2018/02/21
AIネットワーク社会推進会議

資料２

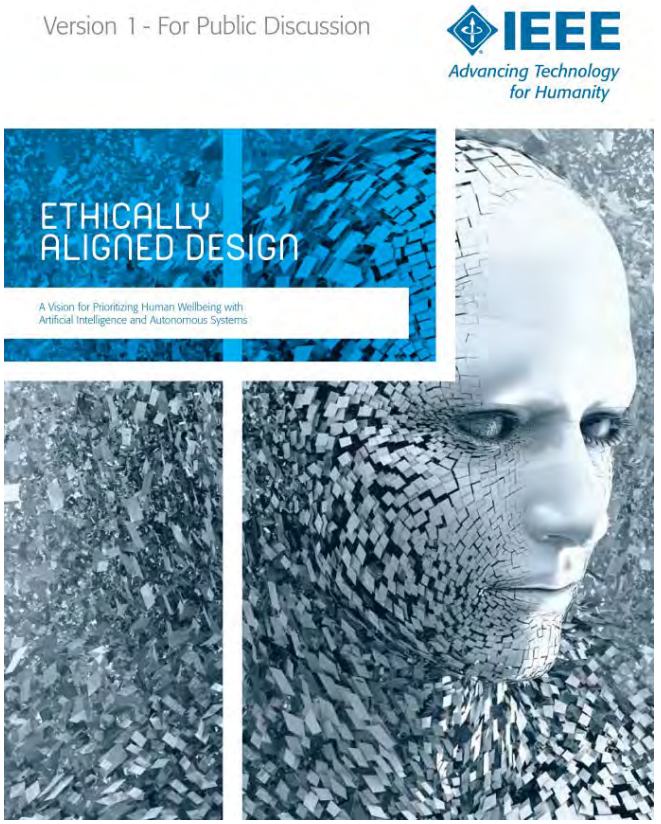
IEEE Global Initiative の活動について

東京大学
江間有沙

IEEE 倫理的に調和したデザイン

IEEE *Ethically Aligned Design* ver1 (Dec, 2016)

1. **一般原則** General Principles
2. **自律知能システムへの価値観の組み込み**
   Embedding Values into Autonomous Intelligence System
3. **倫理的研究と設計を導く方法論**
   Methodologies to Guide Ethical Research and Design
4. **汎用人工知能（AGI）と人工超知能（ASI）の安全性と恩恵**
   Safety and Beneficence of Artificial General Intelligence and Artificial Superintelligence
5. **個人情報と個別アクセス制御**
   Personal Data and Individual Access Control
6. **自律型兵器システムの再構築**
   Reframing Autonomous Weapons Systems
7. **経済的／人道的問題**
   Economic/Humanitarian Issues
8. **法律** Law



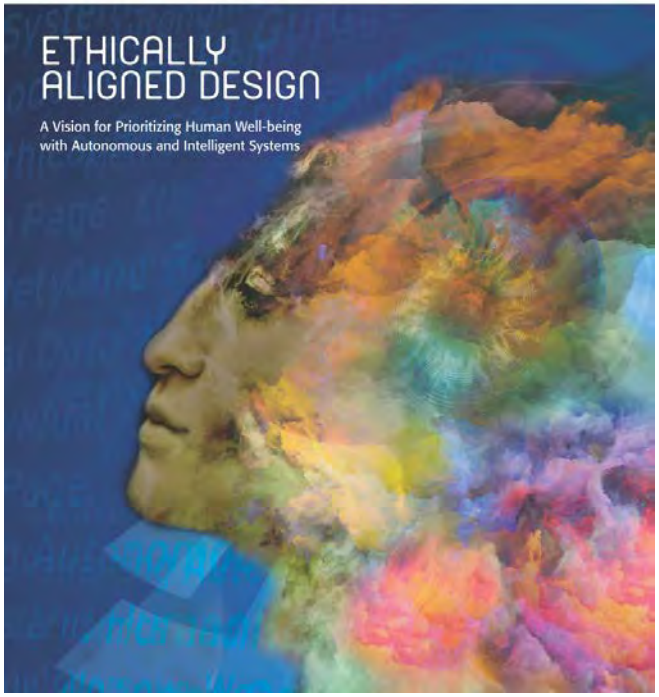
The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (“The IEEE Global Initiative”)

http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

IEEE 倫理的に調和したデザイン

IEEE *Ethically Aligned Design* ver2 (Dec, 2017)

1. 一般原則
2. 自律知能システムへの価値観の組み込み
3. 倫理的研究と設計を導く方法論
4. 汎用人工知能（AGI）と人工超知能（ASI）の安全性と恩恵
5. 個人情報と個別アクセス制御
6. 自律型兵器システムの再構築
7. 経済的／人道的問題
8. 法律
9. アフェクティブコンピューティング
10. 政策
11. ICTにおける伝統的倫理観
12. 複合現実
13. ウェルビーィング　　※Ver2より追加



The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (“The IEEE Global Initiative”)

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

標準化活動も同時並行

- “Ethically Aligned Design”が標準化にはならない
- 執筆メンバーが標準化ドラフトを作成中
  - 「倫理綱領」ではなく技術の「倫理設計」の標準化に関するものはIEEE史上初の試み
  - 「倫理」の標準化ではない

http://jp.ieee.org/activities/files/About_IEEE-SA_July2015.pdf

現在進行中の標準化活動

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ィング測定基準に関する標準（P7010）

今後の予定と戦略

- 2019年に確定版を作成予定
  - 1年かけてコメントを募集、ブラッシュアップ、ウェブセミナーなどを開催予定
  - 今までも世界経済フォーラムやIGF等様々な場所で報告
  - アウトリーチ委員会も組織し、欧米諸国だけではなく、ブラジルやアフリカ、中国など様々な国にも働きかける仕組みの構築
  - デファクトスタンダード化を狙う

  - 中身を考えるだけではなく、様々な人を巻き込み、普及の方法論も共に考えていく戦略

用語集（Glossary）の作成

- EADv2の公開と同時に自律的で知的なシステムに関する用語集の作成も立ち上げ
  - P7000標準でのキーワードの整理
  - STEM教育や学術的な研究にも貢献
- 方法論
  - IEEE P7000標準のリーダーに説明が必要と思われるキーワードを列挙してもらう
  - 辞書や様々な分野のジャーナルから定義をひいてきて、最終的には用語コミッティが決定してEADやP7000に用いて議論の端緒とする
  - 定義がまだないところは空欄とし、フィードバックを募集
    - 参加型での用語集の作り上げ
  - すべてCreative Commonライセンス付き

http://standards.ieee.org/develop/indconn/ec/eadv2_glossary.pdf

Glossary表示方法とリスト

TERM	Ordinary language	Computational Disciplines	Engineering	Government, Policy, and Social Sciences	Ethics and Philosophy
ACCOUNTABILITY	Liability to account for and answer for one's conduct; judgment of blameworthiness; obligation to provide a satisfactory answer to an external oversight agent	A set of mechanisms, practices and attributes that sum to a governance structure which "consists of accepting responsibility for the stewardship of personal and/or confidential data with which it [data organization] is entrusted in a cloud	National Society for Professional Engineers, Fundamental Canon #6, "6. Conduct themselves honorably, responsibly, ethically, and lawfully so as to enhance the honor, reputation, and usefulness of the profession."	"Accountability involves the means by which public agencies and their workers manage the diverse expectations generated within and outside the organization"(Romzek and Dubnik 1987, 228). "Administrative accountability is the concept that officials	Accountability is a component of the state of being responsible, alongside being answerable and being attributable. "To be answerable . . . is to be susceptible for assessment of, and respond to, the reasons one takes to justify one's actions. …To be

用語リスト

Accountability, Affect, Agency, Agent, AIS, Anticipatory ethics, Art, Artificial, AI, Assistive technology, Augmented Reality, Autonomy, Beneficence, Cognition, Cognitive Computing, Computation, Consciousness, Consent, Consensus, Control, Culture, Data, Development, Digital Personal Assistant, Discrimination, Duty, Equality, Ethics, Ethical Theory, Expert system, Evil, Governance, Harm, Health, Human Rights, Humanity, Humanitarian, Impact Assessment, Implementation, Individually Identifiable Data, Information, Intelligence, Intelligent Agent, Law, Legal Personhood, Maleficence, Malfeasance, Methodology, Mind, Mitigation, Mixed Reality, Moral, Moral Agent, Moral Autonomy, Moral Norms, Norms, Normative System, Nudging, Ontology, Patients, Personal Data, Persuasion, Persuasive Technology, Policy, Principles, Privacy, Proprietary, Research, Responsibility, Rights, Risk, Safety, Social Norms, Sociotechnical Systems, Superintelligence, Sustainability, System, Technical Norms, Technology, Test, Training, Transparency, Triple Bottom Line, Trust, Trustworthiness, Values, Validation, Verification, Virtual Reality, Weapon System, Wellbeing.