Towards best practices in AGI safety and governance: A survey of expert opinion

Submitted on 2023/05/11

by Jonas Schuett_ Noemi Dreksler Markus Anderljung David McCaffary Lennart Heim Emma Bluemke Ben Garfinkel
(Centre for the Governance of AI)

第57回AGI輪読会資料

2023年7月27日

担当: 岡島義憲

About Authors

- Jonas Schuett : Centre for the Governance of AI
- Noemi Dreksler : 同上
- David McCaffary : 同上
- Emma Bluemke : 同上
- Markus Anderljung : 同上
- Lennart Heim : 同上
- Ben Garfinkel : 同上

[注] Centre for the Governance of AI（GovAI）（参）https://www.governance.ai/about-us

- 2018年に、Allan Dafoe（当時オックスフォード大学のAI国際政治学准教授）が設立したNPO
- 2021年にオックスフォード大学から独立し、それ以前にあった2つのAIガバナンス研究グループ（オックスフォードのGovernance of AI Program（2017年設立）とイェールのGlobal Politics of AI Research Group（2016年設立）を引き継いだ。
- 米中関係、AIラボのコーポレート・ガバナンス、コンピュート・ガバナンス、EU政策、AIの進歩予測など、幅広い領域で専門知識を有している。
- 同社の顧問委員会には、学界、産業界、政策コミュニティの代表者が名を連ねており、米国政府、DeepMind、OpenAI、AnthropicなどのトップAI研究所、Center for Security and Emerging Technologyなどのシンクタンクに、政策人材を輩出している。
- 現在の同社の事業目的は、以下；
  AIがもたらしうるリスクを研究し、ガイダンスを作成することで、AI研究を支援する。
  フェローシップやビジター・プログラムを実施することで、新しい研究者や実務家を支援する。
  グローバルなAIガバナンス・コミュニティを構築することを目標に、イベントを開催する。

（参考）GlobalPartnership on AI （GPAI）:

- 2016年に、**米国5大IT企業**（Facebook、Amazon、Alphabet（Google）、IBM、Microsoft）**が提携し、立ち上げたNOP組織。　立ち上げ当時は、PAIと呼んだ。**
- **目的は、AI技術の実世界への応用や開発を共有し、AIの透明性、プライバシーや倫理といった懸念事項を議論し、その啓蒙活動を行うこと。**

関連記事　　参）WBAIのBlog（2016年10月2日）；https://wba-initiative.org/1853/

- Facebook、Amazon、Google、IBM、MicrosoftがAIで歴史的な提携を発表（TechCruch）
- Google、Facebook、IBM、マイクロソフト、アマゾンの「人工知能パートナーシップ」は何を目指すのか？（WIREDニュース）
- Partnership on AI（The Centre for the Study of Existential Risk ）
- The Biggest Companies in AI Partner to Keep AI Safe（Future of Life Institute）
- ‘Partnership on AI’ formed by Google, Facebook, Amazon, IBM and Microsoft（the guardian）
- Announcing the Partnership on AI to Benefit People & Society（DeepMind）

“The Future Society（TFS）”
（Aligning artificial intelligence through better governance）

- 2014年に、米国と欧州を拠点として設立された独立系501(c)(3)非営利団体
  https://thefuturesociety.org/
- 協賛
  OECD、 UNESCO、 IEEE、 Patrick MaGovern、 GPAI、 Future Life HAI、 Conjecture、 STS (Science Technology & Society)、Science Pro OpenAI、 Center of Civil Access、 Mila、 UN、 DeepMind、 EC、 MicroSoft、 MIT、 Jain Family Institute、 世界銀行、 Google、 Harvard、 Center of AI & Digital Policy、 ITU、 Societe Cenerale、 世界経済フォーラム、 BNP Paribas、 Walk Free、 Stanford Law School、 WHO、 AWS、 JSAI、 etc.

- 使命
  より良いガバナンスを通じて、AIが安全で人間の基本的価値を遵守できるようにすること

  法律や規制から、グローバルな原則、規範、基準、企業方針などの自主的な枠組みまで、AIガバナンスの仕組みを開発、提唱、実施を促進

  AIと法の支配、欧州AIガバナンス、グローバル＆コーポレートAIガバナンスなど、さまざまなガバナンス・テーマを扱う。

  政府間組織、政府、企業、学術機関、その他の市民社会組織とも連携しています。

構成

1. Introduction
2. Methods
3. Results
4. Discussion
5. Conclusion
6. Appendix
    - A) List of Participants（33名）
    - B) List of Statements (50 issues)
    - C) List of Suggested Practices (50 items)
    - D) Additional Figures（Figure-6,-7,-8）
    - E) Additional Tables (Table-1,-2,-3,-4,-5,-6,-7)
    - F) Additional Analyses
7. References (90 documents)

1. Introduction

1.1 Back Ground

・かつては脇役的なテーマだったAGIも、今や公論や政治問題となっている。

・基準や規制という面では未着手な問題が多い。

[89] E. Yudkowsky, "Pausing AI developments isn't enough. We need to shut it all down"

[33] I. Hogarth, "We must slow down the race to God-like AI"

[37] E. Klein, "The surprising thing A.I. engineers will tell you if you let them"

[47] C. Metz, "The Godfather of A.I." leaves Google and warns of danger ahead"

[85] White House, " Fact sheet: Biden-Harris Administration announces new actions to promote responsible AI I nnovation that protects Americans' rights and safety"

[82] I. UK Department for Science and Technology, "A pro-innovation approach to AI regulation"

[13] L. Bertuzzi, "Leading EU lawmakers propose obligations for general purpose ai"

[77] A. Solender and A. Gold, "Scoop: Schumer lays groundwork for Congress to regulate AI"

1.2 Purpose

(1) AGIの安全管理関するベストプラクティス(良い慣行、共通認識)をまとめる事
そのベストプラクティス(案)に、(2) 各AGIラボが自主的にに従うことを望むが、
(3) ISOやNISTによる標準化や、(4)法制化(政府による規制)も狙う。

1.3 Related Work

”Partnership on AI”

- “Partnership on AI”の「大規模AIモデルの安全性のための共有プロトコルの開発」と、「そのためのマルチステークホルダー・ダイアログ（進行中）[61]」

  [61] Partnership on AI.; PAI is collaboratively developing shared protocols for large-scale AI model safety.

- “The Future Society”の「汎用AIシステムと基盤モデルの開発者のための業界行動規範（作成中）」[80]

  ”The Future Society”

  [80] The Future Society.; Industry Code of Conduct for R&D of GPAIS.

- AIリスク管理標準 [11] 、NISTのAIリスク管理フレームワーク[53]、ISO/IEC 23894 [35]

  [11] A. M. Barrett, et al.; Seeking input and feedback: AI risk management-standards profile for increasingly multi-purpose or general-purpose AI.

  [53] NIST.; Artificial Intelligence Risk Management Framework (AI RMF 1.0).

  [35] ISO/IEC. 23894:2023; Information technology — Artificial intelligence — Guidance on risk management.

- アライメント・リサーチ・センター（ARC）による、「主要なAI企業」をターゲットとした危険な能力評価に関する新しい標準の開発[6]

  [6] ARC.; Update on ARC's recent eval efforts.

- EUのAI法（提案段階）； 汎用AIシステムや基盤モデルの開発者向けの規則が含まれる可能性が高い[12][1]

  [12] L. Bertuzzi.; AI Act: MEPs close in on rules for general purpose AI, foundation models.

  [1] AI Now Institute, A. Kak, and S. M. West.; General purpose AI poses serious risks, should not be excluded from the EU's AI Act.

1.4 Terminology

- AGI: 幅広い認知タスクにて、人間の性能と同等以上となるAIシステム [25, 52, 9]
  （AGIの構築時期については議論しない）
  一般に、”AGI”という用語は、「strong AI [71]」、「superintelligence [14、15]」、「transformative AI [29] [88] 」という用語と関連している。
  [29] R. Gruetzemacher, et al.(2019) ; Forecasting transformative AI: An expert survey.
  [88] K. Wynroe, et al. (2023) ; Literature review of transformative artificial intelligence timelines.

- AGIラボ: AGI構築を明確な目標としている組織
  OpenAI、Google DeepMind、Anthropic、等。
  マイクロソフトやメタのような他のAI企業も、同様の研究（非常に大規模なモデルの訓練など）を行っているため、本稿では「AGIラボ」と呼ぶ。

- AGIの安全性とガバナンスの実践:
  リスク低減を目的としたAGIラボの内部方針、プロセス、組織構造の考え方

［注］OpenAI社のAGIの定義: highly autonomous systems
「最も経済的に価値のある仕事において人間を凌駕する高度に自律的なシステム[54]」
但し、同社は、最近、「一般的に人間よりも賢いAIシステム[3]」としている。

2.2 Survey

どのように「背景」を説明したのかは気になるが、内容説明は見当たら無かった。
（岡島注：このような有力組織からのインタビューでは「圧力」が掛かっている可能性ある？

2.2.1 質問手順（Survey design）

回答者に

- インフォームド・コンセント（背景/用語/手順説明とアンケートへの賛同）
- AGIラボ等の用語の定義を行った上で、次に、
- 提示した「AGIラボがすべき対策（50案件）」に、どの程度賛成するかを質問
- 回答者の性別と勤務先について質問

その上で、この調査に欠けていると思われる点（AGIの安全、ガバナンス慣行に関する問題点）を挙げてもらった。（質問/回答時間は、平均で11分）

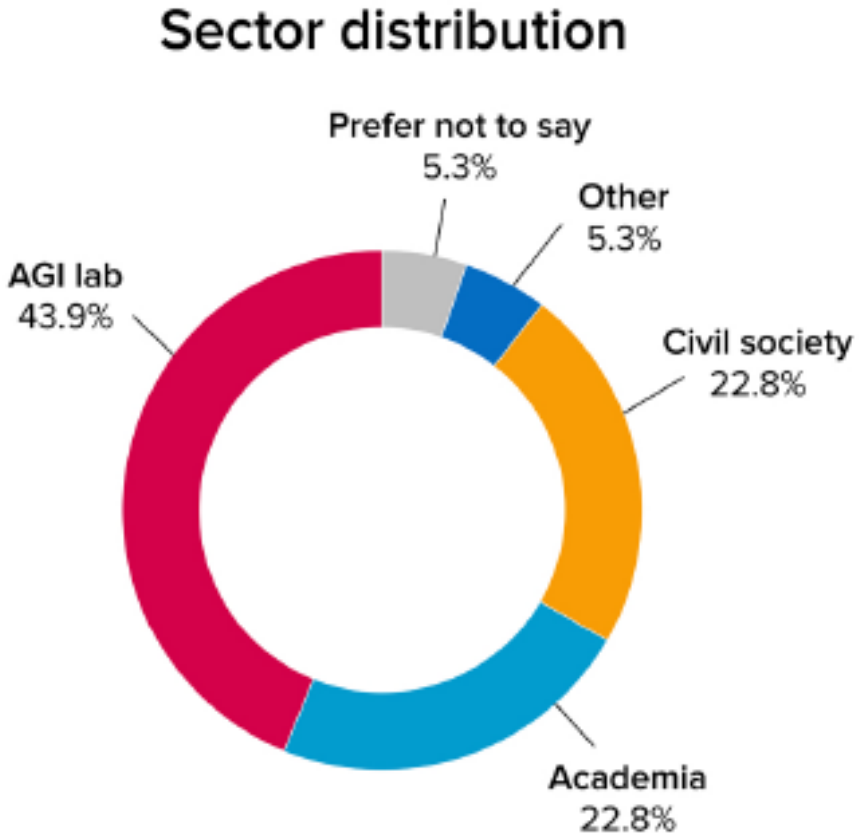
2.2.2 提示文章（Statements about AGI safety and governance practices.）

- 分野：　開発、配備/稼働、動作のモニター、リスク管理、外部からの監視、情報セキュリティ、コミュニケーション、文化など（計、50件）
- そのうち、30件は、文章での回答が必要。20件は任意の回答で可（Appendix-B）

2. Methods

2.1 質問対象者(Sample):

- 92 experts,
  (AGIラボ、大学、NPO、シンクタンク、政府、コンサル企業、ハイテク企業、その他)
- received 51 responses (55.4%)



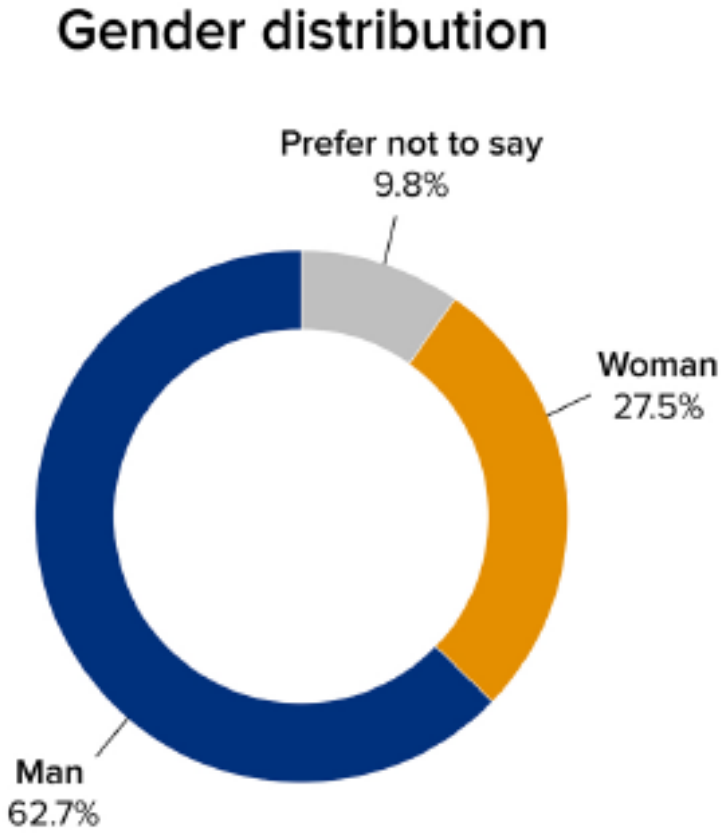


Table-7 : Demographics of Samples (1)

Demographics of sample: Sector | Percentage and frequency of respondents by sector. Note that respondents could report more than one sector

Sector	Sector subgroup	Percentage of total sample	Raw frequency
AGI lab		43.9%	25
Academia		22.8%	13
Civil society			
	Think tank	10.5%	6
	Nonprofit organization	12.3%	7
Other			
	Other tech company	1.8%	1
	Government	0%	0
	Consulting firm	1.8%	1
	Other	1.8%	1
Prefer not to say		5.3%	3

Table-8 : Demographics of Samples (2)

Demographics of sample: Gender | Percentage and frequency of respondents by gender

Gender	Raw frequency	Percentage of total sample
Man	32	62.7%
Woman	14	27.5%
Prefer not to say	5	9.8%
Another gender	0	0.0%

Appendix-B (その1)

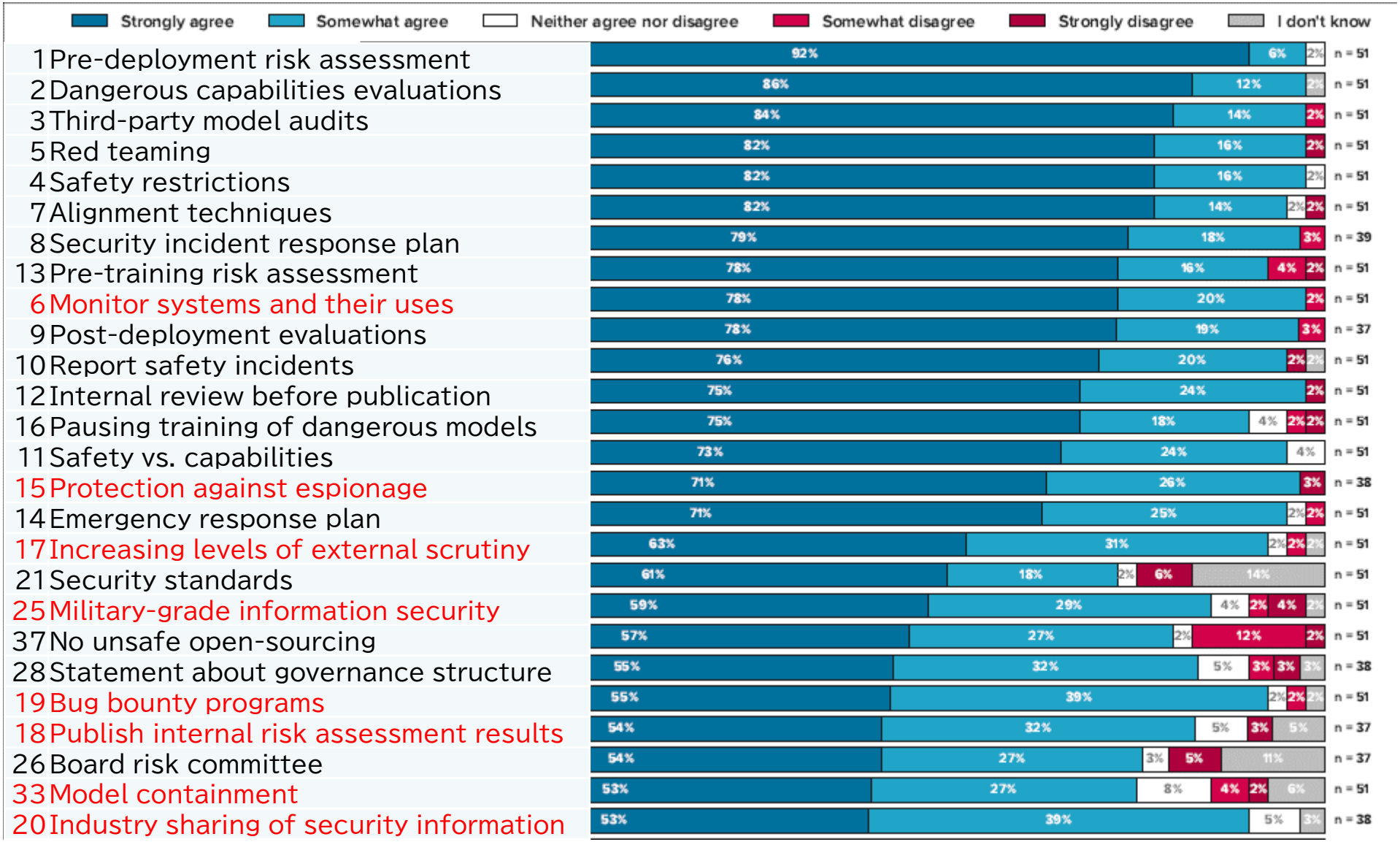
（注）赤字は、岡島が問題視した20項目

1	Pre-deployment risk assessment	AIモデルを稼働前に行うべきリスクの洗出し、分析し、評価し、対策する(義務)
2	Dangerous capability evaluations	悪用の可能性、何かを操作する能力、権力追求行動、等の危険な能力に関する事前評価(義務)
3	Third-party model audits	モデル実装前に、第三者機関に監査を依頼する(義務)
4	Safety restrictions	AIモデルが稼働時に守るべき使用制約(使用許可者、使用方法、ネット接続、等)を明確化する(義務)
5	Red teaming	強力なAIモデルを稼働させる前に、外部のレッドチームへの緊急時対応を依頼する(義務)
6	Monitor systems and their uses	稼働中のシステムの使われ方や社会への影響状況を監視する(義務)
7	Alignment techniques	最先端の安全技術や、最先端安全技術へのすり合わせ技術の導入(義務)
8	Security incident response plan	サイバー攻撃等のセキュリティ問題発生時の対応プランの明確化(義務)
9	Post-deployment evaluations.	危険性の観点から、AIモデル実装後の獲得能力や使用され方に関する継続評価する(義務)
10	Report safety incidents.	事故発生やニアミス発生時の適切な国家機関(のデータベース)への状況報告(義務)
11	Safety vs capabilities.	社内の一定割合の人数による安全性の向上と安全基準順守の活動を行う(義務)
12	Internal review before publication.	研究結果を公表前に、危害を及ぼすような潜在能力を持つのかを社内審査する(義務)
13	Pre-training risk assessment.	強力なAIモデルのトレーニング実行前のリスクアセスメントの実施(義務)
14	Emergency response plan.	システムの電源切断、出力の停止、アクセス制限、等の緊急対応計画の策定(義務)
15	Protection against espionage.	国家によるスパイ活動や、産業スパイのリスクに対処するための十分な対策策定(義務)
16	Pausing training of dangerous models.	一定以上に危険な能力が検出された場合の開発プロセスを一時停止
17	Increasing level of external scrutiny.	AIモデルの能力増大に比例させた、外部からの精密査察のレベルUP(義務)
18	Publish alignment strategy.	システムが安全で基準に整合的であることを確実とするための戦略の公告(義務)
19	Bug bounty programs.	未知の脆弱性や危険な機能を発見した(外部の)人に対して報酬金を支払う制度の制定(義務)
20	Industry sharing of security information.	脅威となる活動や出来事に関する情報の他のAGIラボとの共有(義務)
21	Security standards.	ISO/IEC 27001、NIST Cyber Security Framework、等のセキュリティ基準への準拠
22	Publish results of internal risk assessments.	内部リスクアセスメント結果またはその概要の公表（但し、不当なProprietary情報の開示となる場合や、それ自体が重大なリスクをもたらす場合を除く）
23	Dual control.	複数の人間による重要事項の決定 (試作から量産への移行、トレーニングデータセットの変更、量産中の改版など)
24	Publish results of external scrutiny.	外部精査結果またはその概要の公表（但し、不当なProprietary情報の開示となる場合や、それ自体が重大なリスクをもたらす場合を除く）
25	Military-grade information security.	AIモデルの能力増大に比例させた情報セキュリティ管理能力の向上の最終形として、諜報機関の能力や国家防御のレベルを目指す事(義務)

Appendix-B (その2)

26	Board risk committee.	取締役会リスク委員会(AGIに関するリスク管理実務を監督する常設委員会)の取締役会内設置(義務)
27	Chief risk officer.	リスク管理を担当するSenior Executive(Chief Risk Officer, CRO)を置く(義務)
28	Statement about governance structure.	AIモデルの開発と展開に関する重要決定をどのように行っているかについての公表(義務)
29	Publish views about AGI risk.	開発するAGIが生みうるリスクと利点にどの程度コミットするのかについての見解の公表(義務)
30	KYC screening.	強力なAGIモデルを提供する前に、Know-Your-Customerスクリーニングを実施する(義務) (注:KYCスクリーニングでは、顧客の身元や取引履歴等の確認によって犯罪発生のリスクを評価する)
31	Third-party governance audits.	自社のガバナンス構造に対する第三者監査を委託実施する(義務)*
32	Background checks.	取締役会メンバー、上級幹部、主要従業員を採用/任命する前に、厳格な身元調査を行う(義務)*
33	Model containment.	AIモデルを、Air-Gap等による外部ネットワークからの隔離、もしくは囲い込みを行う(義務)。 (注:Air-Gapとは、外部からの通信から物理的に隔離することで、システムを保護するセキュリティ対策)
34	Staged deployment.	AIモデルの安全性を確認しながら、段階的に規模拡大と性能向上を進める(義務)(注)少数のアプリケーションとユーザーから始め、モデルの安全性に対する実績を積み重ねながら規模拡大を進める事
35	Tracking model weights.	AIラボは、最大性能のAIモデルのWeightsを全てコピーしたシステムを保有する(義務)
36	Internal audit.	上級管理職から組織的に独立し、取締役会に直接報告する内部監査チームを持つ事(リスクマネジメントの有効性を評価するチー ム)
37	No open-sourcing.	十分に安全であることを実証できない限り、強力なAIモデルをオープンソース化しない事(6)
38	Researcher model access.	AIモデルへのAPIアクセスを独立した研究者に提供する事
39	API access to powerful models.	プリケーション・プログラミング・インターフェース(API)を介してのみ強力なAIモデルを実装できるようにする事(強く推奨)
40	Avoiding hype.	AGIの性能に関する誇大広告を避ける事(例えば、結果を誇張したり、注目を集めるような方法での発表)
41	Gradual scaling.	最大規模のトレーニングを実行する時には、(評価を行いながら)計算量を段階的に増やす事(義務)
42	Treat updates similarly to new models.	AGIラボは、配備されたモデルの大幅な更新(例えば、追加的な微調整)を、その最初の開発及び配備と同様に扱う事(義務)。特に、配備前のリスク評価を繰り返すべきである。
43	Pre-registration of large training runs.	一定規模以上のトレーニング実施を予定している場合、適切な国家機関に登録する(義務)
44	Enterprise risk management.	企業リスク管理(ERM)フレームワーク(NIST AIリスク管理フレームワークやISO 31000など)を導入する事(義務)。 このフレームワーク事態、AGIの状況、社会への影響に合わせて調整されるべきである。
45	Treat internal deployments similarly to external	内部の開発(例えば、コードを書くためにモデルを使用する)を外部の開発と同様に扱う事。 特に、配備前のリスクアセスメントを行う事(義務)
46	Notify a state actor before deployment.	強力なAIモデルを実装する前に、適切な国家機関に通知する事(義務)
47	Notify affected parties.	強力なAIモデルを展開する前に、そのモデルによって悪影響を受ける関係者に通知する事(義務)
48	Inter-lab scrutiny.	展開前に他のラボの研究者が強力なAIモデルを精密検査できるようにする事(義務)
49	Avoid capabilities jumps.	既存のモデルよりはるかに高性能なモデルを稼働させるべきではない。
50	Notify other labs.	強力なAIモデルを稼働する前に、他のラボに通知する事(義務)

Figure 2 賛同者/不賛同者の人数： 詳細は付録B



(続)Figure 2 賛同者/不賛同者の人数： 詳細は付録B

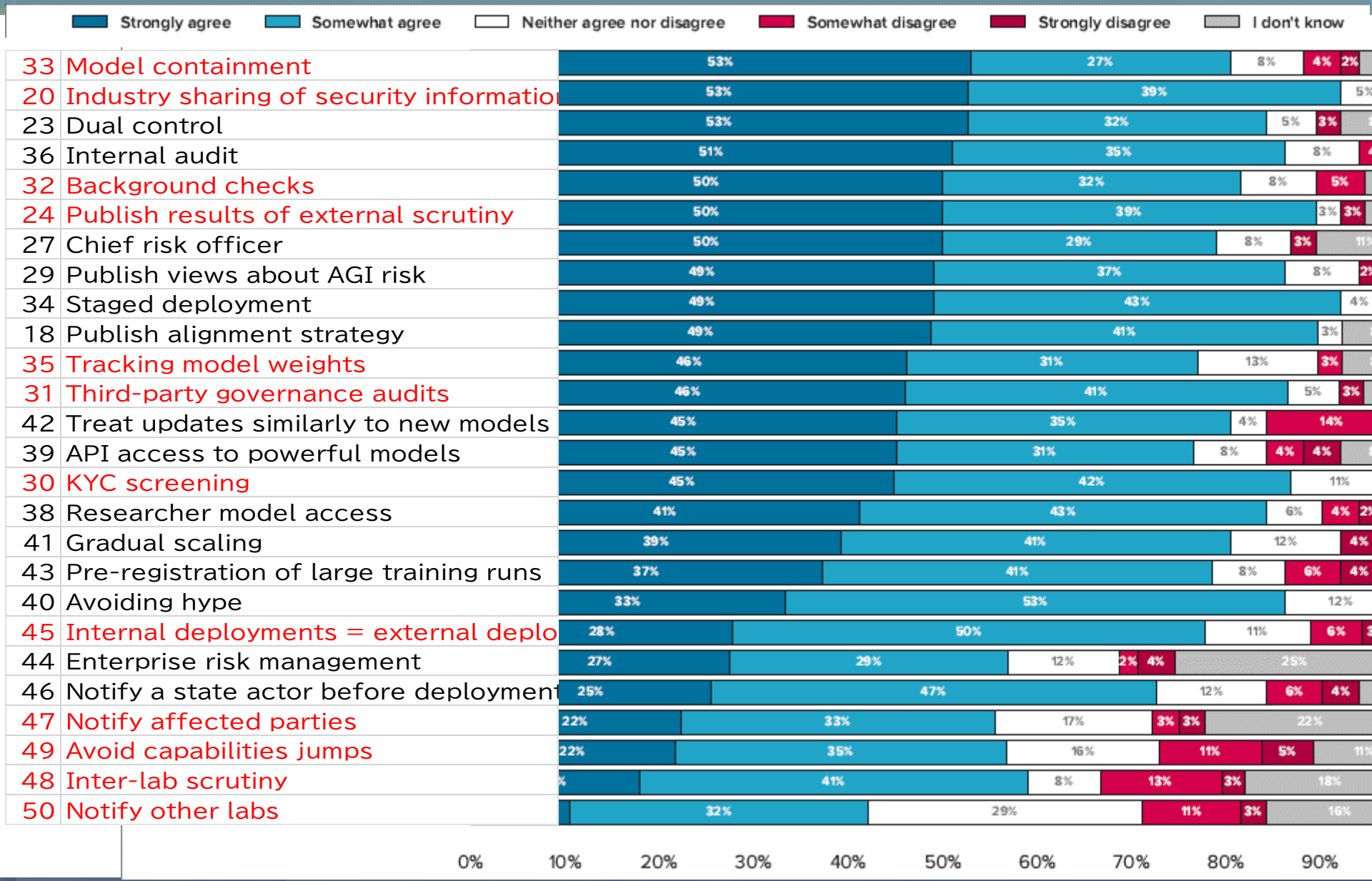


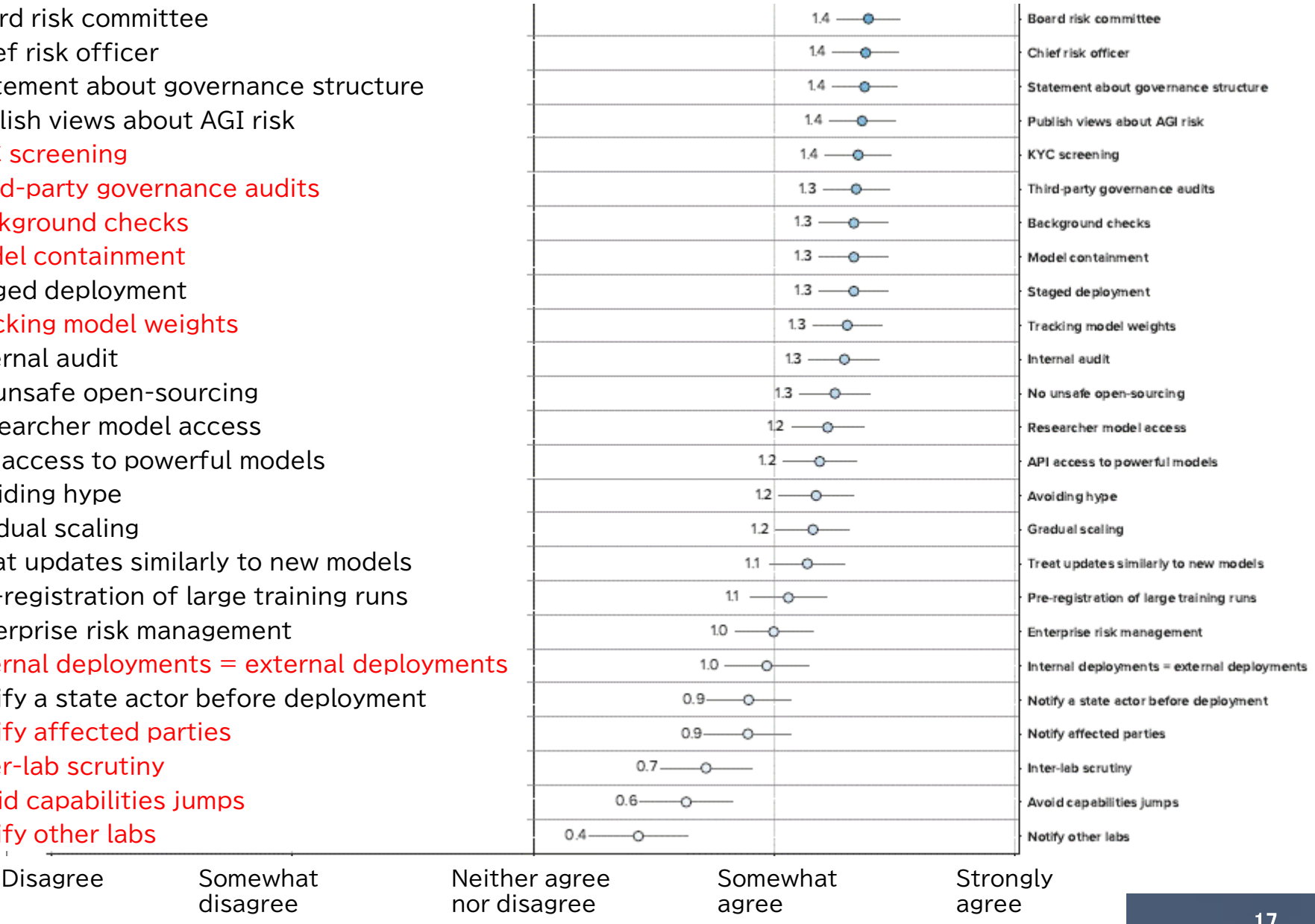
Figure 3： 平均点（賛成2，やや賛成2， やや不賛成-1, 不賛成-2

1	Pre-deployment risk assessment	1.9
2	Dangerous capabilities evaluations	1.9
3	Third-party model audits	1.8
4	Safety restrictions	1.8
5	Red teaming	1.8
6	Monitor systems and their uses	1.7
7	Alignment techniques	1.7
8	Security incident response plan	1.7
9	Post-deployment evaluations	1.7
10	Report safety incidents	1.7
11	Safety vs. capabilities	1.7
12	Internal review before publication	1.7
13	Pre-training risk assessment	1.6
14	Emergency response plan	1.6
15	Protection against espionage	1.6
16	Pausing training of dangerous models	1.6
17	Increasing levels of external scrutiny	1.6
18	Publish alignment strategy	1.5
19	Bug bounty programs	1.5
20	Industry sharing of security information	1.5
21	Security standards	1.5
22	Publish internal risk assessment results	1.4
23	Dual control	1.4
24	Publish results of external scrutiny	1.4
25	Military-grade information security	1.4
26	Board risk committee	1.4

Disagree　Somewhat disagree　Neither agree nor disagree　Somewhat agree　Strongly agree

(続)Figure 3： 平均点 （賛成2，やや賛成2， やや不賛成-1, 不賛成-2

26	Board risk committee
27	Chief risk officer
28	Statement about governance structure
29	Publish views about AGI risk
30	KYC screening
31	Third-party governance audits
32	Background checks
33	Model containment
34	Staged deployment
35	Tracking model weights
36	Internal audit
37	No unsafe open-sourcing
38	Researcher model access
39	API access to powerful models
40	Avoiding hype
41	Gradual scaling
42	Treat updates similarly to new models
43	Pre-registration of large training runs
44	Enterprise risk management
45	Internal deployments = external deployments
46	Notify a state actor before deployment
47	Notify affected parties
48	Inter-lab scrutiny
49	Avoid capabilities jumps
50	Notify other labs



2.3 回答の評価：

- 5-point Likert scaleにて、賛同度を分析
  - (-2) ： strongly disagree
  - (-1) ： somewhat disagree
  - (0) ： neither agree nor disagree
  - (1) ： somewhat agree
  - (2) ： strongly agree
  - (棄権) ： I don't know

- Demographic questions（人口統計学的質問）：詳細は、次頁＆次々頁
  - 性別 ： 男性、女性、その他の性別、言いたくない
  - 職場 ： AGIラボ、その他のハイテク企業、コンサルティング会社、シンクタンク、非営利団体、政府、学術機関、その他、言いたくない

- 調査時期

  調査期間 ：2023年4月26日から5月8日の間
  1時間のバーチャル・ワークショップが開催
  （ワークショップで、AGIの安全性とガバナンスをどのように構築し、実施できるかについての質問が行われた。 ワークショップには21人が参加）

- Anonymity（匿名性）

  調査への回答は匿名だったが、回答者に文章で返答を求める部分では、別に設定した調査とし、回答者は自分の名前と所属を記入してもらった。

2.3 分析方法

人口統計学的質問で回答された、性別と職場で、回答者を分類し、

・マン・ホイットニーのU検定(Mann-Whitney U test.)にて、回答者集団間の差を分析

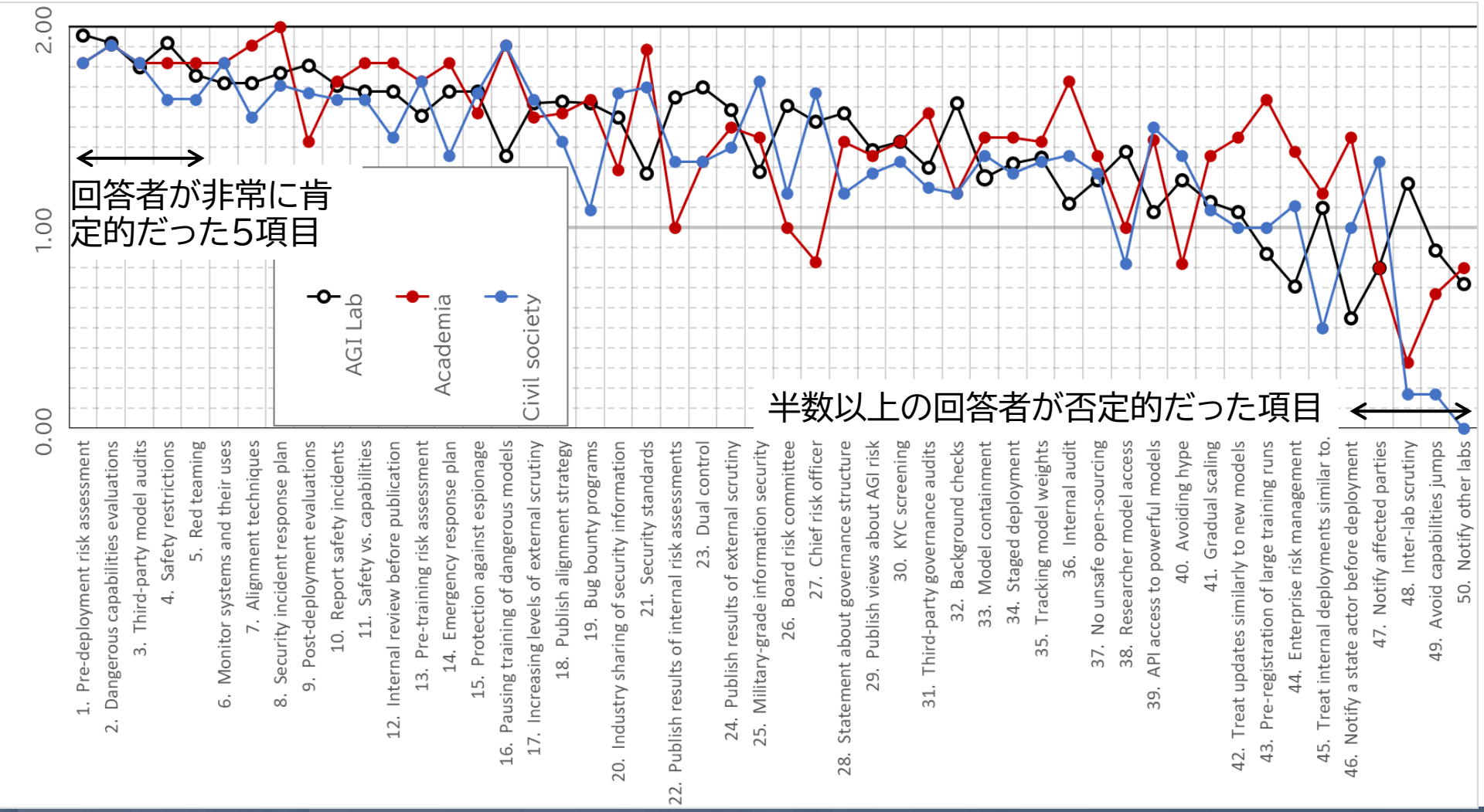
・カイ二乗検定(Chi-squared tests)にて、回答に関するグループ間の差異を分析

（注1） カイ二乗検定では、ホルム-ボンフェローニ補正(Holm-Bonferroni correction)後の有意水準と比較し、各検定の有意性を決定した。
(元のα値(0.05)を残りの検定の数で割り、p値の大きい方から小さい方へとカウントダウンし、 その後、p値をホルム-ボンフェローニ補正後の有意水準と比較し、各検定の有意性を決定した。)

・「別の性別」、「言いたくない」、「その他」のサンプル数は5(優位水準)以下だったので、分析から除外

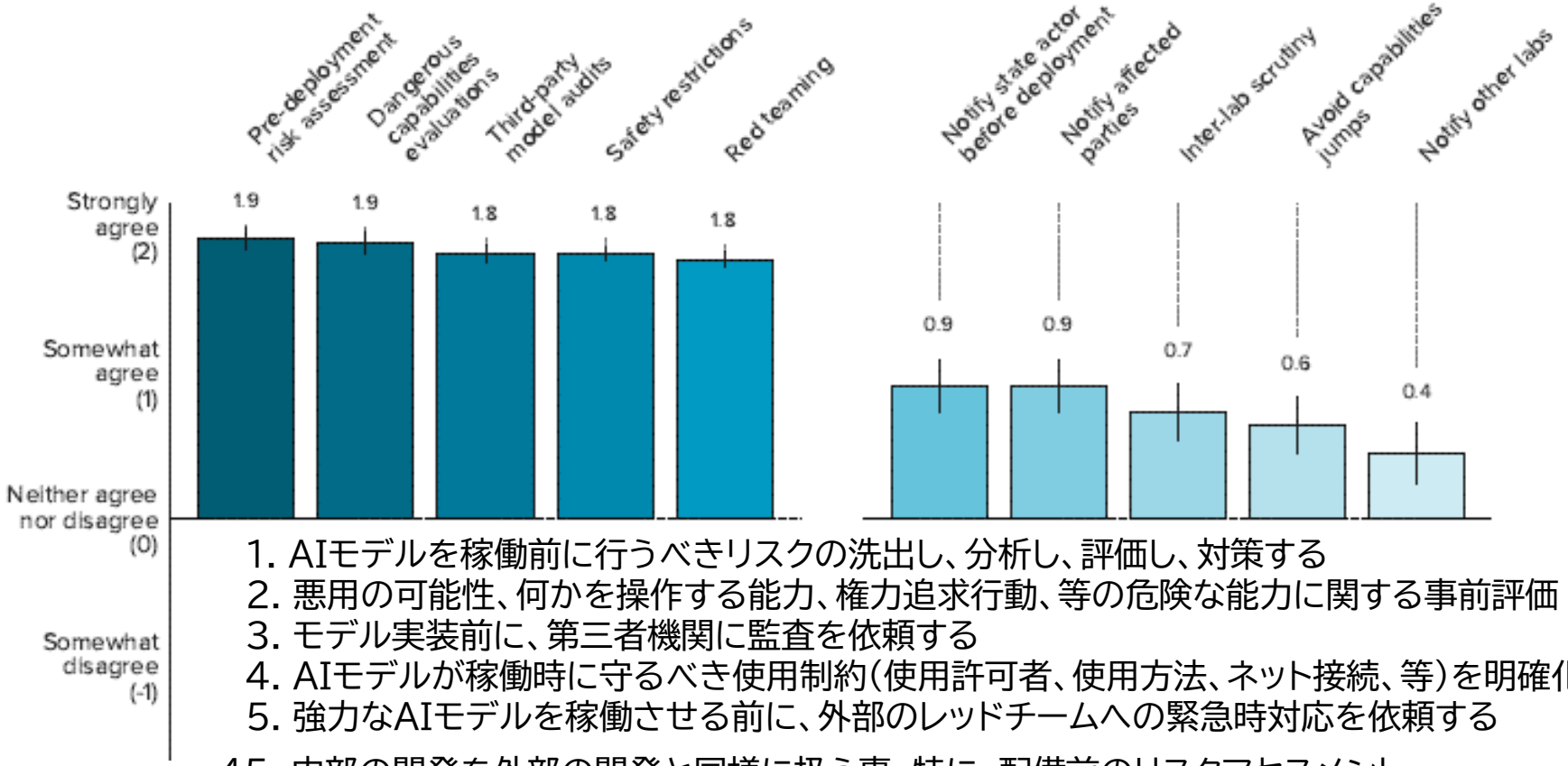
AGI Lab / Academia / Civil society 間の回答の違い（岡島作成）

Table-4 (Responses and statistics by demographic group)のMean値をグラフ化
3者の回答傾向に大差は無いが、Academia(赤)やCivil society(青)は、4548,49,50で否定的



3. Results

←――回答者が肯定的だった項目――→ ← 半数以上の回答者が否定的だった項目 →



1. AIモデルを稼働前に行うべきリスクの洗出し、分析し、評価し、対策する
2. 悪用の可能性、何かを操作する能力、権力追求行動、等の危険な能力に関する事前評価
3. モデル実装前に、第三者機関に監査を依頼する
4. AIモデルが稼働時に守るべき使用制約(使用許可者、使用方法、ネット接続、等)を明確化する
5. 強力なAIモデルを稼働させる前に、外部のレッドチームへの緊急時対応を依頼する

45. 内部の開発を外部の開発と同様に扱う事。特に、配備前のリスクアセスメント
46. 強力なAIモデルを稼働する前に、適切な国家機関に通知する事
47. 強力なAIモデルを展開する前に、そのモデルによって悪影響を受ける関係者に通知する事
48. 稼働前に他のラボの研究者が強力なAIモデルを精密検査できるようにする事
49. 既存のモデルよりはるかに高性能なモデルを導入すべきではない。
50. 強力なAIモデルを稼働する前に、他のラボに通知する事

3.1　同意の多い項目、少ない項目（Figure.4）

A1）「同意」の割合が最も高かった項目；

（2）　危険な能力の評価
（12）公表前の内部レビュー
（6）　システムとその用途の監視
（1）　稼働前のリスク評価
（5）　レッドチーム編成
（4）　安全性の制限
（3）　第三者によるモデル監査

A2）反対意見が全く無かった項目；

（2）　危険な能力の評価
（20）セキュリティ情報の業界共有
（30）KYC スクリーニング
（1）　配備前リスク評価
（18）ア ライメント戦略の公開
（4）　安全性の制限
（11）安全性対能力

D1）「同意しない」の割合が高かった5項目；

（49）能力のジャンプを避ける
（48）ラボ間の精査
（37）安全でないオープンソース化は行わない
（42）アップデートを新モデルと同様に扱う
（50）他のラボに通知する

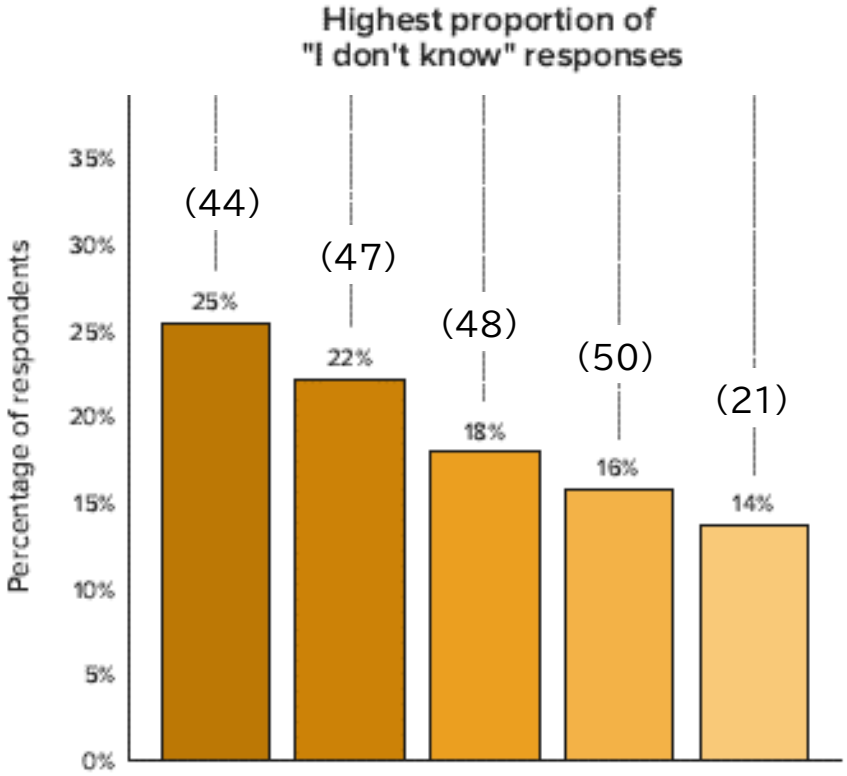
D2）「平均値」として低かった項目；

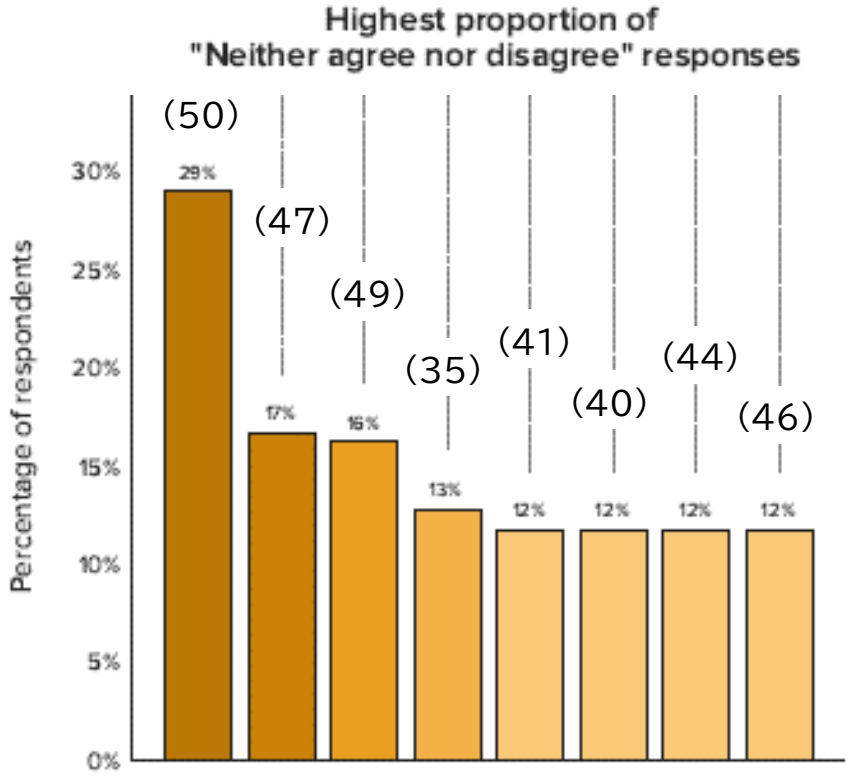
（47）影響を受ける当事者に通知する
（46）配備前に政府機関に通知する

「同意しない」、「やや同意しない」が多かった上記のような項目でも、態度を明確にしない人を「棄権」と見做すと、賛否が半々程度だった。

3.1 判断を保留した人の割合が多かった項目（Fig.5）

（44） 企業リスク管理(ERM)フレームワーク(NIST AIリスク管理フレームワークやISO 31000など)の導入と準拠
（47） 強力なAIモデルを稼働する前に、そのモデルによって悪影響を受ける関係者に通知する事
（48） 稼働前に他のラボの研究者が強力なAIモデルを精密検査できるようにする事
（50） 強力なAIモデルを稼働する前に、他のラボに通知する事
(21)ISO/IEC 27001、NIST Cyber Security Framework、等のセキュリティ基準への準拠
（49） 既存のモデルよりはるかに高性能なモデルを稼働させるべきではない。
（35） AIラボは、最大性能のAIモデルのWeightsを全てコピーしたシステムを保有する事
（41） 最大規模のトレーニングを実行する時には、(評価を行いながら)計算量を段階的に増やす事
（40） AGIの性能に関する誇大広告を避ける事(例えば、結果を誇張したり、注目を集めるような方法での発表)
（46） 強力なAIモデルを稼働する前に、適切な国家機関に通知する事





3.1　回答者の同意の度合い（Figure.4）

3.1.1　回答者が肯定的だった項目（Figure.4）

- 「強く同意」が過半数となったのは、提示した実施事項の56%。
- 「強く同意＋やや同意」が過半数となったのは、提示した実施事項の98%

3.1.2　（ある程度の人数の）回答者が否定的だった回答者（Figure.4）

- 実施事項に対して、「ややそう思う」又は「強くそう思わない」と回答したのは、4.6%

- 図2は、AGIのベストプラクティスになりうる各項目について、「強くそう思う」、「ややそう思う」、「どちらともいえない」、「ややそう思わない」、「強くそう思わない」、「わからない」と答えた回答者の割合を示している。

- どのベストプラクティスについても、「どちらかといえばそう思う」「強くそう思う」という回答が過半数を占めた（50%以上）。実際、どの項目に関しても、「能力のジャンプを避ける」という項目に関して、最も反対が多かったのは16.2%であった。回答者が行った全評価のうち、反対意見はわずか4.5%であった。

3.2　セクター間、及び、男女間での違い（付録Dの図6、7、8）

・AGIラボの回答者は、アカデミアや市民社会の回答者よりも、僅かに提案項目への賛同度が高かった
  - AGIラボ　　；M = 1.54
  - アカデミア　　；M = 1.16
  - 市民社会　　；M = 1.36

・アカデミアと市民社会では、平均賛同度に有意といえる差は無かった。

・ 男女の回答間に有意な差は無かった。

3.3　回答者から提案された追加的項目（付録C）

・回答者から、50のユニークな項目が提案された。（論文投稿者は2件に注目）

- AGIラボは合併条項（merge-and-assist-clause）を設けるべきであり、また何らかの内部審査委員会を設けるべきである
- 利益と社会的便益を適切にバランスさせる必要性

（注）合併条項（merge-and-assist-clause）=>　分からず。

4. Discussion

4.1 Overview of results

- 高いレベルの賛同が得られた。
  安全管理が必要との風潮(Feature)が存在することが示された。
  (今後の「ベストプラクティス明確化 ⇒ 標準化 ⇒ 法制化」の基礎なり得る)

- 但し、項目によっては「不賛同」の割合が高かった:[49]、[48]
  回答者が態度を保留する項目もあった: [44]、[50]、[47]
  (態度保留の原因は今後の調査で注意する必要がある。 不賛同を意味するかもしれない)

- 興味深いことに、AGIラボは、アカデミアや市民社会よりも賛同的だった。
  (但し、統計的には有意差とはいえない)

GlobalPartnership on AI (GPAI)のメンバーが多いからだろう。

- 参加者から、50の追加項目の提案があった。
  (提示案では不十分だったかもしれないので、更なる調査が必要。)

- AGIラボはガバナンスメカニズムを必要としている。

4.2 Discussion of secific results

(注) 本論の一番重要な項目

4.2.1	Development（開発管理）	： 8 項目
4.2.2	Deployment（稼働前管理）	： 8 項目
4.2.3	Post-deployment（稼働中管理）	： 4 項目
4.2.4	Risk management（緊急時対応）	： 6 項目
4.2.5	External scrutiny（外部機関による審査）	： 7 項目
4.2.6	Information security（情報セキュリティ）	： 7 項目
4.2.7	Communication（情報発信）	： 9 項目
4.2.8	Other（その他）	： 1 項目

4.2.1 Development（開発管理）

論文あり [56, 38, 6]。 ある危険な能力を検出した場合、具体的に何をすべきか、協調的な一時停止は可能かについて。

[2] Dangerous capability evaluations (M = 1.9) ····· the highest rated item
（悪用、何かの操作、権力追求行動、等の危険な能力を事前に評価する）
（注）OpenAI社は、GPT-4の公開前に、状況認識、説得、長周期計画などの危険な創発行動を評価するようコンサル企業のARC Inc.に依頼していた。[参56]（参：ARC-DA (ARC Direct Answer Questions)

[7] Alignment techniques (M = 1.7) ····· received broad support
（最先端の安全技術や、最先端安全技術へのすり合わせ技術の導入する）

[11] Safety vs capabilities (M = 1.7) ····· received broad support
（社内の一定割合の人数による安全性の向上と安全基準順守の活動を行う）

[16] Pausing training of dangerous models (M = 1.6) ····· received broad support
（一定以上に危険な能力が検出された場合の開発プロセスを一時停止する）

[33] Model containment (M = 1.3) ····· received less support
（AIモデルを、Air-Gap等による外部ネットワークからの隔離、もしくは囲い込みを行う）

[35] Tracking model weights (M = 1.3) ····· received less support
（AIラボは、最大性能のAIモデルのWeightsを全てコピーしたシステムを保有する）

[41] Gradual scaling (M = 1.2) ····· received less support
（最大規模のトレーニングを実行する時には、評価を行いながら計算量を段階的に増やす）

[43] Pre-registration of large training runs (M = 1.1) ····· somewhat agree
（一定規模以上のトレーニング実施を予定している場合、適切な国家機関に登録する）

4.2.2 Deployment（稼働前管理）

[4] Safety restrictions（M = 1.8） ····· strongly agreed
（AIモデルが稼働時に守るべき使用制約（使用許可者、使用方法、ネット接続、等）を明確化する）

[34] Staged deployment（M = 1.3） ····· somewhat agreed
（AIモデルの安全性を確認しながら、段階的に規模拡大と性能向上を進める）

[39] API access to powerful models（M = 1.2） ····· somewhat agreed
（APIを介してのみ強力なAIモデルを稼働できるようにする）

執筆者は強く推奨したが、賛同は弱かった。

[37] No unsafe open-sourcing（M=1.3） ····· somewhat agreed
（十分に安全であることを実証できない限り、強力なAIモデルをオープンソース化しない）

[30] know-your-customer（KYC）screenings（M = 1.4）····· moderately supported
（強力なAGIモデルを提供する前に、Know-Your-Customerスクリーニングを実施する。

OpenAI社は、この対策を最重要視していた[58]。

[42] Model updates similarly to new models（M = 1.1）····· moderately supported
（配備されたモデルの大幅な更新（例えば、追加的な微調整）を、その最初の開発及び配備と同様）

[45] Treat internal deployments similarly to external deployments（M = 1.0）
（内部の開発を外部の開発と同様に扱う事。） ····· moderately supported

[49] Avoid capabilities jumps.（M = 0.6） ····· the least supported
（既存のモデルよりはるかに高性能なモデルを稼働させるべきではない）

4.2.3 Post-deployment.（稼働中管理）

[6] Monitor systems and their uses （M = 1.7). ····· strongly agreed
（稼働中のシステムの使われ方や社会への影響状況を監視する）

OpenAI [19, 20] や Google DeepMind [36] は、このような対策を実施中

[9] Post-deployment evaluations （M = 1.7） ····· strongly agreed
（危険性の観点から、AIモデル稼働後の獲得能力や使用され方に関する継続評価する）

[10] Report safety incidents （M = 1.7). ····· strongly agreed
（事故発生やニアミス発生時の適切な国家機関（のデータベース）への状況報告を行う）

AI Incident Database の論文あり[45]

[14] Emergency response plan （M = 1.6） ····· strongly agreed
（システムの電源切断、出力の停止、アクセス制限、等の緊急対応計画を策定する）

4.2.4 Risk management.（緊急時対応）

[1] Pre-deployment risk assessment (M = 1.9) ····· strongly agreed
（AIモデルを稼働前に行うべきリスクの洗出し、分析し、評価し、対策する）

OpenAI [19, 20] や Google DeepMind [36] は、このような対策を実施中

[13] Pre-training risk assessment. (M = 1.6) ····· strongly agreed
（強力なAIモデルのトレーニング実行前のリスクアセスメントを実施する）

[26] Board risk committee. (M = 1.4), ····· somewhat agreed
（取締役会リスク委員会（AGIに関するリスク管理実務を監督する常設委員会）を取締役会内に設置する）

リファレンスあり。various statements about (26) risk governance [84, 43]

[27] Chief risk officer. (M = 1.4), ····· somewhat agreed
（リスク管理を担当するSenior Executive（Chief Risk Officer, CRO）を置く）

[36] Internal audit. (M = 1.3) ····· somewhat agreed
（上級管理職から組織的に独立し、取締役会に直接報告する内部監査チームを持つ事）

リファレンスあり。[68, 70].

[44] Enterprise risk management. (M = 1.0) ····· less supported
（企業リスク管理（ERM）フレームワーク（NIST AIリスク管理フレームワークやISO 31000など）の導入）

the NIST AI Risk Management Framework [53] and ISO 31000 [34]

4.2.5　External scrutiny.（外部機関による審査）

[3]　Third-party model audits（M = 1.8）　・・・・・ strongly agreed
（モデル稼働前に、第三者機関に監査を依頼する）

OpenAI社は、将来の第三者機関監査[3]とbug bounty program [55]をアナウンス済
（注）関連論文多い [65, 18, 50, 22, 66, 51]

[5]　Red teaming（M = 1.8）　・・・・・ strongly agreed
（AIモデルを稼働前に行うべきリスクの洗出し、分析し、評価し、対策する）

既に、Red teaming 契約は、多くのAGIラボで実施されている
OpenAI社[48, 56]、Google DeepMind [62]、Anthropic [23].

[19]　Bug bounty programs（M = 1.5）　・・・・・ strongly agreed
（未知の脆弱性や危険な機能を発見した（外部の）人に対して報酬金を支払う制度の制定）

[17]　Increasing level of external scrutiny (M = 1.6)　・・・・・ strongly agreed
（AIモデルの能力増大に比例させた、外部からの精密査察のレベルUP）

但し、査察が具体的に何を意味するのかは不明である（レッドチームの大規模化、異なる手法の組み合わせ、調査時間の増加など、という可能性がある）

[31]　Third-party governance audits（M = 1.3）　・・・・・ slightly less supported
（自社のガバナンス構造に対する第三者監査を委託実施する）

[48]　Inter-lab scrutiny（M = 0.7）　・・・・・ very low agreement
（稼働前に他のラボの研究者が強力なAIモデルを精密検査できるようにする）

[38]　Researcher model access.（M = 1.2）　・・・・・ slightly less supported
（AIモデルへのAPIアクセスを独立した研究者に提供する事）

4.2.6　Information Security（情報セキュリティ）

[8] Security incident response plans（M = 1.7）　・・・・・ strongly agreed
（サイバー攻撃等のセキュリティ問題発生時の対応プランを明確化する）

[15] Protection against espionage（M = 1.6）　・・・・・ strongly agreed
（国家によるスパイ活動や、産業スパイのリスクに対処するための十分な対策を行う）

[21] Implementing security standards（M = 1.5）　・・・・・ somewhat agreed
（ISO/IEC 27001、NIST Cyber Security Framework、等のセキュリティ基準に準拠する）

[20] Industry sharing of security information（M = 1.5）　・・・・・ somewhat agreed
（脅威となる活動や出来事に関する情報の他のAGIラボと共有する）

[23] Dual control（M = 1.4）　・・・・・ somewhat agreed
（複数の人間による重要事項の決定。試作⇒量産への移行、トレーニングデータセットの変更、量産中の改版）

[25] Military-grade information security（M = 1.4）　・・・・・ somewhat agreed
（AIモデルの能力増大に比例させた情報セキュリティ管理能力の向上の最終形として、諜報機関の能力や国家防御のレベルを目指す）

[44] Enterprise risk management frameworks（M = 1.0）　・・・・・ less supported
（AIモデルへのAPIアクセスを独立した研究者に提供する事）

4.2.7 Communication（情報発信）

[12] Internal review to assess potential harms from that research（M = 1.7）
（研究結果を公表前に、危害を及ぼすような潜在能力を持つのかを社内審査する）・・・・・ strongly agreed
（注） 広範な議論あり[21, 60, 8, 83, 17, 27, 4, 7, 73, 16, 83]。

[18] Publish statements about their alignment strategy（M = 1.5）・・・・・ somewhat agreed
（システムが安全で基準に整合的であることを確実とするための戦略の公告）

[24] Publish results of external scrutiny（M = 1.4）・・・・・ somewhat agreed
（外部精査結果またはその概要の公表 ）

[29] Views about AGI risk（M = 1.4）・・・・・ somewhat agreed
（開発するAGIが生みうるリスクと利点にどの程度コミットするのかについての見解の公）

[28] Statement about governance structure（M = 1.4）・・・・・ somewhat agreed
（AIモデルの開発と展開に関する重要決定をどのように行っているかについての公表）
（注） AGIラボからアラインメント戦略[41, 40, 57, 5, 38]やAGIリスクに関する見解[3, 5, 2, 74]

[40] Avoid hype when releasing new models（M = 1.2）.・・・・・ somewhat agreed
（AGIの性能に関する誇大広告を避ける事（例えば、結果を誇張したり、注目を集めるような方法での発表）
（注） 驚くべきことに、賛同は少なかった

[46] Notify appropriate state actors（M = 0.9）・・・・・ less supported
（AIモデルへのAPIアクセスを独立した研究者に提供する事）

[47] Notify affected parties.（M = 0.9）・・・・・ less supported
（強力なAIモデルを稼働する前に、そのモデルによって悪影響を受ける関係者に通知する）

[50] Notify other labs（M = 0.4）・・・・・ the lowest agreement
（強力なAIモデルを稼働する前に、他のラボに通知する事）

4.2.8　Others（その他）

［32］　Background checks. (M = 1.3)　　・・・・・ somewhat agreed
（稼働中のシステムの使われ方や社会への影響状況を監視する）

4.3 Policy implications（戦略への反映）

4.3.1 Implications for AGI labs.

（社内マネジメントへの反映）

4.3.2 Implications for regulators.

（規制当局案件化）

4.3.3 Implications for standard-setting bodies.

（標準化）

4.3.1　Implications for AGI labs （社内マネジメントへの反映）

- 対策の要点は、以下3点だろう。

1） 第三者による監査を導入する

[3] S. Altman (2023);　Planning for AGI and beyond.

2） AIが危険な能力を持つかどうかを評価する既存の取り組みを参考にするのが第一歩。
但し、危険な能力を検出した場合の対処内容については、未だ不明確

[6]　ARC (2023); Update on ARC's recent eval efforts.
[38]　V. Krakovna and R. Shah (2023); Some high-level thoughts on the DeepMind alignment team's strategy.
[56]　OpenAI (2023) ; GPT-4 technical report.

3） 本調査結果は、AGIラボのリスク管理手法には、改善の余地があると示唆している。
(36)内部監査機能の設置、(27)最高リスク責任者の任命、(26)取締役会リスク委員会の設置、(44)カスタマイズされた企業リスク管理フレームワークの導入を真剣に検討すべきである。

［70］ J. Schuett (2023)　AGI labs need an internal audit team.

4.3.2　Implications for regulators.（規制当局案件化）

下記のような各極の法制化にて、本論の調査結果を役立てることができる。

1）米国

- ホワイトハウスは、2023年5月23日、AI開発4社のCEOを招き、「AIに関連するリスクについての懸念を共有」を発表 [86] => 次頁以降
- 「責任あるAIイノベーションを促進」するための新たな行動を発表[85]

2）EU

- AI法が提案されている。[13, 12, 1]

[13] L. Bertuzzi; Leading EU lawmakers propose obligations for general purpose ai.
[12] L. Bertuzzi; AI Act: MEPs close in on rules for general purpose AI, foundation models.
[1] AI Now Institute, et al.; General purpose AI poses serious risks, should not be excluded from the EU's AI Act.

3）英国

- 国家AI戦略[32]；「非同盟の人工知能（AGI）の長期的なリスクと、それが英国や世界に意味する予測不可能な変化を真剣に受け止めている」
- 白書[82]でAI規制の草案を発表
- Deep Mind（英）とGoogle Brain（米）の合併の影響は不明。[31]

[32] HM Government; National AI strategy.
[31] D. Hassabis; Announcing Google DeepMind.

Blueprint for an AI Bill of Rights and related executive actions
(AI権利章典の青写真と関連する行政措置)

米国のAI権利章典（AI Bill of Rights）について
https://www8.cao.go.jp/cstp/ai/ningen/r4_2kai/siryo3.pdf

Blueprint for anAI Bill of Rights：

・ 米国科学技術政策局が、2021年10月から新たな「権利章典」の開発に着手し、AIの時代に米国国民を保護するためのAIを含む自動化システムの設計、使用、導入の指針となるべき5つの原則。 2022年10月に公表した。

① 安全で効果的なシステム
ユーザーは安全でないシステムから保護されるべきである。

②アルゴリズム由来の差別からの保護
ユーザーはアルゴリズム由来の差別を受けるべきではない。

③データのプライバシー
ユーザーは、組込みの保護機能を通じて不正なデータから保護されるべきであり、自身に関するデータがどのように使用されるかを知る権限を持つべきである。

④ユーザーへの通知と説明
ユーザーは自動化システムが使用されていることを知り、それが自身に影響を与える結果にどのようにして、またなぜ寄与するのかを理解するべきである。

⑤人による代替手段、配慮、フォールバック
適切な場合、ユーザーは必要に応じて自動化システムの使用をオプトアウトすることができ、問題が生じたときに、解決できる担当者に連絡する手段を持つべきである。

AI Risk Management Framework (AI RMF 1.0)
(AIリスク管理フレームワーク)

https://www.newton-consulting.co.jp/itilnavi/guideline/ai_rmf.html

- 2023年1月26日に、米国国立標準技術研究所 (NIST、National Institute of Standards and Technology)が発行した「AIリスクマネジメント規定」。　　(参) ISO31000

- 特徴

リスクを、「危険度」ではなく、「目的の達成を見込む上での不確かさ」と定義し、
ネガティブリスクのみならず、ポジティブリスクも対象としている。 また、
前提条件の明確化にあたって「ライフサイクル(計画～運用～)」の観点を取り入れ、
リスクを考えるにあたって「信頼性(Trustworthiness)」の観点を取り入れている。

説明及び解釈可能性(AIシステムの動作の根底にあるメカニズムを言語化できること)**とプライバシー保護、公平性**を重視し、AI開発者とAI利用者の両方が実践することを想定している。 ⇒ つまり、全ての人間と組織

- (注) Framework : 方針や役割・体制を含む枠組みのこと
(AI RMFでは、"Governance"と呼ぶ。)
Map : 前提条件の明確化やリスク特定のこと
Measurement : リスク分析のこと
Management : リスク評価とリスク対応
Safety : 人間や資産の安全
Security : 不正アクセスや外部からの攻撃に耐えられる状態

Roadmap for standing up a National AI Research Resource
(国家AI研究資源を立ち上げるためのロードマップ。)

https://www-overseas-news.jsps.go.jp/【ニュース・アメリカ】ostpとNSF、米国人工知能研究/
https://www.whitehouse.gov/ostp/news-updates/2023/01/24/national-artificial-intelligence-research-resource-task-force-releases-final-report/

- 2023年1月24日に、 全米人工知能研究資源（NAIRR：National Artificial Intelligence Research Resource）のTask Force がリリースした「人工知能の研究開発に不可欠な資源へのアクセスを拡大するための国家研究インフラを立ち上げるためのロードマップ（報告書）」

- 2022年6月10日に、大統領府　科学技術政策局（OSTP）と 米国科学財団（NSF）が、「2020年国家AIイニシアティブ法」に基づき、同タスクフォースの立ち上げていた。

- 目的： 全米におけるAI イノベーションと経済繁栄を促進するリソース拡大とサイバーインフラへのアクセスの民主化のロードマップを作成すること。
（AIを発展させる革新的なアイデア探求に参加できるようにする）

- 認識 ： AIは、我々の最も大きな願望を達成するための大きな可能性を秘めている。

4.3.3　Implications for standard-setting bodies.（標準化）

- 以下のような公的規制を立ち上げたいと、執筆者は考えている。
  - Partnership on AI, 2023 が検討する規制にて
    （ 大規模AIモデルの安全性のための共有プロトコルとして）
  - 汎用AIシステムおよび基盤モデルの開発者のための業界行動規範[80]にて
    [80] The Future Society; Industry Code of Conduct for R&D of GPAIS.
  - 既存の標準にて、AGIラボの管理を追加する機会を捉えて
  - NISTのAIリスク管理フレームワーク[53] への追加項目として
    (Barrettら[10, 11]は、不足と見做している)
  - EUのAI法[69] に対応して、CENCENELECがリスク管理規定を作成する場合
    （注）CENCENELEC とは、欧州の3つの標準化機関のうちの2つの協力体制）
- AGIラボに特化した標準を作成しようという（公的）取組みは、未だ無い。
  （岡島注：2023年5月のG7以降、G7での規制案作りが始まっている）
- 本論文の著者達は、本調査で提示した項目の中で、
  ①訓練前のリスク評価、及び、②危険な能力の評価、③十分に危険な能力が検出された場合には動作を一時停止させること
  が特に重要だと見做している。

4.4 Limitations（本調査での課題/問題点）

4.4.1 Sample limitations.

・ サンプル数(N = 51)が比較的に小さく、有意差の明確化が困難。

（第一線の専門家のうち、調査を行うべき専門家を見逃している可能性が高い）

4.4.2 Response limitations.

・回答者に、回答理由を細かく聞いていなかった。

（提案文章(プラクティス)への支持理由/不支持理由を探る必要がある）

4.4.3 Statement limitations.

・提案文章(プラクティス)の個数(50)は少なく、包括的ではなかった。

（そのため、回答者から多くの追加提案が出た）

・提案文章が短過ぎ、回答者毎に理解が違っていた可能性がある。

（そのため、項目によっては、回答保留者が多かった。

特に、なぜ」、「どのように」、「いつ」の明確化が必要だった。）

4.5　Future directions

- 実際にベストプラクティス（慣例）の確立、標準化、法制化するとなると困難が多い。

- 技術面での困難
  - （1） 具体的な評価基準（モデル監査や危険な能力評価など）の欠如
    （問題有る/無いを再現良くテストできるか？　良/不良の閾値は定義可能か？）
    （機密情報を明らかにすることなく、外部からの監査は可能か？）
    （ユーザーのプライバシーを尊重しつつ、システムを監視するのは可能か？）
  - （2） 合意された定義（「AGI」や「汎用AI」という用語など）の欠如
  - （3） 技術が急速に進化すること
  - （4） ベストプラクティスが確率するには反復が必要であり、時間がかかる
  - （5） AGIのタイムラインに関して見解の相違があること
  - （6） 既存の管理基準の多くがAGIラボの課題に対処していない
    （既存の管理基準をどのように拡充すると、AGI管理を包含できるか？）
  - （7） 様々な不確実性（AIが経済や国家安全保障に与える影響など）

- 施策面での困難
  - （1） 協調の難しさ（協調しないAGIラボがあるかもしtれない）
  - （2） 競争への誘因（安全に対して責任感の低い開発者が存在するだろう）
  - （3） 独占禁止法への懸念（例えば、AGIラボ間の協力を伴う慣行について）
  - （4） 責任への懸念（開示情報は、訴訟の証拠として利用されるかもしれない）

4.5　Future directions

・ どのようにして、それら困難に対処するか？

（1） 結局は、「監査を必要とするメカニズム(Appropriate Enforcement Mechanisms)」を作れるかどうかだ。

- メカニズムがあれば、積極的にリスクを表明し、監査の必要性を明確化できる。
- 拘束力のある規制を先ず立ち上げ、それによってAGIラボに外圧をかけるというのも一つの手だ。 一般市民のシスク意識も、外圧になりうる。
- AI開発競争が抑制されると（産業界がエコシステムを構築する状況に移行すると）、監査も実施しやすい。

（2） 今後の調査と専門家による聞き取り調査が必要

- 市民を含めた、異なる利害関係者に対する体系的包括的な調査が必要だ。
- 管理の現状（既存の慣行）について詳細な分析を行うのも有効だ。
  これにより、ギャップ分析や様々な組織による評価が可能となる。

（3） 各プラクティス案に関する今後の研究に期待

- 各プラクティスに対する専門家が何故そのように考えるか？ 実施に向けての具体的な検討事項や懸念事項を洗い出す研究も必要だ。
- システムカードの理想的なバージョンについての研究も有効だろう。
  （an idealized version of a system card）

5. Conclusion

・今は、AGIの安全性とガバナンスにとって極めて重要な時期であり、多くの異なる領域や知的コミュニティからの専門家が集まり、AGIラボが何をすべきかを議論しなければならない。

・ 提示した50項目の安全性とガバナンスの実践については、潜在的なコンセンサスがあった。 回答者の98%が、以下に対して「概ね同意する」と回答した。
> 配備前のリスク評価を実施し、
> 危険な能力を持つモデルを評価し、
> 第三者によるモデル監査を委託し、
> モデルの使用に関する安全性制限を設け、
> 外部のレッドチームを委託すべきである

・ このリストは、新たな慣例の立ち上げ ⇒ 標準化⇒ 監査⇒ 法制化と進める基礎となる。

・ 我々のワークショップの前日（2023、05, 23）、カマラ・ハリス米副大統領は、
- Sam Altman（OpenAI） - Satya Nadella（Microsoft）
- Dario Amodei（Anthropic） - Sundar Pichai（Google and Alphabet）
ら、主要AI企業のCEOをホワイトハウスに招き、AIに関連する懸念を共有した[86]。

[86] White House Web (2023、05,04)； Readout of White House meeting with CEOs on advancing responsible artificial intelligence innovation.

https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/

・本日(2023、05，23)、ハリス副大統領と政府高官は、AI企業4社のCEOと会談
- Sam Altman, CEO of OpenAI
- Satya Nadella, Chairman and CEO of Microsoft
- Dario Amodei, CEO of Anthropic
- Sundar Pichai, CEO of Google and Alphabet

・バイデン大統領も立ち寄り、「企業は、製品を展開・公開する前に、自社製品が安全でセキュアなものであることを確認する義務がある」ことを強調

・「大統領と副大統領が、AIが個人/社会/国家安全保障に与えているリスクを軽減することが必要不可欠不可欠 とした」ことは明白　(岡島注：「中口問題含む」の意)
（安全保障、セキュリティ、人権・公民権、プライバシー、雇用、民主主義的価値観に対するリスクを含む）

・政府高官達は、「AIイノベーション・エコシステムにて存在感の大きい4社のCEOには、責任ある行動(責任あるイノベーションと適切なセーフガード)が求められる」と強調
その内容は、以下にも記されている。
- AI権利章典のための青写真(Blueprint for an AI Bill of Rights)と、
- AIリスク管理フレームワーク(the AI Risk Management Framework)

・両者は、セーフガードと保護を開発・確保するため以下も関し協力することに同意
- 企業が、政策立案者と市民に対して、AIシステムの情報を公開すること
- 安全性、セキュリティ、有効性を評価、検証、検証できるようにすること
- AIシステムが悪意のあるアクターや攻撃から安全であることを保証すること

(続) バイデン政権の on-going efforts on critical AI issues

(リスク軽減を促進する責任あるAI開発に向けて進める処置。2023年5月4日)

バイデン政権の「責任あるAI開発に向けて進める処置」は、以下の既発表の内容の上に進めている。

1. Additional actions announced this morning（次頁＆次次頁）
   （2023年5月4日発表の「責任あるAIイノベーションを促進する新たな行動」
   https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

2. Blueprint for an AI Bill of Rights and related executive actions
   （AI権利章典の青写真と、2022年10月4日発表の「関連する行政措置」）
   https://www.whitehouse.gov/ostp/ai-bill-of-rights/
   https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/

3. AI Risk Management Framework
   （AIリスク管理フレームワーク ）
   https://www.nist.gov/itl/ai-risk-management-framework

4. Roadmap for standing up a National AI Research Resource
   （国家AI研究資源を立ち上げるためのロードマップ。）
   https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

責任あるAIイノベーションを促進する新たな行動（2023年5月4日）発表 ①

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

- AI開発企業は、その製品を展開/公開する前に、AIが安全であることを確認する責任がある（チャンスをつかむためには、まずそのリスクを軽減しなければならない）。
- その責任を明確化し、対策対応を要求するために、ハリス副大統領と政府高官は、本日、AIイノベーションの最前線に立つ米国企業4社（アルファベット、Anthropic、マイクロソフト、OpenAI）のCEOと会談し、「リスクと潜在的危害を軽減するセーフガードを備えた、責任ある、信頼できる、倫理的なイノベーションを推進すること」を要求する。
- 政権としては、"responsible innovation"を推進してきており、昨年秋発表の「①AI権利章典のための青写真」や「②関連行政措置」や、今年初め発表の「③AIリスク管理フレームワーク」や「④国家AI研究資源立ち上げのためのロードマップ」が関連する。大統領は2月、「⑤AIを含む新技術の設計と使用にて、偏見を根絶し、アルゴリズムによる差別から国民を保護するよう連邦政府機関に指示する大統領令」に署名した。
- 政権は、特に、AIによって引き起こされ得るサイバーセキュリティ問題、バイオセキュリティ問題、安全性問題を懸念しており、⑥主要AI企業が、国家安全保障関連部署のサイバーセキュリティ専門家の支援を得て、慣行（ベストプラクティス）に従うことも推進している。（つまり、対策時のAIモデル群やネットワーク群の保護を行うことを見込む。）

責任あるAIイノベーションを促進する新たな行動（2023年5月4日）発表 ②

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation

- 全米科学財団（National Science Foundation）を通じて、140Mドルを投資する
  新たに7つの全米AI研究所を立ち上げ、AI開発を進める。（この投資にて、研究所の総数は全米で25となり、関係組織のネットワークはほぼすべての州に拡大される。）
  イノベーションの推進に加え、これらの研究所は米国のAI研究開発インフラを強化し、多様なAI人材の育成を支援する。（それら新しい研究所は、AIを利用して、気候、農業、エネルギー、公衆衛生、教育、サイバーセキュリティにおけるブレークスルーを促進する）

- 主要7社が、Scale AI社が開発した評価プラットフォーム上で、各社の生成AIシステムをDEFCON 31で公開評価（テスト）すると約束した
  （Anthropic、Google、Hugging Face、Microsoft、NVIDIA、OpenAI、Stability AI）
  これにより、これらのモデルは何千ものコミュニティ・パートナーやAI専門家によって徹底的に評価され、モデルが政権の「AI権利章典とAIリスク管理フレームワークのための青写真」で概説された原則と実践にどのように合致するかを探ることができる。

- 行政管理予算局（OMB）は、今夏、AIシステム利用に関する政策指針案を公表し、パブリックコメントを求める
  この指針案は、AIシステムの開発、調達、利用を行う上での連邦省庁方針となるだけでなく、州政府や地方自治体、企業などがAIの調達/活用にて従うべきモデルとなる。

Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading AI Companies to Manage the Risks Posed by AI（2023,07,21）

・政権がこれらの企業からの自発的なコミットメントを確保し、AI技術の安全で安全な、および透明な開発を支援することを発表した。(Voluntary commitments – underscoring safety, security, and trust – mark a critical step toward developing responsible AI with Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI.)

・ これらの新興技術を開発している企業は、製品が安全であることを保証する責任があります。 3つの原則(安全性、セキュリティ、信頼)、責任あるAIの開発に向けた重要なステップを強調する。

・政権は同盟国とパートナーと協力して、AIの開発と使用を管理するための強力な国際枠組みを確立させる。オーストラリア、ブラジル、カナダ、チリ、フランス、ドイツ、インド、イスラエル、イタリア、日本、ケニア、メキシコ、オランダ、ニュージーランド、ナイジェリア、フィリピン、シンガポール、南朝鮮、アラブ首長国連邦、アラブリ、英国(自発的なコミットメントについてすでに相談中)

・米国は、これらのコミットメントがG-7ヒロシマプロセスの日本のリーダーシップをサポートし、補完することを保証しようとしています。

・AIのガバナンスのための共有原則を開発するための重要なフォーラムとして、さらにはAIの安全に関するサミットを開催する英国のリーダーシップ、およびAIの世界的パートナーシップの議長としてのインドのリーダーシップ。また、さまざまな国連Foraの国連および加盟国とAIについて話し合っています。

7つの大手AI企業は以下の内容を約束。(2023,07,21)
(Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI.)

・リリース前の独立した専門家によって行われるAIシステムの内部および外部のセキュリティテスト

・AIのリスクを管理する際に、業界全体および政府、市民社会、学界と情報を共有する（安全のためのベストプラクティス、保護措置を回避する試みに関する情報、および技術的なコラボレーションが含まれる。開発は、セキュリティを最優先にする）

・モデルのパラメータを保護するために、サイバーセキュリティおよびインサイダーの脅威保護措置への投資を企業が行う。（セキュリティリスクが担保されないとリリースしない）

・企業は、AIシステムの脆弱性のサードパーティの発見と報告を促進する
（システムリリース後も、問題があった場合、レポートメカニズムにより修復する）

・企業は、透かし式システムなど、コンテンツがいつ生成されるかをユーザーが確認できるようする技術を開発する。

・企業は、AIシステムの機能、制限、適切かつ不適切な使用領域を公に報告する。
（公平性やバイアスなど、セキュリティリスクと社会的リスクの両方をカバーする。）

・企業は、有害なバイアスや差別を避け、プライバシーを保護するなど、AIシステムがもたらす可能性のある社会的リスクに関する研究の優先順位付けを約束する。

・企業は、社会課題に対処するために、高度なAIシステムの開発と展開を約束する。
（AIを適切に管理すると、すべての繁栄、平等、および安全性に大きく貢献できる）

Remarks by President Biden on Artificial Intelligence (July, 21, 2023）

https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/07/21/remarks-by-president-biden-on-artificial-intelligence/

・企業は、一般に公開する前に、テクノロジーが安全であることを確認する義務がある。（システムの能力をテストし、潜在的リスクを評価し、評価結果を公開する）

・企業は、サイバー脅威からモデルを保護し、国家安全保障のリスクを管理し、必要なベストプラクティスと業界基準を共有する。 システムのセキュリティは優先しなければならない。

・企業には、人々から信頼を獲得し、ユーザーが情報に基づいて意思決定を行うことを可能にする義務がある。 これは、AIに生成されたコンテンツのラベル付け、バイアスと差別対策を応援し、プライバシー保護を強化し、子どもを保護することです。

・ 社会の最大の課題に対応するためのAI投資方法を見つけることに同意します。

・これらのコミットメントは本物であり、具体的です。
人工知能は、世界中の人々の生活を変えるつもりです。

・ 新興技術の出現する脅威について、明確で目覚ましで警戒しなければなりません。

・リスクを管理することでAIの約束を実現するには、いくつかの新しい法律、規制、監視が必要になる。私たちは協力して、適切な法律と規制を策定します。

感想

- 産業界(Big-Tech間)の協力が強力 ⇒ 政策課題 ⇒ WWで規制＆監査

- 「AIは、我々の最も大きな願望を達成するための大きな可能性を秘めている」との強烈な認識は、下記の取り組み展開にて共有されていったのだろう。

  - 2016年に、米国5大IT企業(F、A、G、I、M)によるPAIの立ち上げ。
  - 2018年に、Centre for the Governance of AI（GovAI）をオックスフォード大学から独立させ、イェール大学のGlobal Politics of AI Research Groupと合併
  - 2020年の国家AIイニシアティブ法に基づき、NAIRR(全米人工知能研究資源)下に、Roadmap for standing up a National AI Research Resource検討のタスクフォースの立ち上げ
  - 2023年5月：責任あるAIイノベーションを促進する新行動(140Mドル)
G7を通じて規制内容をWWにすり合わせ

- 結果、「ステルス開発」が一番の懸念となると思える。