができる。つまり、これは人間の決断の結果でありながらも、同時に、AI に基づく解決策に基づくものでもありうるのである。他方、リスクとしては、AI システムによって、人間の振舞いの中に計画されていなかった望まざる変化が生じることで、AI システムが人間の自己決定を侵食する、という事態があげられる。AI による絶え間ないナッジという働きは、人間の自己決定に役立ち、社会的結合を促進すべきものなのであり、人間の尊厳や開花繁栄を損なうようなものであってはならないのである。

まとめると、以上の四つの機会とそれに対応するリスクによって、AI が社会や人々に与える影響についての、要素混合的なある見取り図（a mixed picture）が描かれる。つまり、トレードオフの存在を認め、一方で機会を獲得しながら、他方で向かい来るリスクを予期し、避け、最小化することによって、人間の尊厳や開花繁栄を促す AI テクノロジーへのより明るい見通しが得られるのである。AI に対して倫理的に関与することによる、社会全体や個人に対する潜在的な利益を概観したからには、次の章では、このようなアプローチを取ることに由来する、組織にとっての「二重の利益（dual advantage）」について強調をしよう。

3. ＜AI に対する倫理的アプローチ＞による二重の利益

AI がもたらす社会的に望ましい帰結を保証するためには、利益を組み入れることと、AI の潜在的な害悪を緩和することの間にある、緊張関係を解きほぐすことが必要である。このような文脈において、AI 技術に対する倫理的アプローチが持つ価値がより十全に保証されるのである。つまり一方で、倫理は、AI が可能にする社会的価値を組織が利用することを可能にする。これは、社会的に受容可能な、あるいは社会的に望ましい新たな機会を、同定し、活用することによる利益である。他方で、倫理は、組織に犠牲の大きい誤りを予期させ避けさせ、少なくとも最小化させることを可能にする。これは、法的には問題がないのにも関わらず、社会的には受け入れることができず拒絶されることになる、一連の行為を防ぎ緩和することによる利益である。

この倫理による二重の利益は、公共的信頼とより包括的で明晰な責任という状況におい

てのみ、機能を果たす。利益が有意味であるように見えながらも、コストが潜在的で回避可能で最小化可能で、少なくとも、補償や保険と言ったリスクマネジメントを通して防ぐことのできる何かとみなされて初めて、AI 技術の公共的受容や適用が生まれるのである。さらに、このような AI 技術の公共的信頼に基づく受容といった態度が生まれるためには、公衆が AI 技術の発展に関与し、AI 技術の機能の仕方が公開され、そして理解可能で幅広い人々によってアクセスすることのできる規制と矯正のメカニズムが作られることが、必要なのである。こうして、AI に対する倫理的アプローチによって、どのような組織であっても持つことのできる二重の利益という明晰な価値は、関与、公開性、そして異議申し立て可能性（contestability）を、十分に正当化するのである。

4. ＜社会における AI のための原則＞による統合された枠組み

人々のための AI は、AI の倫理的含意を考察する初めての試みなのではない。既に多くの組織が、AI 社会の発展と展開を導く価値あるいは原則についての声明文を発表している。我々はここにおいて、勧告のための倫理的基礎の働きをなす他の一連の原則を作り上げるのではなく、議論を建設的に前に進めて、既存の一連の原則を統合して、原則から政策・最良の実践・具体的勧告へと歩みを進めることにしよう[2]。我々は、以下の 6 つの文書を評価し、その中に見出せる一連の原則の中にある共通性と注目すべき違いに焦点を当てて、20 箇条の勧告を引き出す。

①アシロマ AI 原則（Asilomar AI Principles）。これは、2017 年 1 月におけるアシロマ会議[3]の出席者の協力のもと、生命の未来研究所（Future of Life Institute）の支援によって策定されたものである。（以降、本論文では、「アシロマ」と表記する。）

②責任ある AI に関するモントリオール宣言（The Montreal Declaration for Responsible AI）[4]。これは、2017 年 12 月に開催された＜社会的に責任ある仕方での AI 開発についてのフォーラム＞を受けて、モントリオール大学によって制定されたものである。（以降、本論文

---

[2] 既存の原則の査定に関する選択や方法については、本論文では明示されておらず、これに関する論文が Cowls と Floridi によって現在準備中である旨が記されている。

[3] アシロマ会議とは、「生命の未来研究所」の主催で、2017 年 1 月 5 日から 8 日にかけて開催された、AI のもたらす利益と課題をめぐる学際的な国際会議である。なお、アシロマ AI 原則が、ここに挙げられている文書の中で、最も多くの原則を与えている。詳細に関しては以下のウェブサイトを参照のこと。（https://futureoflife.org/bai-2017/）

[4] この宣言は、モントリオール財団が「ケベック研究基金」の支援を受けて、市民、専門家、政策機関、企業、市民団体、専門家集団と一年以上の協議を重ねて作り上げた、AI の開発に関する一連の倫理的ガイドラインである。詳細については下記のウェブサイトを参照のこと。（https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/）

では「モントリオール」と表記する。）

③『倫理的に整えられたデザイン：自律的で知性的なシステムによって人間の福利に優先順位を付けるための見解（*Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*）』の第二版において提示された一般的原則。これは、自律的で知的なシステムを倫理的に開発したりデザインしたりするために必要な、原則や勧告を発展させるために、250 人のグローバルリーダーが寄稿した、浩瀚な国際的論文集であり、2017 年 12 月に出版された。（以降、本論文では、「IEEE」と表記する。）

④『AI とロボット工学と「自律的」システムについての声明（*Statement on Artificial Intelligence, Robotics and "Autonomous" System*）』において提示された倫理原則。これは、2018 年 3 月に、欧州委員会の「科学と新技術における倫理に関する欧州グループ（European Group on Ethics in Science and New Technologies）[5]」によって、出版された。（以降、本論文では、「EGE」と表記する。）

⑤、2018 年 4 月に出版された、イギリス貴族院 AI 委員会の報告書『イギリスにおける AI：その準備、意志、可能性はあるのか？（*AI in the UK: ready willing and able ?*）』において提示された、「AI コードのための 5 つの包括的原則（five overarching principles for an AI code）[6]」。（以降、本論文では、「AIKU」と表記する。）

⑥＜AI に関するパートナーシップ＞による原則[7]（The Tenets of the Partnership on AI）。＜AI に関するパートナーシップ＞とは、学者、研究者、市民団体、企業といった、多数のステークホルダーによって構成された組織である。（以降、本論文では、「パートナーシップ」と表記する。）

これら全てを合わせると 47 の原則となるが、ここでは、これらの原則を、生命倫理学において一般に用いられる、四原則（善行原則、無危害原則、自律原則、正義原則）と対照させることにしよう。すると、この四原則は、AI によって生じた新たな倫理的課題に驚くほどよく適用されるということが示されるだけでなく、理解可能性（intelligibility）と説明責任（accountability）の両者の複合として理解される新たな原則、つまり説明可能性原則（*explicability*）も追加として必要とされる、ということが明らかになるだろう。

---

[5] この組織は、1991 年に組織された、欧州員会の独立諮問機関であり、EU の法制度や政策に関係して、科学や技術による倫理的問題についての様々な答申を与えている。詳細は下記のウェブサイトを参照のこと。
（https://ec.europa.eu/info/research-and-innovation/strategy/support-policy-making/scientific-support-eu-policies/ege_en）

[6] 当該報告書は下記のウェブサイトにて公開されている。
（https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/）

[7] これは、AI に関するパートナーシップによって提出された 8 つの原則である。下記のウェブサイトを参照のこと。（https://www.partnershiponai.org/tenets/）

4.1. 善行原則（Beneficence）：福利を促進する、尊厳を守る、地球を持続させる

先の文書に見られる AI のための原則を通覧すると、生命倫理学の四原則の中では、この善行原則が、最も容易に見つけ出されることだろう。先の文書の中で、人間に対して善行をなすような AI 技術を作るという原則は、様々な形で表現されているのだが、他方で、これは、諸原則を取りまとめた各々のリストの内部においても、とりわけ最上位に位置づけられている。事実、「モントリオール」と「IEEE」は「福利（well-being）」という言葉を用いており、前者においては、AI の開発は、あらゆる有感な存在者（sentient being）の福利を促進すべきであると定められており、後者においては、人間の福利に優先順位を付ける必要性が述べられている。また、「AIUK」と「アシロマ」は、この原則を「共通善」として特徴づけ、前者は、AI 技術は共通善と人間の利益（beneficence）を促進するべきであると定めている。また、「パートナーシップ」は、AI が出来る限り多くの人に対して利益を与えるものとなることを意図しており、他方で「EGE」は、「人間の尊厳」と「持続可能性（sustainability）」の原則を強調している。この「持続可能性」の原則は、AI 技術が、地球における生命の基本的条件の保証や、人類の継続的な繁栄や、将来世代のための良い環境の保存、といったことと調和したものであるべきだと論じている。そのため、その原則は、恐らく、あらゆる利益についての最も広義の解釈を表明していると言えるだろう。まとめると、善行原則は、確かに、人間と地球の福利を促進する重要性を強調しているのである。

4.2. 無危害原則（Non-maleficence）：プライバシー、安全、そして「能力への警告（Capability Caution）」

この善行原則と無危害原則は論理的に同一のように見えるが、生命倫理学と AI 倫理学の文脈においては異なった原則として考えられなければならない。つまり、先の 6 つの文書における一連の原則は、福利や利益の共有そして共通善の獲得を推し進めているのだが、他方で、それらは、AI 技術を過度に使用したりあるいは誤って使用したりすることから生じる、多くの潜在的に否定的な帰結に対して、警告を与えてもいるのである。とりわけプライバシーの侵害の回避という点は、先の 6 つの文書の中の 5 つの文書の内に現れるものであり、「IEEE」においては「人権」の原則の一部として提示されている。

しかし、プライバシーの侵害だけが避けられるべき危険なのではなく、いくつかの文書は、AI 技術が他の仕方で誤用されることを回避するという重要性も強調している。例えば、「アシロマ」は、未来の AI が持つ能力の上限をめぐって警告を発する必要性を述べているだけでなく、軍事 AI 技術拡大競争（AI arms race）や AI の再帰的自己改善能力（recursive self-improvement of AI）という脅威についても述べている。また、「パートナーシップ」は AI が安全な制約の下で操作される重要性について、「IEEE」は AI の誤用を避ける必要について、「モントリオール」は、「EGE」の影響をうけて、技術革新によって生まれたリスクに対抗する人間の責任について、それぞれ言及している。

AI を発展させるのは人々なのかあるいは技術それ自身なのか、つまり人間と技術のどち

らが害悪をなさないように促されるべきなのかということについて、以上の様々な警告から、完全に明らかであるとは言えないだろう。意図の問題もまた混乱している。無危害原則は、我々が技術の「過度な使用」と呼ぶ偶然的な害悪（accidental harms）を避けることにも、我々が技術の「誤用」と呼ぶ意図的な害悪（deliberate harms）を避けることにも関わっているように見える。いずれにせよ、ここで無危害原則が問題にするのは、人間の意図に基づいたものであれ、機械の予測不可能な振る舞いに基づいたものであれ、生ずる害悪を避けるということなのである。しかし、この行為者性と意図とコントロールをめぐる問題は、次の原則を考慮する際に、いっそう解決しがたいものとなって表れるのだ。

4.3. 自律原則：決定する力（決定するかどうか）

この自律原則も生命倫理学における古典的な原則の一つだが、これは、個人は自身が受ける、あるいは受けない治療に関して、自身で決定する権利を持つ、という考えである。実際の医療において、例えば、患者が決定する心的の応力を失ったりする場合、この原則はもっとも頻繁に傷ついてしまうものであるが、AI を考慮に入れると、状況はより一層複雑なものになる。つまり、AI を受け入れる場合、我々は、意志決定のある部分を厭わず機械に譲り渡してしまう（willingly cede）ということになるのである。かくして、我々自身のために保持する意志決定の力と、AI に委譲するものとの間で、バランスがとられなければならないのだ。

この自律原則は、上記の 6 つの文書中 4 つの文書において明確に述べられている。「モントリオール」は、人間の意志決定と機械に導かれた意志決定の間でバランスを取る必要性を述べているし、「EGE」は、自律的システムは「人間が持つ規則や規範を設定する自由を傷つけてはならない」と論じているし、「AIUK」は、「人間を傷つけ、破壊し、欺くような力は AI に決して与えられてはいけない」と述べているし、「アシロマ」においても、「決断をどのように AI システムに委譲するのか、またそれを AI システムに委譲するのかどうか」といったことを人間が選択すべきであるという限りで、自律原則が掲げられている。

以上の文書は、善行原則と無危害原則の間の区分に関して、多少異なった仕方ではあるが、似通った意見を提示している。つまり、人間の自律が促進されるべきであるだけでなく、機械の自律も制限されるべきあり、航空機のパイロットが自動運転から手動でのコントロールを奪い返す時のように、人間の自律も再構築されるべきであるのだ。まとめるならば、人間の選択の内在的価値を守ることと、機械に対して過剰に委譲してしまうというリスクが当然含まれてしまうということが、ポイントなのである。従って、ここで最も重要になることとは、我々が言うところの「メタ自律（meta-autonomy）」、あるいは「委譲決定（decide-delegate）」モデルである。つまり、人間は常にどの決断を下すべきなのかということを決定する力を保持しているべきなのである。このように、どのような委譲であろうとも、それは、原則的に、人間のコントロールによって覆されうるもの（overridable）であり続けるべきなのだ。

しかしながら、この＜決断するのか委譲するのか決定する能力＞は社会を通して等しく分配されているのではない。生命倫理学によって触発された、四つの内の最後の原則（正義原則）においては、自律が潜在的に不均衡であることから生ずる帰結が問題となる　。

4.4. 正義原則（Justice）: 繁栄を促進し、そして連帯を守る

古典的な生命倫理学の 4 つの原則の最後は正義原則であるが、これは、典型的には、資源の分配に関して問題にされる。この原則も、我々が分析してきた AI のための原則において明確に見て取ることができる。「正義原則」の重要性は、「モントリオール」の中でも明確に述べられているし、「アシロマ」も、AI による「共有された利益」と「共有された繁栄」の両者が必要であるとする。また、「EGE」は、AI は、「グローバル・ジャスティス」と、AI 技術による「利益の平等なアクセスに、貢献するべきである」と述べる。また「AIKU」においては、市民は「AI に伴って、心理的、感情的、経済的に、開花繁栄する」ことが出来るべきであると論じられており、他方、「パートナーシップ」では、「AI の発展によって影響を被る全ての者の利益を尊重する」ことが誓約されている。

他の原則と同様、AI という文脈における倫理的原則としての正義が一体何を意味しているのか、ということについての解釈は、大枠で似通ったものでありながら、わずかながらの違いを含んでもいる。先に挙げた文書において、正義原則は下記のことに関わっている。

a） 不平等な差別といった、不正をただすために AI を使用すること。

b） AI の使用は共有された（少なくとも共有しうる）利益を作り上げると保証すること。

c） 既存の社会構造を衰退させるといった、新たな害悪を作り上げることを妨げること。

また、正義原則に関連して、人々に直面した（*vis-à-vis*）AI の位置づけが、多様であるということには注意が必要である。「モントリオール」と「EGE」では、まさに AI 技術それ自身が、「グローバル・ジャスティスに寄与すべきである」と述べられている一方、「モントリオール」では、「AI の発達」が「正義を促進すべきである」と述べられており、「AIKU」においては、人間は単に AI に「伴って（alongside）」開花繁栄すべきであると述べられているにとどまるのだ。細かな意味合いを細分化していくことをここでは目的としていないのだが、人間と AI の間にある関係が多様であるということは、人間によって作り出された「スマート・エージェンシー」としての AI をめぐって、大きな混乱があることをほのめかしている。

4.5. 説明可能性（Explicability）: 理解可能性と説明責任を通じて他の原則を可能にする

人間と AI の間に横たわる状況は、一様ではない。人間におけるちょっとした変化であったとしても、それは、現在、他の全員の毎日の生活を既に変容させている一連の技術のデザインや発達に関わっているのである。このように、AI の働きは、最も専門的な観察者以外にとっては、しばしば見えなかったり、あるいは、理解されなかったりするのであって、このような現実は、先の文書の著者たちにおいても見失われておらず、AI の意志決定の過程

を理解したり、説明し続けたりする必要性が述べられているのである。つまり、「アシロマ」においては「透明性（transparency）」として、「EGE」においては「説明責任（accountability）」として、「IEEE」においては「透明性」と「説明責任」の両者として、「AIKU」においては「理解可能性（intelligibility）」として、「パートナーシップ」では「理解可能（understandable）」で「解釈可能（interpretable）」なものとして、この原則は表現されているのである。

我々はこのような原則を「説明可能性（explicability）」としてまとめることにするが、これこそが、生命倫理学の枠組みを AI 倫理学に適用した際に導くことのできなかった原則なのである。また、AI が利益をもたらし危害をもたらさないものであるためには（for AI to be beneficent and non-maleficent）、我々は、AI が実際に社会に与える善あるいは悪について、理解する（understand）ことが出来なければならないのであるから、この説明可能性原則は他の四つの原則を補うものでもあるのだ。

まとめると、以上の 5 つの原則は、専門家による先の 6 つの文書に含まれる 47 の原則の中心的な意味をまとめ上げるのであり、以下の勧告が提示される倫理的な枠組みを作り上げるのである。この枠組みは下記の図 2 のようになる。

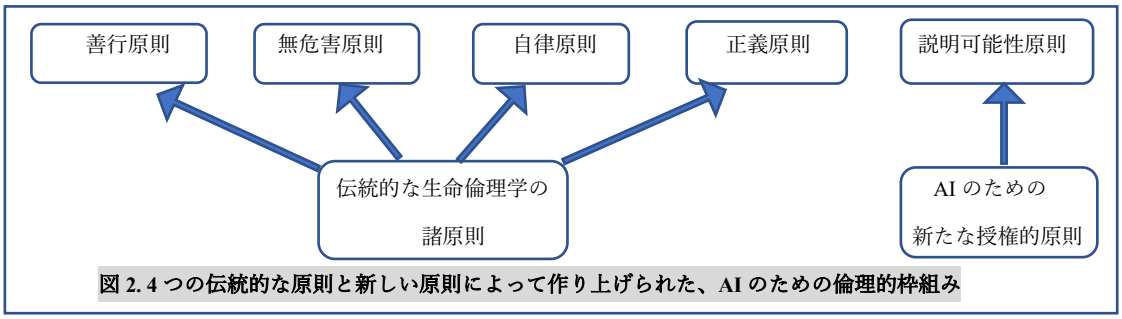

図 2. 4 つの伝統的な原則と新しい原則によって作り上げられた、AI のための倫理的枠組み

5. 良き AI 社会へ向けた勧告

この章は良き AI 社会のための勧告を提示する。以下、序文と 20 箇条の具体的行動（action points）の二つの部分に分かれる。そして、具体的行動も、評価、開発、報奨化、そして支援の 4 種類に分かれる。

5.1 序文

良き AI 社会を作り上げるためには、AI のデフォルトの活動の中に、先の倫理的原則が埋め込まれているべきだ。特に、AI は、不平等を減少させ、人間の自律を社会的に促進し、全ての人に等しく共有される利益を増大させるような仕方で、作り上げられ、発達させられるべきだ。そして、AI が説明可能であるということはとりわけ重要である。説明可能性は、技術に対する公共的信頼や公共的理解を作り上げる、決定的な道具であるからだ。

我々は、また、良き AI 社会を作り上げるためには、マルチ・ステークホルダーによる取り組みが必要であると考える。このようなアプローチは、開発者、使用者、そしてルール制定者が当初から協力することによって、AI が社会の必要に答えることを保証する、最も効

果的な方法なのである。

本稿は、ヨーロッパ的アプローチを提示しているが、我々は、人々からの信頼を守り、公的利益を保持し、共有された社会的責任を強固にする仕方で、AI の開発に関わる。

最後になるが、この一連の勧告は「生きている文書（living document）」として見なされるべきだ。具体的行動は変動的な（dynamic）ものとしてデザインされており、あるひとまとまりの政策を要求しているだけでなく、勧告の効果を保持しようとする一貫し持続した努力をも要求しているのである。

5.2 具体的行動（action points）

5.2.1 評価（assessment）

①民事法廷といった既存の制度が持っている、AI システムによって作られた誤りやそれによって課された害悪をただす能力を、評価すべきである。この評価では、不注意や対立を削減することが目的とされ、前向きな制度設計（design stage onwards）をする責務に対する、持続可能で大多数の人によって支持された基礎があるかどうかが評価されなければならない。（第 5 勧告も参照せよ。）

②社会的価値や公共的意見の理解と連携していることを保証するために、関与するメカニズムの使用を通じて、どの課題やどの意志決定機能が AI システムに対して委譲されるされるべきでないのか、評価すべきである。この評価は、既存の立法を考慮に入れるべきだし、（政府や、産業や、市民社会を含めた）全てのステークホルダーの間の継続した対話によって支えられるべきであり、AI はどのように社会的意見に影響を与えるのか議論することを目的としている。（勧告 17 と共同している。）

③技術的発展と歩調を合わせることができる立法的枠組みを提示するために、現在の規制が十分に倫理的に基礎づけられているかどうか、評価すべきである。このような立法的枠組みは、＜切迫した、そして／あるいは、予期しなかった問題＞へと適用可能な、カギとなる原則による枠組みを含むかもしれない。

5.2.2 開発（development）

④社会的に重要な決定を作り上げる AI システムの説明可能性を増大させる枠組みを、開発するべきである。このような枠組みにとって主要なこととは、とりわけ望ましくない帰結をもたらす場合において、意志決定の過程を、事実に基づき、直接的に、明晰に説明することができる個人の能力である。このことは、恐らく、異なった産業に対して個別的な枠組みが発達することを要求するし、また、科学や、経済学や、法学や、倫理学における専門家と並んで、専門家による学会も、この説明可能性を増大させる枠組みを作り上げる過程に、関与するべきであろう。

⑤法廷においてアルゴリズムに基づいた綿密な調査が可能となるために、適切な法的手続

きを開発させ、司法制度が持つ IT インフラを改良すべきである。このことは、勧告 4 で提示された枠組みのうち、法的システムに個別的に当てはまる説明可能性の枠組みの創造を含むだろう。適切な手続きの過程の中には、IP 訴訟において、センシティヴな商業情報を適切に開示させるといったことなどが含まれるだろう。

⑥不公平な偏見といった望ましくない帰結を同定するために、AI システムの監督メカニズム（auditing mechanisms）を開発するべきであり、また、AI に集中的に関わるセクター（AI-intensive sectors）における厳しいリスクに対処する、連帯メカニズム（solidarity mechanisms）を（例えば、保健部門などと共同して）開発するべきである。このようなリスクは、マルチ・ステークホルダーによる早期の取り組みによって、緩和されるだろう。デジタル化以前の経験が示しているように、場合によっては、社会が技術に追い付くには数十年を要するのであって、ICT に関してなされたように、使用者や政府が早期に関われば関わるほど、この時間的な間隙は短いものとなるだろう。

⑦AI によって引き起こされた不正や不服に対する救済策や補償を与える、矯正プロセスや矯正メカニズムを開発するべきである。AI に対する公共的な信頼をはぐくむためには、社会は、引き起こされた害悪、被ったコスト、あるいは、技術によって引き起こされた他の不服を矯正するための、広くアクセスすることのできる、そして信頼のできるメカニズムを必要とする。このようなメカニズムは、人々や組織に対する、明晰で包括的な説明責任を必然的に持つ。そして、このプロセスの発展は、勧告 1 で素描された既存の能力の評価から生じなければならない。つまり、もしある能力が欠けていることが分かった場合、人々に補償を与えるために、追加的な制度的解決策が、国家的あるいは EU レヴェルで開発しなければならない。このような解決策には下記のようなものが含まれる。

- AI の不公平あるいは不平等な使用を監督する「AI オンブズマン」
- 情報の自由化への要求と同種の、告訴のための（for registering a complaint）規定された手続き。
- 責務保証メカニズム（liability insurance mechanisms）の開発。これは、EU や他の市場で販売される一部のクラスの AI に義務的に伴われるものである。このことは、とりわけロボット工学における、AI を搭載した人工物の相対的な信頼性（reliability）が、保険の値段に反映し、かくして、競合する製品の市場価格に反映することを保証するものである。

⑧新たな組織によっても既存の組織によっても採択される、AI 製品や AI サービスがもつ信頼性を評価する、合意された測定基準を開発するべきである。この測定基準は、あらゆる市場化された AI 製品についての、使用者を起点とした評価基準を作り上げるための基礎として働くだろう。このようにして、AI の信頼性のインデックスは、製品の価格に加えて、発達され示されうるようになるのだ。この AI のインデックスは、より安全でより社会的に利益のある AI についての公共的理解を促進し、その発展をめぐる競争を発生さ

せ、長い目で見ると、適格な製品やサービスを保証する広範囲なシステムのための基礎を形作るだろう。

⑨AI 製品、ソフトウェア、システム、あるいはサービスを科学的に評価し監督することを通じて、公共の福祉を守る責任を負った、新しいＥＵ監督機関を開発するべきである。これは、例えば、欧州医薬品庁（European Medicines Agency）のような、AI に対する「発売後（post-release）」の監督をする機関である。

⑩ヨーロッパ AI 観測機関（European observatory for AI）を開発するべきである。この機関の役割とは、AI の発展を監視し、議論やコンセンサスをはぐくむフォーラムを提供し、（概念や入手可能な情報へのリンクを含んだ）AI に関する情報やソフトウェアについてのリポジトリを提供し、具体的行動に対する勧告やガイドラインを順を追って発表する、といったものである。

⑪労働環境における、人間と機械の間のスムーズで有益な関わり合いの基礎を提供するために、法律文書や契約に関してのひな型（contractual template）を開発するべきである。また、「包摂的な技術革新」という考えを保持する現在のヨーロッパ社会と歩調を合わせながら、新たな種類の仕事への変化をスムーズに果たすために、ヨーロッパ・グローバリゼーション調整基金（European Globalisation Adjustment Fund）と並列して、ヨーロッパ AI 調整基金（European AI Adjustment Fund）が設置されうるだろう。

5.2.3　報奨化（insentivisation）

⑫EU レベルで、（単に受容可能なだけでなく）社会的に望ましくもあり、また、環境に優しい（つまり、単に持続可能であるだけでなく、環境にとって好ましいものでもある）＜EU 内部における AI 技術の使用と発展＞を、金銭的に報奨化するべきである（*incentivise*）。このことは、何らかの AI プロジェクトが、社会的に望ましいのか、また、環境に優しいのかということを評価するための助けとなる、方法論の開発を含むだろう。こうして、倫理的に健全で、共通善による利益に一致した、＜AI による特別な解決策の発展＞が、創造性と競争性をもって、促されるかもしれないのだ。

⑬持続され、強化され、整合的で、AI の個別的な特徴に適合した、ヨーロッパにおける研究への取り組みを、金銭的に報奨化すべきである。これは、AI を社会の善に向けて前進させようという明確な任務に関わっていなければならないし、社会的機会に対してあまり焦点をあてないことによって、AI の流行（AI trends）に対する唯一の均衡力として働くという明確な任務にも関わっていなければならない。

⑭技術と社会的課題と法学と倫理学の交流に関係する領域横断的な協調を、金銭的に報奨化すべきである。技術的な課題についての議論は実際の技術の進展に後れを取るかもしれないが、しかし、もしこれらの議論が様々なマルチ・ステークホルダーによって戦略的に形成されるならば、これらの議論は技術革新を正しい方向へと導き、そして、それを支

えるかもしれないのだ。また、このような議論において、ジェンダー、階級、民族、学問分野、そして、その他関連領域における多様性が、AI のデザインや発達を形作ることが極めて重要である。

⑮AI の研究計画における倫理的、法的、社会的考慮の包摂を金銭的に報奨化すべきである。これに並行して、AI が社会的に肯定的な技術革新を育んでいるのかというテストに従って、立法の定期的な見直しを報奨化すべきである。この両者の制度によって、AI 技術が根本的に倫理的で、また、AI 政策が技術革新に向かうものであることが保証されるのだ。

⑯AI システムの経験的なテストと発展のために、EU 内部において、＜合法的に規制緩和された区画（特区）＞を発展させまた使用することを、金銭的に報奨化すべきである。

⑰＜AI とその応用が公的に受容され理解されることに関する研究＞、及び、＜AI に関係した政策やルールを作り上げる、構造化された公共的な協議システムの実行化（implementation of structured public consultation mechanisms）＞を、金銭的に報奨化するべきである。これは、世論調査といった伝統等的な調査方法だけによるのでなく、AI システムによる倫理的ジレンマについてのシミュレートされた事例の提示といった、より実験的アプローチあるいは社会科学的な実験に従った、世論の直接的な顕在化を含むかもしれない。さらに、この研究指針は、単に世論を測ることだけでなく、政策や規則を、ともに作り上げる（co-creation）ことにも結びついていなければならない。

5.2.4　支援

⑱データや AI に関係した専門家の振る舞いを規制する、特別な倫理的義務を負った、自己規制コードの発展を支援するべきである。これは、人々に確実に AI の利益を理解させ、AI を要求するようにさせる、信頼という印（trust-labels）によって、「倫理的な AI」を確証し、医者や弁護士といった他の専門職（professions）と同列に加わるということである。

⑲ある企業の役員がその企業の AI 技術の倫理的な含意に対して責任をとる能力を、支援するべきである。例えば、既存の役員会に対するトレーニングを改良したり、内部監査役によって倫理委員会を発展させたりすること、などがあげられるだろう。

⑳AI が持つ社会的、法的、倫理的インパクトをめぐる、教育カリキュラムや公共的な意識を向上させる活動を支援するべきである。これには下記のことが含まれるかもしれない。

- コンピューター・サイエンスが、教えられるべき基礎的な教科に包摂されるよう支援する、学校カリキュラム。
- AI 技術に関わるビジネスにおいて、AI と共に働くことの社会的、法的、倫理的インパクトについて、雇用者を教育するための、イニシアティブや評価プログラム。
- データ・サイエンティストや AI・サイエンティストなどの領域における、倫理や人権を包摂するヨーロッパレベルでの勧告。
- 官僚や政治家やジャーナリストといった人々に照準を合わせた、一般大衆に向け

られた同様のプログラム。
・ITU AI for Good[8]として開催されるイベントや、持続可能な開発目標（SDGs）に取り組む NGO といった、より幅広いイニシアティブに参加すること。

6. 結論

現在、ヨーロッパや世界全体は、人間の生活の多くの側面に対するとても刺激的な期待を抱かせながら、同時に大きな脅威をも与える、AI 技術の登場に直面している。本稿では（とりわけ先の勧告において）、AI 技術の発達やデザインや展開から、倫理的そして社会的に望ましい帰結が得られるように努力した。そして、5 つの倫理的原則に加え、AI がもたらす中心的機会とそれに相当するリスクを同定し、AI による最も差し迫った社会的課題に対する具体的で建設的な反応を作り上げことに資する、20 の具体的行動を作り上げた。

技術の変化はとても速いために、現在のリベラル・デモクラシーにおける政治的手続きは、時代遅れで、調和を欠き、価値を保存し社会や人々の利益を促進するという課題に、もはやもう合致していないとみなすことに魅力を感じてしまうかもしれない。しかし我々はこのような考えに反対する。センター、機関、カリキュラムなどの策定といった、ここで我々が掲げた勧告によって、我々は、政策制定や技術革新についての、野心ある、包摂的で、公平なプログラムの一例を作り上げたのであって、そして、このことは、全ての人々や我々が共有する世界が受け取る、AI からの利益を保存し、そしてそのリスクを緩和するということに寄与するだろうと、我々は考えるのである。

（三上　航志）

[8] ITU AI for Good とは、国際電気通信連合（ITU）と XPRIZE 財団によって主催される、国際会議であり、AI が社会的課題を解決し「持続可能な開発目標（SDGs）」の進展を加速させる大きな可能性を秘めていると考え、AI による技術革新を幅広い世界的な問題に結びつけることを理念としている。詳しくは下記のウェブサイトを参照のこと。
（https://aiforgood.itu.int/）