Google

Advancing Responsible AI

制度の在り方検討ワーキンググループ発表資料
2026年11月

- Googleは、AIの大胆かつ責任ある開発と展開、そしてその恩恵を実現するためにすべてのステークホルダーと協力して取り組むことに尽力しています。

- 私たちは、AIが責任を持って構築・展開されれば、最も有益で変革をもたらすテクノロジーの一つになると確信しています。その計り知れない可能性を実現するためには、この重要な発展段階を、慎重に舵取りしていく必要があると考えています。

AI Safety and risk management - Googleの取組

- Googleは、AIの安全性に対して多層的なアプローチをとっています。
- フロンティアAIのリスクに対する当社の綿密なアプローチは、「フロンティア・セーフティ・フレームワーク（Frontier Safety Framework：FSF）」を基盤としています。FSFの中核をなすのは「重大な能力レベル（Critical Capability Levels：CCLs）」であり、これは適切な緩和策を講じなければモデルが深刻なリスクをもたらし得る閾値を指します。
- 当社では、主に以下の 2つのリスクカテゴリーにおいて CCLを追跡・監視しています。:
    - 悪用リスク：サイバー攻撃、CBRN、及び有害な情報操作・誘導（モデルが悪意ある行為者の能力を大幅に引き上げてしまうかどうかの評価）
    - 機械学習の研究開発及びミスアラインメントのリスク：社会が AIリスクを管理・制御する能力を損なう恐れのあるシナリオ
- CCLと合わせて、当社は「追跡対象の能力レベル（ Tracked Capability Levels：TCLs）」も監視しています。これはCCLよりも低い能力の閾値で重大なリスクを捉えるものであり、リスクの度合いに応じた緩和策を適用しています。

- 当社のフロンティアモデルは、リリース前の綿密なテストを受け、ライフサイクル全体にわたる継続的なセキュリティ対策及び展開時のリスク緩和策（モデルの重み流出の防止、安全性のためのファインチューニング、常時モニタリングなど）が講じられています。また、社内外の両チームによる広範なレッドチーミングも実施しています。

- 私たちは、実行時のセーフティ・クラシファイアや思考連鎖（ Chain-of-Thought：CoT）のモニタリングも適用しています。

- リリース後も、悪用がないか能動的に監視し、アカウントレベルでの制御を実施しています。

- 社内で展開されるモデルについては、モデルのアラインメントに次ぐ「第二の防衛線」として、「 AIコントロール・ロードマップ（ AI Control Roadmap）」を運用しています。例えば、最高水準のサイバーセキュリティモデルを用いたプロアクティブなスキャンを通じて、評価環境及びトレーニング環境の堅牢化を積極的に進めています。

AI Safetyに関する制度について

- 私たちは、「世界で最もAIを開発・活用しやすい国」を目指す日本のビジョンを支持します。

- アジャイルかつバランスの取れたアプローチをとる現在の日本モデルAI推進法及びソフトローに基づく枠組み）は、グローバルな道標となっています。

- フロンティアリスクに関する制度的枠組みを検討するにあたっては、以下の要素を考慮することが重要です。
    - 第一に、主要なAI開発事業者によるフロンティアAIリスク管理の様々な取組：Googleは、フロンティアリスクに関する強固な管理及び緩和プロセスを確立しており、それらをモデルカード（Model Cards）、安全性レポート、さらには「広島AIプロセス（HAIP）透明性レポート」などの透明性に関する枠組みを通じて、適時かつ公に共有しています。
    - 第二に、フロンティアAIのリスクは絶えず変化しており、それらを評価する科学的手法はいまだ研究途上の領域にあって、共通の基準やベンチマークが確立されていないという事実です。

- 上記を踏まえ、フロンティアAIの規制において、硬直的なアプローチは避けるのが賢明と考えます。むしろ現段階では、フロンティアAIの共通基準や情報開示を優先し、各国のAIセーフティ・インスティテュート間での基準作りや連携に重点を置くことが、最も効果的なガバナンスのアプローチとなります。

- 私たちは、潜在的な悪影響に対処するための様々な既存法が日本には既に整備されていると考えていますが、日本が制度的枠組みの次のステップを検討する場合、以下の原則を念頭に置くことが重要と考えます：

1. アジャイルでイノベーション志向であること
   - 国際的な連携を図りつつ、AIの導入・普及を妨げる恐れのある提供前の義務の導入は避けること。
   - 営業秘密を侵害したり、プライバシーやセキュリティ上のリスクを生じさせたり、あるいは商業上の機密情報の提供を企業に強制したりするような情報開示要件は避けること。

2. 明確で重複のない規制要件
   - 例えば、AIのインシデント報告要件を設ける場合、既存のサイバーインシデント報告に関する規制をはじめとする既存の報告制度と整合させ、また重複を最小限に抑えるべきです。また、対象は日本国内において確認された重大な被害を実際に引き起こしたインシデントまたはセキュリティや安全性に対する差し迫った脅威をもたらすインシデントに絞るべきであり、そのインシデントの状況を最もよく把握し、制御できる立場にある事業者を報告主体とすべきです。

3. 業界全体にわたるリスクベースのアプローチ
   - 企業やモデルの規模、企業の出身国、あるいはモデルがオープンウェイトかプロプライエタリかを問わず、AIエコシステム全体のすべての事業者に対して安全基準を適用すること。

ご清聴ありがとうございました。

Google