経済産業省 第八回 AI事業者ガイドライン検討会
議事要旨

令和 8 年 2 月 10 日(火)
11:00～13:00
オンライン会議

冒頭挨拶

- AI 事業者ガイドラインの令和 7 年度の更新内容および AI 事業者ガイドライン活用ガイド(案)について、事務局が案をまとめた。委員各位のご意見をいただきたい。
- 前回の検討会では、多義的な用語の整理、主体区分の明確化について意見をいただいた。また、諸外国の法制度との評判・平仄を考慮すべき一方で、我が国特有の実務慣行への配慮をすべきとのご指摘もいただいた。AI 事業者ガイドライン更新にあたっては、様々な観点のバランスを図ることが、重要な視点。
- AI 事業者ガイドライン全体の更新案および AI 事業者ガイドライン普及の観点から作成した活用ガイド案について、御意見を賜りたい。

◆ 全体討議

■ AI 技術の動向の反映

- AI エージェント等最新動向を踏まえており、良い内容と評価している。
- 以下内容はセキュリティの原則論の内容を含むため、本ガイドラインに含めるべきか要議論と考えるが、AI エージェントの特徴である、人が関与せずに自動で処理を実行し続ける点について、意図しない挙動が起きた際の影響を抑えるための実行環境の分離やエージェントに与える権限の最小化等について追記するのも一案である。
- フィジカル AI については、社会実装に当たり、従来よりも高い水準の安全性・フェールセーフ設計が求められる。製造物責任法との関係も含め、記載範囲については要議論である。
- AI エージェントについては、触れること自体は差し支えないと考える。一方で、定義や具体的リスクに関する記載については、次年度以降に追加する方針とすることが適当と考える。OECD においても AI エージェントおよびエージェンティック AI の定義等を整理している段階にあるため、それが公表されてから当該内容を参照する形が望ましいと考える。
- フィジカル AI においては物理的な危害というリスクが増えることに対して、「制御可能性」(Controllability)が重要になると考える。2017 年「国際的な議論のための AI 開発ガイドライン」に示されていたような制御可能性の原則を、努力義務として次年度以降追記いただきたい。こうした原則を含めつつ、日本において罰則のないソフトローが有効に機能していることを国際的に示すことで、日本の AI 製品への信頼性向上や競争力強化に繋がると考える。
- フィジカル AI については次年度以降詳細に記載を追加するのが良いと考える。フィジカル AI と自律型ロボットとの違いを明確に説明することが容易ではないと考

えるためである。また、フィジカル AI の本来の目的は、フィジカル空間において未だ十分にデータ化されていない事象をも学習対象とし得る点にあると考えられるが、その点が曖昧なままフィジカル AI を定義することには一定の注意を払う必要があると考える。

- AI エージェントやフィジカル AI の記載追加は良いと考える。ただ、技術進展の最中で、定義やリスクを決定的なものとして記載することは時期尚早と感じるため、今後も変更され得ることを明記することが重要である。
- AI エージェントの出力の根拠が説明できず、事業者が一般消費者からの問い合わせに十分な説明ができない場面が更に増えると懸念する。一般消費者からはどのような AI エージェントを使用しているか見えにくいため、これまで以上に企業の信頼性確保の重要性が高まると考える。同時に、利用者側の AI リテラシー向上の必要性も重要になる。
- AI エージェントとエージェンティック AI について、一般的な国内のビジネス分野では両者をまとめて AI エージェントと呼んでいると認識しており、現時点で両者を区別するのはあくまで学問的な話に過ぎないのではないかと考えている。そのため、AI エージェントとエージェンティック AI を区別する際は、一定の注意を払う必要があると考える。
- AI エージェントのリスクとして「誤注文」の記載があるが、「誤り」かどうかは目的によって変わり得るため、AI エージェントが誤った注文をしたというより、人間の意図と異なる行動をしたというのがより正確な内容と考える。また、AI エージェントを使用した場合でも最終的には何かしら人間の判断が必要になる場合もある点を踏まえると、AI エージェントが行うのは「判断」でなく「推論」ではないかとも捉えられるため、「判断」という表記は見直すべきと考える。
- フィジカル AI による肖像権やプライバシー侵害のリスクについては、AI が搭載された防犯カメラ・導線解析カメラ等でも同様に生じ得る課題であり、必ずしもフィジカル AI 固有の問題ではないと考える。フィジカル AI の本質的なリスクは物理世界に働きかける点だと考える。
- AI エージェントが内部データを意図せず外部へ送信するという内容は AI エージェントの不正操作によっても生じるため、「機密情報の流出」のみならず「悪用」という面でも記載すべき事項ではないか。
- エージェンティック AI やフィジカル AI の詳細な記述は時期尚早と考える。仮に言及するのであれば、AI のスケーラビリティにより、個々の行為は問題なくとも大量・同時実行によって社会システムとの不整合が生じ得る点に触れるのが良いと考える。

■ AI によるリスクの追加・見直し

- オープンなモデルを追加学習してガードレールを外し攻撃に利用する手法や、プロンプトインジェクションによって既存モデルのガードレールを外す手法も発達してきている。サービス提供時には悪意ある利用者を検知して停止できる一方、追加学習されたモデルはガードレールを外したまま悪意を持って利用できる状態で流通してしまう。そのため、モデルが正規の AI 開発者によって提供されたものかというモデルの真正性、またベリフィケーションについての追記が必要である。

- リスク分類について、技術的・社会的リスクは How ・What の関係であると考えられるため、How・What のマトリクス型で各リスクを整理するのが実態に即しているのではないか。
- 「トリアージにおける差別」を修正することに異論はない。しかし、リスク分類については依然として課題があると考える。例えば、人格権の侵害やディープフェイクポルノといった多くのリスクがある中で、これらがどの分類に該当するのかが不明瞭であり、恐らく全て「悪用」に該当するのではないかと考えている。このように、全体的にリスクの分類を見直すべきだと考える。
- AI によるリスク「生命等に関わる事故の発生」について、自動運転車が例として挙げられているが、自動運転車が悪者であると捉えられかねない記載は見直す必要があると考える。また、「大規模な」事故という表現は、被害者が多いという意味にもとられるため、「深刻な」あるいは「生命・身体に関わる」に変更するのが良いと考える。
- リスクの対策を考えるだけではなく、適切なリスクの許容度（リスクアペタイト）を設定するということも重要だと考える。
- AI の種類（評価/判断に用いる AI、生成 AI、エージェンティック AI、フィジカル AI）ごとに想定されるリスクが大きく異なるため、種類別に生じるリスクや問題を検討する必要性があると考える。
- AI を使わないこと自体がリスクとなり得るため、適切なリスク許容度を設定することの必要性を言及いただけると良いと考える。

■ 主体区分の整理

- ノーコード開発に用いられる SaaS の AI プラットフォームも AI システムの構成要素と紐づけて整理した方が良いと考える。また、AI システムの構成要素の「データ」について、「データは収集・前処理を経てモデルに取り込まれ～」とあるが、「取り込まれ」という表現は見直す余地があると考える。
- RLHF の説明が適切ではないと考える。RLHF で作ったポリシーモデルは、直接 LLM の出力を最適化するわけではなく、あくまで LLM の学習データを評価するものだからである。
- ガードレールの定義に関して、ガードレール自体が説明責任を持つわけではないので、記載を見直す必要があると考える。尚、ガードレールの整理内容については異論はない。
- AI 開発者と AI 提供者の責任分解は今後より複雑になると考える。AI 開発者が開発した AI モデル・AI システムのブラックボックス化が進み、AI 提供者がそれをそのまま受け入れることが多くなると想定されるが、その場合 AI 提供者はどこまで責任を持つのか・法的責任の所在はどうなるのかを、実務と乖離しないよう記載すべきと考える。また、AI 利用者と AI 提供者の関係性が複雑化する中で、責任分解の整理が必要と考えるが、AI 提供者に役割を過度に寄せることには懸念がある。最後に、以上の内容と併せて具体的リスクへの実務的に有効な対応の在り方も検討すべきと考える。
- 主体区分について詳細に整理されており、良い内容と評価している。別添では技術的な対処法等が AI 開発者側に集中して記載されている。そのため、AI 提供者

向けに、技術的な対処法等が別添第 3 部に記載されている旨を追記するのが、読者からも分かりやすくて良いと考える。

- RAG で呼び出すデータ等を推論用データに含めるという整理は分かりやすい。また、RAG の構築は AI 提供者側の責任となっていることも認識した。ここで、現ガイドライン「別添 4.AI 提供者向け」の箇所に「AI システム・サービスの構成及びデータに含まれるバイアスへの配慮」が記載されているものの、データについての言及が抽象度の高い表現にとどまっている。そのため、「推論用データ」には RAG 等を介して参照するデータも含まれる旨を一言でも良いので追記したほうが良いと考える。
- 類似概念の用語を増やし過ぎないよう留意が必要である。

■ ユーザビリティの改善

- 想定する読者層についてはより解像度を上げ、見るべき箇所を整理できると良いのではないかと考えている。具体には、AI ガバナンス、IT ガバナンス、利活用推進、セキュリティのような本社部門と、実際に AI を開発・提供・利用していく事業部門の 2 つの視点でそれぞれどこをどういう観点で見ると良いのかを提示してあげると現場に寄り添った形になると考える。
- AI ガバナンスの構築にあたっては、セキュリティ、データガバナンス等と一体的に取り組む必要があること、活用ガイドは AI ガバナンスを中心に記載していることをガイドライン冒頭で明記すると良い。
- AI 事業者ガイドラインを活用することのメリットを分かりやすく示すことも重要である。活用は義務ではないが、活用促進は積極的に対応していくべきと考える。
- 日本におけるソフトローの位置付けやガイドライン作成の意図を、文字数との兼ね合いになるとは思うが、活用ガイドの導入部で説明すると、よりガイドラインのゲートウェイという位置付けに近づくと感じる。
- AI 技術者や AI 開発者が活用ガイドを参照することを考慮し、本編、別添に技術的対策や具体的な実行例が記載されていること案内いただきたい。1 から 2 行程度で概要の記載があれば分かりやすい。
- 活用ガイドに関して、重要なポイントが親しみやすく取り組みやすい形で記載されていて、大変良いと感じる。
- 活用ガイドのメインターゲットが取り組み始めた方々であるとするならば、表記として「ライトユーザー」は曖昧である。「取り組み始めた導入初期層の方々」等、より具体的な表現の方がよいと感じる。
- AI ガバナンス実践に向けた準備事項の並び順について、「組織内の AI 開発・提供・利用の状況を把握する仕組み」、「ルールの策定」、「リスク評価」、「インシデント」とした方が、ガバナンス実践の要素との対応関係が分かりやすい。
- 「よくある誤解」は「活用における正しい認識」等の前向きな表現に言い換えると良い。
- AI 活用の便益として、「コスト削減・生産性向上」に加え、「新たな価値創出」等の記載を追記すると良い。
- 第 3 章冒頭の説明文「をご説明する第 3 章の構成」は、削除しても問題はない。

- 活用ガイドは、非常にわかりやすくまとめていただき、ガバナンスの取り組みがこれで進むところもあるのではと思う。
- 4 つの要素（状況把握、リスク評価、インシデント対応、ルール策定・運用）が必須ということには同意する。ただ、必ずしも体制を新規に作る必要はなく、中小規模の事業者では既存の体制を組み合わせて活用するのも一案であることを追記すると良い。

■ **AI ガバナンスに関する動向の反映**

- 「人工知能関連技術の研究開発及び活用の推進に関する法律」（以下、AI 法）の制定や AI 戦略本部の設置を踏まえ、AI 事業者ガイドラインが AI 法に基づく横断的な指針の一つであることを冒頭で明確に位置付けを記載することが重要ではないかと考える。
- AI 法など法律との関係性について、更に踏み込んで記載する余地がある。AI 事業者ガイドラインを含めた各ガイドラインに従うことが法的根拠を持ち始めていることを明記することを検討いただけると良い。
- プリンシプル・コードは知的財産保護を主眼としつつ、透明性の観点では AI 事業者ガイドラインと重なる部分がある。両者の関係性の整理を次年度ぜひ議論したい。
- 透明性（技術情報の開示）に関して整理を行う際には、セキュリティ攻撃リスクの観点も踏まえておく必要があると考える。

■ **その他**

- サイバーセキュリティや IT ガバナンスの観点も考えたとき、どこまで AI ガバナンスに含めるのかは企業の中でも判断が難しい。具体的なトピックを記載するよりも、企業内の関連部門が連携して主体的に対応する考え方が恐らく本質的には正しいと考える。
- ISO/IEC 42001 との関係性について、AI 事業者ガイドライン内に記載すべきである。国際に関する記述箇所に、ISO/IEC 42001 や EU AI Act 等と併記することが考えられる。
- 人間中心の考え方について、重要性を理解しつつも分かりにくいと感じる。人間中心を実現するための事項が細分化されてしまっており、結局ユーザーが何をすべきか分かりにくい。具体的な取り組みなどを1から 2 行程度の補足説明を追記いただきたい。
- 大企業を中心に、標準化やガイドラインの Appendix の充実を求める傾向があると認識している。一方で、分野横断的な一律基準の策定は困難であり、過度な標準化は創意工夫や競争力を損なうおそれがある。各企業が工夫しながら取り組む姿勢を促す意識づくりが必要と考える。
- AI 事業者ガイドラインは公表から約 2 年が経過し、認知度は向上しているものの、引き続き周知・認知向上が必要である。特に大学や研究機関では、事業活動に

関するガイドラインと受け取られ、最初から参照対象外と捉えられている節があると認識している。将来の事業化を見据え、研究段階から意識を持ってもらう観点から、注意喚起を含め、対象範囲に関する追記についても検討したいと考える。

- リスクが常に変化するからこそ運用段階で回し続けるというアジャイルガバナンスの考え方が十分に理解されていない。最初から完璧なリスク規定が必要だという意識を変え、アジャイルガバナンスの概念をより浸透させる必要がある。
- 「人間中心」概念の背景には、2019 年に策定された「人間中心の AI 社会原則」が現在でも適用されている。生成 AI の普及等により OECD の AI 原則も変化している中で、政府全体として「人間中心の AI 社会原則」の見直しを検討する余地がある。
- 各事業者によって誰がどこを見るかは変わってくるので、AI 事業者ガイドラインでは、AI ガバナンスでよく問題になる問題であれば、IT ガバナンスやサイバーセキュリティの話でも軽く触れてもよいのではないか。ただ、IT ガバナンスやサイバーセキュリティについてはすでに既存の体制があるので、それと AI ガバナンスをどう統合していくのかも難しいと考える。

以上