\Orchestrating a brighter world NEC

OSS推進フォーラム　ビッグデータ部会
イノベーションテック勉強会

人工知能社会実装に向けた倫理と標準化

2019/07/30

日本電気株式会社 江川尚志
t-egawa@ct.jp.nec.com

自己紹介

江川　尚志（えがわ　たかし）
NEC 技術イノベーション戦略本部 標準化推進部

AIの標準化屋。特にAIの倫理 (透明性、バイアスほか) 標準と品質標準を扱う

経歴
　通信技術の研究者
→ 通信の標準化屋 (特にITU, 国際電気通信連合)
→ AIの標準化屋

ISO/IEC JTC 1/SC 42 (AI) 国内専門委員会幹事
SC 42/WG 3 (trustworthiness) 小委員会主査

IEEE P7001 (transparency of autonomous systems) セクレタリ

IEC SEG10 (Ethics in Autonomous and Artificial Intelligence Applications) 日本委員

イントロダクション：そもそも標準とは

標準化されたサービス

メンバーのスタイルや売り出し方を標準化したアイドルビジネスも

http://stage48.net/wiki/index.php/AKB48

- 同時に複数の場所で公演可能。メンバーが交代してもサービス継続
- ウィーン少年合唱団が本分野の先駆者

© 江藤学先生

標準の9個の機能: ISO設立25周年記念出版物より

1. 単純化
2. 互換性の確保
3. 伝達手段としての標準
4. 記号とコードの統一
5. 全体的な経済の効果;
6. 安全・生命・健康の確保
7. 消費者の利益の保護
8. 消費社会の利益の保護
9. 貿易の壁の除去

T.R.B. Sanders, *The aims and principles of standardization*, 1972

標準とは何か、標準化屋は延々と議論し標準化。
だが、本質は「人々が尊重する」であろう

新しいデファクト標準化手法：オープンソースソフトウェア

議論する暇があったらさっさと作って市場を占有してしまえ

Chainer

- コンピュータのプログラムを作り、無料で公開、誰でも貢献可能
  - 週末プログラマのお勉強レベルから、東京証券取引所の取引システムの基盤まで
  - 最先端の巨大システム向けソフトが無料で公開されている
  - 膨大なソフトが絶えずつくられ、ごく一部が生き残る多産多死モデル
- ユーザへのアピール力は抜群。「動かして試して下さい」「無料です」
- エンジニア的には、紙の標準よりも説得力があり逆らい難い
  - 紙の標準化：英語が酷い、そこカンマが抜けてる、その言い回しは曖昧だ、、、
  - オープンソース：昔、基本ソフトの中核 (Linux kernelのスケジューラ) での論争「xxの方がエレガントだ」「その方式を実装する難しさ、分かっているのか。作って持って来い、そうしたらまた議論しよう」

自然言語による、従来型標準化の長所

自然言語による標準

- 可読性良し
  非専門家は必ず自然言語
  入門編、概念の説明は技術者でも自然言語
- 漠然とした（≒曖昧な）記述可
  「既存仕様との技術的整合性に必要かつ十分な配慮をすること」はコードでは書けない

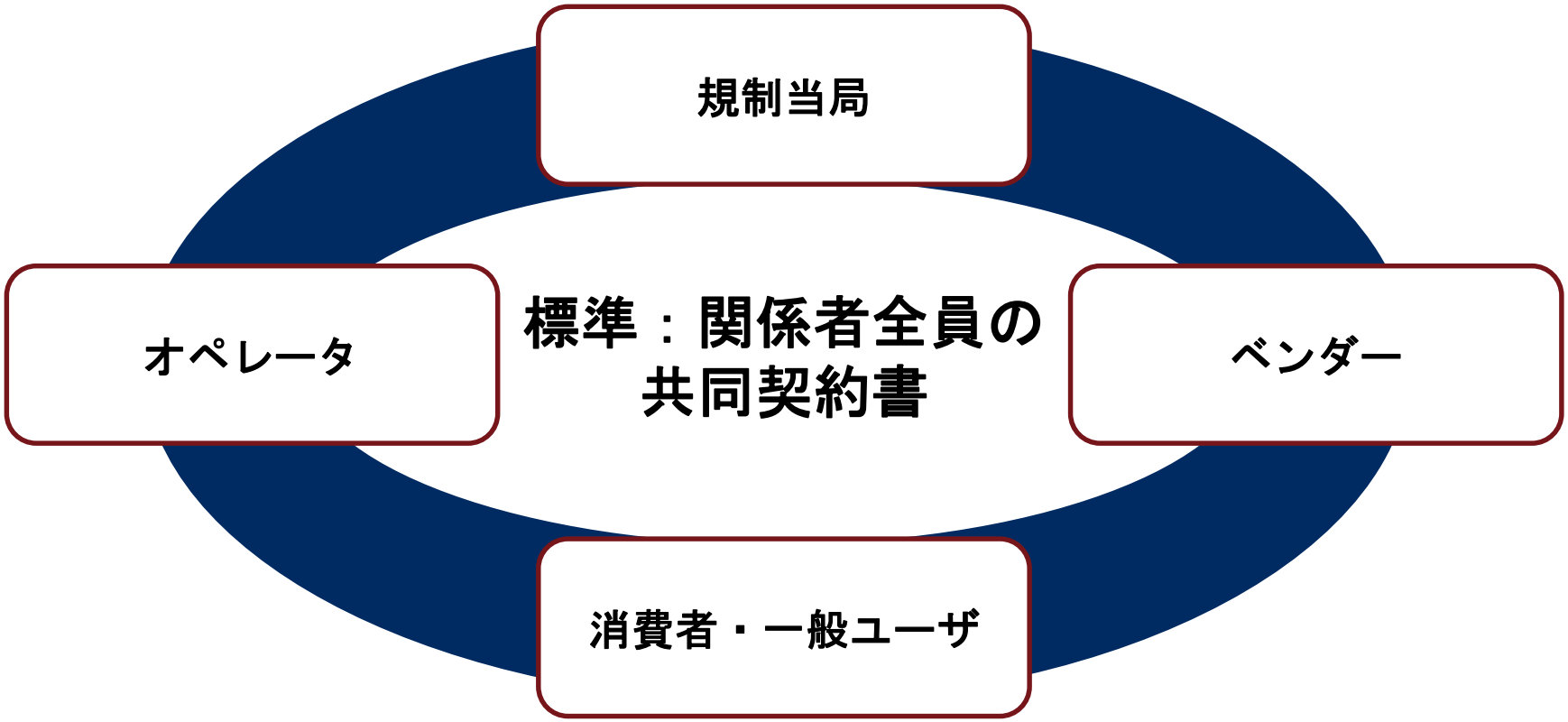
従来型標準化機関による標準化

- 仕様を標準にするのは人々の敬意。権威づけ、は常に一定の効力あり

従来型の標準化機関による標準は残らざるを得ない。
特に強制規格

AIで必要となる標準

標準とは、皆が合意したルール・契約書



- テレコム標準の世界で語り伝えられてきた成功則
- 実は「マルチステークホルダー・アプローチ」そのもの
- ソフトローとして、法というハードローを補完し得る

いまAIで開発者たちが困っていること

IPA「社会実装推進調査報告書」(2018/06) 6)実装課題の整理・分類 より

分類		No.	課題
開発に係る課題	AI人材	1	AI関連人材が不足
	学習環境	2	学習に大量のアノテーターやGPU環境が必要
	学習データ・学習済モデルの流通	3	一般企業の学習データが不足
		4	流通する学習データや学習済モデルの信頼性が不明
		5	将来に向けた学習データ収集が難しい
AIの特性に係る課題	AIシステムの検証性	6	どこまで検証すれば十分かがわからない
		7	AIに欠陥があっても、ユーザには証明できない
	AIシステムの安全性	8	AIが正常であるかが、はた目にはわからない、説明性がない
		9	ハッキングされた場合に、より高度な攻撃が懸念される
	AIの精度	10	AIの精度が100%近くでないという理由で現場が受け入れない
		11	AIの精度が学習してみないとわからない
国際課題	国際競争力	12	米国・中国のAI投資が先行している
		13	輸出先から学習データを入手できるかわからない
法制度に係る課題	AIと法制度	14	法制度がAIを想定していない
		15	学習データや学習済モデルの知財権保護と流通容易性が矛盾
		16	知財権があるデータによる学習が規制されてない
	個人情報・プライバシー	17	ネット上から集めた個人データからプライバシーを侵害しうる
		18	匿名データで学習しても、個人を特定できる可能性がある
ユーザや社会に係る課題	AIの理解	19	一般企業のAIの理解が不十分
		20	一般企業がAI導入に踏み切れない
	社会受容性	21	世論がAIを受け入れない
	AIと人の能力	22	学習内容を人に移転できない
		23	AIが肩代わりすることで、人の能力が低下する

いまAIで開発者たちが困っている、標準化が有効な課題

IPA「社会実装推進調査報告書」(2018/06) 6)実装課題の整理・分類 より

分類		No.	課題
開発に係る課題	AI人材	1	AI関連人材が不足
	学習環境	2	学習に大量のアノテーターやGPU環境が必要
	学習データ・学習済モデルの流通	3	一般企業の学習データが不足
		4	流通する学習データや学習済モデルの信頼性が不明
		5	将来に向けた学習データ収集が難しい
AIの特性に係る課題	AIシステムの検証性	6	どこまで検証すれば十分かがわからない
		7	AI[illegible]があっても[illegible]
	AIシステムの安全性	8	AIが正常であ[illegible]い、説明性がない
		9	ハッキングされた場合に、より高度な攻撃が懸念される
	AIの精度	10	AIの精度が100%近くでないという理由で現場が受け入れない
		11	AIの精度が学習してみないとわからない
国際課題	国際競争力	12	米国・中国のAI投資が先行している
		13	輸出先から学習データを入手できるかわからない
法制度に係る課題	AIと法制度	14	法制度がAIを想定していない
		15	学習データや学習済モデルの知財権保護と流通容易性が矛盾
		16	知財権があるデータによる学習が規制されてない
	個人情報・プライバシー	17	ネット上から集めた個人データからプライバシーを侵害しうる
		18	匿名データで学習しても、個人を特定できる可能性がある
ユーザや社会に係る課題	AIの理解	19	[illegible]理解が[illegible]
		20	一般企業がAI導入に踏み切れない
	社会受容性	21	世論がAIを受け入れない
	AIと人の能力	22	学習内容を人に移転できない
		23	AIが肩代わりすることで、人の能力が低下する

品質や安全：ベンダーと顧客の接点であり
伝統的に標準と関係が深い

ソフトローとしての標準やガイドライン
ハードローを補完

理解や受容性：**AI標準**、**AI倫理標準**をマルチステークホルダー・アプローチで作り解決を試みるべき

ソフトロー：日本政府のAI関連委員会

AIとはそもそも何か、の基本を社会として理解し原則を合意

人工知能と人間社会に関する懇談会	2016年3月	内閣府
国際的な議論のためのAI開発ガイドライン案	2017年7月	総務省
人間中心のAI社会原則	2019年3月	内閣府

応用分野を横断する制度 (一部のみ)

新たな情報財検討委員会	2017年3月	内閣府
知的財産推進計画2017	2017年5月	知的財産戦略本部
産業構造審議会　知的財産分科会　営業秘密の保護・活用に関する小委員会　第四次産業革命を視野に入れた不正競争防止法に関する検討（中間とりまとめ）	2017年5月	経済産業省
AI・データの利用に関する契約ガイドライン	2018年6月	経産省
データと競争政策に関する検討会	2017年6月	公正取引委員会

特定分野への応用を深く検討 (一部のみ)

自動運転の段階的実現に向けた調査研究	2017年3月	警察庁
保健医療分野におけるＡＩ活用推進懇談会	2017年6月	厚労省
自動運転における損害賠償責任に関する研究会	2017年9月	国土交通省
AI を活用した医療診断システム・医療機器等に関する課題と提言 2017	2017年12月	独立行政法人医薬品医療機器総合機構　科学委員会　AI専門部会
平成２９年度次世代医療機器・再生医療等製品評価指標作成事業　人工知能分野　審査WG報告書	2018年3月	厚労省次世代医療機器評価指標検討会

必要に応じ制度を整備へ。細部は標準や業界ガイドラインとする場合も多い

倫理・品質問題と対応する標準の例

「倫理的な仕様」を策定し合意する方法やプロセス？

- そもそも、なにが問題となり得るかを普通のエンジニアが知る方法
  「高校の倫理の授業は寝てました」「差別対策で行うべきことが分かりません」
- 非倫理的な出力とは具体的に何か。それは仕様として明確に記述し顧客と合意できるか

→ IEEE P7000: 設計を倫理に行う方法を規定するプロセス標準

→ CWA(*) 17145: 倫理委員会とその評価基準　等々

頑健なシステムを作り、問題が起きたら修正する方法？

→ ISO/IEC JTC1/SC42 TR: Trustworthiness

→ IEEE P7009: フェールセーフ

→ IEEE P7001: 自律システムの透明性　等々

確かに頑健な、品質を満たしたシステムであることの確認方法？

→ISO/IEC JTC1/SC42 NN頑健性の評価 Part 1: overview
他　多数の品質系、安全系の標準 (ISO/IEC 250xx, IEC 61508, …)

* CEN Workshop Agreement, 標準化の前段階の文書

代表的なAI標準化関連活動 (JTC1, IEC, IEEE)

EU AI-HLEG倫理ガイドライン 評価リスト (2019/04版)

1. Human agency and oversight, 2. Technical robustness and safety, 3. Privacy and data governance, 4. Transparency, 5. Diversity, non-discrimination and fairness, 6. Societal and environmental well-being

7. Accountability

Auditability:

- Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?
- Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative Impact:

- Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?
- Did you provide training and education to help developing accountability practices?
  - Which workers or branches of the team are involved? Does it go beyond the development phase?
  - Do these trainings also teach the potential legal framework applicable to the AI system?
  - Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas?
- Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?
- Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs:

- Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?
- How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

Ability to redress:

- Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?
- Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

ISO/IEC JTC1 SC42 (Artificial Intelligence)

議長: Wael Diab (Huawei US), 幹事国：米国 (ANSI)

Scope

- 1. Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence (JTC1内のAI標準化の集約点かつ推進役)
- 2. Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications (JTC1, IEC, ISOの各委員会のAI標準化をガイド)

会合: 半年に１回

- #1: 2018/04, 北京
- #2: 2018/10, 米国
- #3: 2019/04, アイルランド
- #4: 2019/10/07-11, 東京

1 JWG, 5 WG体制で活動

- JWG1(w/ SC40): Governance implications of AI (コンビーナJanna Lingenfelder, 独)
- WG1: Foundational standards (コンビーナPaul Cotton, カナダ)
- WG2: Big Data (元JTC 1/WG 9のプロジェクトを継承) (コンビーナWo Chang, 米)
- WG3: Trustworthiness (コンビーナDavid Filip, アイルランド)
- WG4: Use cases and applications (コンビーナ 丸山 文宏, 富士通研)
- WG5: Computational approaches and characteristics of artificial intelligence systems (コンビーナ Tangli Liu, 中)

SC42が作成中の文書 (1)

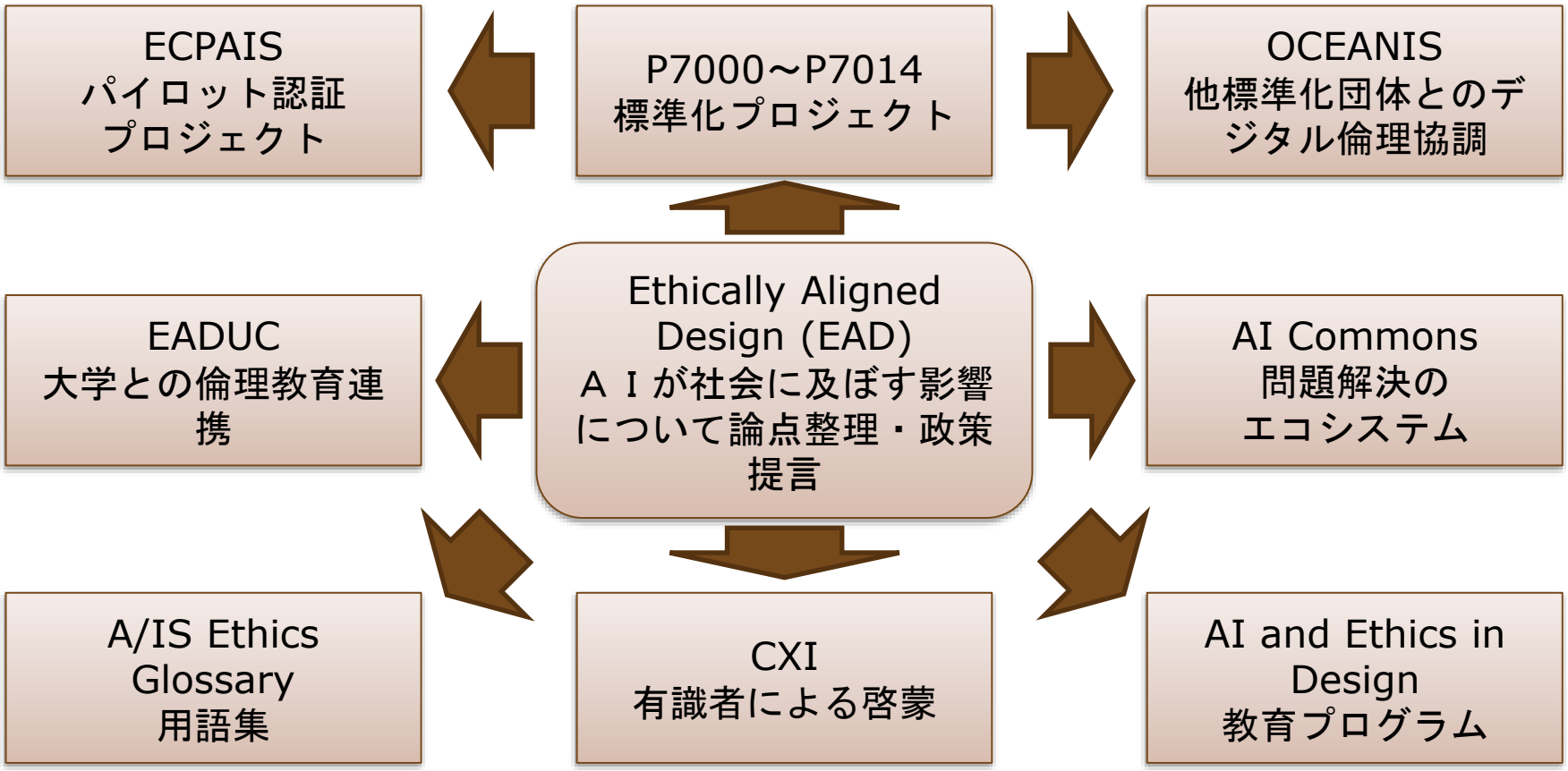
WG	規格		種	タイトル
JWG 1	38507		IS	Governance of IT – Governance implications of the use of Artificial Intelligence by organizations (AIのガバナンス)
WG 1	22989		IS	Artificial Intelligence – Concepts and Terminology (概念と用語)
	23053		IS	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (フレームワーク)
WG 2	20546		IS	Big data – Overview and vocabulary（ビッグデータの用語)
	20547			Big data reference architecture (参照アーキテクチャ)
		-1	TR	Part 1: Framework and Application Process (フレームワークとアプリのプロセス)
		-2	TR	Part 2: Use cases and derived requirements (ユースケースと要求条件)
		-3	IS	Part 3: Reference architecture (参照アーキテクチャ)
		-5	TR	Part 5: Standards roadmap (標準ロードマップ)

SC42が作成中の文書 (2)

WG	規格	種	タイトル
WG 3	24027	TR	Bias in AI systems and AI aided decision making (バイアス)
	24028	TR	Overview of trustworthiness in Artificial Intelligence (信頼感概観)
	24029-1	TR	Assessment of the robustness of neural networks – Part 1: Overview (NNの頑健性、概観)
	23894	IS	Artificial Intelligence – Risk Management (リスク管理)
	24368	TR	Artificial Intelligence (AI) – Overview of ethical and societal concerns (倫理と社会的な懸念の概観)
WG 4	24030	TR	Artificial Intelligence (AI) – Use cases (ユースケース)
WG 5	24372	TR	Overview of computational approaches for AI systems (計算論的アプローチ)

IEEEは政策提言文書（EAD）を中心にAI倫理の活動活発化

- IEEE (Institute of Electrical and Electronics Engineers)
  情報通信や電力を中心とした米国に本部を持つエンジニアの団体
- AI倫理の論点整理として始まったEADが出発点
- 出た論点を基に、デジタル倫理に対象拡大しつつ様々な活動立ち上げ



EAD: IEEE SAによるAIを巡る政策提言文書

- 名称：Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (EAD, 倫理的整合のとれた設計: AIとASにおいて人の幸福を優先するためのビジョン)
- 元々は論点整理の文書 (白書と同等)。最新版は政策提言文書のベースに
  - 最新版全体が政策提言の予定が、文書が大き過ぎレビューが困難と判明、別文書化
- 10のテーマにつき背景、勧告案、参考文献を記述
  - 標準化以外 (教育や立法他) で解決すべきを含めて広範に論点整理し政策提言
- 履歴
  - 2016/04 検討開始
  - 2016/12 for Public discussion 第１版 (Ver.1); テーマ数12
  - 2017/12 for Public discussion 第２版 (Ver.2); テーマ数13
  - 2019/03/27 第１版 (1st edition); テーマ数10
    Reframing Autonomous Weapons Systems (自律兵器の再定義)、Mixed Reality in ICT (ICTでの複合現実)、Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) (汎用AIと超知性の安全性と受益者)の3テーマは将来的に過ぎるテーマとして除外

EAD 1st editionの10テーマ

一般原則

- General Principles (一般原則)

倫理的な基礎

- Classical Ethics (古典的倫理)

影響される分野

- A/IS for Sustainable Development (持続可能な成長のためのA/IS)
- Personal Data Rights and Agency over Digital Identity (デジタルアイデンティティにおける個人情報の権利と代理)
- Legal Framework for Accountability (説明責任のための法的フレームワーク)
- Polices for Education and Awareness (教育および意識向上のための政策)

実施

- Well-being Metrics (幸福の指標)
- Embedding Values into Autonomous Systems (自律システムへの価値組込)
- Methods to Guide Ethical Research and Design (倫理的研究と設計を導く方法)
- Affective Computing (感情を読み表すコンピュータ)

EADから派生した活動

- **OCEANIS** (The Open Community for Ethics in Autonomous and Intelligent Systems)
  AIを出発点にデジタル倫理を議論しイノベーションに標準化が果たすべき役割を議論するフォーラム。2018/06設立。メンバーは第一に標準化団体、が企業も加入可能
- **ECPAIS** (The Ethics Certification Program for Autonomous and Intelligent Systems)
  P70xxシリーズ標準に基づく認証を実現するためのパイロット認証プロジェクト。フィンランドEspoo, ウィーンの2都市が現在進行中のP70xx標準化プロジェクト (特に透明性、バイアス、説明責任) にユースケース他を提供し認証の実施細目策定
- **CXI** (The Council on Extended Intelligence)
  IEEEがMITメディアラボと共に”Extended Intelligence (知性の拡張)” について立ち上げた協議会
- **EADUC** (The Ethically Aligned Design University Consortium)
  大学と立ち上げ。将来のエンジニアに向けた倫理教育プログラムを検討
- **Artificial Intelligence and Ethics in Design**: 実務家向けのEADの教育プログラム
- **A/IS Ethics Glossary**: 用語集
- **AI Commons**: AI技術を持つ人々と、そうした技術にアクセスしたい人とのマッチングシステム

IEEE P70xx標準化プロジェクト一覧 (1/3)

	タイトル	役職者	開始/完成目標
P7000	倫理的設計のモデルプロセス	議長: Ali Hessami (Vega Systems) 副議長: Sarah Spiekermann (ウィーン経済・経営大学) 幹事: Zvikomborero Murahwi (ICTコンサルタント)	2016年9月/2018年末
P7001	自律システムの透明性	議長: Alan Winfield (西イングランド大) 副議長: Nell Watson (シンギュラリティ大) 幹事 江川尚志 (NEC)	2016年12月/2018年1月
P7002	データプライバシーのプロセス	議長: Matthew Silveira (Objective Business Solutions) 副議長: John Wunderlich (John Wunderlichアソシエーション) 幹事: Robert Donaldson (INTAG systems)	2016年12月/2018年1月
P7003	アルゴリズミックバイアス (差別)	議長: Ansgar Koene (ノッティンガム大) 副議長: Christopher Clifton (Purdue大) 幹事: Liz Dowthwaite (ノッティンガム大)	2017年2月/2018年7月
P7004	子供と学生データのガバナンス	議長: Marsali Hancock (DQ Institute, シンクタンク, NPO) 副議長: Jack McArtney (McArtney Group)	2017年3月/2019年2月
P7005	従業員データのガバナンス	議長: Ulf Bengtsson (Sveriges Ingenjorer, スウェーデン大学卒エンジニア協会) 副議長: Christina Colclough (UNI Global Union, 労組の国際組織)	2017年3月/2017年12月

AIならでは: P7001, 03, 08, 13, 14;　個人情報保護系: P7002, 04, 05, 06, 12;
ソフトウェア系: P7000, 09; その他: P7007, 10, 11

開始: 上位委員会 (NESCOM)でプロジェクトが設立承認された日時。完成目標: 親ソサエティでの第1回スポンサー投票目標日時。投票成功=技術的に完成とみなされる。その後議論の公平性等のチェックを経て半年後に正式標準に
* Selfとは、IEEE会合に個人として参加していると申告し議事録等にもそのように記載されていることを示す。

IEEE P70xx標準化プロジェクト一覧 (2/3)

	タイトル	役職者	開始/完成目標
P7006	パーソナルデータAIエージェント	議長: Katryna Dow (Meeco) 副議長: Gry Hasselbalch (DataEthics (シンクタンク)) 幹事: Ken Wallace (self*)	2017年3月/2017年12月
P7007	用語	議長: Edson Prestes (リオグランデ・ド・スル連邦大学) 副議長: Sandro Rama Fiorini (パリ第12大学) 幹事: Paulo Goncalves (カシュテロ・ブランコ工芸学校)	2017年3月/2018年3月
P7008	人を倫理的につき動かすAI	議長: Laurence Devillers (LIMSI, CNRS付属研究所) 副議長: John Sullins (ソノマ州立大学)	2017年7月/2018年12月
P7009	AIのフェールセーフ設計	議長: Danit Gal(北京大学, IEEEアウトリーチ委員会議長) 副議長: Ken Wallace (self*)	2017年7月/2018年12月
P7010	AI時代の幸福の指標	議長: Laura Musikanski (Happiness Alliance, NPO) 副議長: John Havens (IEEEコンサルタント、社会活動家) 幹事: Colleen Chen (self*)	2017年7月/2018年12月
P7011	ニュース源の信頼性の特定と信頼性評価	議長: Joshua Hyman (ピッツバーグ大) 副議長 兼 幹事: Sean La Roque-Doherty (self*)	2018年2月/2020年4月

AIならでは: P7001, 03, 08, 13, 14;　個人情報保護系: P7002, 04, 05, 06, 12;
ソフトウェア系: P7000, 09; その他: P7007, 10, 11

開始: 上位委員会 (NESCOM)でプロジェクトが設立承認された日時。完成目標: 親ソサエティでの第1回スポンサー投票目標日時。投票成功=技術的に完成とみなされる。その後議論の公平性等のチェックを経て半年後に正式標準に
* Selfとは、IEEE会合に個人として参加していると申告し議事録等にもそのように記載されていることを示す。

IEEE P70xx標準化プロジェクト一覧 (3/3)

	タイトル	役職者	開始/完成目標
P7012	機械可読な個人情報の合意	議長: David Reed (元MITメディアラボ) 副議長: Lisa LeVasseur (Wrethinking,The Foundation) 幹事: Sunil Malhorta (ideafarms)	2018年2月/ 2019年7月
P7013	自動顔分析技術の包括的ガイドライン	議長: Marie-Jose Montpetit (MIT)	2018年5月 /2019年7月
P7014	共感のエミュレーション	議長: Ben Bland	2017年3月/ 2018年3月

AIならでは: P7001, 03, 08, 13, 14;　個人情報保護系: P7002, 04, 05, 06, 12;
ソフトウェア系: P7000, 09; その他: P7007, 10, 11

開始: 上位委員会 (NESCOM)でプロジェクトが設立承認された日時。完成目標: 親ソサエティでの第1回スポンサー投票目標日時。投票成功=技術的に完成とみなされる。その後議論の公平性等のチェックを経て半年後に正式標準に
* Selfとは、IEEE会合に個人として参加していると申告し議事録等にもそのように記載されていることを示す。

Questions?