NIST AI リスクマネジメントフレームワークプレイブック

日本語翻訳版 Rev. 1.0.0 （2024 年 4 月 19 日）

当翻訳は、NIST AI RMF Playbook の内容の理解促進の一助となるよう、AI セーフティ・インスティテュート（AISI）が作成したものです。正確には原文を参照してください。

目次

はじめに

プレイブックは、AIリスクマネジメントフレームワーク（AI RMF）コア（AI RMF1.0の表1～表4）に示された成果を達成するための推奨されるアクションを提供する。提案は、AI RMFの4つの機能（Govern、Map、Measure、Manage）の中の各サブカテゴリーに沿っている。

プレイブックは、チェックリストでもなければ、そのまま従うべき一連の手順でもない。

プレイブックの提案は任意である。各組織は、各自の業界の使用事例や関心に当てはまる提案を、いくつでも（あるいは少しでも）借用することによって、この情報を活用することができる。

Govern

Govern 1

AI リスクの Mapping、Measure、Manage に関する組織全体の方針、プロセス、手続き、プラクティスが整備され、透明性があり、効果的に実施される。

Govern 1.1

AI に関わる法的・規制的要件を理解し、マネジメントし、そして文書化されていること。

概要

AI システムは、適用される特定の法的および規制的要件の対象となる場合がある。法的要件の中には、文書化、情報開示、AI システムの透明性の向上を（例えば無差別、データ・プライバシーおよびセキュリティ制御について）義務付けるものがある。これらの要件は複雑であり、適用できない場合や、アプリケーションあるいはコンテクストによって異なる場合がある。

例えば、「格差影響」のようなバイアス測定のための AI システムのテストプロセスは、法的なコンテクストの中で一律に適用されているわけではない。格差影響とは広義には、保護される特性に基づいて、ある集団に不釣り合いに害を及ぼす、表面的には中立的な政策あるいは慣行と定義される。注目すべきは、人口統計学的情報に依拠したモデリング・アルゴリズムや脱バイアス手法の中には、格差待遇（すなわち意図的差別）の法的禁止と抵触する可能性があることである。

さらに、AI システムの想定ユーザの中には、基本的なインターネット技術に対する安定した、あるいは一貫したアクセスを持たない人（「デジタルデバイド」と広く表現される現象）や、障がい（ディスアビリティあるいはインペアメント）のために AI システムとの対話が困難な人がいるかもしれない。このような要因により、異なるコミュニティが AI システムにアクセスしようとする際に、バイアスやその他のネガティブなインパクトを経験するかもしれない。このような設計上の問題に対処しないと、例えば障がい者にかかわる雇用関連の活動において、法的リスクをもたらす可能性がある。

推奨されるアクション

- 業界、業種、事業目的およびデプロイされた AI システムのアプリケーションコンテクストに特有の、適用される法的および規制的な考慮事項や要件を常に認識すること
- リスクマネジメントの取り組みを適用される法的基準に合わせる
- AI 関連の設計、開発、デプロイ活動に影響を及ぼす可能性のある、必要な法的または規制的な考慮事項に関する組織スタッフのトレーニング（および再トレーニング）の方針を維持する

透明性と文書化

組織は以下を文書化することができる

- 法規制における最低限必要な事項を含め、エンティティはどの程度まで規制環境を定義し、文書化しているか？
- システムは、適用される法律、規制、基準、ガイダンスに準拠しているか、見直されたか？
- 法規制における適用要件を含め、エンティティはどの程度まで規制環境を定義し、文書化しているか？
- システムは、適用される法律、規制、基準、ガイダンスに準拠しているかどうか、レビューされたか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

参考文献

Andrew Smith, "Using Artificial Intelligence and Algorithms," FTC Business Blog (2020).

Rebecca Kelly Slaughter, "Algorithms and Economic Justice," ISP Digital Future Whitepaper & YJoLT Special Publication (2021).

Patrick Hall, Benjamin Cox, Steven Dickerson, Arjun Ravi Kannan, Raghu Kulkarni, and Nicholas Schmidt, "A United States fair lending perspective on machine learning," Frontiers in Artificial Intelligence 4 (2021).

AI Hiring Tools and the Law, Partnership on Employment & Accessible Technology (PEAT, peatworks.org).

Govern 1.2

AI のトラストワージネスの特徴が、組織の方針、プロセス、手続きに組み込まれている。

概要

方針、プロセス、手続きは、効果的な AI リスクマネジメントの中心的な構成要素であり、個人と組織のアカウンタビリティの基本である。すべてのステークホルダは、設計と怠慢による危害の防止を義務付ける方針、プロセス、手順からベネフィットを得る。

組織の方針と手順は、利用可能なリソースとリスクプロファイルによって異なるが、AI のライフサイクルを通じて AI アクターの役割と責任を体系化するのに役立つ。このような方針がなければ、リスクマネジメントは組織全体で主観的なものとなり、時間の経過とともにリスクを最小化するよりもむしろ悪化させる可能性がある。方針またはその要約は、関連する AI アクターに理解可能である。方針は、政策と AI システムの設計、開発、デプロイ、利用をサポートするために必要な基礎となるメトリクス、測定およびテストの理解を反映している。

責任と指揮系統に関する明確な情報の欠如は、リスクマネジメントの有効性を制限する。

推奨されるアクション

- 組織の AI リスクマネジメント方針は、以下のように設計されるべきである AI システムに関する主要な用語と概念およびその目的と活用意図の範囲を定義する
- AI ガバナンスを既存の組織ガバナンスおよびリスクコントロールに合わせる
- 広範なデータガバナンスポリシーとプラクティス、特にセンシティブなデータやリスクのあるデータの活用に合わせる
- 実験デザイン、データ品質、モデルトレーニングの基準を詳述する
- リスクマッピングと測定のプロセスおよび基準の概要を説明し、文書化する
- モデルのテストと妥当性確認プロセスを詳述する
- 法的機能とリスク機能のレビュープロセスを詳述する
- モニタリング、監査、レビュープロセスの頻度と詳細を確立する
- 変更管理要件を概説する
- 社内外のステークホルダの参画プロセスを概説する
- AI システムに関する重大な懸念の報告を促進するために、内部通報者にかかる方針を確立する
- インシデント対応計画を詳述しテストする
- 正式な AI リスクマネジメント方針が、既存の法的基準、業界のベストプラクティスや規範に沿ったものであることを検証する
- AI システムのトラストワージ特性に幅広く沿った AI リスクマネジメント方針を確立する

- 正式な AI リスクマネジメント方針が、現在デプロイされている AI システムおよびサードパーティの AI システムを含むことを検証する

透明性と文書化

組織は以下を文書化することができる

- これらの政策は、AI システムの活用に対する国民の信頼と信用をどの程度醸成しているのか？
- AI システムの利用が、そのエンティティが表明している価値観や原則と一致していることを保証するために、その企業はどのような方針を策定してきたか？
- 意図したとおりに AI システムを使用することを奨励するために、エンティティはどのような方針と文書を作成したか？
- モデルのアウトプットは、社会からの信頼と公平性を育むためのエンティティの価値観や原則とどの程度一致しているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

参考文献

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021).

GAO, "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities," GAO@100 (GAO-21-519SP), June 2021.

NIST, "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools".

Lipton, Zachary and McAuley, Julian and Chouldechova, Alexandra, Does mitigating ML's impact disparity require treatment disparity? Advances in Neural Information Processing Systems, 2018.

Jessica Newman (2023) "A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle," UC Berkeley Center for Long-Term Cybersecurity.

Emily Hadley (2022). Prioritizing Policies for Furthering Responsible Artificial Intelligence in the United States. 2022 IEEE International Conference on Big Data (Big Data), 5029-5038.

SAS Institute, "The SAS® Data Governance Framework: A Blueprint for Success". URL

ISO, "Information technology — Reference Model of Data Management, " ISO/IEC TR10032:200.

“Play 5: Create a formal policy,” Partnership on Employment & Accessible Technology (PEAT, peatworks.org).

"National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity.

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0." National Institute of Standards and Technology (NIST), January 16, 2020.

“plainlanguage.gov – Home,” The U.S. Government.

Govern 1.3

組織のリスク許容度に基づき、必要なリスクマネジメント活動のレベルを決定するためのプロセスと手順が整備されている。

概要

どのような組織においても、リスクマネジメントのリソースは有限である。適切な AI ガバナンス方針は、効果的なリスクマネジメントを確保するために、AI システムにとって最も重要な問題にリソースを割り当てるための、リスクのマッピング、測定、優先順位付けを明確にする。方針は、マッピングされ測定されたリスクを標準化されたリスク尺度に割り当てるための体系的なプロセスを規定することができる。

AI のリスク許容度は、無視できるものから重大なものまで、それぞれ、ほとんどリスクがないものから、人的、評判的、財務的、環境的に救いようのない損失をもたらす可能性のあるリスクまである。リスク許容度の評価方針は、さまざまなリスク源（例えば、財務、業務、安全と福祉、ビジネス、評判またはモデル・リスクなど）を考慮する）。典型的なリスク測定の方法は、測定または推定されたインパクトとインパクトの可能性を、掛け算や定性的に組み合わせてリスクスコア（リスク≒影響×可能性）とする。次に、このスコアをリスクの尺度に当てはめる。リスクの尺度は、レッド・アンバー・グリーン（RAG）のような定性的なものである場合もあれば、シミュレーションや計量経済学的アプローチを伴う場合もある。インパクトアセスメントは、マッピングされたリスクの重大性を理解するための一般的なツールである。最も徹底した AI リスクマネジメントアプローチでは、すべてのモデルにリスクレベルが割り当てられる。

推奨されるアクション

- 例えば、AI のライフサイクルの主要な段階における定期的なインパクトアセスメント、システムのインパクトやシステム更新の頻度との関連付けなどを通じて、AI システムの潜在的なインパクトを測定または理解するためのメカニズムを定義する方針を確立する
- AI のライフサイクルの主要な段階において、AI システムのインパクトの可能性とその大きさを測定または理解するためのメカニズムを定義する方針を確立する
- 潜在的な AI システムのインパクトを測定するための評価尺度を定める方針を確立する。評価尺度は、RAG のような定性的なものであってもよいし、シミュレーションや計量経済学的アプローチを伴うものであってもよい
- 例えば、マッピングされたリスクのインパクトと可能性の乗算または組み合わせ（リスク≒影響度×可能性）のような、AI システムまたはその重要な構成要素について、全体的なリスク測定のアプローチを割り当てるための方針を確立する
- 組織の AI ポートフォリオ全体で有効な、統一されたリスク尺度にシステムを割り当てるための方針を確立する（例えば、文書テンプレート）。また、AI システムのライフサイクルの中で、リスク許容度やリスクレベルが変化する可能性があることを認識すること

透明性と文書化

組織は以下を文書化することができる

- システムのパフォーマンスメトリクスは、リスク許容度の決定にどのように反映されるのか？
- AI システムの活用が組織のリスク許容度に合致していることを保証するために、エンティティはどのような方針を策定したか？
- エンティティのデータセキュリティとプライバシーのアセスメントは、リスク許容度の決定にどのように反映されるか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

参考文献

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019).

Brenda Boultwood, How to Develop an Enterprise Risk-Rating Approach (Aug. 26, 2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023.

GAO-17-63: Enterprise Risk Management: Selected Agencies' Experiences Illustrate Good Practices in Managing Risk. URL

Govern 1.4

リスクマネジメントプロセスとそのアウトカムは、透明性のある方針、手順、その他の組織のリスク優先順位に基づいた管理を通じて確立される。

概要

文書化と透明性に関する明確な方針と手順は、AIのライフサイクル全体にわたる「Map」「Measure」「Manage」機能の役割と責任を伝える取り組みを促進し、強化する。標準化された文書化は、組織がAI のリスクマネジメントプロセスを体系的に統合し、アカウンタビリティの取り組みを強化するのに役立つ。例えば、作業成果物の文書に連絡先情報を追加することにより、AI 関係者はコミュニケーションを改善し、作業成果物の所有権を高め、潜在的に作業成果物の品質に関する検討を強化することができる。文書化することで、システムの再現性と堅牢性の向上に関連した、下流におけるベネフィットが生まれるかもしれない。適切な文書の保管とアクセス手順により、ネガティブインシデント発生時に重要な情報を迅速に取り出すことができる。説明可能な機械学習の取り組み（モデルと説明方法）は、AI アクターによるレビューと解釈のための追加情報を導入することで、技術文書作成の実践を強化することができる。

推奨されるアクション

- 特に、以下に関連する情報について、文書化方針を確立し、定期的に見直す
  - AI アクターの連絡先
  - ビジネスの正当性
  - 範囲と用途
  - 予想される潜在的なリスクとインパクト
  - 仮定と限界
  - 学習データの説明と特徴
  - アルゴリズムの方法論
  - 代替アプローチの評価
  - 出力データの説明
  - テストおよび妥当性確認結果（説明的な視覚化および情報を含む）
  - 下流と上流の依存関係
  - デプロイメント、モニタリング、変更管理の計画
  - ステークホルダ参画計画
- AI システムの文書化方針が組織全体で標準化され、最新であることを確認する。
- モデルに関する文書をインベントリするシステムの方針を確立し、その完全性、有用性、有効性を定期的に見直す。
- リスクマネジメントプロセスの有効性を定期的に見直す仕組みを確立する。
- リスクマネジメントプロセスやアプローチの有効性を評価し、その結果に基づいて軌道修正する責任を負う AI アクターを特定する。

- AI の使用や、インパクトのアセスメント、監査、モデルに関する文書、妥当性確認およびテストの結果などのリスクマネジメント資料の公開に関する方針とプロセスを確立する。
- さまざまな種類の透明性ツールの使用と有効性を文書化し、レビューし、モデルが使用される時点の業界標準に従う。

透明性と文書化

組織は以下を文書化することができること

- エンティティは、関連するステークホルダに役割、責任および権限の委任をどの程度明確にしているか？
- AI システムの設計、開発、デプロイメント、アセスメントおよびモニタリングに携わる人員の役割、責任、権限の委任は何か？
- AI のデプロイ後、精度などの適切なパフォーマンスメトリクスをどのようにモニターするのか？ベースライン・パフォーマンスからの分布シフトやモデル・ドリフトはどの程度許容できるか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL

・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011).

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021).

Margaret Mitchell et al., "Model Cards for Model Reporting." Proceedings of 2019 FATML Conference.

Timnit Gebru et al., "Datasheets for Datasets," Communications of the ACM 64, No. 12, 2021.

Emily M. Bender, Batya Friedman, Angelina McMillan-Major (2022). A Guide for Writing Data Statements for Natural Language Processing. University of Washington. Accessed July 14, 2022.

M. Arnold, R. K. E. Bellamy, M. Hind, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development 63, 4/5 (July-September 2019), 6:1-6:13.

Navdeep Gill, Abhishek Mathur, Marcos V. Conde (2022). A Brief Overview of AI Governance for Responsible Machine Learning Systems. ArXiv, abs/2211.13130.

John Richards, David Piorkowski, Michael Hind, et al. A Human-Centered Methodology for Creating AI FactSheets. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

Christoph Molnar, Interpretable Machine Learning, lulu.com.

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD.

OECD (2022), “OECD Framework for the Classification of AI systems”, OECD Digital Economy Papers, No. 323, OECD Publishing, Paris.

Govern 1.5

リスクマネジメントプロセスとそのアウトカムの継続的なモニタリングと定期的なレビューが計画され、定期的なレビューの頻度の決定を含め、組織の役割と責任が明確に定義されている。

概要

AI システムは動的であり、デプロイ時またはデプロイ後に予期せぬ動作をする可能性がある。継続的なモニタリングは、AI システムのライフサイクルを通じて、予期せぬ問題やパフォーマンスの変化をリアルタイムまたは特定の頻度で追跡するためのリスクマネジメントプロセスである。

インシデントレスポンスと「不服申し立て（Appeal）と無効化（Override）」は、情報技術マネジメントでよく使われるプロセスである。これらのプロセスは、潜在的なインシデントをリアルタイムで把握し、システムのアウトカムを人間が判断することを可能にする。

インシデント対応計画を策定し、維持することで、AI インシデント発生時にインパクトが加わる可能性を低減することができる。ガバナンス・プログラムが充実していない可能性のある小規模組織は、システム障害、悪用、誤用に対処するためにインシデント対応計画を活用することができる。

推奨されるアクション

- AI システムが個人、コミュニティ、社会に与えるインパクトをアセスメントするための適切な資源と能力を配分するための方針を確立する
- システムのライフサイクルを通じて、バイアスやセキュリティの問題を含め、AI システムのパフォーマンスとトラストワージネスをモニタリングし、対処するための方針と手順を確立する
- AI システムのインシデント対応方針を策定するか、既存のインシデント対応方針が AI システムに適用されることを確認する
- AI システムのモニタリングとインシデント対応活動に責任を持つ組織機能と人員を定義するポリシーを確立する
- AI システムによるネガティブなインパクトについて、インパクトを受けた個人やコミュニティからのフィードバックを共有できる仕組みを構築する
- インパクトを受けた個人やコミュニティが、問題のある AI システムのアウトカムに異議を唱えるための手段を提供する仕組みを確立する
- オプトアウトの仕組みを確立する

透明性と文書化

組織は以下を文書化することができる。

- システム／エンティティは、明示された目標や目的に対する進捗をどの程度一貫して測定しているか？

- 組織は、特定された AI ソリューションのデプロイに伴うリスク（人的リスクや商業目的の変更など）に対処するために、リスクマネジメントシステムを導入しているか？
- 組織では、ユーザビリティの問題に取り組み、ユーザーインタフェースが意図した目的を果たしているかどうかをテストしているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・WEF Model AI Governance Framework Assessment 2020.

参考文献

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity.

National Institute of Standards and Technology. (2012). Computer Security Incident Handling Guide. NIST Special Publication 800-61 Revision 2.

Govern 1.6

AI システムをインベントリする仕組みが整備されており、組織のリスク優先順位に従ってリソースが確保されている。

概要

AI システムのインベントリとは、AI システムやモデルに関連する成果物の整理されたデータベースのことである。システム文書、インシデント対応計画、データ辞書、実装ソフトウェアやソースコードへのリンク、関連する AI アクターの名前と連絡先またはモデルやシステムの保守、インシデント対応に役立つその他の情報などが含まれる。AI システムのインベントリはまた、組織の AI 資産の全体像を把握することを可能にする。使用可能な AI システムのインベントリは、以下の迅速な解決を可能にする。

- 例えば、"このモデルが最後にリフレッシュされたのはいつか？"といったような、単一のモデルに対する特定のクエリ
- 例えば、"現在、組織内にデプロイされているモデルはいくつあるか？"、"モデルによってインパクトを受けるユーザは何人いるか？"といったような、すべてのモデルにわたるハイレベルなクエリ

AI システムのインベントリは、従来のモデル・リスクマネジメント手法の一般的な要素であり、技術上、ビジネス上およびリスクマネジメント上のベネフィットを提供することができる。部分的なインベントリでは、完全なインベントリの価値が得られない可能性があるため、通常、インベントリはすべての組織モデルまたはシステムを把握する。

推奨されるアクション

- AI システムのインベントリの作成と維持を定義する方針を確立する
- インベントリの維持に責任を持つ特定の個人またはチームを定める方針を確立する
- どのモデルまたはシステムをインベントリ化するかを定義する方針を確立する。すべてのモデルまたはシステムをインベントリ化することを優先するか、あるいは最小限に抑え、リスクの高いモデルまたはシステム、あるいはリスクおよび報酬の高い場面でデプロイされたシステムをインベントリ化することを優先する
- 文書、ソースコードへのリンク、インシデント対応計画、データ辞書、AI アクターの連絡先情報など、インベントリ化すべきモデルやシステムの属性を定義するポリシーを確立する

透明性と文書化

組織は以下を文書化することができる。

- AI システムのインベントリの詳細を文書化し、維持する責任は誰にあるのか？
- データの生成、取得／収集、取り込み、ステージング／ストレージ、トランスフォーメーション、セキュリティ、メンテナンス、そして普及のためにどのようなプロセスが存在するか？

- この AI の目的を考慮した場合、それがまだ正確かどうか、バイアスがないか、説明可能かどうかなどをチェックするための適切な間隔は？このモデルのチェック項目は何か？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.

参考文献

"A risk-based integrity level schema", in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. See Annex B.

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). See "Model Inventory," pg. 26.

VertaAI, "ModelDB: An open-source system for Machine Learning model versioning, metadata, and experiment management." Accessed Jan. 5, 2023.

Govern 1.7

AI システムの廃止と段階的廃止を、リスクを増大させず、組織のトラストワージネスを低下させない方法で安全に行うためのプロセスと手順が整備されている。

概要

モデルまたは AI システムの不規則または無差別な終了または削除は、不適切であり、組織のリスクを増大させる可能性がある。例えば、AI システムは、規制要件の対象となったり、将来のセキュリティ調査や法的調査に巻き込まれたりする可能性がある。信頼を維持するために、組織は、AI システムを体系的かつ意図的に廃止するための方針とプロセスの確立を検討する。一般的に、このような方針は、ユーザやコミュニティの懸念、依存システムやリンクされたシステムのリスク、セキュリティ、法律または規制に関する懸念を考慮したものである。廃止されたモデルやシステムは、アクティブなモデルとともに、モデルのインベントリに一定期間保管される。

推奨されるアクション

- AI システムを廃止するための方針を確立する。そのような方針は、通常、以下の事項を扱う
  - ユーザや地域社会への懸念、風評リスク
  - 事業継続と財務リスク
  - 上流と下流のシステム依存関係
  - 規制要件（データ保持など）
  - 将来起こりうる法的、規制上、セキュリティ上または法医学上の調査
  - 適切であれば、更改システムへの移行
- 廃止されたシステム、モデルおよび関連する成果物の保管場所と期間を定める方針を確立する
- 廃止された AI システムの十分な理解や実行のために保存しなければならない付随的なデータや成果物（予測、説明、中間入力特徴表現、ユーザ名とパスワードなど）に対処する方針を確立する

透明性と文書化

組織は以下を文書化することができる。

- データの生成、取得／収集、取り込み、ステージング／ストレージ、トランスフォーメーション、セキュリティ、メンテナンス、そして普及にどのようなプロセスが存在するか？
- これらの政策は、AI システムの活用に対する国民の信頼と信用をどの程度醸成しているのか？
- AI がもはやこの倫理的枠組みを満たしていないと考える人がいる場合、誰がその懸念を受け止め、必要に応じて問題を調査し、改善する責任を負うのか？彼らは AI の活用を修正するか、制限するか、止める権限を持っているのか？
- 人に関するものであれば、申請／審査／承認の倫理的な審査はあったか？(例えば、治験審査委員会の申請など)

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.

・Datasheets for Datasets.

参考文献

Michelle De Mooy, Joseph Jerome and Vijay Kasschau, “Should It Stay or Should It Go? The Legal, Policy and Technical Landscape Around Data Deletion,” Center for Democracy and Technology, 2017.

Burcu Baykurt, "Algorithmic accountability in US cities: Transparency, impact, and political economy." Big Data & Society 9, no. 2 (2022): 20539517221115426.

“Information System Decommissioning Guide,” Bureau of Land Management, 2011.

Govern 2

AI リスクを Mapping し、Measure し、Manage するために、適切なチームと個人に権限を与え、責任を負わせ、訓練を受けさせることで、アカウンタビリティを果たせる体制を整備する。

Govern 2.1

AI リスクの Mapping、Measure と Manage 関する役割と責任、コミュニケーションラインが組織全体の個人とチームに文書化され明確になっている。

概要

リスクを意識した組織文化の醸成は、責任を明確にすることから始まる。例えば、いくつかのリスクマネジメントの体制ではテストや評価業務を行う専門家は、AI システム開発者から独立し、リスクマネジメント機能を通じてまたは直接経営陣に報告する。このような仕組みは、集団思考やサンクコストの誤謬といった暗黙のバイアスに対抗し、リスクマネジメント機能を強化するのに役立つ可能性がある。

AI システムの設計や導入の決定に対して、権限を与えられた AI アクターが疑問を投げかけ、軌道修正できるような文化を根付かせることで、組織がリスクを予見し、それが定着する前に効果的に管理する能力を高めることができる。

推奨されるアクション

- 取締役会または諮問委員会、上級管理職、AI 監査機能、製品マネジメント、プロジェクトマネジメント、AI 設計、AI 開発、人間と AI の相互作用、AI のテストと評価、AI の取得と調達、インパクトのアセスメント機能、オーバーサイト機能（ただし、これらに限定されない）AI システムに直接的・間接的に関連する役職について、AI リスクマネジメントの役割と責任を定める方針を確立する
- AI リスクマネジメントの取り組みに参加する AI アクター間の定期的なコミュニケーションを促進する方針を確立する
- AI システムの独立した軌道修正を可能にするため、AI システム開発機能のマネジメントと AI システムのテスト機能を分離する方針を確立する
- AI のリスクマネジメントの取り組みにおける利害の衝突を特定し、その透明性を高め、防止するための方針を確立する
- AI のリスクマネジメントの取り組みを妨げる可能性のある確証バイアスや市場インセンティブに対抗するための方針を確立する
- AI のリスクマネジメントの活動において、AI アクターが既存の法律、オーバーサイト、コンプライアンスまたは企業リスク部門と協力するインセンティブを与える方針を確立する

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、関連するステークホルダの役割、責任および権限の委任をどの程度明確にしているか？
- AI の意思決定の最終的な責任は誰にあり、その人物は分析の意図する用途と制限事項を認識しているか？
- さまざまな AI ガバナンス・プロセスに関与する人員の責任は明確に定義されているか？
- AI システムの設計、開発、デプロイ、評価、モニタリングに携わる人員の役割、責任、移譲された権限は何か？
- 組織は、データ管理と保護においてアカウンタビリティに基づく慣行（PDPA や OECD プライバシー原則など）を導入しているか？

AI の透明性に関するリソース

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities

参考文献

Andrew Smith, “Using Artificial Intelligence and Algorithms,” FTC Business Blog (Apr. 8, 2020).

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011).

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021).

ISO, “Information Technology — Artificial Intelligence — Guidelines for AI applications,” ISO/IEC CD 5339. See Section 6, “Stakeholders’ perspectives and AI application framework.”

Govern 2.2

組織の人員とパートナーは、関連する方針、手順、協定に則って職務と責任を遂行できるよう、AIリスクマネジメント研修を受ける。

概要

AI リスクマネジメントの導入と有効性を高めるために、組織は適切な研修カリキュラムを定め、企業の学習要件に組み込むことが奨励される。定期的な訓練を通じて、AI アクターは以下の認識を維持することができる。

- AI リスクマネジメントの目標とその達成における役割
- 組織の方針、適用される法律と規制、業界のベストプラクティスと規範

その他の関連情報については、Map 3.4 および 3.5 を参照のこと。

推奨されるアクション

- 継続的な教育に取り組む人員のための方針を定める
  - 適用できる AI システムの法律および規則
  - AI システムから生じる可能性のあるネガティブインパクト
  - 組織の AI ポリシー
  - トラストワージな AI の特徴
- オーバーサイトの役割を担う AI アクター（法務、コンプライアンス、監査など）と比較し、技術的なタスクを遂行する AI アクター（開発者、オペレーターなど）のサブグループ間で適切なトレーニングが行われるようにする
- 研修が AI リスクマネジメントの技術的および社会技術的側面を包括的に扱うようにする
- 組織の AI ポリシーに、内部の AI 人員が自らの役割と責任を認識し、コミットするための仕組みが含まれていることを確認する
- 組織の方針が変更管理に対応でき、AI システムの大幅な変更についてコミュニケーションをとり、承認する仕組みが含まれていることを確認する
- リスクの懸念を軽減するために、アカウンタビリティを内外にエスカレーションするための経路を定義する

透明性と文書化

組織は以下を文書化することができる

- AI システムを扱う関連スタッフは、AI モデルの出力や決定を解釈し、データのバイアスを検出して管理するための適切な訓練を受けているか？
- エンティティは、AI システムの設計、開発、デプロイ、アセスメント、モニターに必要なスキルや経験をどのように判断するのか？

- エンティティは、人員が割り当てられた責任を果たすために必要なスキル、訓練、リソースおよびドメイン知識を有しているかどうかをどのように評価しているか？
- AI システムのインパクトを受ける地域社会を反映した経歴、経験、視点を持つ人材を採用し、育成し、維持するために、エンティティはどのような取り組みを行ってきたか？

AI の透明性に関するリソース透明性と文書化

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities

参考文献

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021).

"Developing Staff Trainings for Equitable AI," Partnership on Employment & Accessible Technology (PEAT, peatworks.org). URL

Govern 2.3

AI システムの開発とデプロイメントに伴うリスクについて、組織の経営幹部が責任をもって決定する。

概要

AI ポートフォリオを維持する組織のシニア・リーダーシップと C-Suite のメンバーは、AI リスクに対する認識を維持し、そのようなリスクに対する組織の意欲を確認し、それらのリスクをマネジメントする責任を負うべきである。

アカウンタビリティは、特定のチームと個人が AI リスクマネジメントの取り組みに責任を持つことを保証するものである。組織によっては、金融機関の AI ポートフォリオ（例えば予測モデリングや機械学習）の十分なパフォーマンスを保証する指名された役員に権限と（人的および予算的な）リソースを付与している。

推奨されるアクション

組織マネジメントでは以下のことができる

- AI システムを開発または使用する際のリスク許容度を宣言する
- AI リスクマネジメントの取り組みを支援し、その取り組みにおいて積極的な役割を果たす
- 組織文化の一部として、AI ライフサイクル全体を通じてリスクと危害防止の考え方を統合する
- 有能なリスクマネジメントの幹部を支援する
- リスクマネジメントを実行するための実行力、資源、権限を、マネジメントチェーン全体を通じて適切な各レベルに委任する

組織は、AI リスクマネジメントやオーバーサイト機能のための各委員会を設置し、それらの機能を組織の広範な企業リスク管理手法の中に統合することができる。

透明性と文書化

組織は以下を文書化することができる：

- 取締役会および／または上級管理職は、組織の AI ガバナンスを後援し、支援し、参加したか？
- AI システムの設計、開発、デプロイメント、アセスメント、モニタリングに携わる人員の役割、責任、権限の委任は何か？
- AI ソリューションは、人員が十分な情報に基づいた判断を下し、それに従って行動を取るのを支援するのに十分な情報を提供しているか？
- エンティティは、関連するステークホルダの役割、責任および権限の委任をどの程度明確にしているか？

AI の透明性に関するリソース

- WEF Companion to the Model AI Governance Framework- 2020.

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

Govern 3

人材の多様性、公平性、インクルージョン、アクセシビリティのプロセスは、ライフサイクルを通じた AI リスクのマッピング、測定、マネジメントにおいて優先される。

Govern 3.1

ライフサイクルを通じた AI リスクの Mapping、Measure、Manage に関する意思決定は、多様なチーム（例えば、人口統計、専門分野、経験、専門知識、経歴の多様性）によって行われる。

概要

多様な経験を持つ AI アクターを含む多様なチーム、リスクマネジメントを実施するためには、組織の能力やリスクを予測する能力を高めるために、専門分野や経歴を持つ人材がより適している。内部のチームに多様な人生経験や専門分野が不足している場合には、外部の人員との協議が必要になることもある。

ダイバーシティ、エクイティ、インクルージョンのメリットをユーザと AI アクターの双方に広げるために、チームはさまざまな背景、視点、専門知識を反映した多様な人々で構成されることが推奨される。

シニア・リーダーシップのコミットメントがなければ、チームの多様性とインクルージョンの有益な側面が、不注意にも多様な人材の広範な価値観と対立する、明文化されていない組織的インセンティブによって上書きされてしまう可能性がある。

推奨されるアクション

組織マネジメントでは以下のことができる

- AI の取り組みのために、学際的な役割、能力、スキル、能力を促進する方針と雇用慣行を当初から定める
- 人口動態や専門分野の多様性につながる方針と雇用慣行を定め、スタッフに必要なリソースとサポートを与え、報復を恐れることなく、スタッフからの意見や懸念の提供を促進する
- 包括性を促進し、新しい知見を既存の実践に統合するためのポリシーを確立する
- 組織の多様性、公平性、インクルージョン、アクセシビリティを補うため、内部の専門知識が不足している場合は、外部の専門知識を求める
- 既存の差別撤廃、アクセシビリティ、アコモデーション、人事部門、従業員リソースグループ（ERG）、多様性、公平性、インクルージョン、アクセシビリティ（DEIA）イニシアティブと連携するよう、AI アクターにインセンティブを与えるポリシーを確立する

透明性と文書化

組織は以下を文書化することができる

- AI システムを扱う関連スタッフは、AI モデルの出力や決定を解釈し、データのバイアスを検出してマネジメントするための適切な訓練を受けているか？
- エンティティは、潜在的なバイアスや差別を含む意図せざる結果を予測し、緩和するために、AI のライフサイクル全体を通じて、技術的および非技術的なコミュニティからの多様な視点を内包する
- ステークホルダの参加。リスクを軽減するために、AI のライフサイクル全体を通じてステークホルダの多様な視点を取り入れる
- 多様な視点を取り入れるための戦略には、データサイエンス、ソフトウェア開発、公民権、プライバシーとセキュリティ、法律顧問、リスクマネジメントなどの専門家が参加する共同作業プロセスや学際的チームの設立が含まれる
- 確立された手続きは、バイアスや不公平感、その他制度に起因する懸念を軽減する上で、どの程度有効か？

AI の透明性に関するリソース

・WEF Model AI Governance Framework Assessment 2020.

・Datasheets for Datasets.

参考文献

Dylan Walsh, “How can human-centered AI fight bias in machines and people?” MIT Sloan Mgmt. Rev., 2021.

Michael Li, “To Build Less-Biased AI, Hire a More Diverse Team,” Harvard Bus. Rev., 2020.

Bo Cowgill et al., “Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics,” 2020.

Naomi Ellemers, Floortje Rink, “Diversity in work groups,” Current opinion in psychology, vol. 11, pp. 49–53, 2016.

Katrin Talke, Søren Salomo, Alexander Kock, “Top management team diversity and strategic innovation orientation: The relationship and consequences for innovativeness and performance,” Journal of Product Innovation Management, vol. 28, pp. 819–832, 2011.

Sarah Myers West, Meredith Whittaker, and Kate Crawford,, “Discriminating Systems: Gender, Race, and Power in AI,” AI Now Institute, Tech. Rep., 2019.

Sina Fazelpour, Maria De-Arteaga, Diversity in sociotechnical machine learning systems. Big Data & Society. January 2022. doi:10.1177/20539517221082027

Mary L. Cummings and Songpo Li, 2021a. Sources of subjectivity in machine learning models. ACM Journal of Data and Information Quality, 13(2), 1–9

“Staffing for Equitable AI: Roles & Responsibilities,” Partnership on Employment & Accessible Technology (PEAT, peatworks.org). Accessed Jan. 6, 2023.

Govern 3.2

人間と AI の構成や AI システムのオーバーサイトに関する役割と責任を定義し、区別するための方針と手順が定められている。

概要

AI のリスクとインパクトの特定とマネジメントは、技術的なものも含め、AI のライフサイクル全体にわたる広範な視点とアクターによって強化される、法律、コンプライアンス、社会科学、ヒューマンファクターなどの専門知識が必要とされる。

AI アクターには、下流タスクのために AI システムを運用、活用、あるいは AI システムと相互作用する者、あるいは AI システムのパフォーマンスをモニターする者が含まれる。効果的なリスクマネジメントの取り組みには以下がある。

- AI システムのオーバーサイトとガバナンスに関する人間の役割と責任を明確に定義し、区別する
- AI システムのオーバーサイトを行う者と、AI システムを活用する者、あるいは AI システムと相互作用する者の違いを認識し、明確にする

推奨されるアクション

- AI システムを使用、対話、モニタリングする際の人間の様々な役割と責任を定義および区別するポリシーと手順を確立する
- 人間と AI のコンフィギュレーションおよび関連する結果に関連するリスク情報を取得し、追跡するための手順を確立する
- システム運用業務およびシステムオーバーサイト業務を遂行する AI アクターの習熟度基準を策定するためのポリシーを確立する
- システム運用業務やシステムオーバーサイト業務を行う AI アクターに対し、所定のリスクマネジメント訓練プロトコルを確立する
- AI アクターの役割およびデプロイされたシステムのヒューマンオーバーサイトの責任に関するポリシーと手順を確立する
- 組織のリスク許容度に関連して、人間と AI のコンフィギュレーション（AI システムが明示的に指定され、主に人間のチームのチームメンバーとして扱われるコンフィギュレーション）を定義するポリシーと手順および関連文書を確立する
- AI システムの説明性、解釈性および全体的な透明性を高めるポリシーを確立する
- 既知の困難な非人間的 AI 構成、人間と AI のチーム編成、AI システムのユーザーエクスペリエンスとユーザーインタラクション（UI／UX）に関するリスクマネジメントのポリシーを確立する

透明性と文書化

組織は以下を文書化することができる

- エンドユーザ、消費者、規制当局、AI システムの活用によってインパクトを受ける個人を含む外部のステークホルダが、AI システムの設計、運用、制限についてどのような情報にアクセスできるか？
- AI を活用した意思決定における人間の関与の適切なレベルを、エンティティはどの程度まで文書化しているか？
- 敵が AI を混乱させようとしたり、あるいは AI の精度に影響を与える可能性のある、作戦／事業環境の関連性のない変化による精度や正確さの変化に、説明責任者はどのように対処するのか？
- エンティティは、関連するステークホルダの役割、責任および権限委任をどの程度明確にしているか？
- エンティティは、人員が割り当てられた責任を果たすために必要なスキル、訓練、リソースおよび領域知識を有しているかどうかをどのように評価しているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・WEF Companion to the Model AI Governance Framework- 2020.

参考文献

Madeleine Clare Elish, "Moral Crumple Zones: Cautionary tales in human-robot interaction," Engaging Science, Technology, and Society, Vol. 5, 2019.

“Human-AI Teaming: State-Of-The-Art and Research Needs,” National Academies of Sciences, Engineering, and Medicine, 2022.

Ben Green, "The Flaws Of Policies Requiring Human Oversight Of Government Algorithms," Computer Law & Security Review 45 (2022).

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD.

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021).

Govern 4

組織チームは、AI リスクを考慮し、コミュニケーションする文化にコミットしている。

Govern 4.1

ネガティブインパクトを最小限に抑えるため、AI システムの設計、開発、デプロイメント、使用において、クリティカルシンキング（批判的思考）と安全第一の考え方を育成するための組織のポリシーとプラクティスが整備される。

概要

リスク文化とそれに付随するプラクティスは、組織が最も重大なリスクを効果的にトリアージするのに役立つ。ある業界の組織は、3 つ（またはそれ以上）の「防衛ライン」を導入しており、開発、リスクマネジメント、監査など、システムライフサイクルのさまざまな側面について別々のチームが責任を負っている。伝統的な 3 ラインのアプローチは、小規模な組織にとっては現実的でないかもしれないが、リーダーシップは、他の手段を通じて強力なリスク文化を育成することにコミットする。例えば、「効果的なチャレンジ」は、そのような変更を行う権限と地位を有する専門家による、重要な設計と実装の決定に対する批判的思考と質問を奨励する、文化に基づく実践である。

レッドチームは、リスク測定とマネジメントのもう一つのアプローチである。この手法は、AI システムをストレス条件下で敵対的にテストし、システムの故障モードや脆弱性を探し出すものである。レッドチームは、内部の AI アクターから独立した外部の専門家や人員で構成される。

推奨されるアクション

- システム設計プロセスの初期段階から、オーバーサイト機能（法務、コンプライアンス、リスクマネジメント）を含めることを義務付ける方針を確立する 3 つの防衛線、モデル監査、レッドチームなどのメカニズムを通じて、AI システムの設計、実装およびデプロイメントの決定に対する効果的なチャレンジを促進するポリシーを確立し、集団思考などの職場のリスクを最小限に抑える
- 安全第一の考え方と一般的なクリティカルシンキングを奨励する方針を確立し、組織および手続きレベルで見直す
- AI システムに重大な問題があることを報告した内部関係者に対する内部告発者保護を確立する
- AI のライフサイクル全体を通じて、危害とリスク予防の考え方を統合する方針を確立する

透明性と文書化

組織は以下を文書化することができる

- エンティティは、AI システムの開発、テスト方法、メトリクス、パフォーマンスのアウトカムをどの程度文書化されているか？

- 過去に失敗した設計に関連することを回避できるような、組織の情報共有のプラクティスは広く守られ、透明性が確保されているか？
- インシデント対応を実施するための訓練マニュアルやその他のリソースは文書化され、利用可能か？
- オペレータによるインシデントやヒヤリハットの報告プロセスは文書化され、利用可能か？

AI の透明性に関するリソース

・Datasheets for Datasets.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・WEF Model AI Governance Framework Assessment 2020.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Patrick Hall, Navdeep Gill, and Benjamin Cox, “Responsible Machine Learning,” O’Reilly Media, 2020. URL

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

GAO, “Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities,” GAO@100 (GAO-21-519SP), June 2021.

Donald Sull, Stefano Turconi, and Charles Sull, “When It Comes to Culture, Does Your Company Walk the Talk?” MIT Sloan Mgmt. Rev., 2020.

Kathy Baxter, AI Ethics Maturity Model, Salesforce.

Govern 4.2

組織チームは、設計、開発、デプロイ、評価、活用する AI 技術のリスクと潜在的インパクトを文書化し、そのインパクトを広く伝える。

概要

インパクトアセスメントは、責任ある技術開発を推進するための一つのアプローチである。また、特定のユースケースにおいて、これらのアセスメントは、組織が特定のアルゴリズムやデプロイメントのリスクをフレームワーク化するためのハイレベルな構造を提供することができる。また、インパクトアセスメントは、組織がリスクを明確にし、危害が発生した場合のマネジメント・オーバーサイト活動のための文書を作成する仕組みとしても機能する。

インパクトアセスメントは以下のようにしてもよい

- ゴールやアウトカムは時間とともに変化する可能性があるため、プロセスの開始時に適用するだけでなく、反復的かつ定期的に適用する必要がある
- 運用者、活用者、インパクトを受ける可能性のあるコミュニティ（歴史的に周縁化されたコミュニティ、障がい者、デジタルデバイドの影響を受ける個人を含む）を含む AI のアクターからの視点を含む
- AI システムの「実施」または「非実施」の判断を支援する
- アセスメント対象の組織チームに関連する利益相反または不当な影響力を考慮する

インパクトアセスメントに関する詳細は、Map 機能プレイブックガイダンスを参照のこと。

推奨されるアクション

- 組織が活用する AI システムのインパクトアセスメント方針とプロセスを確立する
- 組織のインパクトアセスメント活動を、関連する規制または法的要件と整合させる
- インパクトアセスメント活動が、システムの潜在的なネガティブインパクトとシステムの変化の速さを評価するために適切であり、評価が定期的に適用されていることを検証する
- AI のシステムリスクに関する広範な評価を行うために、インパクトアセスメントを活用する

透明性と文書化

組織は以下を文書化することができる

- エンティティは、不公平あるいは差別的なアウトカムなど、データのバイアスによる潜在的なインパクトをどのように特定し、軽減したか？
  - エンティティは、ソース、起源、変換、拡張、ラベル、依存関係、制約およびメタデータなど、AI システムのデータの出所をどのように文書化しているか？
- エンティティは、AI システムの技術仕様や要件をどの程度明確に定義しているか？

●エンティティは、AI システムの開発、テスト手法、メトリクス、パフォーマンスのアウトカムについて、どの程度文書化し、周知しているか？

●マシンエラーとヒューマンエラーは異なる可能性があることを文書化し、説明しているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.

・Datasheets for Datasets.

参考文献

Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker, “Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability,” AI Now Institute, 2018.

H.R. 2231, 116th Cong. (2019).

BSA The Software Alliance (2021) Confronting Bias: BSA’s Framework to Build Trust in AI.

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

David Wright, “Making Privacy Impact Assessments More Effective." The Information Society 29, 2013.

Konstantinia Charitoudi and Andrew Blyth. A Socio-Technical Approach to Cyber Risk Management and Impact Assessment. Journal of Information Security 4, 1 (2013), 33-41.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, & Jacob Metcalf. 2021. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest”.

Microsoft. Responsible AI Impact Assessment Template. 2022.

Microsoft. Responsible AI Impact Assessment Guide. 2022.

Microsoft. Foundations of assessing harm. 2022.

Mauritz Kop, “AI Impact Assessment & Code of Conduct,” Futurium, May 2019.

Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, “Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability,” AI Now, Apr. 2018.

Andrew D. Selbst, “An Institutional View Of Algorithmic Impact Assessments,” Harvard Journal of Law & Technology, vol. 35, no. 1, 2021

Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study in Healthcare. Accessed July 14, 2022.

Kathy Baxter, AI Ethics Maturity Model, Salesforce

Govern 4.3

AI のテスト、インシデントの特定、情報の共有を可能にする組織的プラクティスがある。

概要

AI システムの限界を特定し、ネガティブなインパクトやインシデントを検知・追跡し、これらの問題に関する情報を適切な AI 関係者と共有することで、リスクマネジメントが改善される。コンセプト・ドリフト、AI のバイアスや差別、ショートカット学習、仕様の不備といった問題は、現在の標準的な AI テストプロセスでは特定が困難である。組織は、そのような問題を特定し管理するために、内部で使用とテストの方針と手順を制定することができる。このような取り組みには、プレアルファ試験やプレベータ試験、あるいは内部で開発されたシステムや製品を組織内で展開するという形をとることができる。テストは、限定的かつ管理された内部または一般に利用可能な AI システムのテストベッドおよび AI システムのインターフェースと出力へのアクセスを必要とする場合がある

一貫したテスト実施を可能にする方針と手順がなければ、リスクマネジメントの実施が迂回されたり無視されたりして、リスクを悪化させたり、一貫性のないリスクマネジメント活動につながる可能性がある。

テストやデプロイメントに検出されたインパクトやインシデントに関する情報共有を行い、以下を実施する。

- AI システムのリスク、障害、悪用、誤用に注意を喚起する
- 組織が幅広い AI のアプリケーションと実装に基づく洞察からベネフィットを得られるようにする
- 組織が既知の故障モードを回避するために、より積極的に行動できるようにする

組織は、AI インシデントのデータベース、AIAAIC、ユーザ、インパクトを受けたコミュニティまたは MITRE CVE リストなどの従来のサイバー脆弱性データベースとインシデント情報の共有を検討するかもしれない。

推奨されるアクション

- AI システムのテストを促進し、整備するためのポリシーと手順を確立する
- AI システムの制限を特定し、制限に関する洞察を適切な AI アクターおよびグループ内で共有するための組織的コミットメントを確立する
- インシデント対応を報告し、文書化するためのポリシーを確立する
- インシデントの公開と情報共有に関するポリシーとプロセスを確立する
- AI システムのリスクとパフォーマンスに関連するインシデント対応のガイドラインを確立する

透明性と文書化

組織は以下を文書化することができる

- 組織は、ユーザビリティの問題に取り組み、ユーザーインタフェースが意図した目的を果たしているかどうかをテストしたか？使用される技術やそのデプロイ方法について透明性を確保するために、開発の初期段階でコミュニティやエンドユーザに相談する組織は、AI ソリューションのデプロイに伴うリスク（人的リスクや事業目的の変更など）に対処するためにリスク管理システムを導入したか？
- AI システムの出力によって影響を受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？

AI の透明性に関するリソース

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

参考文献

Sean McGregor, “Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database,” arXiv:2011.08512 [cs], Nov. 2020, arXiv:2011.08512.

Christopher Johnson, Mark Badger, David Waltermire, Julie Snyder, and Clem Skorupka, “Guide to cyber threat information sharing,” National Institute of Standards and Technology, NIST Special Publication 800-150, Nov 2016.

Mengyi Wei, Zhixuan Zhou (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. ArXiv, abs/2206.07635.

BSA The Software Alliance (2021) Confronting Bias: BSA’s Framework to Build Trust in AI.

“Using Combined Expertise to Evaluate Web Accessibility,” W3C Web Accessibility Initiative. URL

Govern 5

関連する AI アクターとの堅牢なエンゲージメントのためのプロセスが整備されている。

Govern 5.1

AI リスクに関連する潜在的な個人的・社会的インパクトについて、AI システムを開発およびデプロイしたチームの外部からのフィードバックを収集、検討、優先順位付けおよび統合するための組織のポリシーとプラクティスが整備されている。

概要

内部や研究室ベースのシステムテストにとどまらず、組織のポリシーとプラクティスは、意図された使用のコンテクストに関連する AI システムの目的適合性を検討することがある。

ステークホルダの参加型エンゲージメントは、AI アクターがプロジェクトを推進するかどうか、あるいはインパクトを考慮した設計を行うにはどうすればよいかといった疑問に答えるための、定性的な活動の一種である。また、このようなフィードバックは、ドメインのエキスパートの意見を取り入れながら、AI アクターが AI アプリケーションの不確実なシナリオやリスクを特定する助けにもなる。いつ、どのようにグループを招集し、どのような個人、グループ、コミュニティ組織を参加させるかを検討することは、システムの目的とそのリスクのレベルに関連したイテレーションプロセスである。その他の要因としては、単なる場当たり的な訓練に終始することなく、ステークホルダからのフィードバックや有用な知見を、いかに協力的かつ敬意を持って収集するかが挙げられる。

このような活動は、参加型の実践、定性的手法および技術的なオーディエンスに向けたコンテクストに沿ったフィードバックができる専門知識を持つ人員が行うのが最適である。

参加型エンゲージメントは 1 回限りのものではなく、AI システムの試運用の当初からライフサイクルの終わりまで実施するのが最適である。組織は、プロジェクトの開始時やシステムのモニタリングの一環として、エンゲージメントをどのように取り入れるかを検討することができる。エンゲージメントは、しばしばコンサルティングの実践として活用されるが、この視点はともすると"参加型ウォッシング"につながるかもしれない。エンゲージメントの目的とゴールに関する組織の透明性は、その可能性を軽減するのに役立つ。

組織は、参加型調査で得られた結果を補完するものとして、対象分野のエキスパートとのターゲットを絞った協議を検討することもできる。エキスパートは、これまで考慮されなかった潜在的なネガティブインパクトを特定し、概念化することで、内部スタッフを支援できるかもしれない。

推奨されるアクション

- ステークホルダやユーザからのフィードバックを収集、評価および取り入れるための仕組みを明確に取り上げた、以下のような AI リスクマネジメントのポリシーを確立する
  - 欠陥のある AI システムの出力に対する是正メカニズム
  - バグの報奨金
  - 人間中心設計
  - ユーザーインタラクション及びエクスペリエンス研究
  - ネガティブなインパクトを経験する可能性のある個人やコミュニティとの参加型ステークホルダ参画

- 歴史的に排除されてきた人々、障がい者、高齢者、インターネットやその他の基本技術へのアクセスが制限されている人々を含む、意図する利用者全体にわたってステークホルダのフィードバックが考慮され、対処されていることを確認する
- AI システムに適用される組織の原則を明確にし（公に提案されているものも考慮する）、組織の価値観を外部のステークホルダに知らせる。AI 原則の公表または採用を検討する

透明性と文書化

組織は以下を文書化することができる

- エンドユーザ、消費者、規制当局、AI システムの使用によってインパクトを受ける個人を含む外部のステークホルダが、AI システムの設計、運用および制限についてどのような情報にアクセスできるか？
- エンティティは、関連するステークホルダの役割、責任および権限の委任をどの程度明確にしているか？
- 外部のステークホルダが入手可能な情報に、どれだけ容易にアクセスでき、それがどれだけ新しいか？
- 損害の可能性を軽減または低減するために何が行われたか？
- ステークホルダの参加：リスクを軽減するために、AI のライフサイクル全体を通じてステークホルダの多様な視点を取り入れる

AI の透明性に関するリソース

・Datasheets for Datasets.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.
・Stakeholders in Explainable AI, Sep. 2018.

参考文献

ISO, “Ergonomics of human-system interaction — Part 210: Human-centered design for interactive systems,” ISO 9241-210:2019 (2nd ed.), July 2019.

Rumman Chowdhury and Jutta Williams, "Introducing Twitter’s first algorithmic bias bounty challenge,"

Leonard Haas and Sebastian Gießler, “In the realm of paper tigers – exploring the failings of AI ethics guidelines,” AlgorithmWatch, 2020.

Josh Kenway, Camille Francois, Dr. Sasha Costanza-Chock, Inioluwa Deborah Raji, & Dr. Joy Buolamwini. 2022. Bug Bounties for Algorithmic Harms? Algorithmic Justice League. Accessed July 14, 2022.

Microsoft Community Jury , Azure Application Architecture Guide.

“Definition of independent verification and validation (IV&V)”, in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. Annex C,

Govern 5.2

AI アクターが、関連する AI アクターからの裁定されたフィードバックを定期的にシステム設計と実装に取り入れることを可能にするメカニズムを確立する。

概要

AI アクターに、システムデプロイメントに関する共同決定を知らせるために必要なプロセス、知識および専門知識を与える組織のポリシーと手続きは、リスクマネジメントを改善する。これらの決定は、AI システムおよび組織のリスク許容度と密接に結びついている。

組織のリーダーシップによって確立されるリスク許容度は、組織がそのミッションを遂行し戦略を遂行する間に受け入れるリスクのレベルと種類を反映する。リスクが発生すると、評価された当該 AI システムのリスクに基づいて資源が配分される。組織は通常、リスク許容度アプローチを適用し、リスクの高いシステムには多い、リスクの低いシステムには少ないリスクマネジメントリソースを配分する。

推奨されるアクション

- AI システムと AI の利用には、潜在的なベネフィットとともに、内在するコストおよびリスクがあることを明確に認める
- 法律、規制、ベストプラクティスまたは業界標準に基づき、AI システムの合理的なリスク許容度を定義する
- すべての関係する AI アクターに、システムの設計と実施に関するフィードバックを提供する有意義な機会が提供されるようなポリシーを確立する
- システムのインパクトアセスメントとインパクトが発生する可能性を組み合わせて、AI システムを確立されたリスク許容レベルに割り当てる方法を定義するポリシーを確立する。このようなアセスメントには、多くの場合、以下のような組み合わせが必要となる
- AI システムのリスクをアセスメントするために、インパクトとインパクトの可能性を計量経済学的に評価する
- AI システムのリスクをアセスメントするための、インパクトの重大性と可能性に関するレッド・アンバー・グリーン（RAG）スケール

透明性と文書化

組織は以下を文書化することができる

- AI の決定に対する最終的な責任は誰にあり、その人物は分析の意図する用途と制限を認識しているか？
- デプロイされた AI の保守、再検証、モニタリングおよび更新は誰が行うのか？
- AI のライフサイクルの全段階における倫理的配慮について、誰が責任を負うのか？

- 確立された手続きは、バイアスや不公平感、その他制度に起因する懸念を軽減する上で、どの程度有効か？
- AI ソリューションは、人員が情報を得た上で意思決定を行い、それに従って行動を取るのを支援するのに十分な情報を提供するか？

AI の透明性に関するリソース

・WEF Model AI Governance Framework Assessment 2020.
・WEF Companion to the Model AI Governance Framework- 2020.
・Stakeholders in Explainable AI, Sep. 2018.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021).

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019). Retrieved on July 12, 2022.

Govern 6

サードパーティのソフトウェアやデータおよびその他のサプライチェーンの問題から生じる AI のリスクおよびベネフィットに対処するためのポリシーと手続きが定められている。

Govern 6.1

サードパーティの知的財産権やその他の権利を侵害するリスクを含め、サードパーティに関連する AI リスクに対処する方針と手順が整備されている。

概要

リスクの測定とマネジメントは、顧客がサードパーティのデータやシステムをどのように AI 製品あるいはサービスに利用したり統合したりするかによって、特に十分な内部ガバナンス構造や技術的な保護措置がない場合には、複雑になる可能性がある。

組織は通常、外部の専門知識、データ、ソフトウェアパッケージ（オープンソースと商用の両方）および AI ライフサイクル全体にわたるソフトウェアとハードウェアのプラットフォームを得るために複数のサードパーティを利用する。このような取り組みには有益な用途がある一方で、リスクマネジメントの取り組みを複雑化させることもある。

サードパーティの（ポジティブおよびネガティブな）リスクを管理する組織的アプローチは、各システムのリソース、リスクプロファイル、ユースケースに合わせて調整することができる。組織は、オープンソース・ソフトウェア、一般公開されているデータおよび商用利用可能なモデルを含む内部のリソースと同様に、サードパーティの AI システムやデータにもガバナンスアプローチを適用できる。

推奨されるアクション

- サードパーティの AI システムやデータに対応するポリシーを共同で策定する
- 関連するポリシーは以下のものがある
  - 学習データ、学習および推論アルゴリズムに関する知識、前提条件および制限を含むサードパーティのシステム機能の透明性
  - サードパーティの AI システムを全体検証。(詳細は Measure を参照）サードパーティのシステムの使用用途に関する明確かつ完全な指示のための要件。サードパーティのテクノロジーに対するポリシーを評価する
- サードパーティのソフトウェアまたはハードウェアシステムおよびデータの調達と使用に関する法的、倫理的、その他の問題を含む、サプライチェーン、製品ライフサイクル全体および関連プロセスに対処するポリシーを確立する

透明性と文書化

組織は以下を文書化することができる

- AI システムの監査可能性（例えば、開発プロセスのトレーサビリティ、学習データの調達、AI システムのプロセス、結果およびポジティブネガティブのインパクトの記録）を促進するメカニズムを確立したか？
- サードパーティが AI を作成した場合、説明可能性や解釈可能性の度合いをどのように確保するのか？
- AI システムが独立したサードパーティによる監査を受けられることを保証したか？
- サードパーティ（例えばサプライヤ、エンドユーザ、被験者、ディストリビュータ／ベンダ及び従業員）が AI システムの潜在的な脆弱性、リスク、バイアスを報告するプロセスを確立したか？
- 計画が、買収、ベンダからのパッケージソフトの調達、サイバーセキュリティ管理、計算機インフラ、データ、データサイエンス、デプロイの仕組みおよびシステム障害に関連するリスクにどの程度具体的に取り組んでいるか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・WEF Model AI Governance Framework Assessment 2020.
・WEF Companion to the Model AI Governance Framework- 2020.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.
・Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

“Proposed Interagency Guidance on Third-Party Relationships: Risk Management,” 2021.

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021).

Govern 6.2

リスクが高いと判断されたサードパーティのデータあるいは AI システムについて、その障害あるいはインシデントに対処するためのコンティンジェンシープロセスが整備されている。

概要

サードパーティのシステム障害による潜在的な損害を軽減するために、組織はサードパーティの機能をカバーするための冗長性を含むポリシーと手続きを導入するかもしれない。

推奨されるアクション

- 重要なサードパーティ製 AI システムの冗長メカニズムの検討を含め、サードパーティ製システムの障害に対処するためのポリシーを確立する
- インシデントの対応計画がサードパーティの AI システムに対応していることを検証する

透明性と文書化

- 組織は以下を文書化することができる買収、ベンダからのパッケージソフトの調達、サイバーセキュリティ管理、計算機インフラ、データ、データサイエンス、デプロイの仕組みおよびシステム障害に関連するリスクにどの程度具体的に取り組んでいるか？
- サードパーティ（例えばサプライヤ、エンドユーザ、被験者、ディストリビュータ／ベンダ、あるいは従業員）が AI システムの潜在的な脆弱性、リスク、あるいはバイアスを報告するプロセスを確立したか？
- 組織がサードパーティからデータセットを入手した場合、そのようなデータセットを利用するリスクをアセスメントし、マネジメントしたか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
・WEF Model AI Governance Framework Assessment 2020.
・WEF Companion to the Model AI Governance Framework- 2020.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.

参考文献

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

“Proposed Interagency Guidance on Third-Party Relationships: Risk Management,” 2021.

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021).

Map

Map 1

コンテクストが確立され、理解される。

Map 1.1

意図された目的、潜在的に有益な用途、コンテクスト特有の法律、規範と期待および AI システムがデプロイされる見込みのある設定が理解され、文書化される。考慮すべき事項には以下が含まれる。それぞれが期待を持つ特定のユーザのグループあるいはタイプ。システム利用が個人、コミュニティ、組織、社会及び地球に及ぼす潜在的なポジティブ及びネガティブなインパクト。AI システムの目的に関する仮定と関連する制限。開発あるいは製品の AI ライフサイクル全体にわたる用途とリスク。テスト、評価、妥当性確認および検証（Test, evaluation, validation, and verification: TEVV）とシステムメトリクス。

概要

高精度で最適化されたシステムは弊害をもたらすこともある。これに関連して組織は、広くデプロイされた AI ツールが意図に関係なく再利用、別の目的での利用および潜在的に悪用される可能性があることを想定すべきである。

AI アクターは、コミュニティ・グループなどの外部関係者と協力して、許容されるデプロイの範囲を明確にし、望ましい代替案を検討し、想定されるリスクをマネジメントするための原則と戦略を定めることができる。コンテクストマッピングは、この取り組みの最初のステップであり、以下の検討を含む場合がある。

- システム使用の意図された目的とインパクト
- 運用のコンセプト
- 意図したデプロイ設定、予想される将来的なデプロイ設定及び実際のデプロイ設定
- システムデプロイメントと運用のための要件
- エンドユーザと運用者の期待
- 特定のエンドユーザのグループまたはタイプ
- 個人、グループ、コミュニティ、組織及び社会に対する潜在的なネガティブインパクト、あるいは、法的要件や環境へのインパクトといったコンテクスト特有のインパクト
- 予期せぬ、下流域での、あるいはその他の未知のコンテクスト要因
- AI システムの変更がどのようにインパクトを与えるか

これらのタイプのプロセスは、AI 技術のデプロイと使用に関連する制限や制約、その他の現実が、実際の世界で AI 技術のデプロイ後や実社会でデプロイあるいは運用された後にどのようなインパクトをもたらすかを AI アクターが理解するのに役立つ。Govern 機能において確立された方針と手順によって強化され

た組織文化と組み合わせることで、Map 機能は、リスク及びインパクトにアプローチするための新たな視点、活動及びスキルを育成し、浸透させる機会を提供することができる。

コンテクストマッピングには、そのコンテクストが AI とその潜在的なネガティブインパクトをマネジメントできるほど狭いのかどうかに関連して、非 AI または非テクノロジーの代替案の議論および検討も含まれる。非 AI の代替案には、半自律的あるいはほとんど手動の方法を用いて情報を取得し、評価することが考えられる。

推奨されるアクション

- 業界、技術および適用される法的な標準を常に意識すること。
- AI システム設計のトラストワージネスを調査し、非 AI ソリューションを検討する
- ヒューマンファクターや社会技術分野の専門家と協力し、意図された AI システム設計タスクと予期せぬ目的を検討する
- AI システムの課題、目的、最小限の機能、利点を定義し、文書化し、プロジェクトの有用性または欠如に関する考慮事項を伝える
- よりトラストワージな結果をもたらす非 AI または非テクノロジーの代替案があるかどうかを特定する
- システム性能の変化が、意思決定などの下流イベントにどのような影響を与えるかを調査する（例えば、AI モデルの目的関数の変化が、就職面接を受ける候補者と受けない候補者の数にどのようなインパクトを与えるか）
- ビジネス要件と技術要件を含む、エンドユーザと組織の要件を決定する
- 以下を含む、予想され許容できる AI システムの使用のコンテクストを決定し、明確にする

社会規範

  - インパクトを受ける個人、グループおよびコミュニティ
  - 個人、グループ、コミュニティ、組織および社会に対するポジティブおよびネガティブな潜在的インパクト
  - 運用環境

- 意図した設定（または意図された設定に近い条件）の範囲内で時間枠、安全性の懸念、地理的エリア、物理的環境、エコシステム、社会的環境および文化的規範に関するコンテクスト分析を実行する
- システム設計に着手する前に、AI システムのパフォーマンスおよびベネフィットに関する科学的な主張の評価についての認識を獲得し、維持する
- アプリケーションが人間の意思決定を支援するのか、それとも代替するのかなど、人間と AI の相互作用や役割を特定する
- 人間と AI の構成に関するリスクを計画し、デプロイされたシステムを人間がオーバーサイトするための要件、役割および責任を文書化する

透明性と文書化

組織は以下を文書化することができる。

- この計画は、買収、ベンダからのパッケージソフトの調達、サイバーセキュリティ管理、計算機インフラ、データ、データサイエンス、デプロイの仕組みおよびシステム障害に関連するリスクにどの程度具体的に対処しているか？
- どの AI アクターが AI の決定に責任を持ち、その人物は分析の使用目的および制限を認識しているか？
- デプロイ後の AI の保守、再検証、モニタリング、更新を担当するのはどの AI アクターか？
- AI のライフサイクルを通じた倫理的配慮に責任を負うのは誰か？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities,
“Stakeholders in Explainable AI,” Sep. 2018.
・"Microsoft Responsible AI Standard, v2"

参考文献

社会技術システム

Andrew D. Selbst, danah boyd, Sorelle A. Friedler, et al. 2019. Fairness andAbstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness,Accountability, and Transparency (FAccT'19). Association for Computing Machinery,New York, NY, USA, 59–68

問題形成

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices inartificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702.

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. InProceedings of the Conference on Fairness, Accountability, and Transparency(FAccT'19). Association for Computing Machinery, New York, NY, USA, 39–48.

コンテクストマッピング

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicineand healthcare. Joint Research Centre (European Commission).

Sarah Spiekermann and Till Winkler. 2020. Value-based Engineering for Ethics byDesign. arXiv:2004.13676.

Social Impact Lab. 2017. Framework for Context Analysis of Technologies in SocialChange Projects (Draft v2.0).

Solon Barocas, Asia J. Biega, Margarita Boyarskaya, et al. 2021. Responsiblecomputing during COVID-19 and beyond. Commun. ACM 64, 7 (July 2021), 30–32.

有害性の特定

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources ofHarm throughout the Machine Learning Life Cycle. arXiv:1901.10002.

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. OvercomingFailures of Imagination in AI Infused System Development and Deployment.arXiv:2011.13416.

Microsoft. Foundations of assessing harm. 2022.

ML における制限の理解と文書化

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2020. UnderspecificationPresents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395.

Arvind Narayanan. "How to Recognize AI Snake Oil." Arthur Miller Lecture on Scienceand Ethics (2019).

Jessie J. Smith, Saleema Amershi, Solon Barocas, et al. 2022. REAL ML:Recognizing, Exploring, and Articulating Limitations of Machine Learning Research.arXiv:2205.08363.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. 2019. Model Cards for ModelReporting. In Proceedings of the Conference on Fairness, Accountability, andTransparency (FAT* '19). Association for Computing Machinery, New York, NY, USA,220–229.

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, et al. 2019. FactSheets:Increasing Trust in AI Services through Supplier's Declarations of Conformity.

ML における制限の理解と文書化

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul et al. "Measuring the

Predictabilityof Life Outcomes with a Scientific Mass Collaboration." Proceedings of the NationalAcademy of Sciences 117, No. 15 (2020): 8398-8403.

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020.Co-Designing Checklists to Understand Organizational Challenges and Opportunitiesaround Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factorsin Computing Systems (CHI '20). Association for Computing Machinery, New York, NY,USA, 1–14.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. 2021. Datasheets forDatasets. arXiv:1803.09010.

Bender, E. M., Friedman, B. & McMillan-Major, A., (2022). A Guide for Writing DataStatements for Natural Language Processing. University of Washington. Accessed July14, 2022.

Meta AI. System Cards, a new resource for understanding how AI systems work, 2021.

デプロイしない場合

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, ordeploy. In Proceedings of the 2020 Conference on Fairness, Accountability, andTransparency (FAT* '20). Association for Computing Machinery, New York, NY, USA,695.

統計的バランス

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019.Dissecting racial bias in an algorithm used to manage the health of populations.Science 366, 6464 (25 Oct. 2019), 447-453.

AI における科学の評価

Arvind Narayanan. How to recognize AI snake oil. URL

Emily M. Bender. 2022. On NYT Magazine on AI: Resist the Urge to be Impressed.(April 17, 2022).

Map 1.2

学際的な AI アクター、コンピテンシー、スキルおよび能力は、人口統計学的多様性と幅広いドメインおよびユーザーエクスペリエンスの専門性を反映し、彼らの参加は文書化されている。

学際的な協力の機会が優先される。

概要

コンテクストのマッピングを成功させるには、多様な経験、専門知識、能力および背景を持ち、重要な調査に取り組むためのリソースと独立性を備えた AI アクターのチームが必要である。

多様なチームを持つことは、設計・開発する技術の目的や機能に関するアイデアや仮定をより広範かつオープンに共有することに貢献し、これらの暗黙の側面をより明確にする。AI リスクをマネジメントする上で、多様なスタッフがいることの利点は、個々の従業員の信念や推定した信念ではなく、集団的な視点から得られる行動である。批判的な探究心を育む環境は、問題を表出させ、既存および顕在化したリスクを特定する機会を生み出す。

推奨されるアクション

- AI の取り組みに必要な幅広いスキル、コンピテンシーおよび能力を反映させる学際的なチームを設立する。チームメンバーに、人口統計学的多様性、幅広い分野の専門知識、実際の経験が含まれていることを検証する。チーム構成を文書化する
- デプロイされた AI システムと、法律、社会学、心理学、人類学、公共政策、システム設計および工学などの AI プラクティス以外の学問分野からの関連する用語や概念との相互依存性を把握、学習および取り組むための学際的な専門家チームを創設し、支援する

透明性と文書化

組織は以下を文書化することができる。

- AI システムの開発・保守を担当するチームは、多様な意見、背景、経験、視点をどの程度反映しているのか？
- エンティティは、開発プロセスに内在する潜在的なバイアスを把握し、コミュニケーションをとるために AI システムの設計および開発に関与したフォーラムの参加者による人口統計を文書化したか？
- ステークホルダは、AI システムの設計、開発、デプロイ、評価、モニタリングにわたって、具体的にどのような観点を共有し、どのように統合したのか？
- エンティティは、AI システムがエンドユーザや影響を受ける人々に及ぼす潜在的なネガティブのインパクトについて、ステークホルダの視点でどの程度対処したか？
- AI システムの設計、運用、制限に関して、AI システムの使用によって影響を受けるエンドユーザ、消費者、規制当局を含む外部ステークホルダはどのような情報にアクセスできるか？

- 組織は、ユーザビリティの問題に取り組み、ユーザーインタフェースが本来の目的を果たしているかどうかをテストしたか？開発の初期段階でコミュニティやエンドユーザに相談し、使用される技術やその挿入方法について透明性を確保する

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

・AI policies and initiatives, in Artificial Intelligence in Society, OECD,2019.

参考文献

Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machinelearning systems. Big Data & Society 9, 1 (Jan. 2022).

Microsoft Community Jury , Azure Application Architecture Guide.

Fernando Delgado, Stephen Yang, Michael Madaio, Qian Yang. (2021). StakeholderParticipation in AI: Beyond "Add Diverse Stakeholders and Stir".

Kush Varshney, Tina Park, Inioluwa Deborah Raji, Gaurush Hiranandani, NarasimhanHarikrishna, Oluwasanmi Koyejo, Brianna Richardson, and Min Kyung Lee.

Participatory specification of trustworthy machine learning, 2021.

Donald Martin, Vinodkumar Prabhakaran, Jill A. Kuhlberg, Andrew Smart and WilliamS. Isaac. “Participatory Problem Formulation for Fairer Machine Learning ThroughCommunity Based System Dynamics”, ArXiv abs/2005.07572 (2020).

Map 1.3

組織の使命と AI 技術に関連する目標が理解され、文書化されている。

概要

より広範な社会的価値観の中で AI システムの具体的なビジネス目的を定義し、文書化することは、チームがリスクを評価するのに役立ち、デプロイの可否に関する「実施／非実施」の判断をより明確にする。

トラストワージな AI 技術は、暗黙的または明示的なコストを上回る実証可能なビジネス上のベネフィットをもたらし、付加価値を提供し、リソースの無駄につながることはない。暗黙的または明示的なリスクが AI システムの利点を上回る場合や、リスクが潜在的な利点を上回る AI ソリューションを導入しない場合について、組織は安心してリスク回避を行うことができる。

例えば、AI システムをより公平なものにすることは、リスクをよりよくマネジメントすることにつながり、包括的に設計され、利用しやすく、より公平な AI システムを作ることのビジネス価値を検討する助けになる。

推奨されるアクション

- AI システム開発プロセスに透明性のあるプラクティスを組み込む
- 社会技術的観点や社会的価値の考慮からの文書化されたシステム目的を検討する
- 社会的価値観と、定められた組織原則および倫理規範との間で起こりうる不整合を判断する
- マイナスのインパクトの一因となりうる潜在的な誘因を明らかにする
- 潜在的なリスク、社会的価値、定められた組織原則を考慮し、AI システムの目的を評価する

透明性と文書化

組織は以下を文書化することができる。

- AI システムは、エンティティが目標と目的を達成するためにどのように役立つのか？
- 技術仕様と要件は、AI システムの目標と目的にどのように合致しているか？
- 出力は、どの程度まで運用のコンテクストに適切か？

AI の透明性に関するリソース

・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI – 2019, URL. LINK

・Including Insights from the Comptroller General's Forum on theOversight of Artificial Intelligence An Accountability Framework forFederal Agencies and Other Entities, 2021, URL, PDF.

参考文献

M.S. Ackerman (2000). The Intellectual Challenge of CSCW: The Gap Between SocialRequirements and Technical Feasibility. Human–Computer Interaction, 15, 179 - 203.

McKane Andrus, Sarah Dean, Thomas Gilbert, Nathan Lambert, Tom Zick (2021). AIDevelopment for the Public Interest: From Abstraction Traps to Sociotechnical Risks.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, et al. 2022. The Values Encoded inMachine Learning Research. arXiv:2106.15590.

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on ModelRisk Management. (April 4, 2011).

Iason Gabriel, Artificial Intelligence, Values, and Alignment. Minds & Machines 30,411–437 (2020).

PEAT “Business Case for Equitable AI”.

Map 1.4

ビジネスの価値あるいはビジネス利用のコンテクストが明確に定義または（既存の AI システムを評価する場合には）再評価されている。

概要

社会技術的 AI リスクは、技術開発の意思決定と、システムがどのように使用されるか、誰が操作するか、そしてシステムがデプロイされる社会的コンテクストとの相互作用から生じる。このようなリスクへの対処は複雑であり、コンテクストの要因が AI ライフサイクルのアクションとどのように相互作用するかを理解するための取り組みが必要である。そのようなコンテクストの要因のひとつは、組織の使命と決定されたシステム目的が、AI システムの設計、開発およびデプロイタスクの中で、どのようなインセンティブを生み出し、それがポジティブおよびネガティブなインパクトをもたらす可能性があるかということである。AI システムのコンテクストと期待について包括的かつ明示的に列挙することにより、組織はこの種のリスクを特定し、マネジメントすることができる。

推奨されるアクション

- ビジネス価値またはビジネス利用のコンテクストを文書化する
- 組織が掲げる価値観、ミッション・ステートメント、社会的責任の取り組みおよび AI 原則と、ビジネスでの活用のコンテクストにおけるシステムの目的に関する文書化された懸念事項とを照らし合わせる
- 組織の価値観を反映しない潜在的なインパクトを持つ AI システムの設計、実装戦略およびデプロイメントを再考する

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、AI システムの設計、開発および／またはデプロイによって、どのようなゴールおよび目標を達成することを期待しているのか？
- システムの出力は、社会の信頼と公平性を育むためのエンティティの価値観や原則とどの程度一致しているか？
- メトリクスは倫理およびコンプライアンスへの配慮を含む、システム目標、目的および制約条件とどの程度一致しているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・WEF Model AI Governance Framework Assessment 2020. URL

参考文献

Algorithm Watch. AI Ethics Guidelines Global Inventory.

Ethical OS toolkit.

Emanuel Moss and Jacob Metcalf. 2020. Ethics Owners: A New Model ofOrganizational Responsibility in Data-Driven Technology Companies. Data & SocietyResearch Institute.

Future of Life Institute. Asilomar AI Principles.

Leonard Haas, Sebastian Gießler, and Veronika Thiel. 2020. In the realm of papertigers – exploring the failings of AI ethics guidelines. (April 28, 2020).

Map 1.5

組織のリスク許容度が決定され、文書化される。

概要

リスク許容度は、組織がその使命を遂行し、戦略を実行する際に受け入れるリスクのレベルと種類を反映する。

組織は、その組織、ドメイン、専門分野、セクター、専門的要件によって確立されたリスク基準、許容度、対応に関する既存の規制やガイドラインに従うことができる。一部のセクターや業界では、危害の定義が確立されてきており、文書化、報告、開示の要件も確立されているかもしれない。

セクター内では、リスマネジメントは、特定のアプリケーションやユースケースの設定に関する既存のガイドラインに依存する場合がある。確立されたガイドラインが存在しない場合、組織は、様々なリスク源（例えば、財務、オペレーショナル、安全と福祉、ビジネス、評判およびモデルのリスク）および様々なリスクレベル（例えば、無視できるものからクリティカルなものまで）を考慮して、合理的なリスク許容度を定義することが望ましい。

リスク許容度は、「実施／非実施」と呼ばれる、開発あるいはデプロイを続行するかの意思決定に情報を提供し、支援する。AIシステムのリスクに関連する「実施／非実施」の決定では、ステークホルダのフィードバックを考慮に入れることができるが、ステークホルダの既得権益である金銭的利益や風評的利益からは独立している。

リスクのマッピングが法外に困難な場合は、特定のシステムに対して「非実施」"の決定が検討されるかもしれない。

推奨されるアクション

- 組織、ドメイン、専門分野、セクター、あるいは専門的要件によって確立されたリスク基準、許容度および対応に関する既存の規制やガイドラインを活用する
- AI システムのリスク許容度を設定し、各レベルに適切なオーバーサイトリソースを割り当てる
- さまざまなリスク源（例えば、財務、オペレーショナル、安全と福祉、ビジネス、評判およびモデルのリスク）やリスクのレベル（例えば、無視できるものからクリティカルなものまで）を考慮してリスク基準を確立する
- コンテクストやアプリケーションの設定の中でそれを超えるとシステムがデプロイされなくなるまたは早期の廃止が必要となる最大リスク許容範囲を特定する
- 提案された使用コンテクストに関連してトラストワージな特性全体にわたるトレードオフを明確にし、分析する。トレードオフが生じた場合は、経営上の決定を通知するためにそれらを文書化し、追跡可能なアクション（例えば、インパクトの軽減、システム開発または利用の除外）を計画する
- 「適応外」目的での AI システムの使用、特に組織が高リスクとみなした設定を見直す。意思決定、リスクに関連するトレードオフ、システムの制限を文書化する

透明性と文書化

組織は以下を文書化することができる。

- システムリスク許容度の策定において、どのような既存の規制やガイドラインが適用され、エンティティはそれに従っているか？
- システムリスク許容度を策定する際に、エンティティはどのような基準と前提を利用したか？
- エンティティは、最大許容リスク許容度をどのように決定したか？
- どのような条件と目的が、システム使用の「適応外」と見なされるのか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

参考文献

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on ModelRisk Management. (April 4, 2011).

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement.(Nov. 20, 2019).

Brenda Boultwood, How to Develop an Enterprise Risk-Rating Approach (Aug. 26,2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023.

Virginia Eubanks, 1972-, Automating Inequality: How High-tech Tools Profile, Police,and Punish the Poor. New York, NY, St. Martin's Press, 2018.

GAO-17-63: Enterprise Risk Management: Selected Agencies' Experiences IllustrateGood Practices in Managing Risk. (See Table 3).

NIST Risk Management Framework.

Map 1.6

システム要件（例えば「システムは利用者のプライバシーを尊重するものとする」など）は、関連する AI アクターから引き出され、理解される。AIリスクに対処するため、社会技術的な影響を考慮した設計決定を行う。

概要

AI システムの開発要件は、従来のソフトウェアの文書化プロセスを上回る可能性がある。文書化された要件が入手できない、あるいは不完全である場合、AI アクターはビジネスやステークホルダのニーズを見落とし、確証バイアスや集団思考といった人間の暗黙のバイアスに頼りすぎおよび計算要件にのみ焦点を当て続ける可能性がある。

システム要件を引き出し、エンドユーザ向けに設計し、設計の初期段階で社会的インパクトを考慮することは、AI システムのトラストワージネスを高めることができる優先事項である。

推奨されるアクション

- トラストワージな特性をシステム要件に積極的に組み込む
- システム設計やデプロイの決定に関連する、関係する AI アクターと内外のステークホルダとの定期的なコミュニケーションとフィードバックの仕組みを確立する
- AI のライフサイクルのすべての段階において、学際的なエキスパート、AI システムを開発またはデプロイしたチームの外部のアクターおよびインパクトを受ける可能性のあるコミュニティと協力して、潜在的なインパクトを評価するための手法を開発し、標準化する
- インパクトのアセスメント活動における優先順位、定義およびアウトカムの策定には、インパクトを受ける可能性のあるグループ、コミュニティおよび外部団体（例えば、市民社会組織、研究機関、地域コミュニティ団体、業界団体）を含める
- エンドユーザとの定性的なインタビューを実施し、人間と AI の構成およびタスクに関する期待と設計計画を定期的に評価する
- コンテクスト的要因とシステム要件との間の依存関係を分析する。意思決定においてトラストワージネス特性の重要性を十分に考慮しないことから生じる潜在的なインパクトを列挙する
- ソフトウェアエンジニアリング、製品管理、参加型エンゲージメントなどのタスクにおいて、責任ある設計手法に従うこと。ステークホルダの要件を引き出し、文書化するためのいくつかの例としては、製品要件文書（PRD）、ユーザーストーリー、ユーザーインタラクション／ユーザーエクスペリエンス（UI／UX）調査、システムエンジニアリング、エスノグラフィーおよび関連するフィールドメソッドなどがある
- AI によってインパクトを受ける個人、グループおよびコミュニティ、彼らの価値観や背景、系統的・歴史的バイアスの役割を理解するためのユーザーリサーチを実施する。データの選択と表現に関する意思決定に学習内容を統合する

透明性と文書化

組織は以下を文書化することができる。

- AI システムの設計、運用および制限に関して、エンドユーザ、消費者、規制当局、AI システムの使用によって影響を受ける個人を含む外部のステークホルダが、どのような情報にアクセスできるか？
- この情報は、透明性を促進するためにどの程度十分かつ適切か？外部のステークホルダが AI システムの設計、運用および制限に関する情報にアクセスできるようにすることで、透明性を促進する
- AI システムの使用に関して、(a)何のためのシステムか、(b)何のためのシステムでないか、(c)どのように設計されたか、(d)その限界は何か、といった関連情報がどの程度開示されているか？(文書化と外部コミュニケーションは、エンティティが透明性を提供する方法を提供することができる)
- AI アクターは、AI システムを混乱させようとする敵対者の試みにより、正確性あるいは精度が変化することや、ビジネス環境あるいは運用環境における無関係な変化により AI システムの正確性に影響がでるような事象にどのように対処するのか。
- AI システムのパフォーマンスを測定するために、企業はどのような指標を開発したか。
- 存在する場合、エンティティは AI システムの前提、境界線、制限に対して、どのような正当性があるか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・Stakeholders in Explainable AI, Sep. 2018.

・ High-Level Expert Group on Artificial Intelligence set up by theEuropean Commission, Ethics Guidelines for Trustworthy AI. URL,PDF

参考文献

National Academies of Sciences, Engineering, and Medicine 2022. FosteringResponsible Computing Research: Foundations and Practices. Washington, DC: TheNational Academies Press.

Abeba Birhane, William S. Isaac, Vinodkumar Prabhakaran, Mark Diaz, MadeleineClare Elish, Iason Gabriel and Shakir Mohamed. “Power to the People? Opportunitiesand Challenges for Participatory AI.” Equity and Access in Algorithms, Mechanisms,and Optimization (2022).

Amit K. Chopra, Fabiano Dalpiaz, F. Başak Aydemir, et al. 2014. Protos: Foundationsfor engineering innovative sociotechnical systems. In 2014 IEEE 22nd InternationalRequirements Engineering Conference (RE) (2014), 53-62.

Andrew D. Selbst, danah boyd, Sorelle A. Friedler, et al. 2019. Fairness andAbstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness,Accountability, and Transparency (FAT* '19). Association for Computing Machinery,New York, NY, USA, 59–68.

Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From designmethods to systems engineering. Interacting with Computers, 23, 1 (Jan. 2011), 4–17.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices inartificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702.

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design inFuture Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019),125-150. Springer, Cham.

Victor Udoewa, (2022). An introduction to radical participatory design: decolonisingparticipatory design processes. Design Science. 8. 10.1017/dsj.2022.24. URL

Map 2

AI システムの分類を行う。

Map 2.1

AI システムがサポートする特定のタスクおよびそのタスクの実装に使用される方法が定義される（例えば、分類器、生成モデル、レコメンダ）。

概要

AI アクターは、AI システムが達成するように設計された技術的学習または意思決定のタスク、あるいはシステムが提供するベネフィットを定義する。タスクの定義が明確かつ狭いほど、そのベネフィットとリスクのマッピングが容易になり、より充実したリスクマネジメントにつながる。

推奨されるアクション

- AI システムの既存および潜在的な学習課題を、既知の前提条件や制限とともに定義し、文書化する

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、AI システムの技術仕様や要件をどの程度明確に定義しているか？
- エンティティは、AI システムの開発、テスト方法、測定基準、パフォーマンス結果をどの程度文書化しているか？
- 技術仕様と要件は、AI システムの目標と目的にどのように合致しているか？
- 組織は、データ管理と保護においてアカウンタビリティに基づくプラクティス（PDPA や OECD プライバシー原則など）を実行したか？
- AI からの出力であることを明確に示すために、出力にはどのようなマークが付けられるのか？

AI の透明性に関するリソース

・Datasheets for Datasets.
・WEF Model AI Governance Framework Assessment 2020.
・WEF Companion to the Model AI Governance Framework- 2020. URL
・ATARC Model Transparency Assessment (WD) – 2020.

参考文献

Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020.

Leong, Brenda (2020). The Spectrum of Artificial Intelligence - An Infographic Tool.Future of Privacy Forum.

Brownlee, Jason (2020). A Tour of Machine Learning Algorithms. Machine LearningMastery.

Map 2.2

AI システムの知識の限界と、システムの出力がどのように利用され、人間によってオーバーサイトされうるかについての情報が文書化されている。文書化は、関連する AI アクターが情報に基づいて意思決定を行い、その後のアクションをとる際に役立つ十分な情報を提供する。

概要

AI のライフサイクルは、多様なアクターが関与する多くの相互依存的なアクションから構成されており、多くの場合、ライフサイクルの他の部分や関連するコンテクストやリスクを完全に可視化または制御することはできない。このようなアクション間の相互依存や、関連する AI アクターや組織間の相互依存は、AI システムの潜在的なインパクトを確実に予測することを困難にする可能性がある。例えば、AI システムの目的と目標を特定する際の早期の決定が、そのふるまいと能力を変化させる可能性があり、また、（エンドユーザや影響を受ける個人などの）デプロイ環境のダイナミクスが、AI システムの決定によるポジティブあるいはネガティブなインパクトを形成する可能性がある。その結果、AI のライフサイクルのある側面における最善の意図が、その後の他のアクティビティにおける意思決定や条件との相互作用によって損なわれる可能性がある。このような複雑さと可視性のさまざまなレベルは、不確実性をもたらす可能性がある。また、AI システムは、一旦デプロイされ使用されると、時として性能が低下したり、予期せぬネガティブなインパクトが現れたり、法的・倫理的規範に違反したりすることがある。このようなリスクやインシデントは、さまざまな要因から生じる可能性がある。例えば、下流の意思決定は、エンドユーザの過信や不信、AI 支援による意思決定に関連するその他の複雑さによって影響を受ける可能性がある。

AI システムの知識の限界とシステムの出力が人間によってどのように利用およびオーバーサイトされるかを予測、明確化、アセスメント、文書化することは、AI システムのデプロイメントの現実に伴う不確実性を軽減するのに役立つ。厳密な設計プロセスには、システムの知識の限界を定義することが含まれ、それは TEVV プロセスに基づいて確認され、改善される。

推奨されるアクション

AI システムの使用目的外の設定、環境、条件を文書化する。

- エンドユーザのワークフローとツールセット、操作概念、エンドユーザおよび関連した定性のフィードバックと組み合わせた説明可能性と解釈可能性の基準を設計する
- 現実世界に近い条件下で人間と AI の構成を計画およびテストし、結果を文書化する
- ステークホルダのフィードバックプロセスに従い、システムが所与の使用コンテクスト状況内で文書化された目的を達成したかどうか、またエンドユーザがシステムの出力や結果を正しく理解できるかどうかを判断する
- 指定されたシステムが、別の AI システムまたは他のデータに対する上流の依存関係であるかどうかを含め、上流データと別の AI システムとの依存関係を文書化する
- AI システムまたはデータが、ネガティブな外部性の可能性がある外部ネットワーク（インターネットを含む）、金融市場および重要インフラと接続することを文書化する。より広範なリスク閾値を考

慮する一貫としてネガティブなインパクトを特定するとともに、それに続くデプロイの実施／非実施、加えてデプロイの廃止決定を特定し、文書化する

透明性と文書化

組織は以下を文書化することができる。

- AI システムは、人員が十分な情報に基づいた意思決定を行い、それに応じて行動するのに役立つ十分な情報を提供しているか？
- AI システムの設計、運用、制限に関して、エンドユーザ、消費者、規制当局および AI システムの使用によってインパクトを受ける個人を含む外部のステークホルダは、どのような種類の情報にアクセスできるか？
- アセスメントに基づいて、組織は AI を活用した意思決定に適切なレベルの人的関与を導入したか？

AI の透明性に関するリソース

・Datasheets for Datasets. URL

・WEF Model AI Governance Framework Assessment 2020.

・WEF Companion to the Model AI Governance Framework- 2020.

・ATARC Model Transparency Assessment (WD) – 2020. URL

・Transparency in Artificial Intelligence - S. Larsson and F. Heintz –2020.

参考文献

使用のコンテクスト

International Standards Organization (ISO). 2019. ISO 9241-210:2019 Ergonomics ofhuman-system interaction — Part 210: Human-centered design for interactivesystems.

National Institute of Standards and Technology (NIST), Mary Theofanos, Yee-YinChoong, et al. 2017. NIST Handbook 161 Usability Handbook for Public SafetyCommunications: Ensuring Successful Systems for First Responders.

人間と AI の相互作用

Committee on Human-System Integration Research Topics for the 711th HumanPerformance Wing of the Air Force Research Laboratory and the National Academiesof Sciences, Engineering, and Medicine. 2022. Human-AI Teaming: State-of-the-Artand Research Needs. Washington, D.C. National Academies Press.

Human Readiness Level Scale in the System Development Process, AmericanNational Standards Institute and Human Factors and Ergonomics Society, ANSI/HFES400-2021

Microsoft Responsible AI Standard, v2.

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector DecisionMaking: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journalof Public Administration Research and Theory, 2022;, muac007.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or toThink: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assistedDecision-making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (April2021), 21 pages.

Mary L. Cummings. 2006 Automation and accountability in decision support systeminterface design. The Journal of Technology Studies 32(1): 23–31.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M. F. (2020). Government byalgorithm: Artificial intelligence in federal administrative agencies. NYU School of Law,Public Law Research Paper, (20-54).

Susanne Gaube, Harini Suresh, Martina Raue, et al. 2021. Do as AI say: susceptibilityin deployment of clinical decision-aids. npj Digital Medicine 4, Article 31 (2021).

Ben Green. 2021. The Flaws of Policies Requiring Human Oversight of GovernmentAlgorithms. Computer Law & Security Review 45 (26 Apr. 2021). URL

Ben Green and Amba Kak. 2021. The False Comfort of Human Oversight as anAntidote to A.I. Harm. (June 15, 2021). URL

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making withmachine assistance: An experiment on bailing and jailing. Proceedings of the ACM onHuman-Computer Interaction, 3(CSCW), 1-25.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, et al. 2021.Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHIConference on Human Factors in Computing Systems (CHI '21). Association forComputing Machinery, New York, NY, USA, Article 237, 1–52.

C. J. Smith (2019). Designing trustworthy AI: A human-machine teaming framework toguide development. arXiv preprint arXiv:1910.03515.

T. Warden, P. Carayon, EM et al. The National Academies Board on Human SystemIntegration (BOHSI) Panel: Explainable AI, System Transparency, and HumanMachine Teaming. Proceedings of the Human Factors and Ergonomics SocietyAnnual Meeting. 2019;63(1):631-635. doi:10.1177/1071181319631100.

Map 2.3

科学的完全性と TEVV への配慮を決定し、文書化する。これには、実験計画、データ収集と選択（例えば可用性、代表性、適合性など）、システムのトラストワージネス、構成の妥当性に関連するものが含まれる。

概要

標準的な試験・評価プロトコルは、システムが設計・要求どおりに動作していることを確認するための基礎となる。AI システムの複雑さは、静的または分離されたシステム性能のために設計されがちな従来の試験・評価手法に課題をもたらす。リスクにさらされる機会は、設計やデプロイをはるかに超えて、システムの運用やシステムが可能にする意思決定の適用にまで及ぶ。そのため、テストと評価の方法論とメトリクスは、一連の活動に対応する。パフォーマンス、安全性および信頼性に関する主要なメトリクスが、AI システムのパイプラインの境界に限定されることなく、社会技術的なコンテクストで解釈されることにより、TEVV は強化される。

AI のリスクをマネジメントするための他の課題は、大規模なデータセットへの依存に関連しており、データの品質と妥当性の懸念にインパクトを与える可能性がある。「適切な」データを見つけることが困難であるため、AI アクターは、AI システムが支援あるいは情報提供しようとしている現象の運用に適しているかどうかよりも、アクセシビリティと可用性を重視してデータセットを選択する可能性がある。このような決定は、プロセスで使用されるデータが、モデル化されている集団や現象を完全に代表するものではなく、下流にリスクをもたらす可能性がある。データセットの再利用のようなプラクティスは、そのデータセットが作成された社会的なコンテクストや時代から切り離すことにもつながる。これは、モデル内でプロキシ、メジャー、あるいは予測子を提供するための基礎となるデータセットの妥当性の問題に寄与する。

推奨されるアクション

- ヒューマンファクター、創発的特性、動的なコンテクストを伴う、AI のような複雑な社会技術システムのテストに妥当な実験設計と統計手法を特定し、文書化する
- モデル、システムおよびそのサブコンポーネント、デプロイおよび運用に関する TEVV プロトコルを開発し、適用する
- AI システムのパフォーマンスと妥当性のメトリックスが、下流の意思決定タスクに対して解釈可能で曖昧さがなく、使用のコンテクストなどの社会技術的要因を考慮したものであることを実証し、文書化する
- Govern で確立されたデータガバナンスのポリシーに従い、データの収集、選択およびマネジメントのプラクティスに関する実験設計手法を含む、AI のライフサイクル全体のテストと評価に使用される前提条件、手法およびメトリクスを定め、文書化する
- AI ライフサイクル全体に組み込むことができるテストモジュールを特定し、独立した評価者による裏付けを可能にするプロセスを検証する

- 設計とデプロイメントの前提条件の妥当性に関して、AI アクターや内外のステークホルダ間の定期的なコミュニケーションとフィードバックを行うためのメカニズムを確立する
- 潜在的に有害なインパクトを検出しアセスメントするために、ライフサイクル全体にわたる TEVV アプローチの開発に関連する AI アクターと内外のステークホルダとの間で定期的にコミュニケーションとフィードバックを行うためのメカニズムを確立する
- 以下を含む、データの選択、キュレーション、準備、分析に用いた仮定と技法を文書化する
  - コンストラクトとプロキシターゲットの特定
  - 指標の開発、特に本質的に観測不可能な概念（例えば「雇用可能性」「犯罪性」および「貸与性」）を運用化するもの
- データに対処し、AI システムのバイアス、プライバシーおよびセキュリティを構築するためのポリシーの遵守をマッピングし、文書化、オーバーサイトおよびプロセスを検証する
- モデル化されている構成要素とデータセットの属性またはプロキシとの間の因果関係を推論するための透明性のある方法（因果関係発見方法など）を特定し、文書化する
- リスクをマッピングするためのテストとトレーニングのデータ系統、そのメタデータリソースを理解し、トレースするプロセスを特定し、文書化する
- 学習データの収集、選択、ラベル付け、クリーニング、分析（例えば、欠損データ、偽データ、異常値データの処理、偏った推定量）に使用される方法に関係する既知の制限、関連するリスク軽減の取り組みを文書化する
- 緊急の特性など、計画されている能力を超える機能をチェックし、新たな能力に照らして事前のリスク管理手順を再検討するための実践を確立し、文書化する
- デプロイするコンテクストに関する設計上の前提条件が、引き続き正確で十分に完全であることをテストおよび検証するプロセスを確立する
- ドメインのエキスパートや外部の AI アクターと協力して、以下のことを行う
  - 人間の行動、組織の要因とダイナミクス、社会がデータセット、プロセス、モデル、システムの出力にどのような影響を与え、どのように表現されるかについて、コンテクストを認識し、知識を取得・維持する
  - 認知バイアスの原因を考慮し、責任ある人間と AI の構成とオーバーサイトタスクに対する参加型アプローチを特定する
  - 計算モデルとシステムにおけるバイアスの原因（系統的、計算的、人間認知的）と、それらの開発における家庭と意思決定を管理および軽減するための手法を特定する
- 組織の価値観や原則と矛盾する可能性のある、製品ライフサイクル全体および関連プロセスに関連する潜在的なネガティブなインパクトを調査し、文書化する

透明性と文書化

組織は以下を文書化することができる。

- データに既知のエラー、ノイズ源、冗長性はあるか？

- データはどのような期間にわたって収集されたか？収集期間は作成期間と一致しているか？
- 変数の選択と評価プロセスとは何か？
- データはどのように収集されたか？データ収集プロセスには誰が関わったか？データセットが人（属性など）に関連している場合、あるいは人によって作成された場合、その人たちはデータ収集について知らされていたか？(例えば、文章、写真、やり取り、取引などを収集するデータセットなど)
- 時間が経過し、条件が変化しても、学習データは依然として運用環境を代表しているか？
- そのデータセットはなぜ作られたのか？(例えば、特定のタスクが念頭にあったのか、あるいは、埋めなければならない特定のギャップがあったのか？)
- エンティティは、収集したデータが意図された目的に対して適切で関連性があり、過剰でないことをどのように保証するのか？

AI の透明性に関するリソース

・Datasheets for Datasets.
・WEF Model AI Governance Framework Assessment 2020.
・WEF Companion to the Model AI Governance Framework- 2020.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・ATARC Model Transparency Assessment (WD) – 2020.
・Transparency in Artificial Intelligence - S. Larsson and F. Heintz –2020.

参考文献

データセット選択の課題

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. SocialData: Biases, Methodological Pitfalls, and Ethical Boundaries. Front. Big Data 2, 13(11 July 2019).

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Dataand its (dis)contents: A survey of dataset development and use in machine learningresearch. arXiv:2012.05345.

Catherine D'Ignazio and Lauren F. Klein. 2020. Data Feminism. The MIT Press,Cambridge, MA.

Miceli, M., & Posada, J. (2022). The Data-Production Dispositif. ArXiv,abs/2205.11963.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language inNLP. arXiv:1608.07836. URL

AI システム開発におけるデータセットとテスト、評価、妥当性確認および検証（TEVV）プロセス

National Institute of Standards and Technology (NIST), Reva Schwartz, ApostolVassilev, et al. 2022. NIST Special Publication 1270 Towards a Standard forIdentifying and Managing Bias in Artificial Intelligence.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, et al. 2021. AI andthe Everything in the Whole Wide World Benchmark. arXiv:2111.15366.

統計的バランス

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019.Dissecting racial bias in an algorithm used to manage the health of populations.Science 366, 6464 (25 Oct. 2019), 447-453. URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Dataand its (dis)contents: A survey of dataset development and use in machine learningresearch. arXiv:2012.05345.

Solon Barocas, Anhong Guo, Ece Kamar, et al. 2021. Designing DisaggregatedEvaluations of AI Systems: Choices, Considerations, and Tradeoffs. Proceedings ofthe 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association forComputing Machinery, New York, NY, USA, 368–378.

測定と評価

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. InProceedings of the 2021 ACM Conference on Fairness, Accountability, andTransparency (FAccT ‘21). Association for Computing Machinery, New York, NY,USA, 375–385.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, et al. 2022. Evaluation Gaps inMachine Learning Practice. arXiv:2205.05256.

Laura Freeman, "Test and evaluation for artificial intelligence." Insight 23.1 (2020): 27-30.

既存のフレームワーク

National Institute of Standards and Technology. (2018). Framework for improvingcritical infrastructure cybersecurity.

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool forImproving Privacy Through Enterprise Risk Management, Version 1.0." NationalInstitute of Standards and Technology (NIST), January 16, 2020.

Map 3

AI の機能、対象となる使用方法、目標、適切なベンチマークと比較した期待されるベネフィットとコストを理解する。

Map 3.1

意図した AI システムの機能とパフォーマンスの潜在的なベネフィットを検討し、文書化する。

概要

AI システムには、生活の質を向上させ、経済的繁栄とセキュリティコストを高める大きな可能性がある。組織は、システムの目的と有用性および現在知られている性能ベンチマークを超える潜在的なポジティブなインパクトを定義し、文書化することが奨励される。

リスクマネジメントおよびベネフィットとインパクトのアセスメントには、影響を受ける可能性のあるグループやコミュニティと定期的かつ有意義なコミュニケーションを行うプロセスが含まれることが推奨される。これらのステークホルダは、システムのベネフィットと考えられる制限に関する貴重な意見を提供してくれる可能性がある。関与するステークホルダの種類や数は組織によって異なる場合がある。

人間中心設計(HCD)や価値重視設計(VSD)など、他のアプローチは、AI チームが個人やコミュニティと広く関与するのに役立つ。このような関わり方によって AI チームは、特定のテクノロジーが、当初は考慮または意図されていなかったポジティブあるいはネガティブなインパクトをどのように引き起こす可能性があるかを知ることができる。

推奨されるアクション

- 参加型アプローチを活用し、システムのエンドユーザと連携し、AI システムの潜在的な利点、有効性、AI タスクの出力の解釈可能性を理解し、文書化する
- システムの内部および外部のステークホルダを構成する個人、グループまたはコミュニティに対する認識を維持し、文書化する
- ユーザ、運用者、外部からのフィードバックを引き出し、収集し、統合し、それを AI 設計・開発機能に変換するなどの参加型活動を実施するための適切なスキルとプラクティスが社内で利用可能であることを検証する
- システム設計やデプロイの決定に関連する、関係する AI アクターと内外のステークホルダとの定期的なコミュニケーションとフィードバックの仕組みを確立する
- 人間のベースライン指標や他の標準ベンチマークに対するパフォーマンスを検討する
- 認識されているシステムの利点について、エンドユーザ、影響を受ける可能性のある個人やコミュニティからのフィードバックを取り入れる

透明性と文書化

組織は以下を文書化することができる。

- AI システムの利点はエンドユーザに伝わっているか？
- AI システムの適切な使用方法について、適切なトレーニング資料や免責事項がエンドユーザに提供されているか？
- 組織は、決定された AI システムの導入に伴うリスク（例えば、人的リスクや商業目的の変更など）に対処するためのリスク管理システムを実装しているか？

AI の透明性に関するリソース

・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI – 2019. LINK

参考文献

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices inartificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN0004-3702.

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. InProceedings of the Conference on Fairness, Accountability, and Transparency(FAT* '19). Association for Computing Machinery, New York, NY, USA, 39–48.

Vincent T. Covello. 2021. Stakeholder Engagement and Empowerment. InCommunicating in Risk, Crisis, and High Stress Situations (Vincent T. Covello,ed.), 87-109.

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design inFuture Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019),125-150. Springer, Cham.

Eloise Taysom and Nathan Crilly. 2017. Resilience in Sociotechnical Systems: ThePerspectives of Multiple Stakeholders. She Ji: The Journal of Design, Economics, andInnovation, 3, 3 (2017), 165-182, ISSN 2405-8726. URL

Map 3.2

組織のリスク許容度に関連する、予想される、あるいは実現した AI のエラーやシステムの機能やトラストワージネスから生じる、非金銭的なコストを含む潜在的なコストが検討され、文書化される。

概要

AI システムのネガティブなインパクトを予測することは難しい課題である。ネガティブなインパクトは、システムの機能不全や運用制限外での使用など、さまざまな要因に起因する可能性があり、軽微な迷惑行為から、重傷、金銭的損失、規制当局による取り締まりにまで及ぶ可能性がある。AI アクターは、幅広いステークホルダと協力し、システムの潜在的なインパクト、そして、その後のシステムのリスクに及ぼすインパクトを理解する能力を向上させることができる。

推奨されるアクション

- コンテクスト分析を実施し、トラストワージネスの特性が統合されていないことから生じる潜在的な負の悪インパクトをマッピングする。ネガティブなインパクトが直接的または明白でない場合、AI アクターは、AI システムを開発または配備したチームの外部のステークホルダや、影響を受ける可能性のあるコミュニティと協力して、調査および文書化することができる
  - 誰が被害を受けるのか？
  - 何が害を及ぼすのか？
  - どのような場合に害が生じるのか？
  - どのようにして害が生じるのか？
- 内外の AI システム障害の定性的・定量的コストを定期的に評価する手順を特定し、実施する。潜在的なリスクおよび関連するインパクトを予防、検出および／または是正するための措置を策定する。AI システムのライフサイクルを通じて、デプロイの継続／中止を決定するために、障害コストを定期的に評価する

透明性と文書化

組織は以下を文書化することができる。

- システム／エンティティは、明示された目標や目的に対する進捗状況をどの程度一貫して測定しているか？
- AI システムの出力によって影響を受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- マシンエラーとヒューマンエラーは異なる可能性があることを文書化し、説明しているか？

AI の透明性に関するリソース

・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI – 2019. LINK,

参考文献

Abagayle Lee Blank. 2019. Computer vision machine learning and future-orientedethics. Honors Project. Seattle Pacific University (SPU), Seattle, WA.

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. OvercomingFailures of Imagination in AI Infused System Development and Deployment.arXiv:2011.13416.

Jeff Patton. 2014. User Story Mapping. O'Reilly, Sebastopol, CA.

Margarita Boenig-Liptsin, Anissa Tanweer & Ari Edmundson (2022) Data ScienceEthos Lifecycle: Interplay of ethical thinking and data science practice, Journal ofStatistics and Data Science Education, DOI: 10.1080/26939169.2022.2089411

J. Cohen, D. S. Katz, M. Barker, N. Chue Hong, R. Haines and C. Jay, "The FourPillars of Research Software Engineering," in IEEE Software, vol. 38, no. 1, pp. 97-105, Jan.-Feb. 2021, doi: 10.1109/MS.2020.2973362.

National Academies of Sciences, Engineering, and Medicine 2022. FosteringResponsible Computing Research: Foundations and Practices. Washington, DC: TheNational Academies Press.

Map 3.3

システムの機能、確立されたコンテクスト、AI システムの分類に基づいて、対象となるアプリケーションの範囲を指定し、文書化する。

概要

狭い範囲で機能するシステムは、学習または意思決定タスクとシステムのコンテクストにおいて、より良いマッピング、測定およびマネジメントを可能にする傾向がある。適用範囲が狭いことは、組織内の TEVV 機能と関連リソースを軽減することにも役立つ。

例えば、インターネット上で一般の人々と対話する大規模な言語モデルやオープンエンドのチャットボットシステムには、意思決定タスクと運用コンテクストの両方から生じる変動性があるため、マッピング、測定、マネジメントが困難な多数のリスクが存在する可能性がある。代わりに、定義された「ユーザージャーニー」に従うテンプレート化された応答を利用するタスク固有のチャットボットは、より簡単にマッピング、測定およびマネジメントできる範囲である。

推奨されるアクション

- 以下に関する要因を含め、システムのデプロイメントのコンテクストを絞り込むことを検討する
  - 成果がユーザ、グループ、コミュニティ、環境に直接的または間接的にどのような影響を与えるか？
  - 再トレーニングの間にシステムがデプロイされる時間の長さ
  - システムが動作する地理的地域
  - コミュニティの基準や、システムの誤用・悪用の可能性（意図的なものであれ、予期せぬものであれ）に関連するダイナミクス
  - AI システムの特徴や機能を、他のアプリケーションの中で、あるいは他の既存のプロセスの代わりに、どのように活用できるか
- 対象となるアプリケーションの範囲を特定する際に、法務部門や調達部門の AI アクターを関与させる

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、AI システムの技術仕様や要件をどの程度明確に定義しているか？
- 技術仕様と要件は、AI システムの目標と目的にどのように合致しているか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI – 2019. LINK,

参考文献

Mark J. Van der Laan and Sherri Rose (2018). Targeted Learning in Data Science.Cham: Springer International Publishing, 2018.

Alice Zheng. 2015. Evaluating Machine Learning Models (2015). O'Reilly.

Brenda Leong and Patrick Hall (2021). 5 things lawyers should know about artificialintelligence. ABA Journal.

UK Centre for Data Ethics and Innovation, “The roadmap to an effective AI assuranceecosystem”.

Map 3.4

運用者と実務者が AI システムのパフォーマンスとトラストワージネスおよび関連する技術基準と認定に習熟するためのプロセスが定義され、アセスメントされ、文書化される。

概要

人間と AI の構成は、完全自律型から完全手動型まで多岐にわたる。AI システムは自律的に意思決定を行うことも、人間のエキスパートに意思決定を委ねることも、人間の意思決定者が追加の意見として利用することもできる。シナリオによっては、特定の分野の専門知識を持つ専門家が AI システムと連携し、特定の最終目標（例えば、別の個人に関する意思決定）に向かって作業を行う。システムの目的によっては、エキスパートが AI システムと対話することもあるが、システム自体の設計や開発に関わることはほとんどない。これらのエキスパートは、機械学習、データサイエンス、コンピューターサイエンスなど、AI の設計や開発に伝統的に関連付けられているその他分野に必ずしも精通しているわけではなく、用途によっては、そのような専門知識を必要としない可能性も高い。例えば、医療提供にデプロイされる AI システムの場合、エキスパートは医師であり、データサイエンス、データモデリングやエンジニアリング、その他のコンピュータ要素ではなく、医学に関する専門知識をもたらす。このような設定での課題は、AI システムの機能についてエンドユーザを教育することや、実務者の専門知識を置き換えることではなく、活用することである。

AI リスクをマネジメントするために、人間と自動化をどのように構成するかについては疑問が残る。プロの運用者や実務家が使用する AI システムを設計、開発、デプロイする組織が次のことを行うと、リスク管理が強化される

- このような知識の限界を認識し、あらゆるコンテクストにおける人間と AI の相互作用や構成におけるリスクと、その結果生じる可能性のあるインパクトを特定するよう努める
- AI システムを使用するとき、あるいは AI システムと相互作用するときの、人間のさまざまな役割と責任を定義し、区別する
- Map 1 で列挙され、Govern 3.2 で確立された、提案された使用コンテクストにおける AI システム運用の習熟度基準を決定する

推奨されるアクション

- 例えば、システムの特徴や能力が、選択的遵守など、人間と AI の様々な構成における既知のリスクをどのように活性化させる可能性があるかといった、デプロイや運用環境における下流の AI アクターの意思決定に影響を与える可能性のある AI システムの特徴や能力を特定し、宣言する。
- AI システムと相互作用する運用者、実務者、その他のドメインのエキスパートに対するスキルと熟練度の要件を特定する。Map 1 に列挙されている既知のリスク、低減基準、トラストワージな特性に関する情報を含む、デプロイおよび運用環境における AI アクター向けの AI システム運用文書を作成する

- AI システムの使用と既知の制限について、提案されているエンドユーザ、実務者、運用者向けのトレーニング資料を定義し、作成する
- 定義された使用コンテクストの中で AI システムを運用するための認証手順と、運用制限を超えるものについての情報を定義し、作成する
- AI システムのプロトタイピングとテスト活動に運用者、実務者、エンドユーザを参加させ、運用上の境界と許容可能なパフォーマンスを知らせる。デプロイ条件と同様のシナリオの下でテスト活動を実施する
- AI システムの運用者、実務者、エンドユーザに提供されるモデル出力がインタラクティブであり、Map 1 で定義されたコンテクストとユーザ要件に合わせて適合していることを検証する
- AI システムの出力が、下流の意思決定タスクに対して解釈可能で明確であることを検証する
- 問題のレベルとコンテクストの複雑さに合わせて、AI システムの説明の複雑さを設計する
- 意思決定環境において AI アクターが安全に操作できるよう、設計原則が整備されていることを検証する
- 人間と AI の構成、運用者、実務者の成果を追跡し、継続的な改善に統合するためのアプローチを開発する

透明性と文書化

組織は以下を文書化することができる。

- AI システムの利用が、宣言された価値観や原則と一致していることを保証するために、そのエンティティはどのような方針を策定してきたか？
- アカウンタビリティの担当者は、AI を混乱させようとする試み、あるいは AI の精度に影響を与える可能性のある運用／ビジネス環境の無関係な変化による精度や正確さの変化に、どのように対処するのか？
- エンティティは、人員が割り当てられた責任を果たすために必要なスキル、トレーニング、リソースおよび専門分野の知識を有しているかどうかをどのように評価しているか？
- AI システムを扱う関連スタッフは、AI モデルの出力や決定を解釈し、データのバイアスを検出して管理するための適切な訓練を受けているか？
- エンティティは、さまざまなコンポーネントのパフォーマンスを測定するために、どのような指標を開発してきたか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

・WEF Companion to the Model AI Governance Framework- 2020.

参考文献

National Academies of Sciences, Engineering, and Medicine. 2022. Human-AITeaming: State-of-the-Art and Research Needs. Washington, DC: The NationalAcademies Press.

Human Readiness Level Scale in the System Development Process, AmericanNational Standards Institute and Human Factors and Ergonomics Society, ANSI/HFES400-2021.

Human-Machine Teaming Systems Engineering Guide. P McDermott, C Dominguez,N Kasdaglis, M Ryan, I Trahan, A Nelson. MITRE Corporation, 2018.

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector DecisionMaking: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journalof Public Administration Research and Theory, 2022;, muac007.

Breana M. Carter-Browne, Susannah B. F. Paletz, Susan G. Campbell , Melissa J.Carraway, Sarah H. Vahlkamp, Jana Schwartz , Polly O’Rourke, “There is No “AI” inTeams: A Multidisciplinary Framework for AIs to Work in Human Teams; AppliedResearch Laboratory for Intelligence and Security (ARLIS) Report, June 2021.

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022).Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal StudiesResearch Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011.

S Mo Jones-Jang, Yong Jin Park, How do people react to AI failure? Automation bias,algorithmic aversion, and perceived controllability, Journal of Computer-MediatedCommunication, Volume 28, Issue 1, January 2023, zmac029.

A Knack, R Carter and A Babuta, "Human-Machine Teaming in Intelligence Analysis:Requirements for developing trust in machine learning systems," CETaS ResearchReports (December 2022).

SD Ramchurn, S Stein , NR Jennings. Trustworthy human-AI partnerships. iScience.2021;24(8):102891. Published 2021 Jul 24. doi:10.1016/j.isci.2021.102891.

M. Veale, M. Van Kleek, and R. Binns, “Fairness and Accountability Design Needs forAlgorithmic Support in High-Stakes Public Sector Decision-Making,” in Proceedings ofthe 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18.Montreal QC, Canada: ACM Press, 2018, pp. 1–14.

Map 3.5

人間による監視のプロセスは、Govern 機能の組織方針に従って定義、評価、文書化される。

概要

AI システムの精度と正確さが進化するにつれて、コンピューターシステムは、純粋に意思決定を支援するためまたは人間の運用者が明示的に使用し、人間の運用者の制御下で使用されるものから、人間からの限られた入力で自動化された意思決定を行うものへと変化してきた。コンピュータによる意思決定支援システムは、意思決定における、通常は人間の別のシステムを強化するものである。このような構成にすることで、人間がほとんど関与することなく出力が得られる可能性が高まる。

AI システムのガバナンスに関する様々な人間の役割と責任を定義、区別し、AI システムのオーバーサイトと AI システムを利用または操作する者を区別することで、AI リスクマネジメント活動を強化することができる。

クリティカルなシステム、リスクの高い設定、リスクが高いとみなされるシステムでは、AI システムをデプロイする前に、リスクとオーバーサイト手順の有効性を評価することが極めて重要である。

最終的には、AI システムのオーバーサイトは共同の責任であり、オーバーサイトのプラクティスを適切に認可または管理しようとする試みは、例えば Govern 機能で提案されているような組織の同意とアカウンタビリティのメカニズムがなければ、効果的なものとはならない。

推奨されるアクション

- Map 1 で特定された、運用および社会的なコンテクスト、トラストワージな特性およびリスクに関連して、ヒューマンオーバーサイトが必要な AI システムの特徴と能力を決定し、文書化する
- Govern 1 で策定されたポリシーに従って、AI システムのオーバーサイト方法を確立する
- AI システムのパフォーマンス、使用コンテクスト、既知の制限とネガティブなインパクト、推奨される警告表示について、関連する AI アクター向けのトレーニング資料を定義し、作成する
- AI システムのプロトタイピングおよびテスト活動に関連する AI アクターを含めること。デプロイ条件と同様のシナリオ下でテスト活動を実施すること
- AI システムの監視プラクティスの妥当性と信頼性を評価する。オーバーサイトのプラクティスが大幅に更新または適応された場合は、再テストし、結果を評価し、必要に応じて軌道修正する
- モデルに関する文書にシステムのメカニズムに関する解釈可能な記述が含まれていることを検証し、オーバーサイトする人員がシステムリスクについて、情報に基づいたリスクベースの意思決定を行えるようにする

透明性と文書化

組織は以下を文書化することができる。

- AI システムの設計、開発、デプロイ、アセスメント、モニタリングに携わる人員の役割、責任および権限の委任は何か？
- エンティティは、人員が割り当てられた責任を果たすために必要なスキル、トレーニング、リソースおよび専門知識を有しているかどうかをどのようにアセスメントしているか？
- AI システムを扱う関連スタッフは、AI モデルの出力や決定を解釈し、データのバイアスを検出してマネジメントするための適切な訓練を受けているか？
- エンティティは、AI システムの開発、テスト方法、測定基準、パフォーマンス結果をどの程度文書化しているか。

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.

参考文献

Ben Green, “The Flaws of Policies Requiring Human Oversight of GovernmentAlgorithms,” SSRN Journal, 2021.

Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers,Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-JanHouben, Catholijn Jonker, Jeroen van den Hoven, Deborah Forster, & ReginaldLagendijk (2021). Meaningful human control: actionable properties for AI systemdevelopment. AI and Ethics.

Mary Cummings, (2014). Automation and Accountability in Decision Support SystemInterface Design. The Journal of Technology Studies. 32. 10.21061/jots.v32i1.a.4.

Madeleine Elish, M. (2016). Moral Crumple Zones: Cautionary Tales in Human-RobotInteraction (WeRobot 2016). SSRN Electronic Journal. 10.2139/ssrn.2757236.

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022).Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal StudiesResearch Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011.LINK,

Bogdana Rakova, Jingying Yang, Henriette Cramer, & Rumman Chowdhury (2020).Where Responsible AI meets Reality. Proceedings of the ACM on Human-ComputerInteraction, 5, 1 - 23.

Map 4

サードパーティのソフトウェアやデータも含め、AI システムのすべての構成要素についてリスクと利益をマッピングする。

Map 4.1

AI 技術とその構成要素の法的リスクをマッピングするアプローチは、サードパーティのデータやソフトウェアの使用を含む、サードパーティの知的財産権やその他の権利の侵害のリスクと同様に、導入され、遵守され、文書化されている。

概要

サードパーティの技術や人員も、AI のリスクマネジメント活動において考慮すべき潜在的なリスク源である。リスクの優先順位や許容範囲がデプロイを行う組織と同じでない可能性があるため、このようなリスクをマッピングするには困難な場合がある。

例えば、事前にトレーニングされたモデルの使用は、大規模な未修正のデータセットに依存する傾向があり、また多くの場合、起源が非公開であるため、プライバシー、バイアス、予期せぬ影響に関する懸念が生じ、統計的不確実性のレベルの増加、再現性の困難さ、科学的妥当性の問題が生じる可能性がある。

推奨されるアクション

- 価値評価およびリスクマネジメント活動を支援するために、監査報告書、テスト結果、製品ロードマップ、保証書、サービス利用規約、エンドユーザライセンス契約、契約書およびその他のサードパーティ事業体に関連する文書をレビューする
- サードパーティソフトウェアのリリーススケジュールとソフトウェア変更管理計画（ホットフィックス、パッチ、アップデート、上位および下位互換性保証）をレビューし、AI システムのリスクにつながる可能性のある不規則な点がないかどうかを確認する
- システムの実装と保守に必要なサードパーティの素材（ハードウェア、オープンソース・ソフトウェア、基盤モデル、オープンソースデータ、独自ソフトウェア、独自データなど）をインベントリする。
- 適切なサポートの欠如による潜在的なリスクを評価するため、サードパーティの技術および人員に関連する冗長性を見直す。

透明性と文書化

組織は以下を文書化することができる。

- AI システムの潜在的な脆弱性、リスク、バイアスをサードパーティ（例えば、サプライヤ、エンドユーザ、被験者、ディストリビュータ／ベンダ、従業員）が報告するためのプロセスを確立したか？

- 組織がサードパーティからデータセットを入手した場合、そのようなデータセットを使用するリスクを評価し、管理したか？
- 結果はどのように独立して検証されるのか？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・WEF Model AI Governance Framework Assessment 2020.

参考文献

言語モデル

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell.2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? InProceedings of the 2021 ACM Conference on Fairness, Accountability, andTransparency (FAccT '21). Association for Computing Machinery, New York, NY,USA, 610–623.

Julia Kreutzer, Isaac Caswell, Lisa Wang, et al. 2022. Quality at a Glance: An Audit ofWeb-Crawled Multilingual Datasets. Transactions of the Association for ComputationalLinguistics 10 (2022), 50–72.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. 2022. Taxonomy of Risksposed by Language Models. In 2022 ACM Conference on Fairness, Accountability,and Transparency (FAccT '22). Association for Computing Machinery, New York, NY,USA, 214–229.

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model RiskManagement, Version 1.0, August 2021

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. 2021. On the Opportunitiesand Risks of Foundation Models. arXiv:2108.07258.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud,Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, TatsunoriHashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus. “Emergent Abilitiesof Large Language Models.” ArXiv abs/2206.07682 (2022).

Map 4.2

サードパーティの AI 技術を含む AI システムの構成要素の内部リスク管理が決定され、文書化される。

概要

AI アクターは、その作業の過程で、オープンソースや、その他自由に利用できるサードパーティの技術を利用することが多いが、その中にはプライバシー、バイアス、セキュリティリスクがある可能性がある。組織は、このようなテクノロジー・ソースの内部リスク管理を検討し、デプロイ前にサードパーティのマテリアルを評価するプラクティスを構築することができる。

推奨されるアクション

- サードパーティによるリスクマッピングの妨害や阻害を、リスク増大の兆候として追跡する
- サードパーティの技術インベントリ承認活動を支援するためにモデルに関する文書のテンプレートやソフトウェアの安全リストなどのリソースを提供する
- バイアス、データ・プライバシー、セキュリティの脆弱性に関連するリスクについて、サードパーティのマテリアル（データやモデルを含む）をレビューする
- 調達、セキュリティ、データ・プライバシー管理などの従来の技術リスク管理を取得したすべてのサードパーティの技術に適用する

透明性と文書化

組織は以下を文書化することができる。

- AI システムは独立したサードパーティによる監査が可能か？
- これらの政策は、AI システムの利用に対する国民の信頼と信用をどの程度醸成するか？
- AI システムの監査可能性を促進する仕組みが確立されているか（例えば、開発プロセスのトレーサビリティ、学習データの入手、AI システムのプロセス、結果、ポジティブおよびネガティブなインパクトの記録）？

AI の透明性に関するリソース

・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・EF Model AI Governance Framework Assessment 2020.
・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI - 2019. LINK

参考文献

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model RiskManagement, Version 1.0, August 2021. Retrieved on July 7, 2022. URL

Proposed Interagency Guidance on Third-Party Relationships: Risk Management,2021.

Kang, D., Raghavan, D., Bailis, P.D., & Zaharia, M.A. (2020). Model Assertions forMonitoring and Improving ML Models. ArXiv, abs/2003.01668.

Map 5

個人、グループ、コミュニティ、組織および社会へのインパクトを特徴づける。

Map 5.1

予想される使用法、同様のコンテクストにおける AI システムの過去の使用法、公開インシデント報告書、AI システムを開発またはデプロイしたチームの外部からのフィードバックまたはその他のデータに基づいて、それぞれの（潜在的に有益および有害な）インパクトの可能性と規模を特定し、文書化する。

概要

AI アクターは、Map 5.1 で決定された AI システムのインパクトの可能性を評価し、文書化し、トリアージすることができる。その後、可能性の推定値をアセスメントし、AI システムのデプロイの実施／非実施を判断することができる。組織がシステムのデプロイを進めることを決定した場合、可能性と規模の推定値を用いて、リスクレベルに適した TEVV リソースを割り当てることができる。

推奨されるアクション

- AI システムのインパクトを測定するためのアセスメントの尺度を確立する。アセスメントの尺度は、レッド・アンバー・グリーン（RAG）のような定性的なものである場合もあれば、シミュレーションや計量経済学的アプローチを伴う場合もある。評価尺度を文書化し、組織の AI ポートフォリオ全体に統一的に適用する
- AI のライフサイクルの重要な段階で、TEVV を定期的に適用し、システムへのインパクトやシステム更新の頻度を把握する
- トラストワージネスの特性に関連するシステムの利益とネガティブインパクトの可能性と規模を特定し、文書化する

透明性と文書化

組織は以下を文書化することができる。

- AI システムはどの集団に影響を与えるのか？
- エンティティは、AI システムに関連するデータセキュリティやプライバシーへのインパクトなど、トラストワージネスの特性についてどのような評価を実施したか？
- AI システムは独立したサードパーティによってテストできるのか？

AI の透明性に関するリソース

・Datasheets for Datasets.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD,2019. URL
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL

・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI - 2019. LINK

参考文献

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicineand healthcare. Joint Research Centre (European Commission).

Artificial Intelligence Incident Database. 2022.

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke.“Actionable Guidance for High-Consequence AI Risk Management: TowardsStandards Addressing AI Catastrophic Risks". ArXiv abs/2206.08966 (2022)

Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms:Methods, Scaling Behaviors, and Lessons Learned. arXiv.https://arxiv.org/abs/2209.07858

Map 5.2

関連する AI アクターとの定期的な関与を支援し、肯定的、否定的および予期せぬインパクトに関するフィードバックを統合するためのプラクティスと人員が整備され、文書化されている。

概要

AI システムは本質的に社会技術的なものであり、その目的を超えて、肯定的、中立的、あるいは否定的な影響を及ぼす可能性がある。ネガティブなインパクトは広範囲に及ぶ可能性があり、環境や国家安全保障だけでなく、個人、グループ、コミュニティ、組織および社会にも影響を与える。

組織は、システムのモニタリングのベースラインを作成することで、突発的なリスクを検出する機会を増やすことができる。AI システムのデプロイ後、（設計、開発、デプロイ活動に関与している AI アクターが知らないベネフィットやネガティブなインパクトを認識している、あるいは経験している可能性がある）さまざまなステークホルダのグループを関与させることで、組織はシステムのベネフィットおよび潜在的なネガティブなインパクトをより容易に理解し、モニターすることができる。

推奨されるアクション

- AI システムが個人、グループ、コミュニティ、組織および社会に与える潜在的なインパクトを特定するため、システム策定の初期段階でステークホルダの関与プロセスを確立し、文書化する
- バリュー・センシティブ・デザイン(VSD)などの手法を用いて、組織や社会の価値観と、システムの実装とインパクトとの間の整合性を特定する
- 潜在的なインパクトや顕在化したリスクに対する継続的な監視を支援するために、システムのエンドユーザやその他の主要なステークホルダを関与させ、その意見を取り入れ、収集し、組み込むためのアプローチを特定する
- 個人、グループ、コミュニティ、組織、そして社会への潜在的なインパクトの評価と文書化に、定量的、定性的、混合的手法を組み込む
- AI システムのメリット、プラスとマイナスのインパクト、その可能性と大きさを評価するために AI の設計および開発部門から独立したチーム（内部または外部）を特定する
- トラストワージネスの特性や設計アプローチおよび原則の変更に関する実用的な洞察につながる潜在的なインパクトを評価するためにステークホルダのフィードバックを評価し、文書化する
- 社会技術的な要素や手法を取り入れた TEVV 手順を開発し、組織文化全体の正常化を計画する。TEVV プロセスを定期的に見直し、改善する

透明性と文書化

組織は以下を文書化することができる。

- AI システムが人間に関係している場合、特定の社会集団を不当に有利にしたり、不利にしたりすることはないのか？どのような方法で？それはどのように管理されているのか？

- AIシステムが他の倫理的に保護された集団に関連する場合、適切な義務が果たされているか？(例えば、医療データには動物から収集された情報が含まれる場合がある)
- AI システムが人に関係している場合、このデータセットが人を危害や法的措置にさらす可能性はあるか？(例えば、金銭的、社会的、その他) 危害の可能性を緩和または低減するために何が行われたか？

AIの透明性に関するリソース

・Datasheets for Datasets.
・GAO-21-519SP: AI Accountability Framework for Federal Agencies &Other Entities.
・AI policies and initiatives, in Artificial Intelligence in Society, OECD,2019.
・Intel.gov: AI Ethics Framework for Intelligence Community - 2020.
・Assessment List for Trustworthy AI (ALTAI) - The High-Level ExpertGroup on AI - 2019. LINK

参考文献

Susanne Vernim, Harald Bauer, Erwin Rauch, et al. 2022. A value sensitive designapproach for designing AI-based worker assistance systems in manufacturing.Procedia Comput. Sci. 200, C (2022), 505–516.

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources ofHarm throughout the Machine Learning Life Cycle. arXiv:1901.10002. Retrieved from

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. OvercomingFailures of Imagination in AI Infused System Development and Deployment.arXiv:2011.13416.

Konstantinia Charitoudi and Andrew Blyth. A Socio-Technical Approach to Cyber RiskManagement and Impact Assessment. Journal of Information Security 4, 1 (2013), 33-41.

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud,J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020Conference on Fairness, Accountability, and Transparency.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, & JacobMetcalf. 2021. Assemlbing Accountability: Algorithmic Impact Assessment for thePublic Interest. Data & Society. Accessed 7/14/2022

Shari Trewin (2018). AI Fairness for People with Disabilities: Point of View. ArXiv,abs/1811.10670.

Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study inHealthcare. Accessed July 14, 2022.

Microsoft Responsible AI Impact Assessment Template. 2022. Accessed July 14,2022.

Microsoft Responsible AI Impact Assessment Guide. 2022. Accessed July 14, 2022.

Microsoft Responsible AI Standard, v2.

Microsoft Research AI Fairness Checklist.

PEAT AI & Disability Inclusion Toolkit – Risks of Bias and Discrimination in AI HiringTools.

Measure

Measure 1

適切な方法とメトリクスを決め、適用する。

Measure 1.1

Map 機能で列挙された AI リスクを測定するためのアプローチとメトリクスは、最も重要な AI リスクから実施するために選択される。測定されない、あるいは測定できないリスクやトラストワージネスの特性は、適切に文書化される。

概要

トラストワージな AI システムの開発と実用化は、基礎となる技術とその活用に関する信頼できる測定と評価にかかっている。従来のソフトウェアシステムと比較すると、AI 技術は新たな失敗モードと学習データと手法への生来的な依存性をもたらしており、データの品質と代表性に直接結びつく。さらに、AI システムは生来的に社会技術的なものであり、社会の力学や人間の行動に影響される。AI のリスク（および利点）は、システムがどのように使用されるか、他の AI システムとの相互作用、誰が操作するか、そしてそれがデプロイされる社会的コンテクストに関連する社会的要因と技術的側面との相互作用から生まれる。言い換えれば、何を測定すべきかは評価の目的、オーディエンスおよびニーズに依存する。

これら 2 つの要因は、Map 機能で列挙された AI リスクに関する測定のアプローチやメトリクスの選択に影響を与える。AI ランドスケープは進化しており、AI 測定の手法やメトリクスも進化している。メトリクスの進化は、測定の有効性を維持するための鍵である。

推奨されるアクション

- 既知のリスク、エラー、インシデントまたはネガティブなインパクトを検出、追跡および測定するためのアプローチを確立する
- システムが目的に適合し、要求されたとおりに動いているかどうかを実際に確かめるためのテスト手順とメトリクスを定める
- AI システムのトラストワージネスを実際に確かめるためのテスト手順とメトリクスを定める
- システムパフォーマンスの許容限界（例えば誤差の分布）を定義し、パフォーマンスが許容限界を超えたとき／場合の軌道修正の提案を含めること
- 効果的なシステム運用のために、AI アクターの能力を評価するメトリクスを定め、定期的にアセスメントする。
- システム設計、開発、デプロイ、活用および評価に関する必要情報にステークホルダがアクセスできるかどうかを評価するための透明性のメトリクスを定める

- アカウンタビリティのメトリクスを活用して、AI 設計者、開発者およびデプロイヤが明確で透明性のある責任範囲を維持し、問い合わせに対応できるかどうかを判定できるようにする
- 検討されたが未活用のものも含め、メトリクスの選択基準を文書化する
- 学習データ、他のコンテクスト用に開発されたモデル、他のコンテクストから再利用されたシステムコンポーネント、サードパーティのツールやリソースなど、AI システムの外部入力をモニターする
- システムの汎用性と信頼性のアセスメントを周知するためのメトリクスを報告する
- デプロイ前とデプロイ後のシステムパフォーマンスについて、既存のリスクと緊急のリスクを含めてアセスメントし、文書化する
- Map 機能で定めたリスクやトラストワージネスの特徴のうち測定されないもの、測定されない理由の正当性を含めて文書化する

透明性と文書化

組織は以下の項目を文書化することができる

- AI のデプロイ後、AI の精度といった適切なパフォーマンスのメトリクスはどのようにモニターされるのか？
- データの品質、正確性、信頼性および代表性を高めるために、エンティティがどのような是正処置を講じたか？
- 推奨されるデータ分割や評価指標があるか？(例えば学習、開発、テストや、正解率／AUC)
- 組織では、ユーザビリティの問題に取り組み、ユーザーインタフェースが意図した目的を果たしているかどうかをテストしたか？
- エンティティは、エラーや限界を特定するために、AI システムに対してどのようなテスト（すなわち、手動か自動か、敵対的テストやストレステスト）を実施したか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・Datasheets for Datasets.

参考文献

Sara R. Jordan. “Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI.” 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019.

IEEE. “IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation.” IEEE Standards Association. URL

ACM Technology Policy Council. “Statement on Principles for Responsible Algorithmic Systems.” Association for Computing Machinery (ACM), October 26, 2022.

Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv. https://arxiv.org/abs/2212.09251

Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv. https://arxiv.org/abs/2209.07858

David Piorkowski, Michael Hind, and John Richards. "Quantitative AI Risk Assessments: Opportunities and Challenges." arXiv preprint, submitted January 11, 2023.

Daniel Schiff, Aladdin Ayesh, Laura Musikanski, and John C. Havens. “IEEE 7010: A New Standard for Assessing the Well-Being Implications of Artificial Intelligence.” 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.

Measure 1.2

AI メトリクスの適切性と既存のコントロールの有効性は定期的に評価され、エラーや影響を受けたコミュニティへのインパクトの報告も含めて更新される。

概要

ニューラルネットワークや自然言語処理など、異なる AI タスクは異なる評価技法から恩恵を受ける。また、AI システムが使用されるユースケースや特定の条件も、評価手法の適切性に影響を与える。運用条件の変化、データドリフト、モデル・ドリフトなどは、定期的に評価し AI メトリクスの適切性および有効性を更新することが推奨される要因の一つであり、AI システム測定の信頼性を高めることができる。

推奨されるアクション

- すべての測定の外部妥当性を評価する（例えば、あるコンテクストで行われた測定が他のコンテクストに一般化できる度合い）
- AI システムのライフサイクルを通じて、既存のメトリクスと管理の有効性を定期的にアセスメントする
- エラー、インシデントおよびネガティブなインパクトの報告を文書化し、修理やアップグレードのための既存のメトリクスの十分性と有効性を評価する
- 修理やアップグレードを実施する際に、既存のメトリクスで不十分あるいは効果的でない場合には、新しいメトリクスを開発する
- サードパーティツールを含む外部入力をモニターし、特徴付け、追跡するためのメトリクスを開発し、活用する
- ステークホルダや影響を受けるコミュニティとメトリクスや関連情報を共有する頻度や範囲を決定する
- Map 機能で確立されたステークホルダのフィードバックプロセスを活用して、エンドユーザや影響を受ける可能性のあるコミュニティからのフィードバックを収集し、それに基づいて行動するとともに共有する
- バグの発生率や深刻度、対応に要する時間、修復に要する時間などのソフトウェア品質メトリクスを収集し、報告する（Manage 4.3 参照）

透明性と文書化

組織は以下の項目を文書化することができる

- エンティティは、AI システムのパフォーマンスを測定するために、どのようなメトリクスを開発したか？
- そのメトリクスは、どの程度まで正確で有用なパフォーマンスの指標となっているか？
- エンティティは、データの品質、正確性、信頼性および代表性を向上させるためにどのような是正処置を講じたか？
- 正確さあるいは適切なパフォーマンスメトリクスはどのようにアセスメントされるか？

● 選択したメトリクスの正当性はなにか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

ACM Technology Policy Council. “Statement on Principles for Responsible Algorithmic Systems.” Association for Computing Machinery (ACM), October 26, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009.

Harini Suresh and John Guttag. “A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2021.

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006.

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. “Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs.” Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, July 2021, 368–78.

Measure 1.3

システムの最前線の開発者でない内部専門家および／または独立した審査員が、定期的なアセスメントおよび更新に関与する。ドメインの専門家、ユーザ、AI システムを開発またはデプロイしたチーム外部の AI アクターおよび影響を受けるコミュニティは、組織のリスク許容度に従い、必要に応じて、アセスメントをサポートする。

概要

現在の AI システムはもろく、故障モードは十分に説明されておらず、システムは開発されたコンテクストに依存しており、学習時の環境の外へはうまく移行できない。これらのシステムの継続的なモニタリングとともに、局所的な評価に頼ることが必要となる。古典的な（テストケース全体の平均をとる）測定方法の枠を超える方法や、潜在的に大きなコストが発生する可能性のある不具合の部分に焦点を絞った測定方法は、リスクマネジメント活動の信頼性を向上させることができる。AI システムがどのように利用されているかについて影響を受けるコミュニティからのフィードバックを受けることで、AI 評価を目的あるものにすることができる。AI システムの最前線の開発者ではない内部の専門家や、独立した評価者が AI システムを定期的にアセスメントすることは、AI システムの性能とトラストワージネスを十分に評価するための大きな助けになる。

推奨されるアクション

- リスクとインパクトを特定するためのインセンティブに関する TEVV プロセスを評価する
- Govern 機能（2.1 および 4.1）において設置された個別のテストチームを活用し、AI システムの独立した意思決定と軌道修正を可能にする。プロセスと測定を追跡し、パフォーマンスの変化を文書化する
- AI システムのプロトタイプを、AI ライフサイクルの初期段階から継続的にエンドユーザと共に評価するよう計画する。テスト結果を文書化し、軌道修正する
- TEVV とオーバーサイト AI アクターの独立性と地位を評価し、保証、コンプライアンスおよびフィードバックのタスクを効果的に実行するために必要なレベルの独立性とリソースを有していることを確認する
- Map 1.2 で確立された、学際的かつ人口統計学的に多様な社内チームを評価する
- 特に、多様なグループからのインプットを引き出し、評価し、統合するプロセスに関する外部ステークホルダからのフィードバック・メカニズムの有効性を評価する
- AI システムのリスクとトラストワージネスに関する AI アクターの可視性と意思決定を強化するための、外部ステークホルダのフィードバック・メカニズムの有効性を評価する

透明性と文書化

組織は以下を文書化することができる

- AI システムの設計、開発、デプロイ、評価およびモニターに関わる人員の役割、責任およびデリゲーションは何か？
- 外部ステークホルダが情報にアクセスするとき、どれだけ容易であるか、またどれだけ最新であるか？
- エンティティは、ステークホルダのコミュニティに AI の戦略的ゴールと目標をどの程度伝えているか？
- AI システムのアウトプットに影響を受けるユーザや関係者が、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- 透明性を促進するための情報がどの程度十分で適切か。外部ステークホルダは、AI システムの設計、運用および制限に関する情報にアクセスできるか？
- エンドユーザ、消費者、規制当局および AI システムの使用によって影響を受ける個人を含む外部ステークホルダが、AI システムの設計、運用および制限についてどのような情報にアクセスできるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011.

“Definition of independent verification and validation (IV&V)”, in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. Annex C,

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. “Participation Is Not a Design Fix for Machine Learning.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2022.

Rediet Abebe and Kira Goldner. “Mechanism Design for Social Good.” AI Matters 4, no. 3 (October 2018): 27–34.

Measure 2

AI システムは、トラストワージ特性について評価される。

Measure 2.1

テスト、評価、妥当性確認および検証（Test, evaluation, validation, and verification: TEVV）中に使用したツールのテストセット、メトリクスおよび詳細が文書化される。

概要

測定アプローチ、テストセット、メトリクス、使用したプロセスや材料、関連する詳細を文書化することは、有効で信頼性の高い測定プロセスを構築する基盤となる。文書化することで再現性と一貫性が備わり、AI リスクマネジメントの意思決定を強化することができる。

推奨されるアクション

- 測定のあらゆる可能な側面の透明性と文書化のために、既存の業界のベストプラクティスを活用する。例えば、データセットのデータシート、モデルカード、[コメント提供者の例] などがある
- 測定アプローチ、テストセット、メトリクス、使用したプロセスや材料を文書化するために使用するツールの有効性を定期的に評価する
- 必要に応じてツールを更新する

透明性と文書化

組織は以下を文書化することができる

- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？
- エンティティは、AI システムの開発、テスト方法、メトリクス、パフォーマンスのアウトカムをどの程度文書化しているか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework WEF Companion to the Model AI Governance Framework, 2020.

参考文献

Emily M. Bender and Batya Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” Transactions of the Association for Computational Linguistics 6 (2018): 587–604.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” FAT *19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 220–29.

IEEE Computer Society. “Software Engineering Body of Knowledge Version 3: IEEE Computer Society.” IEEE Computer Society.

IEEE. “IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation.” IEEE Standards Association.

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011.

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85.

Jeanna Matthews, Bruce Hedin, Marc Canellas. Trustworthy Evidence for Trustworthy Technology: An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems. IEEE-USA, September 29 2022.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. “Hard Choices in Artificial Intelligence.” Artificial Intelligence 300 (November 2021).

Measure 2.2

人間の被験者を対象とする評価は、適用要件（被験者保護を含む）を満たし、関連する集団を代表するものである。

概要

AI システムの測定や評価には、人間を対象としたテストや、人間から取得したデータを使用することが多い。被験者の保護は、連邦政府が資金を提供する研究を実施する際に法律で義務付けられており、分野によっては固有の要件となっている。標準的な被験者保護手順には、被験者の福祉と利益の保護、被験者へのリスクを最小化するための評価の設計および法的要件と期待に関し義務となる研修を修了することが含まれる。

被験者や被験者データを利用した AI システムのパフォーマンス評価は、その利用状況における母集団を反映したものでなければならない。非代表的なデータを利用した AI システム活動は、不正確な評価や否定的で有害な結果をもたらす可能性がある。AI システムの全運用範囲を反映したデータを収集したり、評価タスクを実行したりすることは、困難なことが多く、時には不可能なこともある。これらのデータを収集、注釈付け、使用するための方法も、課題の一因となりうる。このような課題に対処するため、組織は、被験者のデータ収集やデータセットの実践を AI システムのコンテクストや目的に結び付け、関連領域の AI アクターと緊密に連携してそれらを行うことができる。

推奨されるアクション

- データセット収集活動中は、インフォームド・コンセントや補償を含む、組織および規律に定められた被験者研究の要件に従う
- エラー、インシデントまたはネガティブなインパクトの類似性を含め、利用者またはデータ対象者について実際の集団が意図したものとどう違うかを分析する
- 細分化された評価方法を活用する（例えば人種、年齢、性別など民族性、条件、地域ごと）ことにより、実世界にデプロイされる AI システムのパフォーマンスを向上させることができる
- 利用コンテクストにおけるデータセットの代表性の閾値と警告手順を確立する
- 使用背景を熟知した専門家と緊密に協力してデータセットを構築する
- データに含まれる主体を含みデータセットおよびその使用に関する知的財産権およびプライバシー権に従う
- 以下の手順でデータの代表性を評価する
    - 既知の故障モードの調査
    - データ品質と多様な調達の評価
    - 公的ベンチマークの適用
    - 従来のバイアステスト
    - カオスエンジニアリング

- ステークホルダのフィードバック
- システムのテストと評価に使用するデータを提供する個人には、インフォームド・コンセントを使用する

透明性と文書化

組織は以下を文書化することができる

- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？
- エンティティは、不公平な結果や差別的な結果など、データのバイアスによる潜在的なインパクトをどのように特定し、軽減したか？
- 確立された手続きは、バイアス、不公平感およびシステムに起因する他の懸念を軽減する上で、どの程度有効か？
- エンティティは、潜在的なバイアス（データに含まれる統計的、コンテクスト的および歴史的バイアス）をどの程度特定し、軽減しているか？
- 人に関するものであれば、そのデータセットがどのような用途に使われるかを説明し、同意を得たか。人々のコミュニケーションから収集されたデータについて、どのようなコミュニティ規範が存在するか。同意が得られた場合、その方法は何か？
- 将来的に、あるいは特定の用途について、同意を撤回する仕組みを提供したか？
- AI システムの開発またはテストに人間の被験者が関わった場合、その安全と福利を促進するためにどのような保護措置が講じられたか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework WEF Companion to the Model AI Governance Framework, 2020.

・Datasheets for Datasets.

参考文献

United States Department of Health, Education, and Welfare's National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Volume II. United States Department of Health and Human Services Office for Human Research Protections. April 18, 1979.

Office for Human Research Protections (OHRP). “45 CFR 46.” United States Department of Health and Human Services Office for Human Research Protections, March 10, 2021. URL Note: Federal Policy for Protection of Human Subjects (Common Rule). 45 CFR 46 (2018)

Office for Human Research Protections (OHRP). “Human Subject Regulations Decision Chart.” United States Department of Health and Human Services Office for Human Research Protections, June 30, 2020.

Jacob Metcalf and Kate Crawford. “Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide.” Big Data and Society 3, no. 1 (2016).

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. "Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing." arXiv preprint, submitted April 20, 2021.

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. "Resolving the Human Subjects Status of Machine Learning's Crowdworkers." arXiv preprint, submitted June 8, 2022.

Office for Human Research Protections (OHRP). “International Compilation of Human Research Standards.” United States Department of Health and Human Services Office for Human Research Protections, February 7, 2022.

National Institutes of Health. “Definition of Human Subjects Research.” NIH Central Resource for Grants and Funding Information, January 13, 2020. URL

Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of the 1st Conference on Fairness, Accountability and Transparency in PMLR 81 (2018): 77–91. URL

Eun Seo Jo and Timnit Gebru. “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, 306–16.

Marco Gerardi, Katarzyna Barud, Marie-Catherine Wagner, Nikolaus Forgo, Francesca Fallucchi, Noemi Scarpato, Fiorella Guadagni, and Fabio Massimo Zanzotto. "Active Informed Consent to Boost the Application of Machine Learning in Medicine." arXiv preprint, submitted September 27, 2022.

Shari Trewin. "AI Fairness for People with Disabilities: Point of View." arXiv preprint, submitted November 26, 2018.

Andrea Brennen, Ryan Ashley, Ricardo Calix, JJ Ben-Joseph, George Sieniawski, Mona Gogia, and BNH.AI. AI Assurance Audit of RoBERTa, an Open source, Pretrained Large Language Model. IQT Labs, December 2022.

Measure 2.3

AI システムの性能または保証基準が定性的または定量的に測定され、デプロイ設定に類似した条件下で実証されている。測定が文書化されている。

概要

現在のリスクとインパクトの環境は、AI システムのパフォーマンス推定が不十分であり、デプロイのコンテクストをより深く理解する必要があることを示唆している。計算に焦点を当てたパフォーマンステストと評価スキームが使えるのは、テストデータセットとシリコンの技術に限定される。これらのアプローチでは、実世界環境におけるリスクやインパクトを直接評価することはできず、AI 利用がどうなるかの予測に基づき、何がインパクトを与えるかを推定することしかできない。リスクを適切にマネージするためには、より直接的な情報が必要であり、デプロイされた AI がどのような状況でどれくらいのインパクトをもたらすのか、誰が最もインパクトを受けやすいのか、そしてその経験がどのようなものなのかを理解する必要がある。

推奨されるアクション

- インパクトを受ける可能性のあるコミュニティとの定期的かつ継続的な関与を行う
- 人口統計学的に多様で、学際的かつ協力的な内部チームを維持する
- 最適化されていない条件でのシステムのテストと評価を定期的に行い、ユーザーインタラクションやユーザーエクスペリエンス（UI／UX）の役割を担う AI アクターと協力する
- ヒューマンファクターや社会技術の専門家と協力し、ステークホルダの参加活動からのフィードバックを評価する
- 社会技術、ヒューマンファクターおよび UI／UX の専門家と協力し、システムテストのシナリオに変換可能な、使用コンテクストにおける注目すべき特性を特定する
- デプロイ前の AI システムを、予想されるシナリオと同様の条件で測定する
- 妥当性（偽陽性率、偽陰性率等）や効率（学習時間、予測遅延時間等）といったデプロイされる条件での真の値に関係する性能指標を測定し、文書化する
- AI アクターのコンピテンシーや経験といった保証基準を測定する
- 測定条件とデプロイ環境の違いを文書化する

透明性と文書化

組織は以下を文書化することができる

- このデータセットで最初にどのような実験が行われたのか？AI システムの実験はどこまで文書化されているのか？
- システム／エンティティは、定められたゴールや目標に対する進捗をどの程度一貫して測定しているか？
- AI のデプロイ後、精度といった適切な性能指標をどのようにモニターするのか。ベースライン・パフォーマンスからの分布シフトやモデル・ドリフトはどの程度まで許容可能か？

- 時間が経過し、状況が変化しても、学習データは依然として運用環境を代表しているか？
- エンティティは、エラーや制限を特定するために、AI システムに対してどのようなテスト（すなわち敵対的テストやストレステスト）を実施したか？

AI の透明性に関するリソース

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework WEF Companion to the Model AI Governance Framework, 2020.

・Datasheets for Datasets.

参考文献

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009.

Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman. "Model Selection's Disparate Impact in Real-World Deep Learning Applications." arXiv preprint, submitted September 7, 2021

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. “The Fallacy of AI Functionality.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 959–72.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. “Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research.” Patterns 2, no. 11 (2021): 100336.

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006.

Md Johirul Islam, Giang Nguyen, Rangeet Pan, and Hridesh Rajan. "A Comprehensive Study on Deep Learning Bug Characteristics." arXiv preprint, submitted June 3, 2019.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. "DQI: Measuring Data Quality in NLP." arXiv preprint, submitted May 2, 2020.

Doug Wielenga. "Paper 073-2007: Identifying and Overcoming Common Data Mining Mistakes." SAS Global Forum 2007: Data Mining and Predictive Modeling, SAS Institute, 2007.

ソフトウェア・リソース

・Drifter library (performance assessment)
・Manifold library (performance assessment)
・MLextend library (performance assessment)
・PiML library (explainable models, performance assessment)
・SALib library (performance assessment)
・What-If Tool (performance assessment)

Measure 2.4

Map 機能で特定された AI システムとそのコンポーネントの機能と動作は、実運用時にモニターされる。

概要

AI システムは、実運用中に時間経過とともに環境が変化するにつれて、新たな問題やリスクに遭遇する可能性がある。この影響は、しばしば「ドリフト」と呼ばれ、AI システムが当初の設計の前提や制限を満たさなくなることを意味する。定期的なモニタリングにより AI アクターは、Map 機能で定められた AI システムとそのコンポーネントの機能と動作をモニターし、必要なシステム介入のスピードと有効性を高めることができる。

推奨されるアクション

- 実運用環境で観察されるメトリクスとパフォーマンス指標が、デプロイ前のテストで収集された同じメトリクスとどのように異なるかをモニターし、文書化する。差異が観察された場合は、エラーの伝播とフィードバック・ループのリスクを考慮する
- 仮説検定または専門家の知識を活用して、新しい入力または出力データのテスト環境に対するモニターされた分布の違いを測定する
- 管理限界値、信頼区間、完全性制約、ML アルゴリズムといったアプローチを用いて異常をモニターする。異常が観測された場合、エラー伝播とフィードバックのループリスクを考慮する
- 新しい入力データの分布や、実運用環境で観測された推定値がデプロイ前のテスト結果と異なる場合または異常が検出された場合にアラートが出ることを検証する
- 生成された出力の精度と品質を、新たに収集された真の値の情報と照らし合わせて評価する
- 予期せぬデータの処理と生成された出力の信頼性を追跡するために、人間によるレビューを活用し、出力が信頼できないときはシステムの利用者に警告する。これらのプロセスに責任を持つヒューマンオーバーサイトの責任者が、当該タスクについて明確に定義された責任を持ち、訓練を受けていることを確認する
- システムテストおよび活動のモニターのために、組織の方針および規制あるいは規律上の要件（インフォームド・コンセント、組織審査委員会の承認、被験者研究の保護等）に従い、運用環境からユースケースを収集する

透明性と文書化

組織は以下を文書化することができる

- 各コンポーネントのアウトプットは、どの程度まで運用コンテクストに適しているか？
- AI システムの仮定、境界および制限に対して根拠がある場合、エンティティはどのような正当性を示したか？
- AI のデプロイ後、精度などの適切なパフォーマンス指標はどのようにモニターされるのか？
- 時間が経過し、状況が変化しても、学習データは依然として運用環境を代表しているか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Luca Piano, Fabio Garcea, Valentina Gatteschi, Fabrizio Lamberti, and Lia Morra. “Detecting Drift in Deep Learning: A Methodology Primer.” IT Professional 24, no. 5 (2022): 53–60.

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub.

Measure 2.5

デプロイされる AI システムは、妥当性と信頼性が実証されている。その技術が開発された条件における汎用性の限界を文書化している。

概要

妥当性確認されていないまたは妥当性確認に失敗した AI システムは、不正確または信頼性に欠ける可能性があり、その制約を超えたデータや設定に対してうまく汎化できない可能性があるため、AI リスクを生み出し増大させ、トラストワージネスを低下させる。AI アクターは、システムの限界を探り、文書化するプロセスを構築することで、システムの妥当性を向上させることができる。これには、システム設計時になかった目的や用途を幅広く考慮することが含まれる。

妥当性確認リスクには、直接的な観測や測定が不可能な現象（公平性、雇用可能性、誠実さ、犯罪傾向等）を運用するために AI 開発チームが構築するプロキシタスクやその他の指標の使用が含まれる。チームは、メトリクスが測定すべきものを測定しているかどうかを確かめるための指標を導入することで、こうしたリスクを軽減することができる（構成概念妥当性とも呼ばれる）。この妥当性確認やその他の妥当性確認を行わないと、交絡変数の存在、潜在的な疑似相関、エラーの伝播、他の相互接続されたシステムへの潜在的なインパクトなどを含む様々な負の特性やインパクトが検出されない可能性がある。

推奨されるアクション

- AI システムが妥当性確認される動作条件と社会技術的コンテクストを定義する
- システムの運用条件と制約を確立するためのプロセスを定義し、文書化する
- 妥当性を測定するための以下のようなアプローチを確立または特定し、文書化する
  - 構成概念妥当性（テストが測定しようとする概念を測定していること）
  - 内部妥当性（テストされる関係が他の要因や変数に影響されないこと）
  - 外部妥当性（結果が訓練条件を超えて一般化可能であること）
  - 実験計画原則と統計分析およびモデリングの使用
- システムの分散を評価し、文書化する。標準的なアプローチには、信頼区間、標準偏差、標準誤差、ブートストラップあるいはクロスバリデーションが含まれる
- 堅牢性の尺度を確立または特定し、文書化する
- 信頼性の尺度を確立または特定し、文書化する
- 測定モデルの基礎となる仮定を特定し、文書化することで、プロキシタスクが測定される概念を正確に反映していることを確認するためのプラクティスを確立する
- 標準的なソフトウェアテストアプローチを活用する（例えばユニットテスト、統合テスト、機能テスト、カオステスト、自動生成テストケースなど)
- 標準的な統計手法を用いて、バイアス、推測的な関連性、採用された測定モデルにおける相関および共分散を検証する

- 標準的な統計手法を用いて、システムのアウトカムの分散と信頼性を検証する
- システムパフォーマンスが定義された条件を逸脱していないか、動作条件をモニターする
- 運用環境とは異なるテストシナリオを含め、AI システムの限界を探るための TEVV アプローチを定める。特定の使用状況についてその分野の専門家に相談する
- アラート後の行動を定義する。考えられる行動には以下が含まれる
  - 行動を起こす前に、関連する他の AI アクターに警告を発する
  - 続いて人間によるレビューを要請する
  - システムが定義された有効範囲外で動作していることを、下流のユーザとステークホルダに警告する。
  - 起こりうるエラー伝播を追跡し、軽減する
  - アクションを記録する
- 明確に定義されたシステムの有効範囲を超えてシステムを使用しようとする場合は、必ず入力データと関連するシステム構成情報をログに記録する
- 時間をかけてシステムを修正し、システムの有効範囲を新しい運転条件まで拡大する

透明性と文書化

組織は以下を文書化することができる

- エンティティは、エラーや制限を特定するために、AI システムに対してどのようなテスト（敵対的テストやストレステストなど）を実施したか？
- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？
- エンティティは、このモデルのチェック項目は何か？不公平な結果や差別的な結果など、データのバイアスによる潜在的なインパクトをどのように特定し、軽減したか？
- 確立された手続きは、偏見や不公平感、その他制度に起因する懸念を軽減する上で、どの程度有効か？
- エンティティは、AI システムの設計、開発および／またはデプロイによって、どのようなゴールと目標を達成しようとしているのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

参考文献

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85.

Debugging Machine Learning Models. Proceedings of ICLR 2019 Workshop, May 6, 2019, New Orleans, Louisiana.

Patrick Hall. “Strategies for Model Debugging.” Towards Data Science, November 8, 2019.

Suchi Saria and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint, submitted April 15, 2019.

Google Developers. “Overview of Debugging ML Models.” Google Developers Machine Learning Foundational Courses, n.d.

R. Mohanani, I. Salman, B. Turhan, P. Rodríguez and P. Ralph, "Cognitive Biases in Software Engineering: A Systematic Mapping Study," in IEEE Transactions on Software Engineering, vol. 46, no. 12, pp. 1318-1339, Dec. 2020

ソフトウェアのリソース

・Drifter library (performance assessment)
・Manifold library (performance assessment)
・MLextend library (performance assessment)
・PiML library (explainable models, performance assessment)
・SALib library (performance assessment)
・What-If Tool (performance assessment)

Measure 2.6

AI システムは、Map 機能で特定された安全性リスクについて定期的に評価される。デプロイされる AI システムは安全であることが実証され、残ったネガティブなリスクはリスク許容度を逸脱せず、特に、システムの知識の限界を超えて動作させた場合、安全に故障することができる。安全性の指標は、システムの信頼性と堅牢性、リアルタイムのモニターおよび AI システムの故障に対する応答時間に関するものである。

概要

製造業や安全保障の分野で、故障がさまざまな物理的・環境的被害をもたらす可能性がある分野において多くの AI システムが交通機関などの現場に導入されている。人命、健康および資産を危険にさらす可能性のある AI システムをデプロイ前に全体的に評価し、システムが通常運転中およびシステムの知識の限界を逸脱した条件においても安全であることを定期的に評価して確認する。

安全のための測定活動はしばしば、開発およびデプロイ環境における徹底的なテストを行い、それはシステムの信頼性、堅牢性および安全動作の限界、システム性能の様々な側面のリアルタイムモニタリングを行う。これらの活動は通常、過去の失敗設計の回避、システム問題への迅速な対応を可能にする事故対応計画の策定とリハーサル、失敗をカバーするための冗長機能の導入、透明で説明責任のあるガバナンスなどが挙げられ、他のリスクマッピング、管理およびガバナンスのタスクとともに実施される。システム安全に関するインシデントや失敗は、組織の力学や文化に関連していることがしばしば報告されている。独立監査人は、AI システムの安全性の証拠を検証するために、重要な独立した視点をもたらす可能性がある。

推奨されるアクション

- 開発およびデプロイ環境やストレス条件下で、システムの性能を徹底的に測定する
  - 本番テストに進む前に、テストデータの評価とシミュレーションを行う。複数のパフォーマンス品質とエラーメトリクスを追跡する
  - ドメインの専門家との相談のもと、想定されるシナリオ（例えばコンセプト・ドリフトおよび高負荷）および既知の限界を超えた条件の下で、システムの性能をストレステストする
  - 過去に発生した事故やヒヤリハットに関連した条件下でシステムをテストし、システムの性能と安全特性を測定する
  - カオスエンジニアリングアプローチを応用し、極限状態でのシステムテストを行い、予期せぬ反応を測定する
  - システムがテストされ、安全に故障することが実証された条件の範囲を文書化すること
- システムのパフォーマンスをリアルタイムで測定・モニターし、AI システムのインシデントが検出された場合の迅速な対応を可能にする。

- 影響を受ける地域社会との情報共有の可能性を想定してまたは AI システムのオーバーサイトの人員の要求に応じて共有できるよう、適切な安全統計（例えばアウトオブレンジ・パフォーマンス、事故対応時間、システム停止時間、負傷者数など）を収集する。
- 測定を継続的改善のゴールに合わせる。生産中のテストやイベントに応じてシステムを修正することにより、システムがフェイルセーフに機能する条件の範囲を拡大する
- AI システムインシデントに対し対応やダウンタイムの記録を含むインシデント対応計画を文書化し、実践し、測定する
- 文書化された安全性試験とモニタリング情報を、現在決められたリスク許容範囲と継続的に比較する
- 安全リスクのマネジメントに関する詳細は、Manage を参照すること。

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、エラーや制限を特定するために、AI システムに対してどのようなテスト（敵対的テストやストレステストなど）を実施したか？
- エンティティは、AI システムの開発、テスト方法、メトリクス、パフォーマンスのアウトカムをどの程度文書化しているか？
- AI システムの監査可能性（例えば、開発プロセスのトレーサビリティ、学習データの調達、AI システムのプロセス、結果およびポジティブ・ネガティブなインパクトの記録）を促進するメカニズムを確立したか？
- AI システムが独立したサードパーティによって監査されることを保証したか？
- サードパーティ（例えば、サプライヤ、エンドユーザ、被験者、ディストリビュータ、ベンダおよび従業員）が AI システムの潜在的な脆弱性、リスク、バイアスを報告するための手続きを確立したか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

AI Incident Database. 2022.

AIAAIC Repository. 2022.

Netflix. Chaos Monkey.

IBM. “IBM's Principles of Chaos Engineering.” IBM, n.d.

Suchi Saria and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint, submitted April 15, 2019.

Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model assertions for monitoring and improving ML models." Proceedings of Machine Learning and Systems 2 (2020): 481-496.

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub.

McGregor, S., Paeth, K., & Lam, K.T. (2022). Indexing AI Risks with Incidents, Issues, and Variants. ArXiv, abs/2211.10384

Measure 2.7

Map 機能で定められた AI システムのセキュリティとレジリエンスが評価され、文書化される。

概要

AI システムだけでなく、AI システムがデプロイされるエコシステムも、予期せぬ不都合な出来事や、環境や用途の予期せぬ変化に耐えることができれば、あるいは、内部や外部の変化に直面しても機能や構造を維持し、必要な場合には安全かつ優雅に劣化させることができれば、レジリエントであると言えるかもしれない。一般的なセキュリティ上の懸念は、敵対的サンプル、データポイズニング、AI システムのエンドポイントを介したモデル、学習データおよびその他の知的財産の抽出（Exfiltration）に関するものである。不正なアクセスや利用を防ぐ保護メカニズムによって機密性、完全性および可用性を維持できる AI システムはセキュアであると言えるかもしれない。

セキュリティとレジリエンスは関連しているが、異なる特性である。レジリエンスとは、予期せぬ不都合な出来事の後に正常な機能に戻る能力のことである。セキュリティにはレジリエンスも含まれるが、攻撃を回避、防御、対応、回復するためのプロトコルも含まれる。レジリエンスは堅牢性に関連し、モデルやデータの予期せぬまたは敵対的な使用（あるいは濫用や誤用）を含む。

推奨されるアクション

- AI システムのセキュリティテストとメトリクス（例えばレッドチームの活動、異常イベントの頻度と発生率、システム停止時間、インシデント対応時間、バイパスまでの時間など）を確立し、追跡する。
- レッドチーム演習を活用して、敵対的またはストレス条件下でシステムを積極的にテストし、システムの応答を測定し、故障モードを評価し、あるいは予期せぬ有害事象が発生した後にシステムが正常な機能に復帰できるかどうかを判断する
- 継続的な改善活動の一環として、セキュリティテストの範囲と結果を含むレッドチーム演習の結果を文書化する
- 対策指標（例えば認証、スロットリング、ディファレンシャル・プライバシー、ロバストな ML アプローチなど）を使って、システムが正常状態に復帰できるセキュリティ条件の範囲を広げる
- テストや実運用で経験した事象に応じてシステムのセキュリティ手順と対策を変更し、攻撃に対する堅牢性とレジリエンスを高める
- エラーや攻撃パターンに関する情報が、インシデントデータベース、同様のシステムを持つ他の組織およびシステムユーザやステークホルダ内で共有されていることを検証する（Manage 4.1 参照）
- 他組織の AI アクターと情報共有の方法を開発・維持し、共通の攻撃から学ぶ
- サードパーティの AI リソースと人員がセキュリティ監査および審査を受けていることを確認する。サードパーティが関連するセキュリティ情報を提供しないこともリスクインジケータに含まれる
- データやモデル抽出攻撃に対する抑止力として電子透かし技術を活用する

透明性と文書化

組織は以下を文書化することができる

- ベンダからのパッケージソフトの取得および調達、サイバーセキュリティ管理、計算インフラ、データ、データサイエンス、デプロイの仕組み、システム障害に関連するリスクにどの程度具体的に対処しているか？
- エンティティは、AI システムに関連するデータセキュリティとプライバシーへのインパクトについて、どのようなアセスメントを実施したか？
- データの生成、取得、収集、セキュリティ、メンテナンスおよび普及のためにどのようなプロセスが存在するか？
- エンティティは、エラーや限界を特定するために、AI システムに対してどのようなテスト（敵対的テストやストレステストなど）を実施したか？
- サードパーティが AI を作成した場合、説明可能性や解釈可能性をどのように確保するのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Matthew P. Barrett. “Framework for Improving Critical Infrastructure Cybersecurity Version 1.1.” National Institute of Standards and Technology (NIST), April 16, 2018.

Nicolas Papernot. "A Marauder's Map of Security and Privacy in Machine Learning." arXiv preprint, submitted on November 3, 2018.

Gary McGraw, Harold Figueroa, Victor Shepardson, and Richie Bonett. “BIML Interactive Machine Learning Risk Framework.” Berryville Institute of Machine Learning (BIML), 2022.

Mitre Corporation. “Mitre/Advmlthreatmatrix: Adversarial Threat Landscape for AI Systems.” GitHub, 2023.

National Institute of Standards and Technology (NIST). “Cybersecurity Framework.” NIST, 2023.

ソフトウェアのリソース

・adversarial-robustness-toolbox

・counterfit
・foolbox
・ml_privacy_meter
・robustness
・tensorflow/privacy

Measure 2.8

Map 機能で特定された透明性とアカウンタビリティに関連するリスクは、調査され文書化される。

概要

透明性を確保することで、AI パイプライン、ワークフロー、プロセスおよび組織すべてをよく可視化できるとともに、AI 開発者やオペレータと他の AI アクターや影響を受けるコミュニティとの間の情報の非対称性を低下させることができる。透明性は効果的な AI リスクマネジメントの中心的要素であり、AI システムがどのように機能しているかの洞察を与え、リスクが顕在化した場合にそれに対処することを可能にする。システムユーザ、個人または影響を受けたコミュニティが、AI システムの不正確な結果や問題のある結果に対して是正を求める能力は、透明性アカウンタビリティを確保するためのコントロールのひとつである。より高次のプロセスは通常、説明可能性と解釈可能性の機能における、より低次の実装努力によって可能となる（Measure 2.9 参照）。

AI リスクを低減するためには、組織やプロセス全体の透明性とアカウンタビリティが不可欠である。アカウンタビリティのあるリーダーシップ（個人であれグループであれ）、透明性のある役割、責任、コミュニケーションラインは、組織内の品質保証とリスクマネジメント活動を促進し、インセンティブを与える。

透明性の欠如は、トラストワージネスの測定を複雑にし、AI システムや組織が様々な個人や集団のバイアスや設計の盲点の影響を受けるかどうかの測定を複雑にし、ユーザ、組織、コミュニティの信頼を低下させ、システム全体の価値を低下させる可能性がある。説明責任と透明性のある組織構造を確立し、システムリスクを文書化することで、システムの改善とリスク管理の取り組みが可能になり、ライフサイクルに沿った AI アクターがエラーを特定し、改善を提案し、AI システムの特徴と結果をコンテクスト化し一般化する新しい方法を見つけ出すことができる。

推奨されるアクション

- 例えば、履歴、監査ログ、その他 AI アクターが使用できる情報といった、エラー、バイアスあるいは脆弱性の原因となりうる項目をレビューし評価するために使用できる情報を維持することによって、測定と追跡のためのシステムを構築する
- ユーザーインタラクションおよびユーザーエクスペリエンス（UI／UX）、ヒューマンコンピュータインタラクション（HCI）および／または人間と AI のチーム編成の専門家と緊密に協力しながら、ユーザのためのコントロールを調整する
- 結果的にエンドユーザや対象者に影響を与えるシステムの意思決定に対する是正を可能にするために、オペレータ、エンドユーザ、意思決定者、意思決定対象者（意思決定を行う個人）を含む様々なオーディエンスを対象とした校正に関する説明を提供するためのテストを行う
- AI システムに対するヒューマンオーバーサイトを測定し、文書化する

- AI システムの出力に関して、特定の AI アクターが提供するオーバーサイトの程度を文書化する
- システムのオーバーライドなど、エンドユーザやオペレータによる下流のアクションに関する統計を維持する
- 報告されたエラーや苦情、対応に要した時間、対応の種類に関する統計を維持し、文書化する
- 裁決活動に関する統計を管理し、報告する

● ポリシーの例外やエスカレーションを通じて、AI システムに関する組織のアカウンタビリティを追跡、文書化、測定し、実施および非実施の意思決定を文書化する。

● AI リスク管理に関する以下のような組織の仕組みの有効性を追跡し、監査する
- AI 関係者、経営幹部、ユーザおよび影響を受けるコミュニティの間のコミュニケーションライン
- AI アクターと経営幹部の役割と責任
- 組織的なアカウンタビリティの役割、例えば、主任モデル・リスク・オフィサー、AI オーバーサイト委員会、責任あるまたは倫理的な AI ディレクターなど

透明性と文書化

組織は以下を文書化することができる

● エンティティは、関連するステークホルダの役割、責任および委任された権限をどの程度明確にしているか？

● AI システムの設計、開発、デプロイ、評価およびモニタリングに携わる人員の役割、責任、権限の委任は何か？

● AI のライフサイクルの全段階における倫理的配慮について、誰が責任を負うのか？

● デプロイされた AI の保守、再検証、モニタリング、更新は誰が行うのか？

● さまざまな AI ガバナンスのプロセスに関与する人員の責任は明確に定義されているか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. 2022.

Inioluwa Deborah Raji and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles." arXiv preprint, submitted January 8, 2020.

Andrew Smith. "Using Artificial Intelligence and Algorithms." Federal Trade Commission Business Blog, April 8, 2020.

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011.

Joshua A. Kroll. “Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 758–71.

Jennifer Cobbe, Michelle Seng Lee, and Jatinder Singh. “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 598–609.

Measure 2.9

AI モデルは責任ある使用とガバナンスのあり方を伝えるため、説明され、妥当性確認され、文書化される。AI システムの出力は、Map 機能で特定されたそのコンテクストの中で解釈される。

概要

説明可能性と解釈可能性は、AI システムを運用またはオーバーサイトする人々や AI システムの利用者が、その出力を含むシステムの機能性とトラストワージネスをより深く理解することを支援する。

説明可能で解釈可能な AI システムは、エンドユーザが AI システムの目的と潜在的な影響を理解するのに役立つ情報を提供する。説明可能性の欠如によるリスクは、AI システムがどのように機能するかを、ユーザの役割、知識、スキルレベルなどの個人差に合わせて説明することで管理することができる。説明可能なシステムは、より簡単にデバッグやモニターを行うことができ、より徹底した文書化、監査、ガバナンスに適している。

解釈可能性に対するリスクは、多くの場合、AI システムが特定の予測やレコメンデーションを行った理由の説明を伝えることで対処できる。

透明性、説明可能性、解釈可能性は、互いに支え合う別個の特徴である。透明性は「何が起こったか」という問いに答えることができる。説明可能性は、システムにおいて「どのように」決定がなされたかという問いに答えることができる。解釈可能性は、システムによって「なぜ」決定がなされたのか、そしてその意味やコンテクストがユーザにとってどうなのかという疑問に答えることができる。

推奨されるアクション

- ベリファイシステムは、説明可能なモデル、ポストホックな説明、監査ログを作成するために開発される
- 可能または利用可能な場合は、従来の一般化線形モデルや罰則付き一般化線形モデル、決定木、最近傍モデルやプロトタイプベースのアプローチ、ルールベースのモデル、一般化加法モデルなど、説明可能なブースティングマシンおよびニューラル加法モデルなど、本質的に説明可能なアプローチを利用する
- 説明の正確性、明確性、理解可能性について、関連する AI アクター、エンドユーザ、インパクトを受ける可能性のある個人またはグループからフィードバックを得るために、デプロイ前に説明方法とその結果の説明をテストする
- モデルの種類（畳み込みニューラルネットワーク、強化学習、決定木、ランダムフォレストなど）、データの特徴、学習アルゴリズム、提案された用途、決定しきい値、学習データ、評価データ、倫理的配慮など、AI モデルの詳細を文書化する

- デプロイ環境に関連する人口統計グループやその他のセグメントにわたるパフォーマンスとエラーのメトリクスを確立し、文書化し、報告する
- 視覚化、モデル抽出、特徴の重要性など、さまざまな方法を用いてシステムを説明する。説明は複雑なシステムを正確に要約しているとは限らないので、忠実性、一貫性、堅牢性、解釈可能性などの特性に従って説明をテストする
- 忠実度（ローカルおよびグローバル）、曖昧さ、解釈可能性、双方向性、一貫性および攻撃や操作へのレジリエンスなどの特性に従って、システム説明の特性を評価する
- エンドユーザやその他のグループとシステム説明の品質をテストする
- モデル開発プロセスのセキュアを確保し、説明プロセスのゲーム化など、外部からの操作に対する脆弱性を回避する
- 実運用データへの反応を調整するモデルを含め、時間の経過に伴うモデルの変化をテストする
- データのステートメントやモデルカードなどの透明性ツールを使用して、説明や妥当性確認情報を文書化する

透明性と文書化

組織は以下を文書化することができる

- AI の目的に照らして、AI がどのように判断したかについて、どの程度の説明可能性や解釈可能性が求められるか？
- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？
- エンティティは、ソース、起源、変換、補強、ラベル、依存関係、制約およびメタデータを含む、AI システムのデータ出所をどのように文書化しているか？
- エンドユーザ、消費者、規制当局、AI システムの使用によって影響を受ける個人を含む外部のステークホルダが、AI システムの設計、運用、制限についてどのような情報にアクセスできるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework WEF Companion to the Model AI Governance Framework, 2020.

参考文献

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. "This Looks Like That: Deep Learning for Interpretable Image Recognition." arXiv preprint, submitted December 28, 2019.

Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." arXiv preprint, submitted September 22, 2019.

David A. Broniatowski. "NISTIR 8367 Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST), 2021.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI.” Information Fusion 58 (June 2020): 82–115.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. “Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems.” IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces, March 17, 2020, 454–64.

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. "NISTIR 8312 Four Principles of Explainable Artificial Intelligence." National Institute of Standards and Technology (NIST), September 2021.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” FAT *19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 220–29.

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. “A Nutritional Label for Rankings.” SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data, May 27, 2018, 1773–76.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." arXiv preprint, submitted August 9, 2016.

Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4, 2017, 4768-4777.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods.” AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 180–86.

David Alvarez-Melis and Tommi S. Jaakkola. "Towards robust interpretability with selfexplaining neural networks." NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, December 3, 2018, 7786-7795.

FinRegLab, Laura Biattner, and Jann Spiess. "Machine Learning Explainability & Fairness: Insights from Consumer Lending." FinRegLab, April 2022.

Miguel Ferreira, Muhammad Bilal Zafar, and Krishna P. Gummadi. "The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems." arXiv preprint, submitted October 31, 2016.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. "Interpretable & Explorable Approximations of Black Box Models." arXiv preprint, July 4, 2017.

ソフトウェアのリソース

・SHAP
・LIME
・Interpret
・PiML
・Iml
・Dalex

Measure 2.10

Map 機能で定められた AI システムのプライバシーリスクは調査され、文書化される。

概要

プライバシーとは一般に、人間の自律性、アイデンティティおよび尊厳を守るための規範および慣行を指す。これらの規範および慣行は通常、侵入からの自由、観察の制限または個人のアイデンティティ（身体、データ、評判など）の側面に関する開示や管理に同意する個人の主体性に対処するものである。

匿名性、機密性、管理といったプライバシーの価値観は、一般的に AI システムの設計、開発およびデプロイにおける選択の指針となるべきである。プライバシー関連のリスクは、安全性、バイアス、透明性に影響を与える可能性があり、これらの他の特性とのトレードオフを伴う。安全性やセキュリティと同様に、AI システムの特定の技術的特徴は、プライバシーを促進することもあれば、低下させることもある。AI システムはまた、推論によって個人を特定したり、個人に関する以前は非公開だった情報を特定したりすることで、プライバシーに新たなリスクをもたらす可能性もある。

AI のためのプライバシー強化技術（Privacy-enhancing technologies, PETs）は、特定のモデル出力に対する非特定化や集約などのデータ最小化手法と同様に、プライバシー強化 AI システムの設計をサポートすることができる。データのスパース性など特定の条件下では、プライバシー強化技術は精度の低下を招く可能性があり、特定のドメインにおける公平性やその他の価値に関する判断にインパクトを与える。

推奨されるアクション

- エンドユーザやインパクトを受ける可能性のあるグループやコミュニティとの直接的な関わりを通じて、利用状況において適用可能なプライバシー関連の価値観、フレームワーク、属性を特定する
- プライバシーポリシーおよびデータガバナンスポリシーに従い、データセットの個人の機微な情報の収集、使用、マネジメント、開示を文書化する個人またはグループを特定する能力など、プライバシーレベルのデータ側面を定量化する（例：k-匿名性メトリクス、l-多様性、t-近接性）
- プライバシーおよびデータガバナンスポリシーに従い、個人の機微な情報を含むトレーニングセットまたは本番データのプロトコル（権限、期間、種類）およびアクセス制御を確立し、文書化する
- 個人記録を分離するパターンを検出するために、本番データへの内部クエリをモニターする
- PSI による機密または法的に保護された属性の開示と推論をモニターする
    - 過度にカスタマイズされたコンテンツから不正操作のリスクを評価する。個人間の違いの様々な軸に沿って抽出された情報を評価する（例えばさまざまな年齢、ジェンダー、人種、政治的信条を持つ人々）
- データセット情報を公開する際には、差分プライバシーのようなプライバシー強化技術を使用する

- プライバシーのエキスパート、AI のエンドユーザやオペレータ、その他のドメインのエキスパートと協力して、使用コンテクストにおける最適なプライバシーのメトリクスを決定する

透明性と文書化

組織は以下を文書化することができる

- 組織は、データ管理と保護においてアカウンタビリティに基づく慣行（例えば PDPA や OECD プライバシー原則）を導入したか？
- エンティティは、AI システムに関連するデータセキュリティとプライバシーへのインパクトについて、どのような評価を実施したか？
- 組織は、特定された AI ソリューションのデプロイに伴うリスクに対処するために、リスク管理システムを導入したか（例えば人的リスクや事業目的の変更）？
- そのデータセットには，機微または機密とみなされる可能性のある情報が含まれているか(例えば個人を特定できる情報）？もしそれが人に関するものであれば、データセットは人を危害や法的行為にさらす可能性があるか(例えば、金銭的、社会的あるいはその他）？危害の可能性を軽減または低減するために何が行われたか？

AI の透明性に関するリソース

・WEF Companion to the Model AI Governance Framework- WEF Companion to the Model AI Governance Framework, 2020.

・Datasheets for Datasets.

参考文献

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0." National Institute of Standards and Technology (NIST), January 16, 2020.

Latanya Sweeney. “K-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 5 (2002): 557–70.

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. “L-Diversity: Privacy beyond K-Anonymity.” 22nd International Conference on Data Engineering (ICDE'06), 2006.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "CERIAS Tech Report 2007-78 t-Closeness: Privacy Beyond k-Anonymity and -Diversity." Center for Education and Research, Information Assurance and Security, Purdue University, 2001.

J. Domingo-Ferrer and J. Soria-Comas. "From t-closeness to differential privacy and vice versa in data anonymization." arXiv preprint, submitted December 21, 2015.

Joseph Near, David Darais, and Kaitlin Boeckly. "Differential Privacy for PrivacyPreserving Data Analysis: An Introduction to our Blog Series." National Institute of Standards and Technology (NIST), July 27, 2020.

Cynthia Dwork. “Differential Privacy.” Automata, Languages and Programming, 2006, 1–12.

Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. "Differential Privacy and Machine Learning: a Survey and Review." arXiv preprint, submitted December 24,2014. URL Michael B. Hawes. "Implementing Differential Privacy: Seven Lessons From the 2020 United States Census." Harvard Data Science Review 2, no. 2 (2020).

Harvard University Privacy Tools Project. “Differential Privacy.” Harvard University, n.d.

John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Matthew Spence Sexton and Pavel Zhuravlev. "The 2020 Census Disclosure Avoidance System TopDown Algorithm." United States Census Bureau, April 7, 2022.

Nicolas Papernot and Abhradeep Guha Thakurta. "How to deploy machine learning with differential privacy." National Institute of Standards and Technology (NIST), December 21, 2021.

Claire McKay Bowen. "Utility Metrics for Differential Privacy: No One-Size-Fits-All." National Institute of Standards and Technology (NIST), November 29, 2021. URL Helen Nissenbaum. "Contextual Integrity Up and Down the Data Food Chain." Theoretical Inquiries in Law 20, L. 221 (2019): 221-256.

Sebastian Benthall, Seda Gürses, and Helen Nissenbaum. “Contextual Integrity through the Lens of Computer Science.” Foundations and Trends in Privacy and Security 2, no. 1 (December 22, 2017): 1–69.

Jenifer Sunrise Winter and Elizabeth Davidson. “Big Data Governance of Personal Health Information and Challenges to Contextual Integrity.” The Information Society: An International Journal 35, no. 1 (2019): 36–51. https://doi.org/10.1080/01972243.2018.1542648

Measure 2.11

Map 機能で特定された公平性とバイアスを評価し、結果を文書化する。

概要

AI における公平性には、有害な偏見や差別といった問題に取り組むことで、平等性や公平性に配慮することも含まれる。公平性のスタンダードは複雑で定義が難しい場合があるが、それは公正性に対する認識が文化によって異なり、また適用によって変化する可能性があるからである。このような違いを認識し考慮することで、組織のリスクマネジメントの取り組みは強化される。有害なバイアスが緩和されたシステムは、必ずしも公平ではない。例えば、人口統計学的なグループ間で予測がある程度均衡しているシステムは、障害を持つ人やデジタルデバイドの影響を受けている人にとっては依然としてアクセスしにくいかもしれないし、既存の格差や制度的なバイアスを悪化させるかもしれない。

バイアスは、人口統計学的バランスやデータの代表性よりも幅広い。NIST は、AI バイアスを考慮しマネージすべき 3 つの主要なカテゴリーとして、系統的、計算および統計的、そして人間の認知的なものを挙げている。これらはいずれも、偏見や不公平、差別的意図がない場合に起こりうる。

- AI データセット、AI ライフサイクル全般にわたる組織の規範、プラクティス、プロセス、そして AI システムを利用する広範な社会には、系統的なバイアスが存在する可能性がある
- AI データセットやアルゴリズム処理には、計算上および統計上のバイアスが存在する可能性があり、多くの場合、非代表的サンプルに起因する系統的エラーに起因する
- 人間の認知的バイアスとは、個人や集団が AI システムの情報をどのように認識して意思決定を行うか、あるいは不足している情報を補うかあるいは、人間が AI システムの目的や機能をどのように考えるかに関係する。人間の認知的バイアスは、設計、実装、運用、保守を含め、AI のライフサイクルやシステム使用における意思決定プロセスに存在している

バイアスは様々な形で存在し、私たちの生活に関する意思決定を支援する自動化システムに染み込む可能性がある。バイアスは必ずしも否定的な現象ではないが、AI システムはバイアスのスピードと規模を増大させ、個人、グループ、コミュニティ、組織および社会への危害を永続化、増幅させる可能性がある。

推奨されるアクション

- 公平性アセスメントを実施し、以下のステップを含むバイアスの計算および統計的形式を管理する
  - 配分的、表象的、サービスの質、ステレオタイプ化、消失など、危害の種類を特定する
  - グループ間、グループ内、グループと交差するグループとの間で、危害を受ける可能性があるものを特定する

- 適切であれば、一般的な公平性のメトリクス（例えば人口統計学的平等、確率の均等化、機会の均等化、統計的仮説検定）と影響を受けるコミュニティと協力して開発された、コンテクストに沿ったカスタムメトリクスの両方を用いて危害を定量化する
- グループ間、グループ内、交差するグループ間のコンテクスト上有意な差異について、定量化された有害性を分析する。
- グループ内格差と交差するグループ間格差の特定を精緻化する
  - 定量化された有害性の分析において、基礎となるデータ分布を評価し、感度分析を採用する
  - 偽陽性率や偽陰性率などの品質指標を評価する
  - 小集団、集団内または部門間のコミュニティ、あるいは一個人に影響するバイアスを考慮する

- 学習データと TEVV データのバイアスの原因を理解し、考慮する
  - 交差するグループを含むグループ間およびグループ内におけるアウトカムの分布の違い
  - データソースの完全性、代表性およびバランス
  - AI システムにおいて、人口統計学的グループメンバーシップのプロキシ（例えばクレジットスコア、郵便番号など）となり、そうでなければバイアスが増加してしまうような入力データの特徴を特定する
  - 画像、テキスト（または単語埋め込み）、その他複雑あるいは非構造化データ系統的バイアスの形態
- インパクトのアセスメントを活用し、潜在的にインパクトを受けるかもしれないコミュニティのエンドユーザやその他の個人、集団から情報を得て、システムが与えるインパクトや危害を特定および分類する
- インパクトを受ける可能性のあるコミュニティとの直接的な関わりを通じて、インパクトを受ける可能性のある個人、グループまたはエコシステムの分類を特定する
- バイアスのテストにおける障がいの有無の考慮、非包括的な設計やデプロイの決定から生じる可能性のある差別的スクリーンアウトのプロセスなど、障害者の包摂に関するシステムを評価する
- 集団と集団外の力学や系統的なバイアスを考慮した目的関数を開発する
- コンテクスト固有の公平性メトリクスを使用して、システムのパフォーマンスがグループ間、グループ内および／または交差するグループ間でどのように異なるかを調べる。メトリクスには、統計的公平性、エラー率公平性、統計的公平性差、均等機会差、平均絶対オッズ差、標準平均差、パーセンテージのポイント差などがある
- 公平性のメトリクスを特定のコンテクストに合わせてカスタマイズし、システムのパフォーマンスと潜在的な有害性がコンテクストの規範の中でどのように変わるかを検証する
- 確立された組織統治方針、ビジネス要件、規制遵守、法的枠組み、倫理基準に従って、使用コンテクストにおけるパフォーマンスの違いの許容レベルを定義する
- 格差レベルが許容値を超えた場合に取るべきアクションを定義する

- インパクトを受けるコミュニティと協力して、予想される集団の中で差別に関する分析を必要とする可能性のあるグループを特定する
- 特定のコンテクストに精通したエキスパートを活用し、実質的な測定の違いを調査し、その違いの根本原因を特定する
- システムの出力を監視し、設定された許容レベルを超えるパフォーマンスやバイアスの問題がないか確認する
- 定期的なモデルの更新を確実に行い、許容範囲内の差異に収まるよう、更新されたより代表的なデータを用いてテストと再較正を行う
- 人口統計学的バランスとデータの代表性に関連する要因に対処するために、前処理としてデータ変換を適用する
- インプロセッシングを適用して、モデルのパフォーマンス品質とバイアスのバランスをとる
- インパクトの審査員、社会技術のエキスパート、そのほか使用されるコンテクストの専門知識を有する AI アクターと緊密に協力して、モデルの結果の後処理に数学的／計算技術を適用する
- バイアスのマネジメントやその他のトラストワージネスの特性を透明かつ慎重に考慮したモデル選択アプローチを適用する
- 特定されたグループのアウトカムの違いに関する情報を収集し、共有する
- 特に、過去の不公正またはバイアスのある人間の意思決定パターンに起因するような相違を緩和するための介入を検討する
- 人間中心設計のプラクティスを活用し、AI ライフサイクルの中で、社会へのインパクトに深く焦点を当て、人間の認知的バイアスに対抗する
- 可用性、観察、確認に関するバイアスなど、認知バイアスの潜在的な原因を特定し、暗黙の意思決定プロセスをより明確にし、調査の対象とするために、ライフサイクルに沿った実践を評価する
- ヒューマンファクターの専門家と協力して、エンドユーザ、オペレータ、実務家に対するシステム出力を提示する際のバイアスを評価する
- 多様な社内スタッフやステークホルダの参画など、状況認識を高めるプロセスを活用する

透明性と文書化

組織は以下を文書化することができる

- 確立された手続きは、バイアスや不公平感、その他制度に起因する懸念を軽減する上で、どの程度有効か？
- もしそれが人々に関するものであれば、それは特定の社会集団に不当に有利あるいは不利か？どのような方法で？それはどのように緩和されたか？
- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？
- エンティティは、不公平あるいは差別的なアウトカムを含む、データのバイアスによる潜在的なインパクトをどのように特定し、軽減したか？

- エンティティは、データに潜在的に含まれる（統計的、コンテクスト的、歴史的な）バイアスをどの程度特定し、軽減しているか？
- バイアスを測定するために、敵対的な機械学習アプローチ（例えばプロンプトエンジニアリング、敵対的モデル）が検討されたかまたは使用されたか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework WEF Companion to the Model AI Governance Framework, 2020.

・Datasheets for Datasets.

参考文献

Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. “Algorithmic Bias and Risk Assessments: Lessons from Practice.” Digital Society 1 (2022).

Richard N. Landers and Tara S. Behrend. “Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models.” American Psychologist 78, no. 1 (2023): 36–49.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning.” ACM Computing Surveys 54, no. 6 (July 2021): 1–35.

Michele Loi and Christoph Heitz. “Is Calibration a Fairness Requirement?” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 2026–34.

Shea Brown, Ryan Carrier, Merve Hickok, and Adam Leon Smith. “Bias Mitigation in Data Sets.” SocArXiv, July 8, 2021. URL

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. "NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." National Institute of Standards and Technology (NIST), 2022.

Microsoft Research. “AI Fairness Checklist.” Microsoft, February 7, 2022. URL Samir Passi and Solon Barocas. “Problem Formulation and Fairness.” FAT* '19:

Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 39–48.

Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. “An Ontology for Fairness Metrics.” AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, July 2022, 265–75.

Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. https://arxiv.org/pdf/1801.07593.pdf

Ganguli, D., et al. (2023). The Capacity for Moral Self-Correction in Large Language Models. arXiv. https://arxiv.org/abs/2302.07459

Arvind Narayanan. “Tl;DS - 21 Fairness Definition and Their Politics by Arvind Narayanan.” Dora's world, July 19, 2019.

Ben Green. “Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness.” Philosophy and Technology 35, no. 90 (October 8, 2022)

Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” Big Data 5, no. 2 (June 1, 2017): 153–63.

Sina Fazelpour and Zachary C. Lipton. “Algorithmic Fairness from a Non-Ideal Perspective.” AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 57–63.

Hemank Lamba, Kit T. Rodolfa, and Rayid Ghani. “An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings.” ACM SIGKDD Explorations Newsletter 23, no. 1 (May 29, 2021): 69–85.

ISO. “ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) —Bias in AI systems and AI aided decision making.” ISO Standards, November 2021.URLShari Trewin. "AI Fairness for People with Disabilities: Point of View." arXiv preprint,submitted November 26, 2018.

MathWorks. “Explore Fairness Metrics for Credit Scoring Model.” MATLAB & Simulink, 2023.

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Quantifying and Reducing Stereotypes in Word Embeddings." arXiv preprint, submitted June 20, 2016.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” Science 356, no. 6334 (April 14, 2017): 183–86.

Sina Fazelpour and Maria De-Arteaga. “Diversity in Sociotechnical Machine Learning Systems.” Big Data and Society 9, no. 1 (2022).

Fairlearn. “Fairness in Machine Learning.” Fairlearn 0.8.0 Documentation, n.d. URL Safiya Umoja Noble. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press, 2018.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science 366, no. 6464 (October 25, 2019): 447–53.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A Reductions Approach to Fair Classification." arXiv preprint, submitted July 16, 2018.

Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." arXiv preprint, submitted October 7, 2016.

Alekh Agarwal, Miroslav Dudik, Zhiwei Steven Wu. "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms." Proceedings of the 36th International Conference on Machine Learning, PMLR 97:120-129, 2019.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. “Fairness and Abstraction in Sociotechnical Systems.” FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29, 2019, 59–68.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations.” CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, April 18, 2015, 3819–28.

ソフトウェアのリソース

・aequitas

・AI Fairness 360:

・Python

・R

・algofairness

・fairlearn

・fairml

・fairmodels
・fairness
・solas-ai-disparity
・tensorflow/fairness-indicators
・Themis

Measure 2.12

Map 機能で定められた、AI モデルのトレーニングとマネジメント活動の環境への影響と持続可能性が評価され、文書化される。

概要

AI システムがトレーニングや運用に使用する大規模で高性能な計算リソースは、環境へインパクトを与える一因となりうる。これらのプロセスによる環境への直接的なネガティブな影響は、エネルギー消費、水消費、温室効果ガス（GHG）排出に関連する。OECD は、各タイプの負の直接的なインパクトに対する評価メトリクスを定めている。

環境に対する間接的なネガティブな影響は、人間の行動、社会経済システムおよび環境間の相互作用の複雑さを反映しており、消費の誘発や「リバウンド効果」を含むことがある、そこでは、効率向上は資源消費の加速によって相殺される。

AI に関連するその他の環境へのインパクトは、コンピュータ機器や通信網の製造（例えば原材料の採掘や抽出など）、ハードウェアの輸送、電子廃棄物のリサイクルや廃棄から生じる可能性がある。

推奨されるアクション

- AI システムの設計・開発計画に消費抑制や効率改善を含む環境へのインパクトを示す指標を盛り込む
- AI システムのエネルギーと水の消費と効率、GHG 排出量の主要指標を特定し、実施する
- 組織の方針、コンプライアンス、法的枠組み、環境保護および持続可能性の規範に準拠した、持続可能な AI システム運用のための測定可能なベースラインを確立する
- 組織の原則と方針、コンプライアンス、法的枠組み、環境保護と持続可能性に関する規範に従い、AI システムのパフォーマンスと持続可能なオペレーションとのトレードオフを評価する
- 許容可能な資源の消費と効率、GHG 排出量レベルと、指標が許容レベルを超えた場合に取るべき措置を定め、確立する
- AI のライフサイクル全体を通して、AI システムの排出レベルを推定する

透明性と文書化

組織は以下を文書化することができる

- 温室効果ガスの排出量、エネルギーと水の消費量と効率は組織内で追跡されているか？
- デプロイされた AI システムは、上流および下流における潜在的な環境へのインパクト（例えば消費量の増加や排出量の増加）について評価されているか？
- デプロイされた AI システムが、大気汚染、水質汚染、有毒物質の流出、火災および爆発などの環境インシデントを引き起こす可能性はあるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・Datasheets for Datasets. URL

参考文献

Organisation for Economic Co-operation and Development (OECD). "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint.” OECD Digital Economy Papers, No. 341, OECD Publishing, Paris.

Victor Schmidt, Alexandra Luccioni, Alexandre Lacoste, and Thomas Dandres. “Machine Learning CO2 Impact Calculator.” ML CO2 Impact, n.d.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. "Quantifying the Carbon Emissions of Machine Learning." arXiv preprint, submitted November 4, 2019.

Matthew Hutson. “Measuring AI’s Carbon Footprint: New Tools Track and Reduce Emissions from Machine Learning.” IEEE Spectrum, November 22, 2022.

Association for Computing Machinery (ACM). "TechBriefs: Computing and Climate Change." ACM Technology Policy Council, November 2021.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. “Green AI.” Communications of the ACM 63, no. 12 (December 2020): 54–63.

Measure 2.13

Measure 機能で採用された TEVV メトリクスとプロセスの有効性が評価され、文書化される。

概要

メトリクスの開発は、客観的であると考えられがちなプロセスであるが、人間や組織主導の取り組みであるため、暗黙のバイアスや系統的バイアスが反映されることがあり、対象機能とは無関係な要因が反映されることがある。測定アプローチは、過度に単純化され、ゲーム化され、批判的なニュアンスを欠き、予期せぬ方法で使われあるいは依存され、影響を受けるグループやコンテクストの違いを考慮できなくなる可能性がある。

Measure 2.1 から Measure 2.12 において選択したメトリクスを、継続的な改善プロセスの中で再検討することは、AI アクターがメトリクスの有効性を評価および文書化し、必要な軌道修正を行うのに役立つ。

推奨されるアクション

- 選択したシステムメトリクスと関連する TEVV プロセスをレビューし、エラーの特定と除去を含むシステム改善を維持できるかどうかを判断する
- システムのメトリクスの有用性を定期的に評価し、過度に複雑な手法の代わりに記述的なアプローチを検討する
- 選択されたシステムのメトリクスについて、エンドユーザおよび影響を受けるステークホルダのコミュニティ内で受け入れられるかどうかを検討する
- リスクを特定し、測定するためのメトリクスの有効性を評価する

透明性と文書化

組織は以下を文書化することができる

- システムおよびエンティティは、明示されたゴールや目標に対する進捗をどの程度一貫して測定しているか？
- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？
- エンティティは、データの品質、正確性、信頼性および代表性向上のためにどのような是正処置を講じたか？
- モデルのアウトプットは、社会からの信頼と公平性を育むためのエンティティの価値観や原則とどの程度一致しているか？
- 正確さや適切なパフォーマンス指標はどのように評価されるのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community

参考文献

Arvind Narayanan. "The limits of the quantitative approach to discrimination." 2022

James Baldwin lecture, Princeton University, October 11, 2022.

Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. “A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System.” CHI ‘20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, April 23, 2020, 1–15.

Rachel Thomas and David Uminsky. “Reliance on Metrics Is a Fundamental Challenge for AI.” Patterns 3, no. 5 (May 13, 2022): 100476.

Momin M. Malik. "A Hierarchy of Limitations in Machine Learning." arXiv preprint, submitted February 29, 2020. https://arxiv.org/abs/2002.05193

Measure 3

特定された AI リスクを長期にわたって追跡する仕組みが整えられる。

Measure 3.1

デプロイされた状況における意図されたパフォーマンスと実際のパフォーマンスなどの要因に基づき、既存の、予期されなかったおよび出現した AI リスクを定期的に特定し、追跡するためのアプローチ、人員、文書が整備されている。

概要

トラストワージな AI システムについては、組織のガバナンス方針、AI アクターの役割と責任、継続的な改善の文化の中で、定期的なシステムのモニタリングが実施される。新たなリスクや複雑なリスクが発生した場合は、定期的なモニタリングなど、内部のリスクマネジメント手順を適応させ、軌道修正する必要があるかもしれない。文書化、リソースおよびトレーニングは、AI システムのエラー、インシデントおよびネガティブなインパクトを調査し対応する AI アクターをサポートする全体的な戦略の一部である。

推奨されるアクション

- AI システムのリスクを以下の項目と比較する
  - よりシンプルな、あるいは伝統的なモデル
  - 人間のベースラインとなるパフォーマンス
  - その他のマニュアルなパフォーマンスのベンチマーク
- デプロイされた AI システムに関するエンドユーザやコミュニティからのフィードバックを、システム性能の内部測定値と比較する
- 顕在化するリスクを特定し測定するためのメトリクスの有効性をアセスメントする
- エラーの応答時間を測定し、応答品質を追跡する
- ユーザーサポートの役割を担う AI アクターから、システムの問題をある程度解決するために必要なメトリクス、説明、その他のシステム情報の種類に関するフィードバックを引き出し、追跡する。以下の項目を検討する
  - 考えられるエラーの原因を調査したり、対応策を特定したりするのに、説明が不十分な場合
  - システムのエラーの原因を特定し診断するためのシステムのログと説明を含むシステムのメトリクス
- インシデント対応や支援の役割を担っている AI 関係者から、効果的かつタイムリーに職務を遂行するための人員や資源の妥当性に関するフィードバックを引き出し、追跡する

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework Implementation and Self-Assessment Guide for Organizations.

透明性と文書化

組織は以下を文書化することができる

- 組織は、特定された AI ソリューションの導入に伴うリスクに対処するために、リスク管理システムを導入したか（例えば人的リスクや事業目的の変更）？
- AI システムのアウトプットによって影響を受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- エンティティは、エラーロギングを含め、AI システムのパフォーマンスを測定するためにどのような指標を開発したか？
- その指標は、どの程度まで正確で有用なパフォーマンスの指標となっているか？

参考文献

ISO. "ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centered design for interactive systems." 2nd ed. ISO Standards, July 2019.

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub.

Measure 3.2

リスク追跡のアプローチは、AI リスクが現在利用可能な測定技術で評価することが困難であるかまたはメトリクスがまだ利用可能でない設定のために検討される。

概要

Map 機能において定められたリスクは、複雑であったり、時間の経過とともに出現したり、測定が困難な場合がある。新規の測定アプローチを含むリスク追跡のための体系的な方法は、定期的なモニタリングと改善プロセスの一環として確立することができる。

推奨されるアクション

- 現在のアプローチでは測定できない可能性のある新たなリスクを追跡するためのプロセスを確立する。プロセスには以下のようなものがある
  - 欠陥のある AI システムの出力に対する是正メカニズム
  - バグの報奨金
  - 人間中心設計のアプローチ
  - ユーザーインタラクションおよびユーザーエクスペリエンスの研究
  - インパクトを受けるまたは受ける可能性のある個人およびコミュニティとの参加型ステークホルダ参画
- 顕在化したリスクの追跡とインベントリ方法を担当する AI アクターを特定する
- 複雑なリスクや測定が困難なリスクについて、発生率と重大度レベルを決定し、文書化する
  - デプロイタスクに対する新しい測定アプローチの優先順位付け
  - AI システムのリスクマネジメントリソースの割り当て
  - AI システムの改善評価
  - 後続システムのイテレーションについて実施あるいは非実施の決定

透明性と文書化

組織は以下を文書化することができる

- AI の決定に対する最終的な責任は誰にあり、その人物は分析の意図する用途と限界を認識しているか？
- デプロイされた AI の保守、再検証、モニタリング、更新は誰が行うのか？
- エンティティは、ステークホルダのコミュニティに AI の戦略的なゴールおよび目標をどの程度伝えているか？
- この AI の目的を考慮した場合、それがまだ正確か、バイアスがないか、説明可能かどうか等をチェックするための適切な間隔はどの程度か？このモデルのチェック項目は何か？

- AI がもはやこの倫理的枠組みを満たしていないと考える人がいるならば、誰がその懸念を受け止め、必要に応じて問題を調査し、改善する責任を持つのか？その人たちは、AI の使用を修正、制限あるいは停止する権限を持っているのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community AI Transparency Resources

参考文献

ISO. "ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centered design for interactive systems." 2nd ed. ISO Standards, July 2019.

Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. “Capability Maturity Model, Version 1.1.” IEEE Software 10, no. 4 (1993): 18–27.

Jeff Patton, Peter Economy, Martin Fowler, Alan Cooper, and Marty Cagan. User Story Mapping: Discover the Whole Story, Build the Right Product. O'Reilly, 2014.

Rumman Chowdhury and Jutta Williams. "Introducing Twitter’s first algorithmic bias bounty challenge." Twitter Engineering Blog, July 30, 2021.

HackerOne. "Twitter Algorithmic Bias." HackerOne, August 8, 2021.

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Bug Bounties for Algorithmic Harms?" Algorithmic Justice League, January 2022.

Microsoft. “Community Jury.” Microsoft Learn's Azure Application Architecture Guide, 2023.

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. "Overcoming Failures of Imagination in AI Infused System Development and Deployment." arXiv preprint, submitted December 10, 2020.

Measure 3.3

エンドユーザや影響を受けるコミュニティが、問題やシステムのアウトカムを報告するためのフィードバックプロセスを確立し、AI システムの評価メトリクスに組み込む。

概要

インパクトのアセスメントは双方向の取り組みである。多くの AI システムの成果や影響は、AI ライフサイクルの開発・展開の次元にまたがる AI アクターには見えなかったり認識できなかったりする可能性があり、エンドユーザや影響を受けるグループの視点からシステムの成果に関する直接的なフィードバックが必要になる場合がある。

フィードバックは、エンドユーザやオペレータからのエラーやその他のフィードバックを収集するために機械化されたシステムを介して、間接的に収集することができる。

このサブカテゴリーで開発されたメトリクスと洞察は、Manage 4.1 および 4.2 に反映される。

推奨されるアクション

- エンドユーザとオペレータのエラー報告プロセスの有効性を測定する
- エンドユーザからの異議申し立て要求とその結果の種類と割合を分類し、分析する
- フィードバック活動の参加率と、フィードバック活動の可用性の認知度を測定する
- 測定アプローチを分析し、その後の行動方針を決定するためにフィードバックを活用する
- 測定アプローチを評価し、実世界のインパクトに関する組織の理解を促進するための有効性を判断する
- エンドユーザやコミュニティからのフィードバックを、ドメインのエキスパートと密接に連携しながら分析する

透明性と文書化

組織は以下を文書化することができる

- AI システムのアウトプットによって影響を受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- 組織は、ユーザビリティの問題に取り組み、ユーザーインタフェースが意図した目的を果たしているかどうかをテストしたか？
- 社外のステークホルダが入手可能な情報に、どれだけ容易にアクセスでき、最新であるか？
- エンドユーザ、消費者、規制当局、AI システムの使用によって影響を受ける個人を含む外部のステークホルダが、AI システムの設計、運用、制限についてどのような情報にアクセスできるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・WEF Companion to the Model AI Governance Framework Implementation and Self-Assessment Guide for Organizations

参考文献

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020.

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021.

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021.

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125.

Batya Friedman, Peter H. Kahn, Jr., and Alan Borning. "Value Sensitive Design: Theory and Methods." University of Washington Department of Computer Science & Engineering Technical Report 02-12-01, December 2002.

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021.

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84.

Measure 4

測定の有効性に関するフィードバックを収集し、アセスメントを実施する。

Measure 4.1

AI リスクを特定するための測定アプローチは、デプロイのコンテクストと関連付けられ、そのドメインのエキスパートや他のエンドユーザとの協議を通じて知らされる。アプローチは文書化される。

概要

TEVV タスクを遂行する AI アクターは、システム利用のコンテクストの中で影響を評価することが困難な場合がある。AI システムのリスクと影響は、エンドユーザや、アウトプットとその後の決定によって影響を受ける可能性のある人々によって説明されるのが最も適切であることが多い。AI アクターは、Govern 5.1 および 5.2 で確立され、Map 1.6、5.1 および 5.2 で実施される参加型エンゲージメントのプロセスを通じて、インパクトを受ける個人やコミュニティからフィードバックを引き出すことができる。

Measure 機能で説明した活動により、AI アクターは、影響を受けた個人やコミュニティからのフィードバックを評価することができる。洞察の認識を高めるために、フィードバックは、インパクトのアセスメント、ヒューマンファクター、ガバナンスのオーバーサイトを担当する AI アクターや、他の社会技術のエキスパートや研究者と緊密に連携して評価することができる。評価のアウトカムを解釈するためのより広範な専門知識を得るために、組織は、アドボカシ団体や市民社会団体との協力を検討することができる。

この種の分析に基づく洞察は、メトリクスや関連する行動方針について、TEVV に基づく決定に反映させることができる。

推奨されるアクション

- システムエンドユーザ（ドメインのエキスパート、オペレータ、実務者を含む）からのフィードバックを得るためのメカニズムをサポートする。よいアプローチを以下に示す
  - エンドユーザが、AI システムの出力に関する疑問や洞察をオープンに共有できるような環境で、特定の使用状況に関連して実施される（設定やタスク固有の調査ラインを含む）
  - ヒューマンファクターと社会技術のエキスパートと研究者によって開発され、実施される
  - インタビュアーとエンドユーザの主観とバイアスを確実にコントロールするように設計されている
- アプローチを特定し、文書化する
  - 継続的な改善プロセスを積極的に伝える
  - システムのエンドユーザから得られたフィードバックを評価し、統合する
  - ヒューマンファクターおよび社会技術領域のエキスパートと協力
- エンドユーザからのフィードバックを、影響を受けたコミュニティからの評価されたフィードバックとともに評価する（Measure 3.3 参照）

- エンドユーザからのフィードバックを活用し、選択したメトリクスや測定手法が組織や業務のコンテクストとどのように相互作用するかを調査する
- 収集したエンドユーザからのフィードバックと比較しながら、システム内部の測定プロセスを分析し、文書化する
- エンドユーザのエリシテーション技術の有効性と満足度を測定するアプローチを特定し、実施し、結果を文書化する

透明性と文書化

組織は以下を文書化することができる

- 組織では、ユーザビリティの問題に取り組み、ユーザーインタフェースが意図した目的を果たしているかどうかをテストしたか?
- モデル開発プロセスや定期的なパフォーマンスのレビューの実施に、ユーザやピア・エンゲージメントをどのように組み込むのか？
- AI システムのアウトプットによって影響を受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- 確立された手続きは、バイアスや不公平感、その他制度に起因する懸念を軽減する上で、どの程度有効か？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Frameworkfor Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework Implementation and Self-Assessment Guide for Organizations

参考文献

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019.

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125.

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96.

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82.

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11– 40.

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022.

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no.2 (2021): 56–61.

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8.

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021.

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78.

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020.

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022.

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306.

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25.

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021.

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021.

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84.

Measure 4.2

AI システムが意図したとおりに一貫して機能しているかどうかを妥当性確認するために、AI システムのデプロイメントコンテクストや AI のライフサイクル全体におけるトラストワージネスに関する測定結果が、ドメイン専門家やその他の関連する AI アクターからの情報によって知らされる。結果は文書化される。

概要

関連する AI アクターから取得したフィードバックは、Measure2.5 から 2.11 の出力と組み合わせて評価することで、AI システムが妥当性と信頼性、安全性、セキュリティとレジリエンス、プライバシー、バイアスと公平性、説明可能性と解釈可能性および透明性とアカウンタビリティに関して事前に定義された運用限界内で機能しているかどうかを判断することができる。このフィードバックは、意図された設定以外での誤用や再利用の可能性など、AI システムの性能に関するさらなる洞察を提供する。

この種の分析に基づく洞察は、メトリクスや関連する行動方針について、TEVV に基づく決定に反映させることができる。

推奨されるアクション

- AI システムのトラストワージネス特性を評価するためのインプットとして、Map 機能からのエンドユーザ、オペレータ、影響を受ける個人やコミュニティからのフィードバックを統合する肯定的なフィードバックと否定的なフィードバックの両方がアセスメントされていることを確認する
- Measure 2.5 から 2.11 の AI システムのトラストワージネスの特徴に関連してフィードバックを評価する
- 出力が妥当で信頼できると考えられているか、説明可能で解釈可能であるかなど、AI システムの性能に対するエンドユーザの満足度や信頼度に関するフィードバックを評価する
- AI システムの出力（例えばレコメンデーション）を確認／サポートするメカニズム）およびそのアウトプットに関するエンドユーザの視点を特定する
- AI システムによる判断のオーバーライド頻度を測定し、結果を評価して文書化し、継続的な改善プロセスに洞察をフィードバックする
- インパクトのアセスメント、ヒューマンファクター、社会技術に関する AI アクターに分析および結果の解釈を支援する技術的作業に関し相談する

透明性と文書化

組織は以下を文書化することができる

- システム／エンティティは、明示された目標や目的に対する進捗をどの程度一貫して測定しているか？
- AI システムの利用が、そのエンティティが表明している価値観や原則と一致していることを保証するために、そのエンティティはどのような方針を策定してきたか？

- モデルのアウトプットは、社会からの信頼と公平性を育むためのエンティティの価値観や原則とどの程度一致しているか？
- AI の目的に照らして、AI がどのように判断したかについて、どの程度の説明可能性や解釈可能性が求められるか？
- AI システムのアウトプットによって影響を受けるユーザや関係者が、どの程度 AI システムをテストし、フィードバックを提供することができるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019.

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125.

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96.

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82.

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11– 40.

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022.

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no.2 (2021): 56–61.

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8.

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of

Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021.

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78.

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018.

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022.

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306.

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25.

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021.

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021.

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84.

Measure 4.3

影響を受ける地域社会を含む関連する AI アクターとの協議や、状況に関連するリスクやトラストワージネスの特性に関するフィールドデータに基づき、測定可能なパフォーマンスの改善または低下が特定され、文書化される。

概要

AI システムのライフサイクルを通じて実施される TEVV 活動は、トラストワージネス特性のベースラインとなる定量的尺度を提供することができる。Measure 2.5 から 2.11、Measure 4.1 および 4.2 の結果と組み合わせると、以下のように説明できる。TEVV アクターは、システムのパフォーマンスを包括的に把握することができる。これらの措置は、影響を受ける可能性のあるコミュニティとの参加型エンゲージメントや、AI システムの影響に関するステークホルダの意見聴取の他の形態によって補強することができる。これらのソースにより、AI アクターはシステムの構成要素に対する潜在的な調整を検討し、運転条件を調整し、パフォーマンスを改善することができる。

推奨されるアクション

- トラストワージネス特性のベースライン定量指標を作成する
- ベースラインの操作値と状態を定義し、特徴付ける
- インパクトのアセスメント、ヒューマンファクターおよび社会技術の AI アクターと緊密に連携し、定量的ベースライン測定を補強・補完するために定性的アプローチを活用する
- 継続的改善の一環として測定値をモニター・アセスメントし、潜在的なシステムの調整や修正を特定する
- システムまたは手順の更新を適用した後のパフォーマンスにおける実際の差異と予想される差異を特徴付けるための感度分析を実施し、文書化する
- 感度分析に関する決定を文書化し、システム性能および特定されたリスクに対する予想される影響を記録する

透明性と文書化

組織は以下を文書化することができる

- モデルのアウトプットは、社会からの信頼と公平性を育むためのエンティティの価値観や原則とどの程度一致しているか？
- コンプライアンスの対象となる可能性のあるセンシティブな変数（例えば人口統計学的、社会経済学的カテゴリーなど）は、モデリング目的のためにどのように特別に選択されたのか、あるいは選択されなかったのか？
- 組織は、特定された AI ソリューションの導入に伴うリスクに対処するために、リスク管理システムを導入したか（例えば人的リスクや事業目的の変更）？

- 敵対する主体が AI を混乱させようとしたり、あるいは関係なく運用／ビジネス環境が変化したりすることにより精度や正確さが変化した際に、アカウンタビリティの担当者はどのように対処するのか？
- モデル開発プロセスや定期的なパフォーマンス・レビューの実施に、ユーザやピア・エンゲージメントをどのように組み込むのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019.

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125.

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for

Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96.

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82.

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11– 40.

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no. 2 (2021): 56–61.

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces Companion, April 14, 2021, 7–8.

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis.“Human-Centered Design of Artificial Intelligence.” In Handbook of

Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021.

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78.

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018.

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020.

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022.

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306.

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25.

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021.

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021.

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84.

Manage

Manage

リスクは優先順位付けされ、予測されるインパクトに基づいて対処される。

Manage 1

Map および Measure 機能からのアセスメントおよびその他の分析出力に基づいて導出された AI リスクは優先順位付けされ、対応され、マネジメントされる。

Manage 1.1

AI システムがその意図した目的および明示された目標を達成しているかどうか、開発やデプロイメントを進めるべきかどうかを判断する。

概要

AI システムは、与えられたビジネスタスクや課題に対して必ずしも適切なソリューションとは限らない。標準的なリスクマネジメントは、AI システムのネガティブリスクとベネフィットを形式的に比較検討し、AI システムが適切なソリューションであるかどうかを判断することである。例えば、システムのパフォーマンスとシステムの透明性の比較に基づいてシステムの導入を決定するなど、トラストワージネスの特性間のトレードオフには、AI のライフサイクル全体を通じて定期的なアセスメントが必要な場合がある。

推奨されるアクション

- AI システムのネガティブリスクとベネフィットを評価する際に、トラストワージネスの特性を考慮する
- リスク処理を検討する際には、Map と Measure 機能から得られるテスト、評価、妥当性確認および検証（Test, evaluation, validation, and verification: TEVV）出力を活用する
- デプロイ後のモニタリングも含め、AI システムのライフサイクルを通じて、ネガティブリスクとベネフィットを定期的に追跡・モニタリングする
- トラストワージネスの特性と、ネガティブなリスクと機会との間のトレードオフに関するシステムのパフォーマンスを定期的にアセスメントし、文書化する
- Map 機能の出力に列挙されているような、実際の使用例やインパクトに関連したトレードオフを評価する

透明性と文書化

組織は以下を文書化することができる。

- 技術仕様と要件は、AI システムのゴールと目標にどのように合致しているか？

- メトリクスは、倫理やコンプライアンスへの配慮を含め、システムのゴールと目標、制約とどの程度一致しているか？
- エンティティは、AI システムの設計、開発および／またはデプロイによって、どのようなゴールと目標を達成することを期待しているのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations.

参考文献

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021).

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation(September 30, 2021).

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022).

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021.

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695.

Manage 1.2

文書化された AI リスクの取扱いは、インパクトと頻度または利用可能なリソースや方法に基づいて優先順位付けされる。

概要

リスクとは、ある事象の発生確率と、それに対応する事象がもたらす結果の大きさ（または程度）の複合的な尺度を指す。AI システムのインパクト、すなわちその結果はポジティブな場合もあるし、ネガティブな場合、あるいはその両方である場合もあり、機会やリスクをもたらす可能性がある。

組織のリスク許容度は、多くの場合、既存の業界プラクティス、組織の価値観、法的または規制上の要件など、いくつかの内外の要因によって決定される。リスクマネジメントのリソースは限られていることが多いため、組織は通常、リスク許容度に基づいてリソースを割り当てる。しかし、より深刻と判断された AI リスクには、より多くのオーバーサイトが実施され、リスクマネジメントリソースが投入される。

推奨されるアクション

- 確立されたリスク許容度に応じたリスクマネジメントリソースを割り当てる。リスク許容度が低い AI システムには、より大きなモニタリング、緩和、マネジメントリソースを割り当てる
- AI リスク許容度の決定方法とリソースの決定を文書化する
- リスク許容度を定期的に見直し、必要に応じて、AI システムのモニタリングおよび評価からの情報にしたがって再調整する

透明性と文書化

組織は以下を文書化することができる。

- 組織は、特定された AI ソリューションの導入に伴うリスクに対処するために、リスクマネジメントシステムを導入したか（例えば人的リスクや商業目的の変更）？
- AI システムに関連するデータセキュリティとプライバシーへのインパクトについて、エンティティはどのような評価を実施したか？
- 組織には、組織の AI 利用をオーバーサイトするために活用できるガバナンス構造があるか？

AI の透明性に関するリソース

・WEF Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations.

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

参考文献

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021).

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation(September 30, 2021).

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022).

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021.

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA,695.

Manage 1.3

Map 機能によって特定された優先度が高いと考えられる AI リスクへの対応は、策定され、計画され、文書化される。リスク対応オプションには、軽減、移転、回避、許容が含まれる。

概要

Govern 1、Map 5 および Measure 2 のアウトカムは、確立されたリスク許容度に基づき、特定されたリスクに対処し、文書化するために活用することができる。組織は、組織的、ドメイン的、分野的、部門的または専門的な要件によって確立された、リスク基準、許容範囲、対応に関する既存の規制やガイドラインに従うことができる。そのようなガイダンスの代わりに、組織は、許容されたモデル・リスクマネジメント、企業リスクマネジメント、情報共有と開示プラクティスなどの戦略に基づくリスク対応計画を策定することができる。

推奨されるアクション

- 組織内のリスク許容値を適用するための規制および確立された組織、部門、分野または専門的な基準および要件を遵守する
- トラストワージネスの特性に関連する AI システムリスクに対処するための手順を文書化する。
- 物理的安全性、法的責任、規制遵守、個人・集団・社会へのネガティブインパクトを含むリスクを優先する
- リスク対応計画、対応機能を実行するためのリソースと組織チームを決定する。
- 関連する AI アクターと適切な人員がアクセスできる、整理されたセキュアなリポジトリにリスクマネジメントとシステムの文書を保管する

透明性と文書化

組織は以下を文書化することができる。

- AI システムが関連する法律、規制、基準、ガイダンスに準拠していることを確認するために、システムの見直しが行われたか？
- エンティティは、法規制における最低限必要な事項など、規制環境をどの程度定義し、文書化しているか？
- 組織は、特定された AI ソリューションの導入に伴うリスクに対処するために、リスクマネジメントシステムを導入したか（例えば人的リスクや商業目的の変更）？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Datasheets for Datasets.

参考文献

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021).

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021).

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022).

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021.

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695.

Manage 1.4

AI システムの取得者とエンドユーザの双方に対するネガティブな残留リスク（軽減されていないリスクの総和として定義）は文書化される。

概要

組織は、Map、Manage 1.3 および 2.1 から文書化されたリスクの一部を受け入れるかまたは移転することを選択することができる。このようなリスクは残留リスクと呼ばれ、システムの調達や利用に携わる者など、下流の AI アクターにインパクトを及ぼす可能性がある。透明性のあるモニタリングと残存リスクの管理により、費用対効果分析と、AI システムの潜在的価値と潜在的ネガティブインパクトの検討が可能になる。

推奨されるアクション

- リスク対応計画において残留リスクを文書化する。残留リスクには許容されたリスク、移転されたリスクおよび最小限のリスク軽減措置の対象となったリスクが含まれる
- 下流の AI アクターに残存リスクを開示する手順を確立する
- Map 3.4 で特定された、安全運用のための要件、既知の制限および推奨される警告ラベルについて、関連する下流の AI アクターに知らせる

透明性と文書化

組織は以下を文書化することができる。

- AI システムの設計、開発、デプロイ、評価、モニタリングに携わる人員の役割、責任、権限の委任は何か？
- デプロイされた AI の保守、再検証、モニタリング、更新は誰が行うのか？
- 更新／改訂はどのように文書化され、伝達されるのか。誰が、どのくらいの頻度で行うのか？
- 社外のステークホルダが入手可能な情報に、どれだけ容易にアクセスでき、最新であるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・Datasheets for Datasets.

参考文献

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021).

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021).

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022).

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021.

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695.

Manage 2

AI のベネフィットを最大化し、ネガティブインパクトを最小化するための戦略は、関連する AI アクターからの情報に基づいて、計画、準備、実施、文書化される。

Manage 2.1

AI リスクをマネジメントするために必要なリソースは、潜在的なインパクトの大きさや可能性を低減するために、AI 以外の代替可能なシステム、アプローチ、方法とともに考慮される。

概要

組織的リスクへの対応には、代替的なアプローチ、方法、プロセスまたはシステムを特定し、分析し、トラストワージネスの特性とそれらが組織の原則および社会的価値とどのように関連するかとの間のトレードオフのバランスをとることが必要となる場合がある。このようなトレードオフの分析は、学際的な組織チームや独立した専門家、個人やコミュニティ・グループとの協議を経て行われる。これらのプロセスには十分なリソース配分が必要である。

推奨されるアクション

- 確立された組織のリスク許容範囲に従って、リスクマネジメントのプラクティスを計画し、実施する
- リスクマネジメントチームが機能を遂行するためのリソースを確保されていることを確認する
- AI 機能に関して、自動化されていない方法、半自動化された方法、その他の手続き的な代替方法を検討するプロセスを確立すること
- AI チームのために、AI システムの透明性メカニズムを強化する
- AI チームによる AI システムの限界の探求を可能にする
- 過去に失敗した設計や、既知の失敗モードを回避するためのネガティブインパクトや結果を特定、評価、カタログ化する
- システムにおけるリスクマネジメントのための資源配分アプローチを特定する
- （システムが）ハイリスクとみなされる
- （システムが）自己更新（適応学習、オンライン学習、強化自己教師あり学習など）する
- （システムに）不確実性が高い場合や、リスクマネジメントが不十分な場合
- 組織の多様性（例えば、人口動態、分野）を補完するために、外部の専門知識や視点を定期的に探し、統合する
- 公平性、インクルージョン、アクセシビリティが内部的に不足している
- システム設計やデプロイの決定に関連する、AI アクターや社内外のステークホルダ間での定期的でオープンなコミュニケーションとフィードバックを可能にし、奨励する
- 継続的なモニタリングとフィードバック・メカニズムの計画を作成し、文書化する

透明性と文書化

組織は以下を文書化することができる。

- 社内のチームがリスクマネジメント機能を効果的に遂行するための権限とリソースを与えられているかどうかを評価する仕組みが整備されているか？
- ユーザやその他のステークホルダのエンゲージメントをリスクマネジメントプロセスにどのように組み込むのか？

AI の透明性に関するリソース

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・Datasheets for Datasets.

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

参考文献

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

David Wright. 2013. Making Privacy Impact Assessments More Effective. The Information Society, 29 (Oct 2013), 307-315.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA,220–229.

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. 2021. Datasheets for Datasets. arXiv:1803.09010.

Manage 2.2

デプロイされた AI システムの価値を維持するためのメカニズムが整備され、適用されている。

概要

AI システムがデプロイされ運用が開始されると、システムの性能とトラストワージネスは時間とともに進化し、変化する可能性がある。この現象は一般に「ドリフト」と呼ばれる。ドリフトは、組織にとって AI システムの価値を低下させ、ネガティブインパクトを及ぼす可能性を高める。AI システムのパフォーマンスとトラストワージネスを定期的にモニタリングすることで、組織がドリフトを検知して対応する能力を高め、導入後の AI システムの価値を維持することができる。定期的なモニタリングのためのプロセスと仕組みは、システムの機能性と挙動だけでなく、インパクトや、特定の使用状況における価値観や規範との整合性にも対処する。

例えば、個人または公共の安全やプライバシーへのインパクトに関する配慮には、自律走行車の運行時の高速走行の制限や、未成年者への違法コンテンツの推奨制限などが含まれる。

定期的なモニタリング活動により、組織は、出現したリスクを体系的かつプロアクティブに特定し、確立されたプロトコルと測定基準に従って対応することができる。組織の対応の選択肢には、1）リスクの回避、2）リスクの許容、3）リスクの軽減、4）リスクの移転がある。これらの対応にはそれぞれ計画とリソースが必要である。組織は、トラストワージネスの特性、デプロイメントのコンテクスト、現実世界へのインパクトを考慮したリスクマネジメン手順を確立することが推奨される。

推奨されるアクション

- トラストワージネスの特性を考慮したリスクコントロールを確立する。以下を含む
- 組織のデータガバナンスポリシーの一環としてのデータマネジメント、品質、プライバシー（例えば、最小化、修正、削除要求等）管理
- 機械学習とエンドポイントセキュリティ対策（例えば、ロバストモデル、差分プライバシー、認証、スロットリング等）
- 特定のコンテクスト内で AI システムの出力を補強、制限または制限するビジネスルール
- トラストワージネスの特性を考慮したリスクコントロールを確立する。以下を含む
- 組織のデータガバナンスポリシーの一環としてのデータ管理、品質、プライバシー（例えば、最小化、修正、削除要求等）管理
- 機械学習とエンドポイントセキュリティ対策（例えば、ロバストモデル、差分プライバシー、認証、スロットリング等）
- 特定のコンテクスト内で AI システムの出力を補強、制限または制限するビジネスルール
- AI のライフサイクルを通じた継続的な改善と TEVV のために、デプロイしたコンテクストに関連するドメインの専門知識を活用する
- 人間と AI のチーム編成の開発と定期的な追跡
- モデル評価とテスト、評価、検証、妥当性確認（TEVV）プロトコル

- 標準化された文書と透明性メカニズムの使用
- AI のライフサイクルにおけるソフトウェア品質保証の実践
- システムの限界を探り、過去の失敗設計やデプロイを回避するメカニズム
- システムのエンドユーザやインパクトを受ける可能性のあるグループからのフィードバックを収集するメカニズムを確立する
- 保険証書、保証書、契約書を見直し、リスク移転手続きに関する法的またはモニタリング要件を確認する
- リスク許容度の決定とリスク許容手順を文書化する

透明性と文書化

組織は以下を文書化することができる。

- AI システムのアウトプットによってインパクトを受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- AI システムが人々に危害やネガティブインパクトを及ぼす可能性はあるのか？
- 損害の可能性を軽減または低減するために何が行われたか？敵対勢力が AI を混乱させようとしたり、運用やビジネス環境が変化したりした場合、アカウンタビリティ担当者はどのように精度や精度の変化に対処するのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

安全性、妥当性、信頼性リスクマネジメントのアプローチとリソース

AI Incident Database. 2022. AI Incident Database.

AIAAIC Repository. 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2020. Under specification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395.

Andrew L. Beam, Arjun K. Manrai, Marzyeh Ghassemi. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. Jama 323, 4 (January 6,2020), 305-306.

Anthony M. Barrett, Dan Hendrycks, Jessica Newman et al. 2022. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. arXiv:2206.08966.

Debugging Machine Learning Models, In Proceedings of ICLR 2019 Workshop, May 6,2019, New Orleans, Louisiana.

Jessie J. Smith, Saleema Amershi, Solon Barocas, et al. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research.arXiv:2205.08363.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, et al. 2020. Improving Reproducibility in Machine Learning Research (A Report from the Neur IPS 2019Reproducibility Program) arXiv:2003.12206.

Kirstie Whitaker. 2017. Showing your working: a how to guide to reproducible research. (August 2017).

Netflix. Chaos Monkey.

Peter Henderson, Riashat Islam, Philip Bachman, et al. 2018. Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence.32, 1 (Apr. 2018).

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204.

Kang, Daniel, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model assertions for monitoring and improving ML models." Proceedings of Machine Learning and Systems2 (2020): 481-496.

リスク・バイアスの管理

National Institute of Standards and Technology (NIST), Reva Schwartz, Apostol Vassilev, et al. 2022. NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.

バイアステストと改善アプローチ

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, et al. 2018. A Reductions Approach to Fair Classification. arXiv:1803.02453.

Brian Hu Zhang, Blake Lemoine, Margaret Mitchell. 2018. Mitigating Unwanted Biaseswith Adversarial Learning. arXiv:1801.07593.

Drago Plečko, Nicolas Bennett, Nicolai Meinshausen. 2021. Fairadapt: Causal Reasoning for Fair Data Pre-processing. arXiv:2110.10200.

Faisal Kamiran, Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems 33 (2012),1–33.

Faisal Kamiran; Asim Karim; Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In Proceedings of the 2012 IEEE 12thInternational Conference on Data Mining, December 10-13, 2012, Brussels, Belgium. IEEE, 924-929.

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, et al. 2017.Optimized Data Pre-Processing for Discrimination Prevention. arXiv:1704.03354.

Geoff Pleiss, Manish Raghavan, Felix Wu, et al. 2017. On Fairness and Calibration.arXiv:1709.02012.

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, et al. 2020. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. arXiv:1806.06055.

Michael Feldman, Sorelle Friedler, John Moeller, et al. 2014. Certifying and Removing Disparate Impact. arXiv:1412.3756.

Michael Kearns, Seth Neel, Aaron Roth, et al. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. arXiv:1711.05144.

Michael Kearns, Seth Neel, Aaron Roth, et al. 2018. An Empirical Study of Rich Subgroup Fairness for Machine Learning. arXiv:1808.08166.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016, Barcelona, Spain.

Rich Zemel, Yu Wu, Kevin Swersky, et al. 2013. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning 2013, PMLR28, 3, 325-333.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh & Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Peter A. Flach, Tijl De Bie,Nello Cristianini (eds) Machine Learning and Knowledge Discovery in Databases.

European Conference ECML PKDD 2012, Proceedings Part II, September 24-28,2012, Bristol, UK. Lecture Notes in Computer Science 7524. Springer, Berlin, Heidelberg.

セキュリティとレジリエンスのリソース

FTC Start With Security Guidelines. 2015.

Gary McGraw et al. 2022. BIML Interactive Machine Learning Risk Framework. Berryville Institute for Machine Learning.

Ilia Shumailov, Yiren Zhao, Daniel Bates, et al. 2021. Sponge Examples: Energy-Latency Attacks on Neural Networks. arXiv:2006.03463.

Marco Barreno, Blaine Nelson, Anthony D. Joseph, et al. 2010. The Security ofMachine Learning. Machine Learning 81 (2010), 121-148.

Matt Fredrikson, Somesh Jha, Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the22nd ACM SIGSAC Conference on Computer and Communications Security (CCS'15), October 2015. Association for Computing Machinery, New York, NY, USA, 1322–1333.

National Institute for Standards and Technology (NIST). 2022. Cybersecurity Framework.

Nicolas Papernot. 2018. A Marauder's Map of Security and Privacy in Machine Learning. arXiv:1811.01134.

Reza Shokri, Marco Stronati, Congzheng Song, et al. 2017. Membership Inference Attacks against Machine Learning Models. arXiv:1610.05820.

Adversarial Threat Matrix (MITRE). 2021.

解釈可能性と説明可能性のアプローチ

Chaofan Chen, Oscar Li, Chaofan Tao, et al. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. arXiv:1806.10574.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. arXiv:1811.10154.

Daniel W. Apley, Jingyu Zhu. 2019. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv:1612.08468.

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, et al. 2021.Manipulating and Measuring Model Interpretability. arXiv:1802.07810.

Hongyu Yang, Cynthia Rudin, Margo Seltzer. 2017. Scalable Bayesian Rule Lists.arXiv:1602.08610.

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, et al. 2021. Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8312. National Institute of Standards and Technology, Gaithersburg, MD.

Scott Lundberg, Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874.

Susanne Gaube, Harini Suresh, Martina Raue, et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digital Medicine 4, Article 31 (2021).

Yin Lou, Rich Caruana, Johannes Gehrke, et al. 2013. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), August 2013.Association for Computing Machinery, New York, NY, USA, 623–631.

プライバシー・リソース

National Institute for Standards and Technology (NIST). 2022. Privacy Framework.

データガバナンス

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, Tomasz Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly, Volume 37, Issue 3, 2020, 101493, ISSN 0740-624X.

ソフトウェア・リソース

・PiML (explainable models, performance assessment)
・Interpret (explainable models)
・Iml (explainable models)
・Drifter library (performance assessment)
・Manifold library (performance assessment)
・SALib library (performance assessment)
・What-If Tool (performance assessment)
・MLextend (performance assessment)
・AI Fairness 360:
・Python (bias testing and mitigation)
・R (bias testing and mitigation)
・Adversarial-robustness-toolbox (ML security)
・Robustness (ML security)
・tensorflow/privacy (ML security)

・NIST De-identification Tools (Privacy and ML security)
・Dvc (MLops, deployment)
・Gigantum (MLops, deployment)
・Mlflow (MLops, deployment)
・Mlmd (MLops, deployment)
・Modeldb (MLops, deployment)

Manage 2.3

これまで知られていなかったリスクが特定された場合、そのリスクに対応し、回復するための手順が守られている。

概要

AI システムは、他のテクノロジーと同様に、機能しなかったり、故障したり、予期せぬ異常な挙動を示すことがある。また、攻撃やインシデント、その他の誤用や悪用が発生する可能性もあり、その原因は必ずしも事前に判明しているわけではない。組織は、以前には特定されていなかったリスクを認識し、それに対抗し、軽減し、管理するための処置手順を確立し、文書化し、伝達し、維持することができる。

推奨されるアクション

- AI システムのパフォーマンス、トラストワージネス、コンテクストの規範や価値観との整合性を継続的にモニタリングするためのプロトコル、リソース、メトリクスが整備されている
- インシデント、ネガティブインパクト、アウトカムに対する処置および対応計画を策定し、定期的に見直す
- システムコンポーネントのドリフト、脱コンテクスト化、その他の AI システムの動作要因を定期的にモニタリングする手順を確立し、維持する
- ネガティブインパクトに関するフィードバックを収集するための手順を確立し、維持する
- ミッションクリティカルな AI システムに関連するネガティブインパクトを処理し、システムを停止するための不測の事態への対応プロセスを検証する
- 関連する AI アクターのグループによる、AI システムの制限の予防的および事後的な調査を可能にする
- リスク許容範囲を超えたシステムを廃止する

透明性と文書化

組織は以下を文書化することができる。

- デプロイされた AI の保守、再検証、モニタリング、更新は誰が行うのか？
- 様々な AI ガバナンス・プロセスに関与する人員の責任（AI システムを廃止する責任も含む）は明確に定義されているか？
- データの生成、取得／収集、取り込み、ステージング／保存、変換、セキュリティ、メンテナンスおよび配信のためにどのようなプロセスが存在するか？
- AI の導入後、精度などの適切なパフォーマンスメトリックはどのようにモニターされるのか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF - Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations.

参考文献

AI Incident Database. 2022. AI Incident Database.

AIAAIC Repository. 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged.

Andrew Burt and Patrick Hall. 2018. What to Do When AI Fails. O'Reilly Media, Inc. (May 18, 2020). Retrieved October 17, 2022.

National Institute for Standards and Technology (NIST). 2022. Cybersecurity Framework.

SANS Institute. 2022. Security Consensus Operational Readiness Evaluation (SCORE) Security Checklist [or Advanced Persistent Threat (APT) Handling Checklist].

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204.

Manage 2.4

意図した使用方法と整合しないパフォーマンスや結果を示す AI システムを停止、解除、無効化するための仕組みが整備され、適用され、責任が割り当てられ、理解されている。

概要

意図した活用方法と整合しないパフォーマンスが、常にリスクを増大させたり、ネガティブインパクトをもたらせたりするとは限らない。厳格な TEVV の実施は、意図した活用に関わらず、ネガティブインパクトから保護するために有用である。ネガティブなインパクトが発生した場合、モデル、AI システムコンポーネントまたは AI システム全体のスーパーシング（バイパス）、切り離し、非アクティブ化／デコミッショニングが必要な場合がある。

- システムが寿命を迎える
- 検出または特定されたリスクが許容閾値を超える
- 適切なシステム緩和措置が組織の能力を超える
- 実現可能なシステム緩和措置が、規制、法律、基準と合致していない
- 継続的なモニタリング中に差し迫ったリスクが検出され、そのリスクに対して実行可能なリスク低減策が適時に特定または実施できない場合

このようなシナリオのもとで、AI システムを一時的または永続的に安全に運用から外すには、運用の中断と下流へのネガティブインパクトを最小限に抑える標準的なプロトコルが必要である。

プロトコルは、確立されたシステムガバナンスポリシー（Govern1.7 を参照）、規制遵守、法的枠組み、ビジネス要件、使用するアプリケーションのコンテクスト内の規範や基準に整合して開発される冗長システムやバックアップシステムを含むことができる。システムコンポーネントをバイパスまたは非アクティブ化するためのアクションの判断閾値とメトリックは、継続的なモニタリング手順の一部である。バイパス／非活性化の決定が下されたインシデントについては、根本原因、インパクト、緩和と再デプロイの潜在的な機会を理解するための文書化とレビューが必要である。組織は、一時的および／または恒久的な廃止の上流および下流におけるインパクトを考慮し、予測すること、併せて不測の事態に備えた選択肢を提供するためのリスクおよび変更マネジメントプロトコルを開発することが奨励される。

推奨されるアクション

- AI システムのバイパス処置のための確立された手順を定期的に見直す、運用および／またはビジネス機能の継続性を確保するための冗長システムまたはバックアップシステムの計画を含む
- バイパスを作動させるまたは非アクティブ化対応するためのシステムインシデントのしきい値を定期的に見直す
- 変更マネジメントプロセスを適用し、AI システムまたは AI システムのコンポーネントをバイパスまたは停止した場合の上流および下流へのインパクトを理解する

- 優先順位を決定するためのプロトコル、リソース、メトリクスを適用する、AI システムまたは AI システムコンポーネントをバイパスまたは無効化する
- フォレンジック、規制および法的のレビューのために資料を保存する
- 社内で根本原因分析を行い、バイパスまたは非アクティブ化イベントのプロセスのレビューを行う
- 再デプロメールの基準を満たすために更新できないシステムコンポーネントは廃止し、保全する
- トラストワージ特性を考慮し、更新されたシステムコンポーネントを再デプロイするための基準を確立する

透明性と文書化

組織は以下を文書化することができる。

- AI システムの設計、開発、デプロイ、評価、モニタリングに携わる人員の役割、責任、権限の委任は何か？
- 組織は、特定された AI ソリューションの導入に伴うリスクに対処するために、リスクマネジメントシステムを導入したか（例えば、人的リスクや商業目的の変更）？
- エンティティは、（もしあれば）エラーや限界を特定するために、AI システムに対してどのようなテスト（すなわち、敵対的テストやストレステスト等）を実施したか？
- エンティティは、AI システムが不要になった場合、そのシステムを終了するための手順をどの程度確立しているか？
- エンティティは、システムの拡大、継続、廃止を決定するために、アセスメントや評価をどのように活用したか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

参考文献

Decommissioning Template. Application Lifecycle And Supporting Docs. Cloud and Infrastructure Community of Practice.

Develop a Decommission Plan. M3 Playbook. Office of Shared Services and Solutions and Performance Improvement. General Services Administration.

Manage 3

サードパーティのエンティティによる AI リスクとベネフィットはマネジメントされる。

Manage 3.1

サードパーティリソースからもたらされる AI リスクとベネフィットは定期的にモニタリングされ、リスクコントロールが適用され、文書化される。

概要

AI システムは、サードパーティのデータ、ソフトウェア、ハードウェアシステムなど、外部のリソースや関連プロセスに依存する場合がある。サードパーティが、AI システムの設計、開発、デプロイ、利用のためのツール、ソフトウェア、専門知識を含むコンポーネントやサービスを組織に提供することで、効率性と拡張性を向上させることができる。しかしまた、複雑性や不透明性を増大させ、ひいてはリスクを増大させる可能性もある。サードパーティの技術、人員、リソースを文書化することは、リスクマネジメントに役立つ。何よりもまず、物理的な安全性、法的責任、企業コンプライアンス、そして、個人、グループ、あるいは社会へのネガティブインパクトを含むリスクに焦点を当てることが推奨される。

推奨されるアクション

- 法的要件に対処したか？
- サードパーティの AI システムに組織のリスク許容度を適用する
- サードパーティの AI 技術、人員、その他のリソースに対して、組織のリスクマネジメント計画およびプラクティスを適用し、文書化する
- サードパーティの AI システムおよびコンポーネントのドキュメントを特定し、維持する
- 独自のアルゴリズムを公開することなく、透明性のニーズに対応するサードパーティの AI システムのテスト、評価、妥当性確認、検証プロセスを確立する
- 一貫性のないソフトウェアのリリーススケジュール、不十分な文書、不完全なソフトウェア変更マネジメント管理（例えば、上位互換性または下位互換性の欠如等）などの、サードパーティのシステムまたはコンポーネントにおける有益な活用とリスク指標を特定するプロセスを確立する。
- 組織は、供給されたリソースの既知および潜在的な脆弱性、リスクまたはバイアスをサードパーティが、報告するためのプロセスを確立することができるミッションクリティカルなサードパーティの AI システムに関連するネガティブインパクトを処理するための不測の事態への対応プロセスを検証する
- サードパーティの AI システムをモニタリングし、トラストワージネスの特性に関連する潜在的なネガティブインパクトやリスクをモニタリングする
- リスク許容範囲を超えるサードパーティのシステムを廃止する

透明性と文書化

組織は以下を文書化することができる。

- サードパーティが AI システムやその構成要素の一部を作成した場合、説明可能性や解釈可能性のレベルをどのように確保するか？文書化しているか？
- 組織がサードパーティからデータセットを入手した場合、そのようなデータセットを使用するリスクを評価し、管理したか？
- サードパーティ（例えば、サプライヤ、エンドユーザ、被験者、ディストリビュータ／ベンダ、労働者など）が AI システムの潜在的な脆弱性、リスク、バイアスを報告するプロセスを確立したか？
- 法的要件に対処したか？

AI の透明性に関するリソース

・Artificial Intelligence Ethics Framework for the Intelligence Community.

・WEF - Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations.

・Datasheets for Datasets.

参考文献

Office of the Comptroller of the Currency. 2021. Proposed Interagency Guidance on Third-Party Relationships: Risk Management. July 12, 2021.

Manage 3.2

AI システムの定期的なモニタリングとメンテナンスの一環として、開発に使用される事前学習されたモデルはモニタリングされる。

概要

AI 開発における一般的なアプローチは、既存の事前学習済みモデルを、異なるが関連するアプリケーションで使用するために適合させる転移学習である。開発タスクの AI アクターは、画像分類、言語予測、エンティティ認識などのタスクに、サードパーティエンティティが事前に学習したモデルを使用することが多い。事前学習されたモデルは、通常、非常に大規模なデータセットと計算集約的なリソースを使用して、様々な分類または予測問題に対処するために学習される。事前に学習されたモデルの活用は、システムのネガティブなアウトカムやインパクトを予測することを困難にする可能性がある。文書化または透明性ツールの欠如は、事前に学習されたモデルを導入デプロイする際の困難性と一般的な複雑性を増大させ、原因究明を妨げる。

推奨されるアクション

- リスク追跡のために、AI システムのインベントリの中から事前に学習されたモデルを特定する
- サードパーティのリスクの追跡の一環として、事前に学習されたモデルのパフォーマンスとトラストワージネスを独立的かつ継続的にモニタリングするプロセスを確立する
- サードパーティのリスクの追跡の一環として、事前に学習されたモデルに接続された AI システムコンポーネントのパフォーマンスとトラストワージネスをモニタリングする
- 組織のリスクマネジメント手順に従い、またサードパーティのリスク追跡の一環として、AI システムコンポーネントや事前に学習されたモデルから生じるリスクを特定し、文書化し、是正する
- サードパーティのリスクの追跡の一環として、リスク許容範囲を超える AI システムコンポーネントと事前学習済みモデルを廃止する

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、ソース、起源、変換、拡張、ラベル、依存関係、制約、メタデータを含む AI システムのデータの出所をどのように文書化しているか？
- このデータセットの収集／処理手順は、本データシートの最初のセクションに記載されているデータセット作成の動機を達成しているか？
- エンティティは、収集したデータが適切であることをどのように保証するのか、関連性があり、意図した目的に照らして過剰ではないか？
- データセットが古くなった場合、そのことをどのように伝えるのか？

AI の透明性に関するリソース

・Artificial Intelligence Ethics Framework For The Intelligence Community.

・WEF - Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations.

・Datasheets for Datasets.

参考文献

Larysa Visengeriyeva et al. “Awesome MLOps, “GitHub. Accessed January 9, 2023.

Manage 4

特定および測定された AI リスクに対する対応と回復、コミュニケーション計画を含むリスク処置は文書化され、定期的にモニタリングされる。

Manage 4.1

デプロイ後の AI システムのモニタリング計画には、ユーザやその他の関連する AI アクターからのインプットを収集・評価、不服申し立てと無効化、廃止措置、インシデント対応、復旧、変更マネジメントなどのメカニズムが含まれる。

概要

AI システムのパフォーマンスとトラストワージネスは、さまざまな要因によって変化する可能性がある。AI システムの定期的なモニタリングは、デプロイヤがパフォーマンスの低下、敵対的攻撃、予期せぬ異常な行動、ヒヤリハット、インパクトを特定するのに役立つ。AI システムのパフォーマンスに関する導入前後の外部からのフィードバックを含めることで、ポジティブおよびネガティブなインパクトに関する組織の認識を高め、リスクや被害に対応する時間を短縮することができる。

推奨されるアクション

- トラストワージネスの特性に関連するリスクやネガティブおよびポジティブなインパクトについて，AI システムのパフォーマンスをモニタリングする手順を確立し、維持する
- AI システムの妥当性と信頼性、バイアスと公平性、プライバシー、セキュリティとレジリエンスを評価するために、デプロイ後の TEVV タスクを実行する
- デプロイ前に AI システムのトラストワージネスをデプロイの活用コンテクストと同様の条件で AI システムのトラストワージネスを評価する
- 所定の頻度でレッドチーム演習を確立・実施し、その効果を評価する
- データ削除や修正要求など、データセットの変更を追跡する手順を確立する
- システムのパフォーマンス、トラストワージネス、インパクトに関する情報を収集するために、関連する AI アクターと社内外のステークホルダとの間で定期的にコミュニケーションを行い、フィードバックする仕組みを確立する
- エラー、ヒヤリハット、攻撃パターンに関する情報を、インシデントデータベース、同様のシステムを持つ他の組織、システムのユーザやステークホルダ内で共有する
- AI システムのパフォーマンスとトラストワージネスにおいて、検出または報告されたネガティブインパクトや問題に対応し、文書化する
- 確立されたリスク許容範囲を超えるシステムを廃止する

透明性と文書化

組織は以下を文書化することができる。

- エンティティは、導入後の AI システムのテスト方法、メトリクス、パフォーマンスアウトカムをどの程度文書化しているか？
- 組織外のステークホルダが入手可能な情報は、どれだけ容易にアクセス可能で、最新であるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Datasheets for Datasets.

参考文献

Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing." Information 11, no. 3 (2020): 137.

Manage 4.2

継続的な改善のための測定可能な活動は、AI システムの更新に組み込まれ、関係する AI アクターを含むステークホルダとの定期的な関与が含まれる。

概要

定期的なモニタリング・プロセスにより、規制や法的枠組み、組織やコンテクストの価値観や規範に従って、パフォーマンスと機能性を高めるためのシステム更新が可能になる。これらのプロセスはまた、根本原因、システムのデグレード、ドリフト、ヒヤリハット、故障の分析、インシデント対応、文書化を促進する。

AI アクターには、ライフサイクル全体にわたって、システムパフォーマンス、制限、インパクトに関する外部からのフィードバックを取り込み、継続的な改善を実施する多くの機会がある。改善は、必ずしもパイプラインやシステム・プロセスをモデル化するものではなく、精度やその他の品質パフォーマンス測定基準を超えるメトリクスに基づく場合もある。このような場合、改善は、ビジネスまたは組織の手順やプラクティスへの適応を伴うかもしれない。組織は、開発者、エンドユーザ、監査人および関連する AI アクターのために、トレーサビリティと透明性を維持する改善を開発することが奨励される。

推奨されるアクション

- トラストワージネスの特性を、継続的な改善のために使用されるプロトコルやメトリクスに統合する
- フィードバックを評価し、AI システムの改善に反映させるプロセスを確立する
- 提案された改善策と関連する規制や法的枠組みとの整合性を評価・査定する
- 活用コンテクスト内の価値観や規範に関連する提案された改善案の整合性を評価・査定する
- トラストワージネス特性、システムリスク、システム機会の間のトレードオフに関する決定の根拠を文書化する

透明性と文書化

組織は以下を文書化することができる。

- ユーザやその他のステークホルダの関与は、モデル開発プロセスやデプロイ後の定期的なパフォーマンスのレビューにどのように組み込まれるのか？
- AI システムのアウトプットによってインパクトを受けるユーザや関係者は、どの程度 AI システムをテストし、フィードバックを提供することができるか？
- エンティティは、法規制における最低限必要な要件を含め、規制環境をどの程度定義し、文書化しているか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

・Artificial Intelligence Ethics Framework for the Intelligence Community.

参考文献

Yen, Po-Yin, et al. "Development and Evaluation of Socio-Technical Metrics to Inform HIT Adaptation."

Carayon, Pascale, and Megan E. Salwei. "Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout." Journal of the American Medical Informatics Association 28.5 (2021): 1026-1028.

Mishra, Deepa, et al. "Organizational capabilities that enable big data and predictive analytics diffusion and organizational performance: A resource-based perspective. "Management Decision (2018).

Manage 4.3

インシデントやエラーは、インパクトを受けるコミュニティを含む関連する AI アクターに伝達される。インシデントやエラーの追跡、対応、回復のプロセスが守られ、文書化される。

概要

特定され報告されたエラーについて、正確かつ透明性のある説明を定期的に文書化することで、AI のリスクマネジメント活動を強化することができる。例として以下が含まれる。

- エラーがどのように特定されたか
- エラーに関連したインシデント
- エラーが修復されたかどうか
- インパクトを受けるすべてのステークホルダとユーザに、どのように修復版を配布するか

推奨されるアクション

- エラー、インシデント、ネガティブインパクトに関する情報を、関連するステークホルダ、事業者、運用者、ユーザ、影響を受ける当事者と定期的に共有する手順を確立する
- 報告日、報告件数、インパクトと重大性の評価および対応などのネガティブインパクトのデータベースを維持する
- システムの変更点、変更理由、変更方法、テスト、デプロイの詳細に関するデータベースを管理する
- バージョン履歴情報とメタデータを管理し、継続的な改善プロセスを可能にする
- 複雑、緊急なリスクの特定を担当する AI アクターに適切なリソースと権限が与えられていることを確認する

透明性と文書化

組織は以下を文書化することができる。

- データの質、正確性、信頼性、代表性を高めるために、エンティティはどのような是正処置をとったか？
- エンティティは、ステークホルダのコミュニティに、AI の戦略的目標と目的をどの程度伝えているか。外部のステークホルダは、そのデータにどの程度容易にアクセスでき、最新の情報を入手できるか？
- AI システムの活用によってインパクトを受けるエンドユーザ、消費者、規制当局、個人を含む外部のステークホルダは、AI システムの設計、運用、制限に関してどのような情報にアクセスできるか？

AI の透明性に関するリソース

・GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.

参考文献

Wei, M., & Zhou, Z. (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. arXiv, abs/2206.07635.

McGregor, Sean. "Preventing repeated real world AI failures by cataloging incidents: The AI incident database." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 17. 2021.

Macrae, Carl. "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk." Risk analysis 42.9 (2022):1999-2025.