The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

AIはAI技術者を倫理的な設計に巻き込むことができるか？

Can AI Involve AI Engineers in Ethical Design Activities?

関口　海良 *1　　堀　浩一 *1*2
Kaira Sekiguchi　　Koichi Hori

*1東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

*2理化学研究所革新知能統合研究センター
Center for Advanced Intelligence Project, RIKEN

Although the importance of AI ethics has been increasingly recognized, it is hard to say that the results of the discussion is incorporated into research and development of AI technologies, so there is a gap between them. In this research, we provide an organic and dynamic AI ethics library with the aim of resolving this gap by supporting the practice of ethical design by AI engineers. Here, organic means that it deals with complex relations among different AI ethics discourses. And dynamic means that, in interaction with users, it dynamically adopts new issues and helps users think in their own contexts.

For example, AI ethics library visualizes a structure of each AI ethics discourse in a standard form, presents the distance and relationships among the discourses, and seamlessly connects them on the extension of AI technologies. Furthermore, AI ethics library is expected that the AI ethics side can also be reconstructed as a more practical one through the practices.

In this paper, with clarifying the framework of AI ethics by applying the ethical design theory, we introduce the overview and cases of AI ethics library and evaluate its effects.

1.　序論

AI 倫理の重要性は近年益々大きく認識されており，学会，財団，行政機関等によって多くの指針やケースが提供されている [IEEE 16, IEEE 17, FLI 17, AI ネ 17a, AI ネ 17b]．しかし，AI 技術者がこれら成果を自身の研究開発に取り入れているとは言い難く，両者の間にギャップが存在している．理由としては，AI 倫理は多様で情報量が多く，AI 技術者としては工学的な仕事と同時に AI 倫理と自身の研究開発との関連を理解するための時間的余裕がないためと考えられる．さらに時間が取れたとしても，提供される AI 倫理の議論は抽象度が高く，上記関連の理解には困難を伴うものと考えられる．

そこで本研究では，AI 技術者が自身の研究開発の延長線上に倫理的な課題を手軽に把握できるようにするための支援を行うこととした．またギャップを埋めることを通じて，倫理的な AI 及び社会的な価値をより一層実現することを目指す．

実現手段として筆者らは，有機的で動的な AI 倫理ライブラリを実装し，提供を開始した [Sekiguchi 17]．ここで有機的とは，AI 倫理ライブラリが異なる AI 倫理間の複雑な関連を踏まえて支援することを意味する．また動的とは，AI 倫理ライブラリが AI 技術者とのインタラクションの中で，AI 技術者が有する／生み出す新たな文脈を考慮しながら支援することを意味する．

具体的には，AI 倫理ライブラリは各種の AI 倫理の言説が持つ構造の違いを明確にしたり，AI 倫理の言説相互の意味的な距離を提示したり，また，AI 技術の延長線上にシームレスに AI 倫理を繋げるようなシナリオパスの推薦を行う．

さらに，AI 倫理ライブラリによって AI 技術と AI 倫理を繋ぐことを通じて，AI 倫理の側もより実践的なものへと再構築され得るものと考えている．

次章より，AI 倫理ライブラリの概要を紹介し，具体例を紹介した上で，上記点について評価を行う．

連絡先: 関口 海良，東京大学大学院工学系研究科航空宇宙工学専攻 知能工学研究室，kaira @ dfrome.com

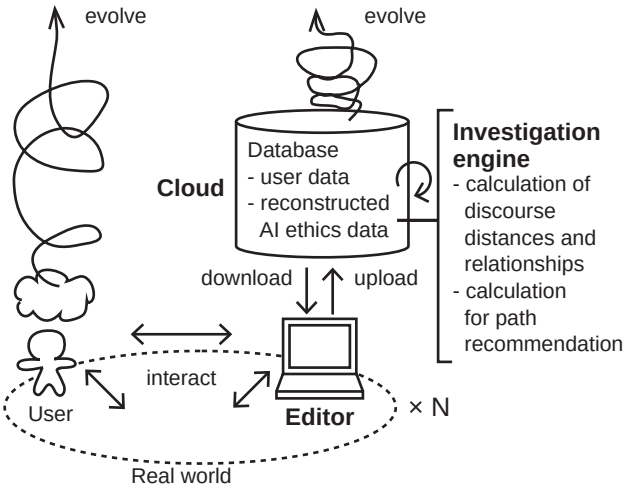

図 1: Overview of the system for the organic and dynamic AI ethics library

2.　AI 倫理ライブラリの概要

初めに，AI 倫理ライブラリの概要を述べる．

2.1　AI 倫理ライブラリの構成

AI 倫理ライブラリの構成を示したのが図 1 である．

図 1 が示す様に，AI 倫理ライブラリは大きく三つの技術から成る．エディタ（Editor），クラウド（Cloud）環境と，探求用エンジン（Investigation engine）である．

エディタは背景として倫理的な視点を提供しながら，編集作業の負荷を下げることでユーザが設計行為により没入できるようにする．次に，クラウド環境はアプリのインストールの手間を省くと共に，設計アイデアを端末に依存せずに管理可能とし，また，ユーザ間で容易に共有できるするようにする．最後に，探求用エンジンは AI 倫理や設計アイデア間の距離を計算

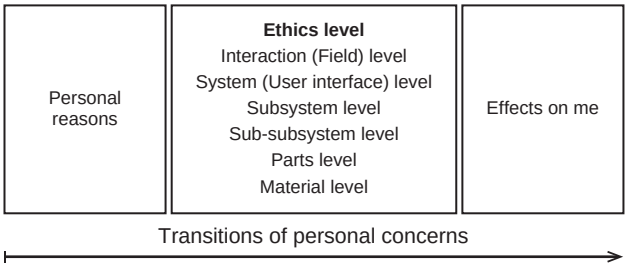

図 2: Redefined version of the hierarchical representation of artifacts [Sekiguchi 10]

する機能と，設計者の文脈において考慮すべきシナリオパスを計算する機能を備えている．

本研究では上記技術に基づき，AI 技術者と AI 倫理ライブラリとのインタラクションを通じて，AI 技術者が持つアイデアとライブラリに蓄積されたコンテンツが動的に発展（evolve）しながら，より倫理的な価値を踏まえた設計が生まれていくものと仮定している．

2.2 倫理的設計学の適用による表現の標準化

AI 倫理ライブラリに蓄積されるデータは倫理的な設計学に基づき標準化された形式に従う．ここで倫理的な設計学とは，筆者らが「倫理レベルからの設計」と呼ぶ視点と，「言説による設計」と呼ぶ方法から成るものである．

2.2.1 倫理的な設計のための視点：倫理レベルからの設計

「倫理レベルからの設計」とは，人工物の階層表現 [Simon 96, 吉川 81] を再定義したもので，工学的に倫理を扱えるようにしたものである．筆者らは，システムのレベル（System level）の上に，時間や流れ，変化を扱うインタラクション [中小路 07] のレベル（Interaction level）を位置付け，さらにその上に社会的な意味や価値を記述するレベルとして倫理レベル（Ethics level）を位置付けた [Sekiguchi 09]．

続いて，これら人工物の階層表現を客観的事象の軸とすると，設計者の主観的な視点の軸を設定することができる [Sekiguchi 10]．図 2 は再定義した階層表現の全体像である．

図 2 において，中央の面における縦方向が階層表現の軸で，全体の横方向が主観的な軸である．両者を直行させているのは，設計者が各レベルに対して知識やアイデアを同様に持つことができることと対応させている．また設計者にとっての時間推移を考えることで，個人的な理由（Personal reasons）と，個人としての影響（Effects on me）の二つの面を階層表現の両側に分けて配置することができる．

関連する議論として，JSAI2017 の公開討論にて江間は「倫理と社会の（ざっくりした）関係」を「研究（者）倫理」,「AI の倫理」,「倫理的な AI」の三つに分類した [人工 17]．本研究との関連では，主観的な軸に関する議論が江間の研究（者）の倫理に対応し，階層表現における議論が AI の倫理に対応し，考察の対象となる人工物のひとつが倫理的な AI に対応する．

また，ELSI との関連では，E の Ethics は上述の通りであり，倫理レベルの主体のひとつとして文節されるのが S の Society である．L の Low はそれら倫理的な意味や価値を実現ないし制約するための人工物として位置付けられる．

2.2.2 倫理レベルからの設計のための方法：言説による設計

「言説による設計」に関して最も重要な特徴は，人工物が生み出す変化をつなげて記述する点である．図 3 はこの記述を可視化したイメージである．

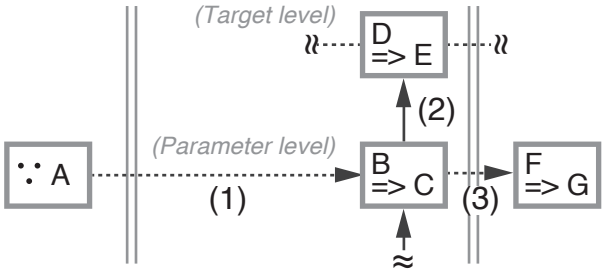

図 3: Visualization of the grammar of design with discourse: (1) Since A is a personal reason, I/we generate a design that B will change to C in the hierarchical representation of artifacts, (2) If B is changed to C at the parameter level, then D will change to E at the target level, (3) If B is changed to C, then F will change to G as the effects on me [Sekiguchi 10]

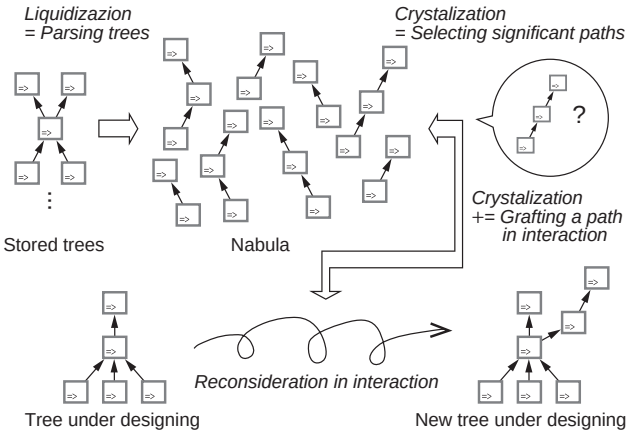

図 4: Overview of the approach for the path recommendation based on knowledge liquidization and crystallization

図 3 に示す通り，設計は三種類の記述から構成されるものとしている．

関連研究との位置付けでは，Knowledge graph の記述は Entity をノードとする [Nickel 16] のに対して，筆者らは Entity の Attributes の変化をノードとしている．倫理的な設計を扱う記述法としてはスピーカーマンのものがあるが，これは客観的な軸と主観的な軸の区別をしていないため，サイモンや吉川らの階層表現等の工学的な視点を生かせていない [Spiekermann 16]．

2.3 知識の液状化・結晶化モデルの適用

創造活動を支援するに際して，AI 倫理ライブラリは知識の液状化・結晶化モデルを適用している [Hori 04]．このモデルでは，知識は静的に蓄積されるものではなく，人と AI がインタラクションをする中で新たな文脈に応じながら動的に発展していくものと考える．AI 倫理ライブラリの中でこれを最も陽に取り入れているのがシナリオパスの推薦である．図 4 はそのアプローチの概要である．

設計者は手元のツリー内のひとつの記述をクエリとして，蓄積されたパスの中から類似したものや発散させる様なものの推薦を受ける．そして，そのインタラクションの中で設計者の視点や思考，設計アイデアが更新される．さらにこれら新しい設計アイデアが保存されるとデータベースも更新され，推薦候補群も更新されることになる．以下，これが繰り返される．

表 1: Overview of Stored Active and Public Trees

Tag	Document	Details	N
EAD	Ethically Aligned Design	Principles of Version 1	4
EAD2	Ethically Aligned Design	Principles of Version 2	5
AAP	Asilomar AI Principles	Ethics and Values	14
RDP	AI R&D Guidelines	Principles	10
RDU	AI R&D Guidelines	Use cases	10
ML	Machine Learning	Overview of the algorithm	5
EIT	Ethical IT Innovation	RFID case	1
OTH	Others	Authors' design etc.	5
			54

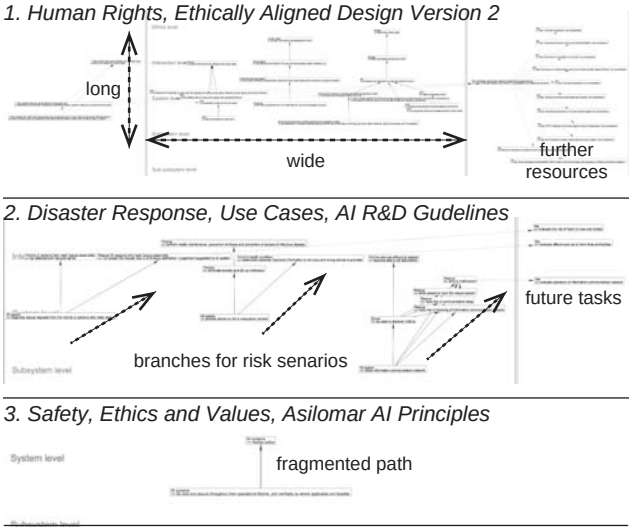

図 5: Visualized images of AI ethics discourses [IEEE 17, AI ネ 17b, FLI 17]

3. AI 倫理ライブラリによる支援の具体例

AI 倫理ライブラリによる支援の具体例を紹介する．実験は 2018 年 3 月 8 日時点で公開しているツリーを対象に，ローカル環境にて行なった．表 1 はデータの内訳である [IEEE 16, IEEE 17, FLI 17, AI ネ 17a, AI ネ 17b, Spiekermann 16]．

3.1 AI 倫理ライブラリによる表現例

初めに，AI 倫理の各言説が持つ構造を可視化した具体例として，三つのツリーを図 5 に示す．

ひとつ目の IEEE の原則（Principle）は広い課題を列挙しておりツリーは広くなり，また長く繋がる傾向がある．総務省から出ているユースケースは，リスクシナリオを併記する形になっているため上向きの分岐（branch）を持つことになる．アシロマの原則はひとつひとつは短文であるため，断片的なパスとなる．

重要なことは，各種の AI 倫理の言説が持つ構造の違いは，いずれも筆者らの倫理的設計学の枠組みの中における表現のされ方の違いとして理解できることである．

表 2: Results of calculation of distances and relationships of Trees ("R" at the top of the left column signifies the Rank)

R	Query tree	
	Human Rights, EAD2	Paragraph vectors, ML
1	Human Rights, EAD2	Paragraph vectors, ML
2	Human Benefit, EAD	Convolutional neural networks, ML
3	Human Values, AAP	RFID case, EIT
4	Ethics, RDP	Knowledge graph, ML
5	Safety, AAP	Education and Human Resource Development, RDU
50	Public Works and Infrastructure, RDU	Value Alignment, AAP
51	Retail and Logistics, RDU	Safety, AAP
52	Failure Transparency, AAP	Fix examples, AAP
53	Liberty and Privacy, AAP	Failure Transparency, AAP
54	AI Arms Race, AAP	Responsibility, AAP

3.2 距離と関係性の計算例

続いて，ツリー間の意味的な距離を計算した例として，Ethically Aligned Design の Human Rights [IEEE 17] をクエリとしたケースと，AI 技術の例として Paragraph vectors [Le 14] をクエリとしたケースの結果を表 2 に示す．

表 2 が示す様に，対象やカテゴリが同じ等，類似したツリーを提示できている．Human Rights のケースでは一般的な倫理的価値に関するものが近く計算されており，Paragraph vectors でも機械学習の手法が近く，特にニューラルネットワークに関するものが自身を除き一番上に判定されている．遠い方のツリーでは個別性が強くなる傾向があり，個別の解釈が必要となるが,「なぜそれが遠いか？」との解釈を促すことで，AI 技術者が AI 倫理を考えるきっかけになるものと考えられる．

3.3 シナリオパスの推薦例

最後に，シナリオパスの推薦例として，Paragraph vectors [Le 14] の記述をクエリとした例を図 6 に示す．

クエリの図中の記述（a）で,「Paragraph vectors => provide state-of-art results on several text classification」である．データベースには二千個程度のパス候補が存在しており，実装したクラスタリング手法及び評価関数に基づき推薦される．今回はトップ 10 位以内の結果の中から筆者が選択して「接木」した．まずは，クエリから上方に向かうパス（A）の推薦を受け，最上部の (d) を次のクエリとしてさらに上方へのパス（B）の推薦を受けた．この時点で，Paragraph vectors の研究は安心安全の実現を通じて，倫理レベルにおいて社会の自由（freedoms）の促進（e）に繋がる可能性が示唆される．さらに，今度はここから下方に向けて検索をすることでパス（C）によりガバナンスの枠組み（governance framework）が並列する解として示唆された．並列の関係は，元のアイデアとの類似点や相違点を考察することで理解を深めるきっかけとなる．

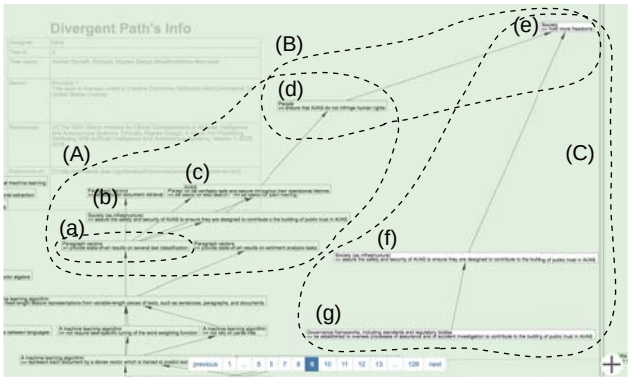

図 6: A case of the path recommendation: (a) Paragraph vectors => provide state-of-art results on several text classification; (b) Society (as infrastructure) => assure the safety and security of AI/AS to ensure they are designed to contribute to the building of public trust in AI/AS; (c) AI/AS => be verifiably safe and secure throughout their operational lifetime; (d) People => assure that AI/AS do not infringe human rights; (e) Society => hold more freedoms; (f)=(b); (g) Governance framework, including standards and regulatory bodies => be established to oversee processes of assurance and of accident investigation to contribute to the building of public trust in AI/AS [Le 14, IEEE 16, FLI 17]

重要なことは，論理の繋がりを示すことで容易には思い付かない様な長いパスや大きな広がりを関連付けられる様にし，高位のレベルを含めて暗黙とせず俎上に乗せていることである．

4. 結論と今後の課題

AI 倫理ライブラリによって，AI 技術者が自身の研究開発と AI 倫理を繋げることが容易になった．AI はこれを通じて AI 技術者を倫理的な設計に巻き込むことができる．さらに AI 倫理も工学的に議論の俎上に載せられたことから，AI 倫理の側もより実践的なものへと再構築される可能性が確認できた．

今後は，AI 倫理ライブラリの透明性を高めるため等，基盤技術の改善を継続すると共に，具体例を増やしながら，AI 倫理の専門家との対話も進めていく予定である．

参考文献

[AI ネ 17a] AI ネットワーク社会推進会議：報告書 2017 別紙 1 – 国際的な議論のための AI 開発ガイドライン案, http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000067.html (2017)

[AI ネ 17b] AI ネットワーク社会推進会議：報告書 2017 別紙 3 – AI ネットワーク化が社会・経済にもたらす影響 – 先行的評価, http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000067.html (2017)

[FLI 17] FLI, : Future of Life Institute, Asilomar AI Principles, https://futureoflife.org/ai-principles/ (2017)

[Hori 04] Hori, K.: Do Knowledge Assets Really Exist in the World and Can We Access Such Knowledge?; Knowledge Evolves Through a Cycle of Knowledge Liquidization and Crystallization, in Grieser, G. and Tanaka, Y. eds., *Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets*, Lecture Notes in Artificial Intelligence, pp. 1–13, Springer (2004)

[IEEE 16] IEEE, : The IEEE Global Institute for Ethical Consideration in Artificial and Autonomous Systems, Ethically Aligned Design: A Vision for Prioritizing Wellbeing with Artificial Intelligence and Autonomous Systems, Version 1, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (2016)

[IEEE 17] IEEE, : The IEEE Global Institute for Ethical Consideration in Artificial and Autonomous Systems, Ethically Aligned Design: A Vision for Prioritizing Wellbeing with Artificial Intelligence and Autonomous Systems, Version 2, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (2017)

[Le 14] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1188–II–1196, JMLR.org (2014)

[Nickel 16] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E.: A Review of Relational Machine Learning for Knowledge Graphs, *Proceedings of the IEEE*, pp. 11–33 (2016)

[Sekiguchi 09] Sekiguchi, K., Tanaka, K., and Hori, K.: "Design with Discourse" to Design from the "Ethics Level", in Družovec, T. W., Jaakkola, H., Kiyoki, Y., Tokuda, T., and Yoshida, N. eds., *Volume 206: Information Modelling and Knowledge Bases XXI:*, Frontiers in Artificial Intelligence and Applications, pp. 307–314, IOS Press (2009)

[Sekiguchi 10] Sekiguchi, K.: The Fifth Rule of "Design with Discourse" for the Orthogonal Representation of Moral Concerns in Design from the Ethics Level, http://www.ethics-level.com/ (2010)

[Sekiguchi 17] Sekiguchi, K.: dfrome; website for Design FROM the Ethics level, https://www.dfrome.com (2017)

[Simon 96] Simon, H. A.: *The Sciences of the Artificial, Third Edition*, MIT Press (1996)

[Spiekermann 16] Spiekermann, S.: *Ethical IT Innovation: A Value-Based System Design Approach*, p. 220, CRC Press (2016)

[吉川 81] 吉川 弘之：一般設計過程, 精密機械, Vol. 47, No. 4, pp. 405–410 (1981)

[人工 17] 人工知能学会倫理委員会：2017 年度人工知能学会全国大会「公開討論人工知能学会　倫理委員会」開催報告（詳細版）, http://ai-elsi.org/archives/583 (2017)

[中小路 07] 中小路 久美代：インタフェースからインタラクションへ — ヒューマンインタフェース研究会（研究会千夜一夜）, 情報処理学会誌, pp. 202–203 (2007)