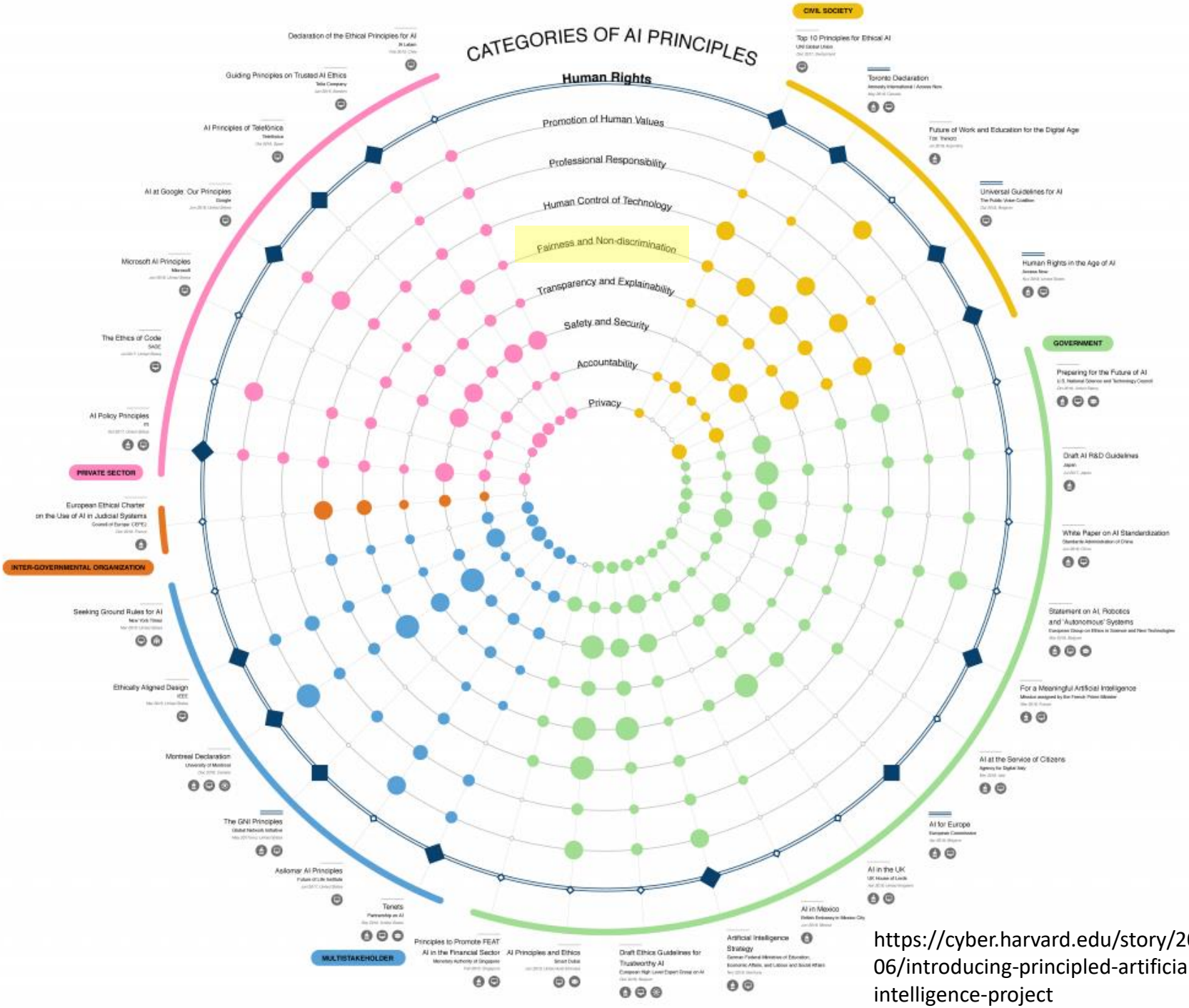
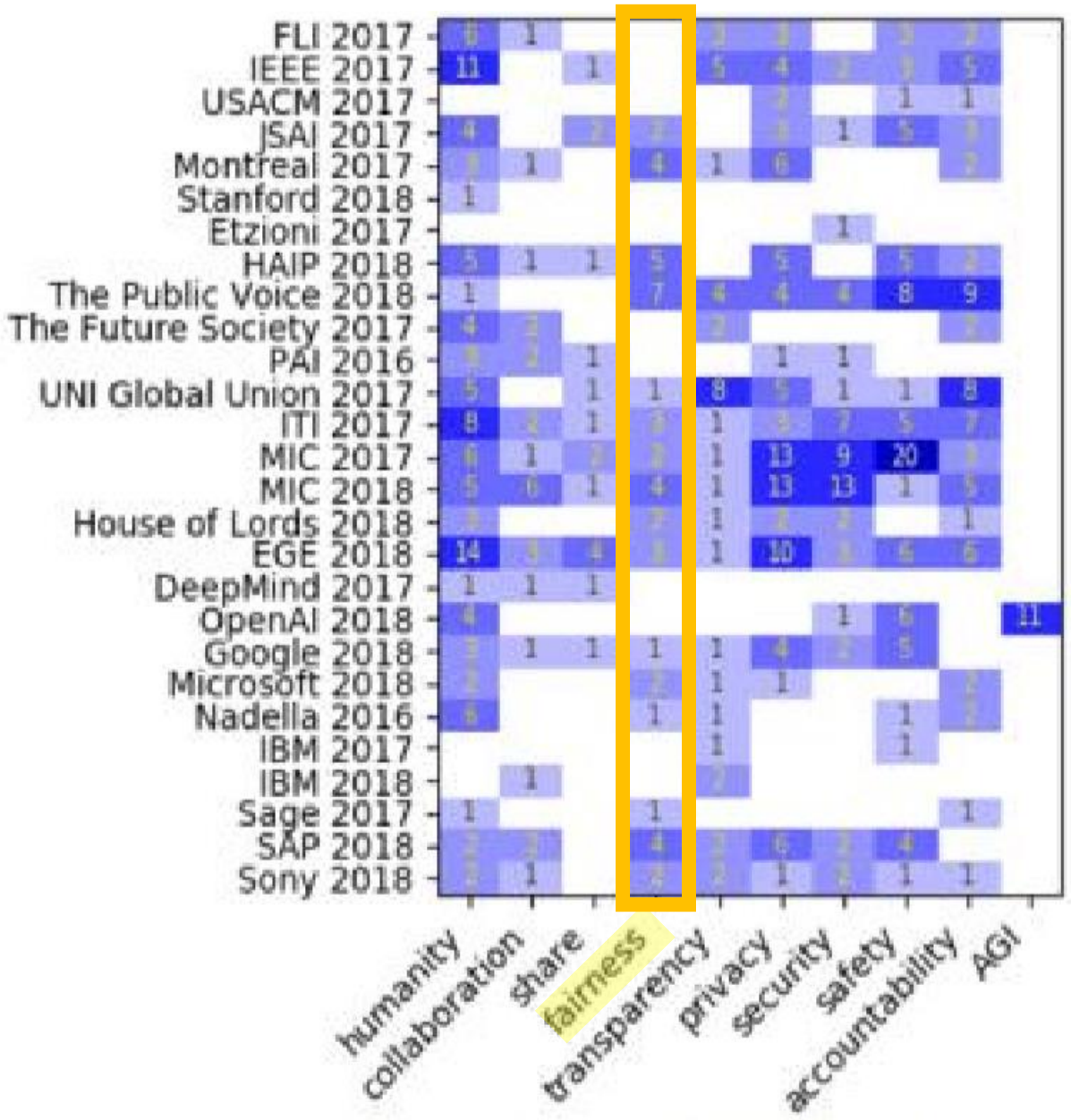
AI公平性に対する研究者コミュニティの社会的責任

東京大学 未来ビジョン研究センター 特任講師
理化学研究所 革新知能統合研究センター 客員研究員
人工知能学会 倫理委員会 副委員長
日本ディープラーニング協会 理事

江間有沙



https://cyber.harvard.edu/story/2019-06/introducing-principled-artificial-intelligence-project



Linking Artificial Intelligence Principles, Yi Zeng, Enmeng Lu, Cunqing Huangfu
https://arxiv.org/abs/1812.04814

FLI: アシロマAI原則
Asilomar AI Principles（2017）

研究課題　Research Issues (1-5)

研究目標、研究資金、研究と政策のリンク、研究文化、競争の回避
Research Goal, Research Funding, Science-policy link, Research Culture, Race Avoidance

倫理と価値　Ethics and Values (6-18)

安全性、障害への透明性、司法透明性、責任、価値との調和、人間の価値、個人のプライバシー、自由とプライバシー、共有された利益、共有された繁栄、人間による制御、非破壊、AI軍備競争
Safety, Failure Transparency, Judicial Transparency, Responsibility, Value Alignment, Human Values, Personal Privacy, Liberty and Privacy, Human Control, Non-subversion, AI Arms Race

長期的な課題　Long-term Issues (19-23)

性能に対する注意、重要性、リスク、再帰的な自己改革、公共の利益
Capability Caution, Importance, Risks, Recursive Self-Improvement, Common Good

https://futureoflife.org/ai-principles/

人工知能学会　倫理指針（2017）

1. 人類への貢献
2. 法規制の遵守
3. 他者のプライバシーの尊重
4. 公正性
   - 人工知能学会会員は、人工知能の開発と利用において常に公正さを持ち、人工知能が人間社会において**不公平や格差**をもたらす可能性があることを認識し、開発にあたって差別を行わないよう留意する。人工知能学会会員は人類が**公平、平等**に人工知能を利用できるように努める。
5. 安全性
6. 誠実な振る舞い
7. 社会に対する責任
8. 社会との対話と自己研鑽
9. 人工知能への倫理遵守の要請

http://ai-elsi.org/archives/471

Partnership on AI: Thematic Pillars

1	安全性　AIの重要性
2	AIにおける公平性・透明性・責任
3	人とAIの連携
4	AIと労働と経済について
5	社会とAIの社会的影響
6	AIと社会的利益
7	特別な取り組み

amazon.com

DeepMind Google

facebook IBM Microsoft

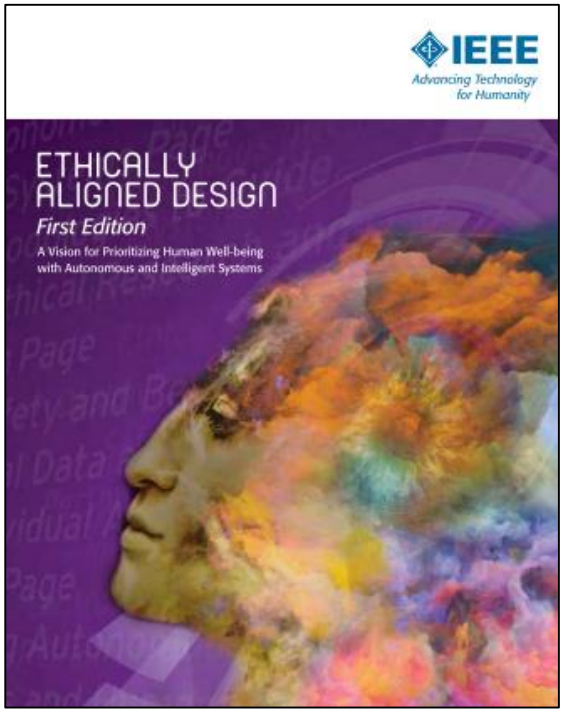
https://www.partnershiponai.org/

IEEE 倫理的に調和したデザイン（2019）

IEEE *Ethically Aligned Design 1st edition*

EADv1サマリーより
ここでいう「倫理的」とは、道徳的な概念を超えており、**社会的な公平性**、持続可能な環境と自己決定権の要求を包括しています。
*We understand "ethical" to go beyond moral constructs and include **social fairness**, environmental sustainability, and our desire for self-determination (p.3).*

1. **原則から実践へ**
2. **一般原則**
3. **ICT における伝統的倫理観**
4. **ウェルビーィング**
5. **アフェクティブコンピューティング**
6. **個人情報とエージェンシー**
7. **倫理的研究と設計を導く方法論**
8. **持続可能な発展のための自律知能システム**
9. **自律知能システムへの価値観の組み込み**
10. **政策**
11. **法**



https://ethicsinaction.ieee.org/

内閣府「人間中心のAI社会原則」（2019）

- 基本理念
  - (1) 人間の尊厳が尊重される社会
  - (2) 多様な背景を持つ人々が多様な幸せを追求できる社会
  - (3) 持続性のある社会

- 人間中心のAI社会原則
  1. 人間中心の原則
  2. 教育・リテラシーの原則
  3. プライバシー確保の原則
  4. セキュリティ確保の原則
  5. 公正競争確保の原則
  6. 公平性、説明責任及び透明性の原則
  7. イノベーションの原則

https://www8.cao.go.jp/cstp/aigensoku.pdf

人工智能北京共识

BEIJING AI PRINCIPLES (2019)

R& D Principles

- Do Good
- For Humanity
- Be Responsible
- Control Risks
- Be ethical
  - Making the system as fair as possible, reducing possible discrimination and biases …
- Be Diverse and Inclusive
- Open and Share

Use

- Use wisely and properly
- Informed-consent
- Education

Government

- Optimizing Employment
- Harmony and cooperation
- Adaption and moderation
- Subdivision and implementation
- Long-term planning

https://www.baai.ac.cn/blog/beijing-ai-principles

（2019）

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

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

国連人権高等弁務官事務所データに対する人権ベースのアプローチ（2018）

1. データ収集や処理プロセスへ様々な人を参加させているか（Participation）
2. 社会的弱者のコミュニティに対する**不平等**などがないかを確認できるかのように個人の属性が適切に分類されているか（data disaggregation）
3. データ収集される人たちにはどのようなデータを開示したりするかや、自分たち自身をどのように定義するかを自己決定できるか（self-identification）
4. データ収集の透明性は担保されているか（transparency）
5. データのプライバシーは保護されているか（privacy）
6. データが人権の観点から説明責任/答責性を負っているか（accountability）

https://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf

価値の議論

- 今まで通ってきた道
  - 結果の公正と機会/手続きの公正
  - 公平性と正確性
  - プライバシーとセキュリティ
  - プライバシーと利便性
  - etc

- そもそもトレードオフにかけていいのか

- 社会として価値の議論が必要
  - 研究者コミュニティの役割と責任とは

共催

- 人工知能学会 倫理委員会
- 日本ソフトウェア科学会 機械学習工学研究会
- 電子情報通信学会 情報論的学習理論と機械学習研究会

後援・協賛

- 情報処理学会
- 日本統計学会
- 特定非営利活動法人 横断型基幹科学技術研究団体連合
- 応用哲学会
- 科学技術社会論学会
- 科学基礎論学会
- 情報法制学会
- 日本社会心理学会
- 法と経済学会
- IEEE Computer Society Tokyo/Japan Joint Chapter, Kansai Chapter, Fukuoka Chapter
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- The Future Society
- 一般社団法人 AIビジネス推進コンソーシアム
- 一般社団法人 日本ディープラーニング協会
- 一般社団法人データサイエンティスト協会
- 一般財団法人 情報法制研究所
- 国立研究開発法人 産業総合研究所 人工知能研究センター
- 国立研究開発法人 情報通信研究機構 知能科学融合研究開発推進センター
- 国立研究開発法人 理化学研究所 革新知能統合研究センター
- 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
- 大学共同利用機関法人 情報・システム研究機構 統計数理研究所

学術団体
情報科学・情報技術

学術団体
人文・社会科学

国際学術団体

一般社団法人

研究機関