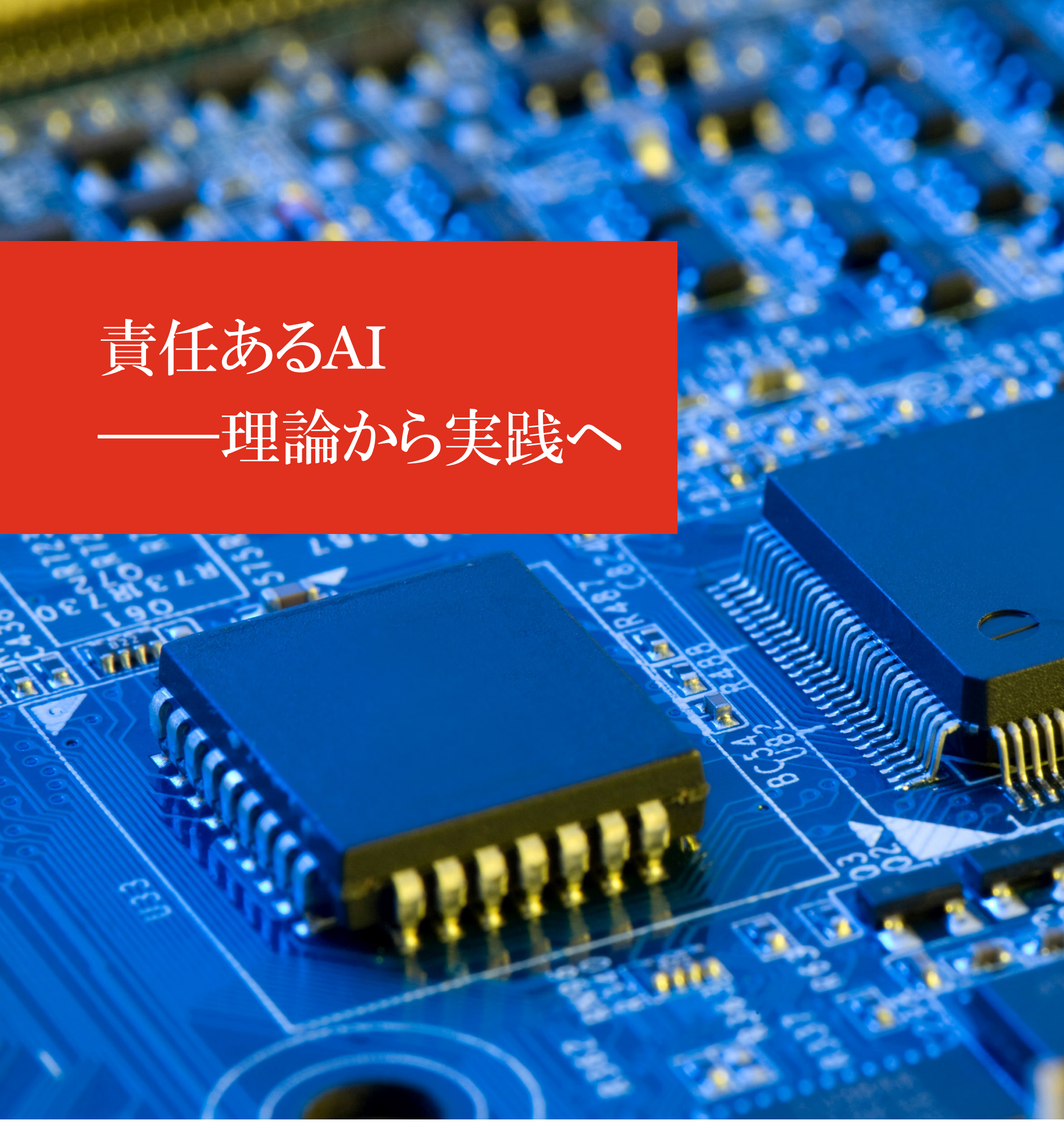


責任あるAI
——理論から実践へ

pwc

www.pwc.com/jp

はじめに

　世界中のあらゆる業界において、AIに関わるリスクの認識が高まるとともに、責任あるAIへの注目が集まっています。国内においても、IT、金融、ヘルスケア、自動車をはじめとしてAIの利活用が進んでいる業界において、AIに関わるリスクをどのように認識し、コントロールするかが喫緊の課題となっています。

　本稿では、PwCの調査で明らかになった、グローバル企業におけるAI倫理原則の実践状況や、AIに関するリスクの認識状況など、責任あるAIの実践状況について紹介します。

目次

責任あるAIの実践状況

AIは、人間の生産性や意思決定の向上に役立つものとして、業界を問わず不可欠になりつつありますが、そのメリットは、社会に及ぼす潜在的な悪影響を凌ぐものでしょうか。私たちはAIが潜在的に持つ破壊的な側面や、不慣れな活用、誤用、濫用によって生じる好ましくない影響も目にしてきました[1]。消費者やメディアも、偏った求人[2] や金融取引[3]、差別に関する懸念[4] などに注目しているため、AIの実装を巡るモラル・ジレンマに関心が高まっています[5]。開発者、製品のユーザー企業は、AIの実社会における問題や、特定された全ての道徳的影響に責任を持って対処するために、明確な指針と原則を必要としています。

PwCが行った広範かつ豊富な調査では、AIの倫理原則について共通性が認められました。調査では国際機関や企業の倫理原則を100個以上を取り上げ、これを9つのコアなAI倫理原則に集約しました。

図表1 – AI倫理原則

認識上の原則

解釈可能性（説明可能性、透明性）
AIシステムを、意思決定モデル全体および個別予測の根拠について、さまざまな利害関係者に説明可能にします。

信頼性／堅牢性／セキュリティ
AIシステムを、適切なモデルとデータセットを用いて長期にわたり、確実かつ安全に機能するように開発します。

一般的なAI倫理原則

説明責任
AIシステムの全ての利害関係者は、AIシステムの使用と誤用に関わる道徳的影響に対して責任を負います。また、個人または企業のいずれであれ、明確に特定可能な責任者が存在します。

データプライバシー
各個人は、AIシステムの学習や運用に用いられるデータを管理し、また、データがその他の目的にどのように再利用されるかを管理する権利を有します。

合法性／コンプライアンス
全ての利害関係者は、AIシステムの設計において、常に法や関連する全ての規制制度に従い行動します。

ベネフィシャルAI
AIの開発では、サステナビリティ、協調、開放性などの公益を促進し、反映します。

人間の介入
AIシステムの意思決定や運用に必要な人間の介入度合いは、認識される倫理リスクの重大性の基準に基づき決定されます。

安全性
AIシステムの運用期間において、人の物理的安全性や精神的完全性を損ないません。

公平性
AIの開発では、最終的にどの集団においても、個人が差別をうけることなく、かつ過程や結果として損害を受けることなく、公平に扱われます。さらに、AIはデータの背後に存在する個人を常に尊重し、差別的なバイアスが含まれるデータセットを使用しません。

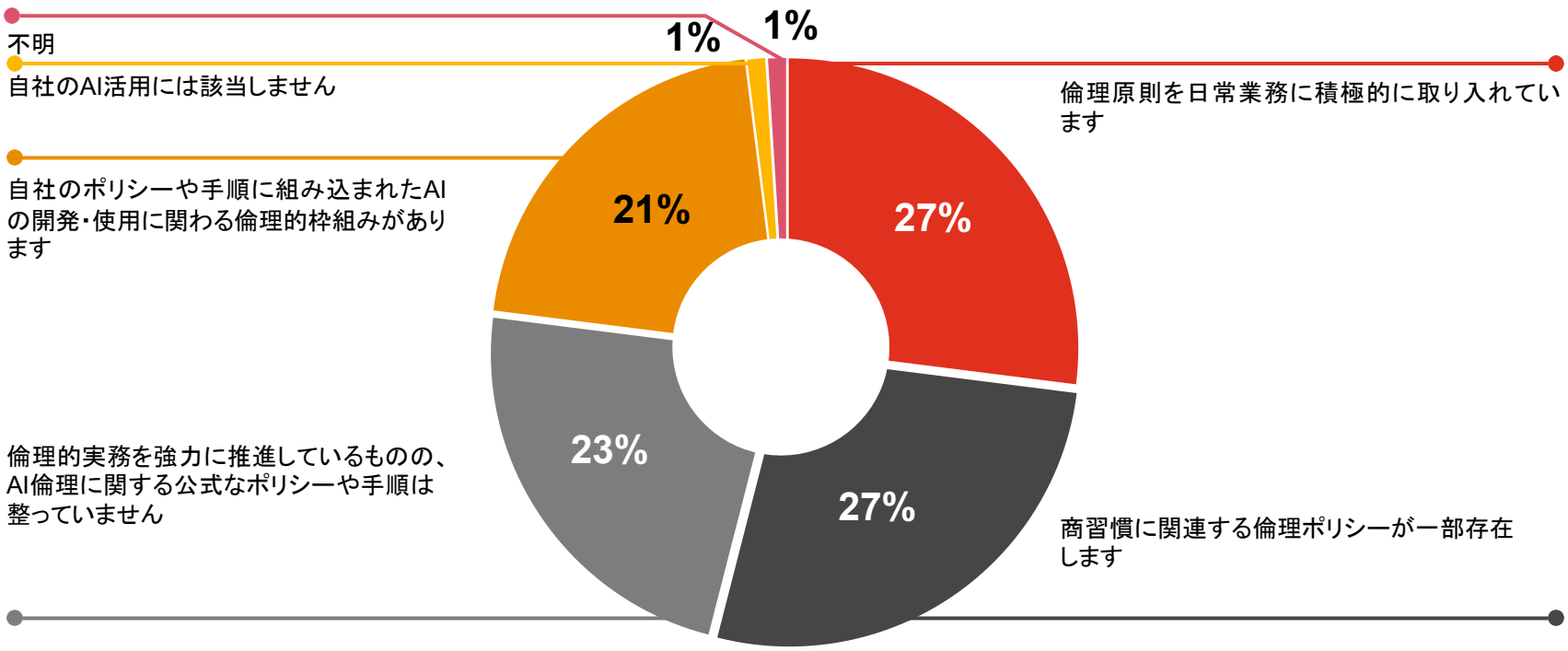
[1] https://doi.org/10.1007/s11023-018-9482-5
[2] https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias
[3] https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair
[4] https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/
[5] https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

「原則」というものは対外的に公表されることが何よりの第一歩ですが、社内においても実践されなければ意味がありません。倫理原則に関する期待が高いにもかかわらず、これを実践に移すための一貫したアプローチはほとんど存在しません。そこでPwCは、2020年に米国、英国、日本、インドの1,000人を超える企業幹部を対象に調査を行いました。

調査の結果、多くの企業がAI倫理原則の実践を試みていることが判明しました。調査対象企業の半数以上がAIの使用に起因する倫理問題に対処する正式なポリシーや原則を複数有しており、さらにその4分の1近くの企業が指針やポリシーを整備しています。これはPwCが調査した国々全体に見られる傾向です。5社に1社がAIの開発と使用に関する倫理的枠組みを整備しており、心強いことに、組織においてAIを全面的に容認している企業では、正式な倫理的枠組みを整備している割合がほぼ2倍に達します。

図表2 – AI倫理原則の実践状況

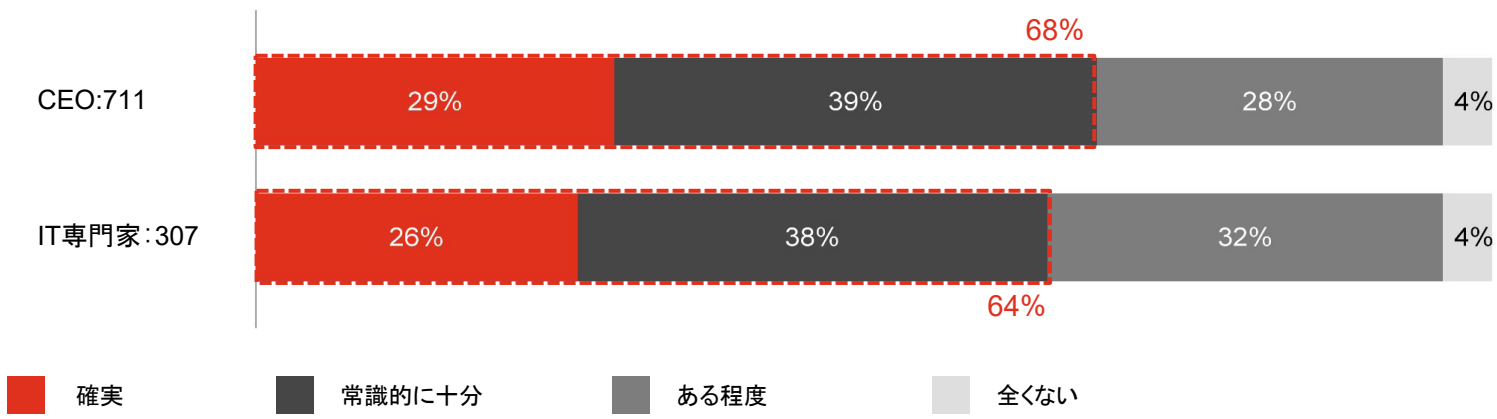
現在、AI の倫理原則についてどの程度実践ができていますか。



投資段階においては、経営陣やエンジニアが、AI倫理原則の実践に関して事前にかつ持続可能な状態でサポートすることが重要です。PwCの調査結果から、CEOの68%、IT専門家の64%が、倫理を考慮したAIへの投資にかなり自信を持っていることが分かりました。経営幹部層、非経営幹部層のいずれも、その大半がAI倫理原則は組織の価値観に合致すると確信しており、AIを全面的に推進している組織ではその比率がさらに高くなります。

図表3 – AI倫理原則の実践に関わる職掌別サポート

AI倫理原則の実践に関して、どの程度サポートができていると思いますか。



しかしながら、企業におけるAI倫理原則の実践は、まだ十分とはいえません。AI倫理原則を実践する際の最大の障壁として、一貫性のないアプローチが挙げられます。多くの場合、開始される新たな取り組み（AIの行動規範の策定から倫理委員会の設置、倫理的枠組みへの具体的な取り組みまで）は個別で検討されるため、効果的に機能する力が限られてしまいます。

行動規範と影響評価は、組織の規模にかかわらず、経営陣に人気のツールである一方、倫理研修の提供、倫理委員会の活用、その他の介入手段については、組織の規模やAI導入の成熟度に応じて利用状況が著しく異なります。実際、AI活用の成熟度が高い大企業は倫理委員会を設置、影響評価を実施、倫理研修を提供している傾向が極めて高く、これらが責任あるAIの実現に必要な要素であることが分かります。

あらゆる規模のセクターや組織において見られた前向きな流れに基づけば、AI倫理原則の実践には、AIで責任ある成果を上げるべくリソースを投資し、インセンティブを提供することが必要であると認識できます。さらに、倫理原則の実践は、AI導入の加速化や投資リターン（ROI）にも大きく貢献します。

図表4 – 倫理的なAI介入のエコシステム

取り組み	詳細
バリュー ステートメント	AIの確固たる倫理ビジョンを経営幹部層が主導して設定し、これにより、AIの成果が公平、透明、有益、安全、堅牢であることを表明します。
行動原則／行動規範	倫理原則を組織の価値観で定義し、その原則を運用できるように、組織のポリシー、行動規範、枠組みに適応させる必要があります。
社外の倫理委員会	倫理委員会は倫理的な意思決定の一部を構成し、これを通じて倫理的な問題の報告、緊張関係の管理、先例の提示を行うことができます。
倫理カルチャー	カルチャーは変化の核心であり、倫理に関わるスキル・知見・行為が、認識・報奨・評価される場です。倫理的な行為を促すため、インセンティブや報奨に関わる適切なスキームを整備します。
教育／研修	従業員に対して、倫理研修のプログラムやカリキュラムを採用する必要があります。コミュニティでの実践、イベント、読書会、チームディベート、ハッカソンなどがこれに含まれます。
報告／助言チャネル	従業員が倫理的ジレンマに関して助言を受けたり、AIやデータに関する違反を報告したりするための適切な手段や方法を備えます。これは、倫理の潜在的問題を特定し、その問題が拡大する前に解決する一助となります。
製品開発／設計	倫理的な意思決定や行動は、製品レベルで行われる必要があるため、開発プロセスを調整し、倫理的なチェックポイントをプロセスの各段階に組み込みます。これにより、原則を基準に置き換え、基準を設計要件やガバナンス要件に置き換えることができます。
定期的評価	AIの性能を公平性、安全性、信頼性に基づき評価し、関連領域に関わる社内外の基準の遵守状況を評価するため、定期的な監査が必要です。

結論

- AI倫理はAIシステムの目的、使用、デプロイを取り巻く倫理的ビジョンに関係するものですが、責任あるAIは、そのビジョンを実践的な指針に置き換えるために必要な学際的領域です。
- 倫理的なAIの枠組みは、道徳的かつ法的な説明責任のみならず、「公益」を目的とした「人間中心」のAI開発も支援するものとして、国際的な人権法[6] に整合していなければなりません。
- 行動規範、倫理委員会、倫理研修、影響評価など個々の取り組みにフォーカスした断片的なAI倫理アプローチではなく、さまざまな倫理的なAI介入を想定した体系的なアプローチが必要です（図表4参照）。

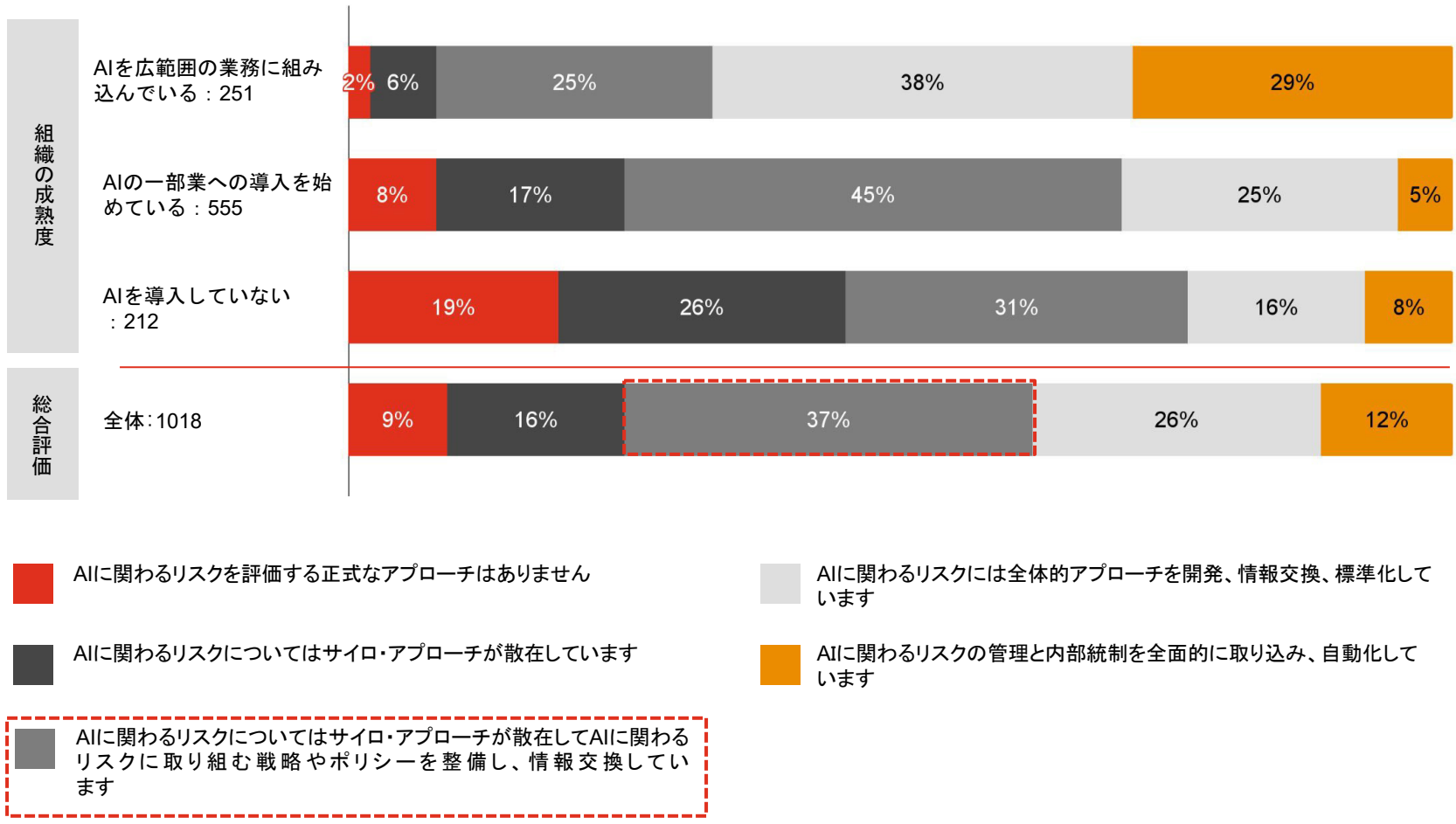
[6] https://tech.humanrights.gov.au/

AIに関わるリスクの特定と対応状況

AI倫理に関する評価は、AIに関わるリスクの認識と並行して高まっています。AIに関わるリスクの多くは倫理リスクでもあるため、当然のことといえます。PwCの調査では、組織が、AIに関わるリスクの軽減に全社的なアプローチを採用することを主な手段として、AIに関わるリスクを特定し、特定したリスクを説明する責任の優先順位を高めていることが示されました。実際、企業の3分の1以上（37%）がAIに関わるリスクに取り組む戦略やポリシーを整備（および情報交換）しており、以前に比べて大幅に増加しています。この他にも、企業の4分の1が、AIに関わるリスクについて情報交換のみならず標準化を含む全社的なアプローチを採用しています。

図表5 – AIに関わるリスク特定状況（組織の成熟度別）

現在、AI に関わるリスクをどのように特定していますか。



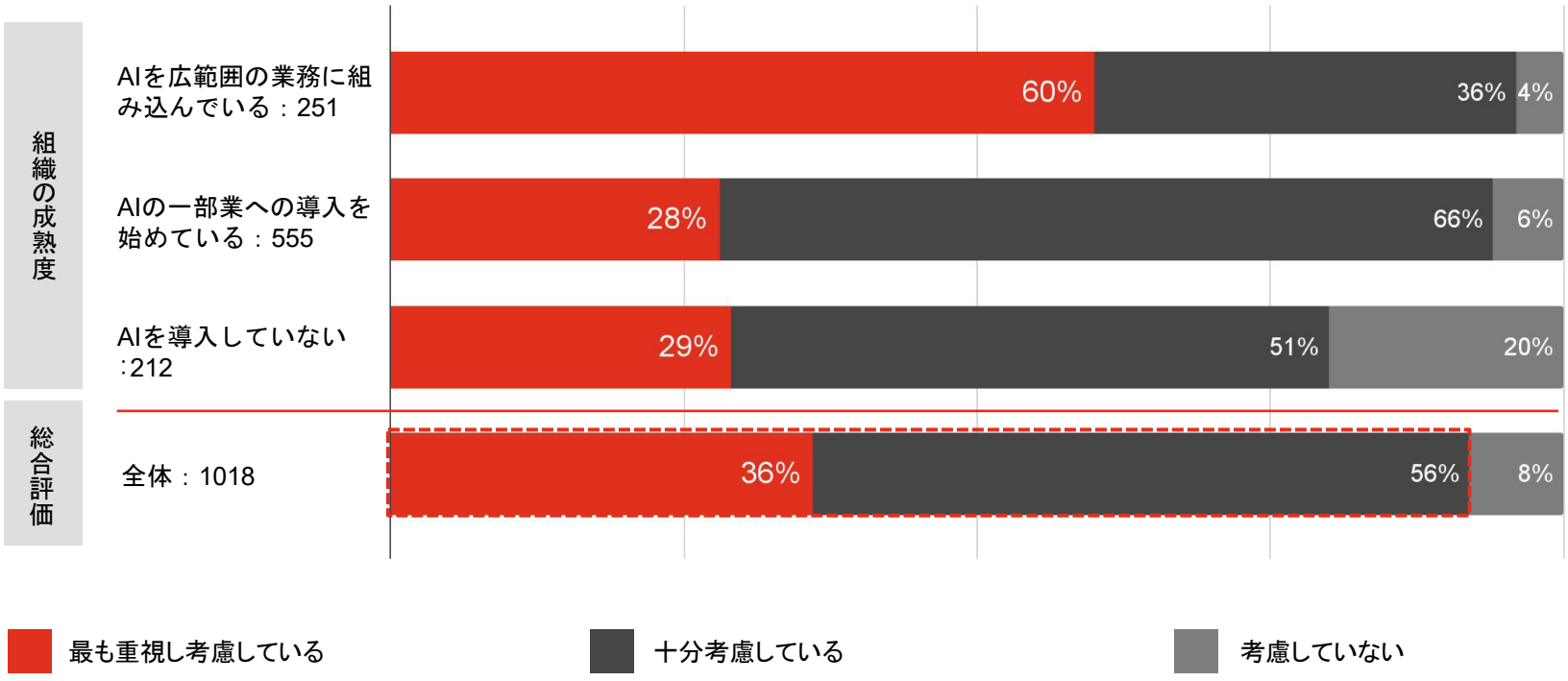
AIに関わるリスクの分類方法には、アプリケーションレベルのもの（性能リスク、統制上のリスク、セキュリティリスクなど）と広範なエコシステムレベルのもの（全社的リスク、社会的リスク、経済的リスクなど）が含まれます。重要性と可視性の高いリスクには性能に影響するものがありますが、これには、精度の低さ、低品質なデータ、バイアス、過学習、不十分なテスト手順に起因するエラーの存在が含まれます[7]。

[7] https://www.techuk.org/resource/with-great-power-comes-great-responsibility-the-importance-of-proactive-ai-risk-management.html

バイアスは多くの組織が最も懸念する事項ですが、これは、新たな規制の枠組みの出現や、メディアおよび消費者の差別への関心、結果として高まったレピュテーションリスクや、「適切な行動をしたい」という欲求にある程度起因しています。PwCの調査においても、回答者の36%がバイアスを主要なリスクフォーカスエリアとし、同56%がバイアスリスクに十分対処できると考えています。成熟企業は、AI開発とAIリスクに関する認識において、より具体的な経験を重ねるため、バイアスを最重要課題として受け止めています（AIを広範囲の業務に組み込んでいる企業の約60%）。また、成熟した組織は、公平性の原則を重視しています。

図表6 – バイアスの重視状況（組織の成熟度別）

過去12カ月間、AIソリューションに関してバイアス（性別、人種、民族性などにおいて、ある集団を別の集団より優遇するなど不公平を生じるシステム）を特に考慮しましたか。



PwCの調査では、バイアスのテーマに対する感度は国によって異なることも示しています。バイアスを最も重視すると表明した企業は、バイアスに関するオープンな議論を重要視し、かつ人口構成が人種的、民族的に複雑なインドや米国で高くなっています。

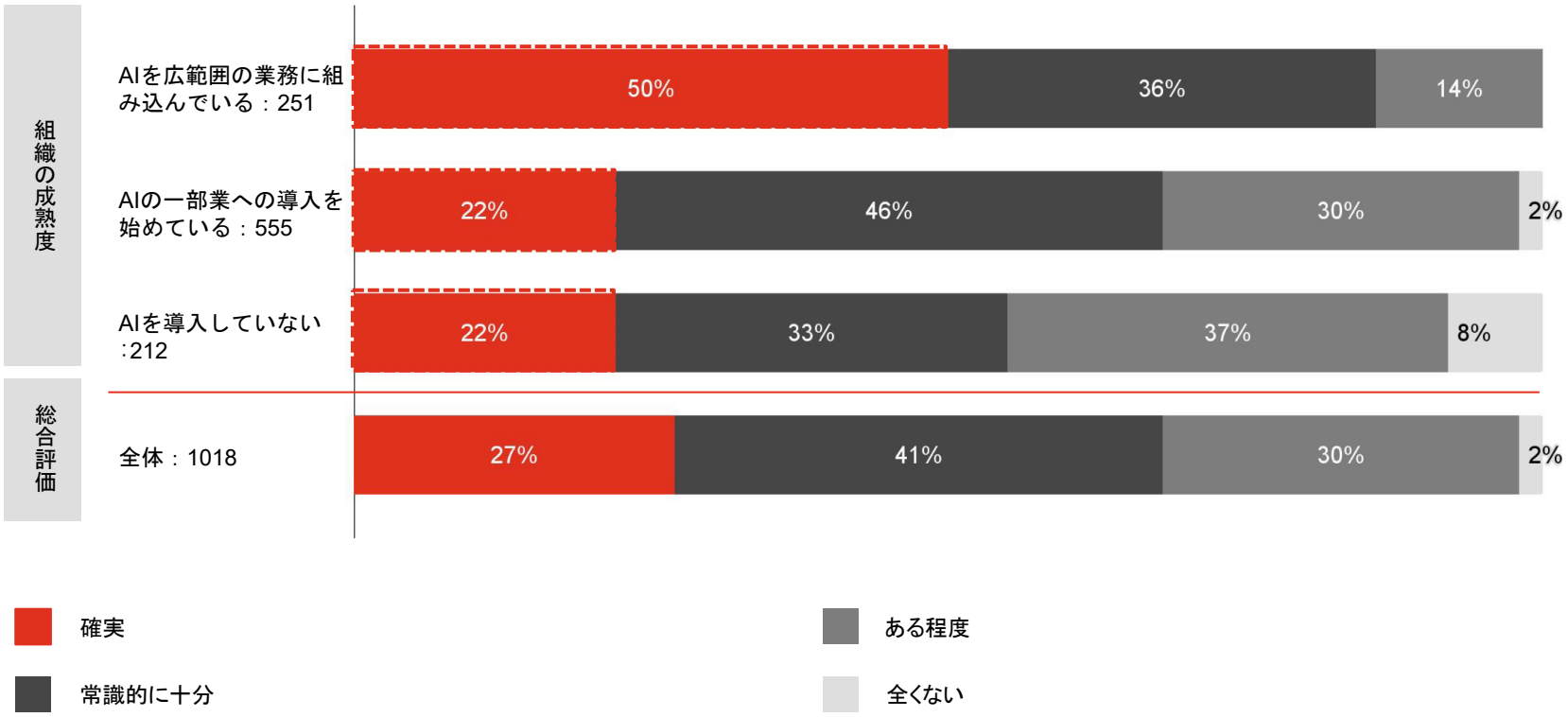
バイアスはしばしばAIの採用や、その責任ある使用を妨げる主な懸念事項として、システムがどのように決定を下すのかを理解できない不透明さとともに生じます。不透明性は、アルゴリズムや手法の複雑さが増すことから生じます。その結果、アプリケーションがどのように機能するか、その最も重要な特性や、それらの因果関係についての理解が不十分になる可能性があります。透明性の欠如と情報交換の不足は、消費者をミスリードし、組織の信用を脅かす点においてリスクを生じさせる可能性があります。

バイアスがもたらす課題への取り組みは、技術ツールや堅牢な処理、ときには性能と説明可能性のトレードオフを伴う複雑なプロセスとなります。これについては、今回の調査における回答者も同意しています。AIの意思決定を説明または正当化する能力を確実に有するとの回答が27%にすぎないのに対し、41%は合理的に説明できるとし、30%がある程度説明できると回答しました。

ただし、AI導入の成熟度による違いを分析すると、状況は一変します。AIを広範囲の業務に組み込んでいる企業の半数がその意思決定を確実に説明できるとしたものの、成熟度の低い組織においては5分の1にすぎません。業界別にみると、ヘルスケア企業の説明可能性成熟度は平均をはるかに上回っていますが、これは、ヘルスケア業界では、誤った意思決定がもたらす潜在的な損害のインパクトが高いためと考えられます。

図表7 – 説明可能性の重視状況（組織の成熟度別）

担当事業部門においてAI が行った意思決定について尋ねられた場合、その意思決定を説明または正当化することはできますか。



開発とテストを経たAIシステムの性能が、その目的に適していると判明した場合、次は長期的な性能の安定性を実現することが現実的な課題となります。今回の調査回答者も、そのことを明確にしています。「信頼性、堅牢性、セキュリティ」が、業界や成熟度を問わず、倫理原則の1位または2位にランクされ、AIの採用を妨げる2番目の要因と特定されました。

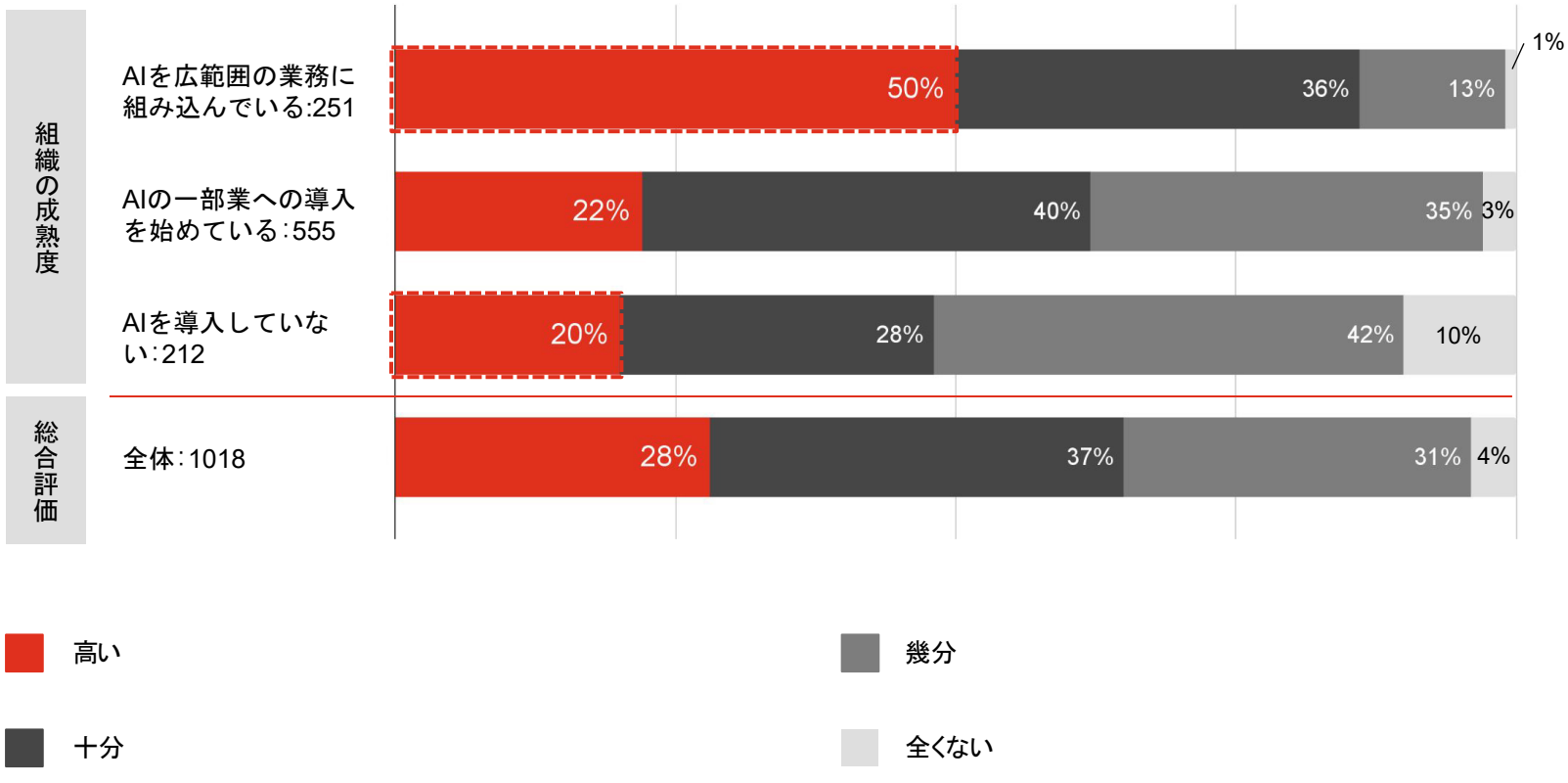
この結果に伴い、AIシステムには「セキュリティリスク」という別のリスクカテゴリーが生じます。これらのリスクの一部には、別のITシステムと同程度のものもありますが、AIではその確率と重大性がともに高まります。セキュリティリスクの一部は、用いられるAI技術に直接起因して生じる可能性もあります。例えば、機械学習モデルに対する悪意ある攻撃は、AIシステムが何かを誤って分類または予測するよう故意に誘導する場合があります（実際にはバナナを見ているにもかかわらず、トースターを見ているとコンピューターに信じ込ませる、など）[8]。また、データ汚染は、学習に用いるデータソースを故意に漏洩して、AIシステムに予期せぬ動作を開始させる場合があります[9]。

さらに、オープンソースソフトウェアの利用で生じるリスクを軽視してはなりません。オープンソースシステムの多くは信頼できるツールですが、ほぼ誰でもツールに貢献し、変更できるように設計されているため、時間の経過とともに安定性と信頼性に問題が生じる場合があります。したがって、ソフトウェアが頻繁に異なって見える可能性があり、組織がこれらのツールのリスクを評価する方法に影響する可能性があります。

調査の回答者は安全性を最大の関心事と明らかにしたものの、システムの誤作動を検知して、システムを停止する能力を有すると報告したのは、成熟度の高い組織に限られました。実際、成熟した組織ではAIリーダーの半数がこうした能力について高い確信を持っているものの、成熟度の低い組織では、AIリーダーのうち、同様に感じているのはわずか20%程度にすぎません。

図表8 – 安全性の重視状況（組織の成熟度別）

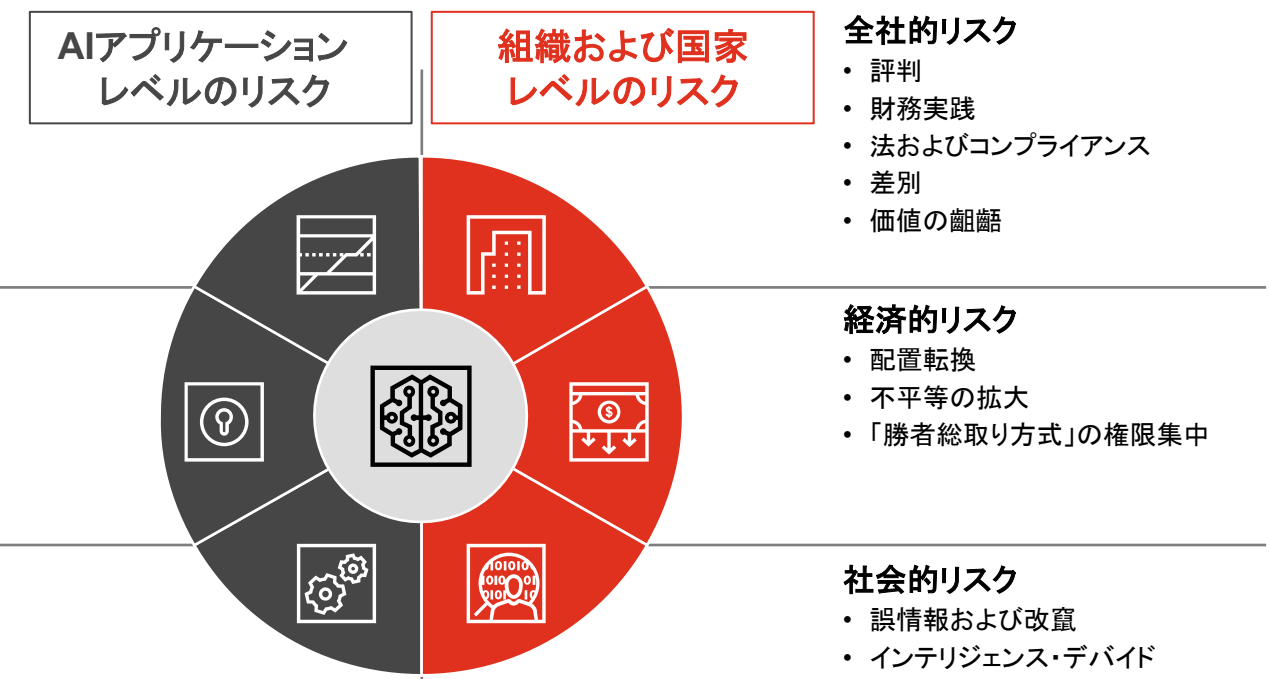
現在、AIシステムの誤作動を適時に（すなわち、深刻な問題の発生前に）検知して、これを停止する能力が貴社にあると、どの程度確信していますか。



[8] https://www.vox.com/future-perfect/2019/4/8/18297410/ai-tesla-self-driving-cars-adversarial-machine-learning
[9] https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

AIで最も重視すべきリスク領域は、性能だけではありません。例えば、今回の調査では、AIリーダーが現在の労働力におけるスキルのミスマッチを最大の懸念事項と認識していることが明らかになりました。地域、業界、成熟度の如何を問わず、技術面、経営面においてAIに適切な人材の不足は、常に懸念事項のトップ5に入ります。スキルギャップは、配置転換や不適格な作業を招くという観点において労働力に重大なリスクを生じるのみならず、質の低いAIシステムやサードパーティの管理の難しさに拍車をかけることにもなります。

図表9 – AIに関わるリスクのカテゴリー

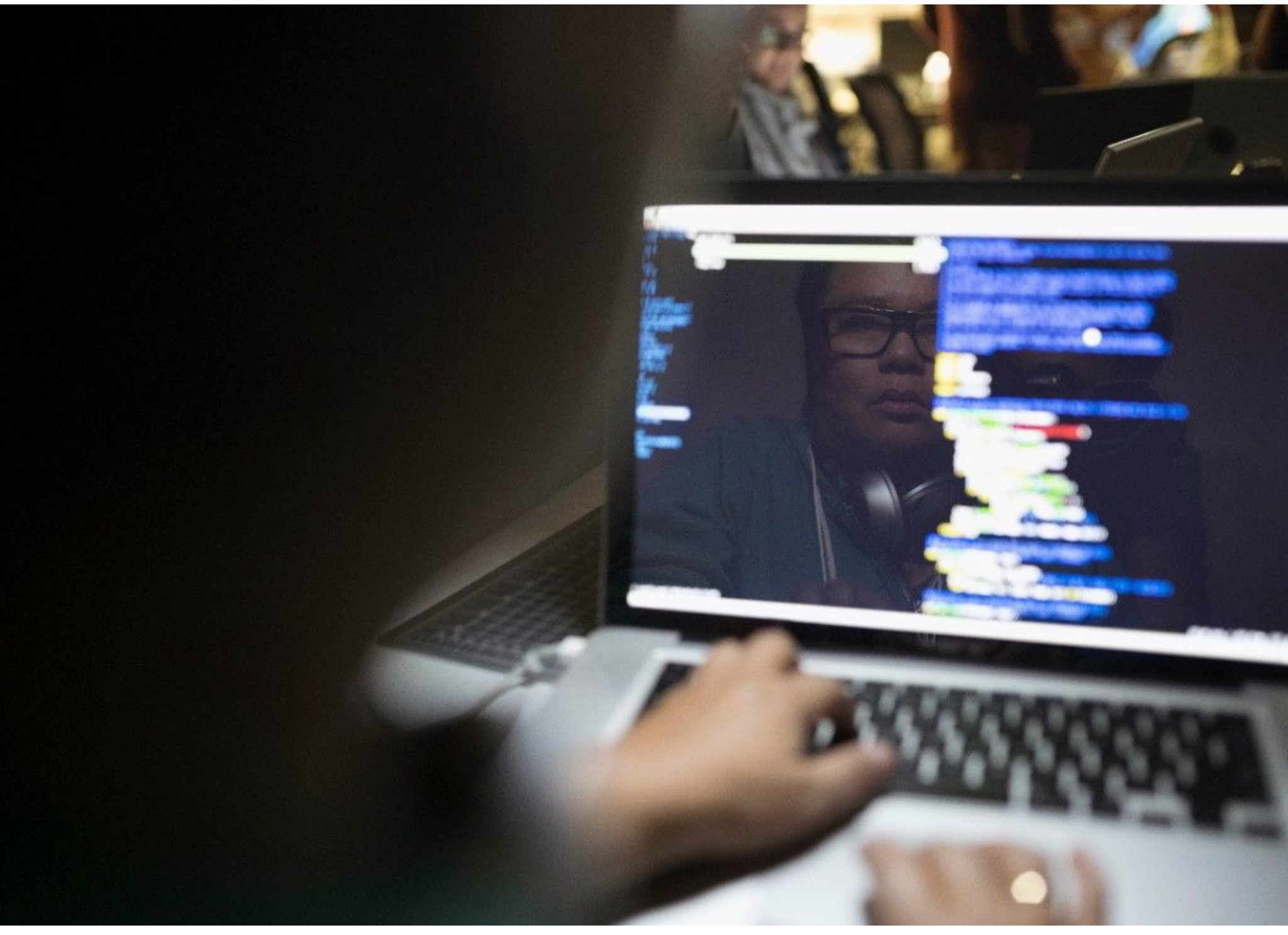


広範な組織レベルのリスクは、自動化や経済面での優位性に対するAIのポテンシャルおよび前述のAIアプリケーションレベルのリスクから生じます。これには、評判や財務上の損失などの全社的リスク、法的要件の不遵守、差別、企業価値と社会的価値との齟齬およびサードパーティやパートナーの管理が含まれます。

最終的には、AIにおけるリスクの優先順位は企業、業種、ユースケースの種類によって異なります。例えば、ヘルスケア業界では信頼性や安定性の経時的な低下から生じるリスクが極めて重要である一方、公共セクターでは人権や高度なコンプライアンス基準の遵守を特に懸念する必要があります。組織は、全社レベルのアプローチでAIリスクを特定し、実際のAIアプリケーションに関わる内容を考慮しつつ柔軟に管理、低減しなければなりません。

結論

- AIがもたらす可能性のある特定のリスクは、そのAIアプリケーションの内容に直接関係します。例えば、どのようなデータが使用されているか、どのような種類の意思決定を誰がしているか、どのAI技術が採用されているかなどを確認する必要があります。
- 組織には、AIに関わるリスクを経時的に特定、評価、軽減、モニタリングする全社的アプローチとリスク管理手順が必要となります。
- AIに関わるリスクについて、それがどのように発生し、どのように軽減できるかを組織全体で認識する必要があります。
- バイアスと解釈可能性のリスクには、特に留意する必要があります。



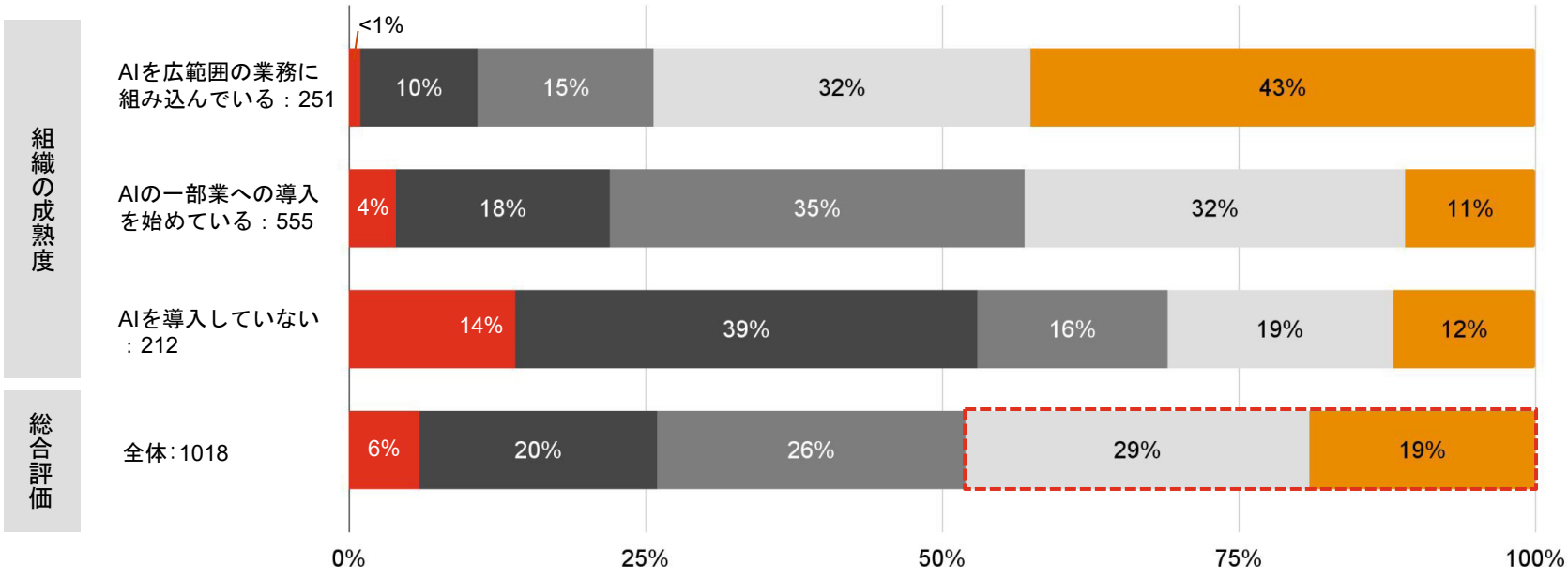
包括的なAIガバナンスの構築

変化する規制状況へのキャッチアップ、リスクの効率的軽減、組織のコンテキスト化されたデータやAI倫理に整合するポリシーの策定には、堅牢なガバナンスと説明責任が必要です。これは事実上、組織がAI開発を監督するツールや構造を特定するのみならず、AIの使用、開発、監視について責任者を定める必要があることを意味します。バイアスの低減、説明可能性の向上、堅牢性の経時的モニタリングなどを行う技術的ソリューションに注目する組織はますます増えています。しかし、これらのツールは成熟レベルが異なるため、使用したとしても倫理原則で描いたニーズに十分には見合わない場合があることに留意する必要があります。ガバナンスへの包括的なアプローチでは、単に「技術優先」ではなく、「技術を可能」にするためのプロセス、ポリシー、基準、包括的なガバナンスを応用します。

PwCの調査では、企業のAIに関わるリスクの特定と説明責任に対する取り組みが依然として初期段階にあることを示しています。全ての利害関係者に報告を行う正式に文書化されたプロセスを備えている参加企業はわずか19%で、特別な事象の発生時に限り正式なプロセスを行う企業が29%、その他の企業においては、正式なプロセスを備えていない、または明確なプロセスを一切有していませんでした。

図表10 – AIに関わる説明責任状況（組織の成熟度別）

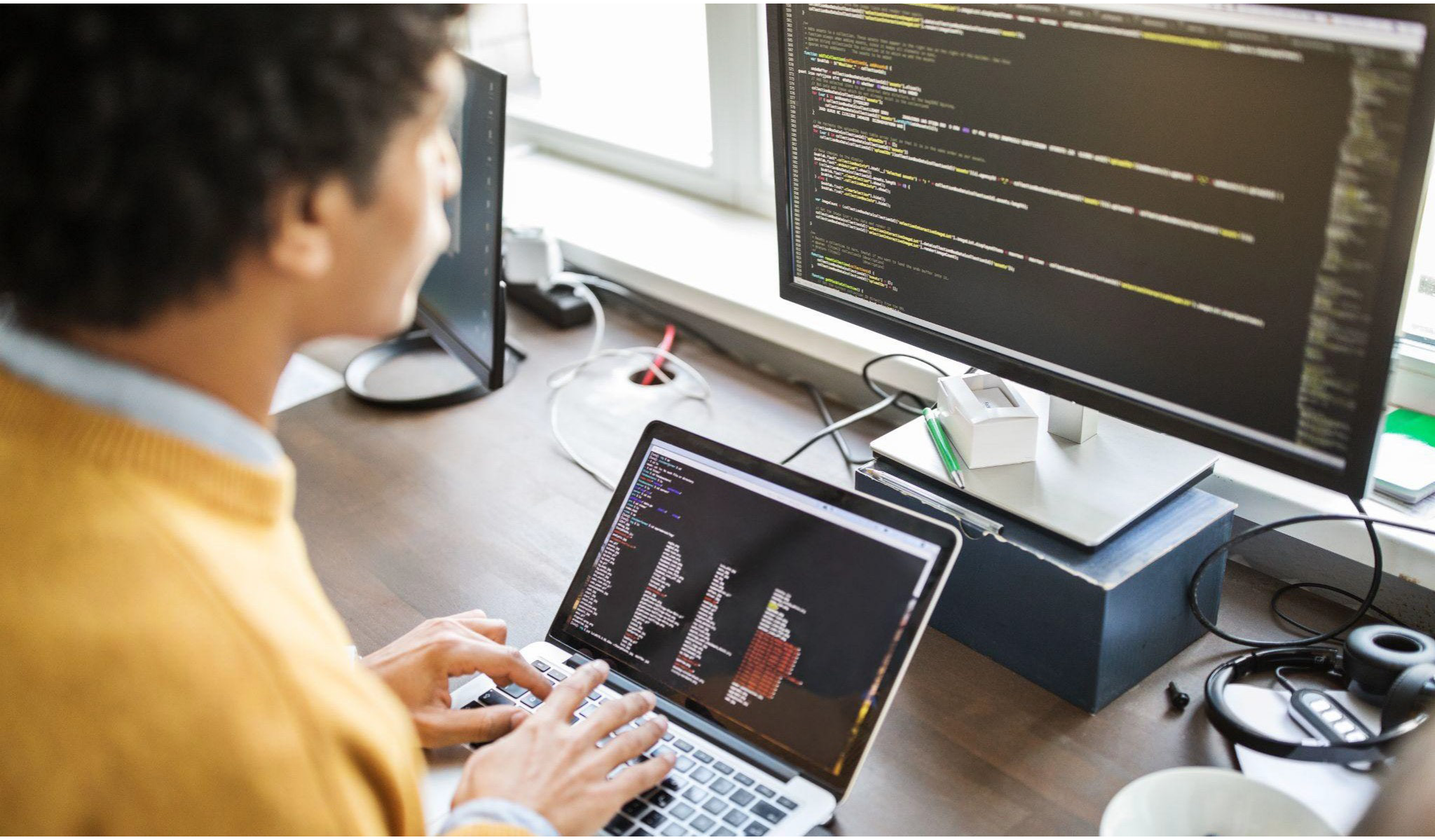
現在、AIの説明責任が組織でどのように特定されているとお考えですか。



効果的なリスク軽減やガバナンスには、AIやデータガバナンスに対する組織横断的なアプローチが必要です。このアプローチはエンドツーエンドであるとともに、リスクの管理責任と説明責任が明確に表現された3つのディフェンスライン全てにわたって広範に及ぶものでなければなりません。

エンドツーエンドのガバナンス

包括的なAIガバナンスは組織の戦略からスタートしますが、これには、データ、解析、AIの望ましい用途と期待が含まれなければなりません。戦略段階では、組織が優先順位を定める必要があります。計画段階は、組織がモデル開発とデータ使用のプログラムを立ち上げる時期であり、戦略フェーズで設定した目標の達成に必要な技術と人員を調達するエコシステム段階がこれに続きます。ガバナンスはAIの活用推進とのバランスが取れていないといけません。つまり、アプリケーションコンテキストそのものに関連付け[10]、過度に負担となるタスクを開発チームに課すことや、技術革新を阻害することを回避します。システム自体のリスク、使用データのプライバシー、システムの新規性、ガバナンスの新たな仕組みの必要性など、いくつかの要因がガバナンスの要件に影響を与える可能性があります。



[10] https://www.pwc.com/jp/ja/knowledge/thoughtleadership/comprehensive-ai-governance-needed-now.html

図表11 – AI開発のライフサイクル

個々のAIシステムは、9つのステップから成るモデルの開発とデプロイプロセスを繰り返して開発されます。これらの段階において、データサイエンティストと開発者は、ビジネスのニーズと優先順位を、詳細に調査したモデルとソフトウェアプロセスに置き換えなければなりません。データは、アプリケーションニーズに応じて取得、変換、処理する必要があります。モデルは、最適なソリューションが決定されるまで繰り返し構築、学習、テストが行われます。このソリューションは正式なデプロイの前に、ユーザーの期待と既存のプロセスに対して個々に検証され、その後、有効性が継続的にモニタリングされます。

ソフトウェア企業やクラウド企業の技術ソリューションは、自らのガバナンスソリューションについてこれら9つのステップを対象としています。各ステージゲートでは、開発チームとビジネスチーム（いわゆる第1線）が、リーダー陣や品質保証の承認者（第2線）と協働でサインオフの獲得を目指します。システムの全てのステージゲート通過は絶対条件ではありません。事実、ステージゲートを通過するためにはテスト、ドキュメンテーション、期待値の調整が必要です。開発プロセスには遵守すべきポリシー、基準、手順が伴います。何よりも、このライフサイクルがAIの全ガバナンスプロセスを構成しているわけではありません。かかるプロセスはさらに大規模であり、スタート地点は組織全体の戦略です[11]。第3線である内部監査では、統制の有効性を評価します。

これら5つのステージゲートは、3つのディフェンスラインをAI開発ライフサイクルの異なる地点に配することで、具体的な要件に基づいてアプリケーションを進める決定を、各地点で意識的に行わせる設計になっています。

[11] https://towardsdatascience.com/top-down-and-end-to-end-governance-for-the-responsible-use-of-ai-c67f360c64ba

結論

- ガバナンスについて、事業ユニットが導入可能な組織の指針や基準を定義します。
- AI開発に3つのディフェンスラインを取り入れ、機密性の高い用途のレビューはクロスファンクショナルチームに報告する必要があります。
- 透明性向上のため、ドキュメンテーションに一貫性のあるテンプレートと基準を適用します。
- ガバナンスツールを意思決定の手段として使用します。

日本のお問い合わせ先

PwC Japanグループ
www.pwc.com/jp/ja/contact.html

PwCコンサルティング合同会社

藤川 琢哉（Takuya Fujikawa）
パートナー

深澤 桃子（Momoko Fukasawa）
マネージャー

www.pwc.com/jp

PwC Japanグループは、日本におけるPwCグローバルネットワークのメンバーファームおよびそれらの関連会社（PwCあらた有限責任監査法人、PwC京都監査法人、PwCコンサルティング合同会社、PwCアドバイザリー合同会社、PwC税理士法人、PwC弁護士法人を含む）の総称です。各法人は独立した別法人として事業を行っています。
複雑化・多様化する企業の経営課題に対し、PwC Japanグループでは、監査およびアシュアランス、コンサルティング、ディールアドバイザリー、税務、そして法務における卓越した専門性を結集し、それらを有機的に協働させる体制を整えています。また、公認会計士、税理士、弁護士、その他専門スタッフ約10,200人を擁するプロフェッショナル・サービス・ネットワークとして、クライアントニーズにより的確に対応したサービスの提供に努めています。
PwCは、社会における信頼を構築し、重要な課題を解決することをPurpose（存在意義）としています。私たちは、世界152カ国に及ぶグローバルネットワークに約328,000人のスタッフを擁し、高品質な監査、税務、アドバイザリーサービスを提供しています。詳細は www.pwc.com をご覧ください。

本報告書は、PwCメンバーファームが2021年に発行した『Responsible AI – Maturing from theory to practice』を翻訳したものです。翻訳には正確を期しておりますが、英語版と解釈の相違がある場合は、英語版に依拠してください。
オリジナル（英語版）はこちらからダウンロードできます。
https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html

発刊年月：2022年12月　　管理番号：I202204-14