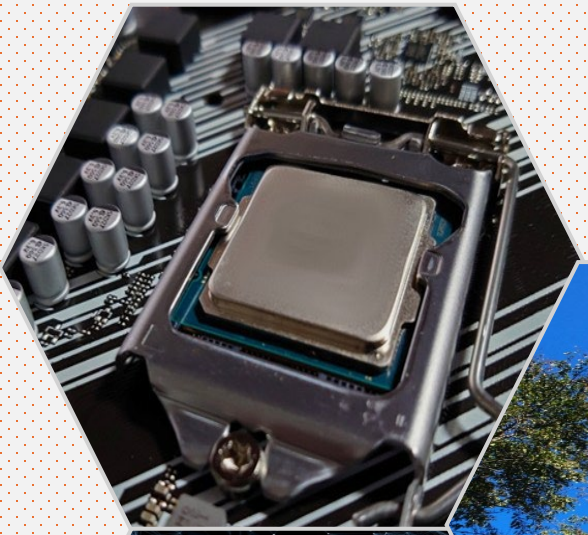
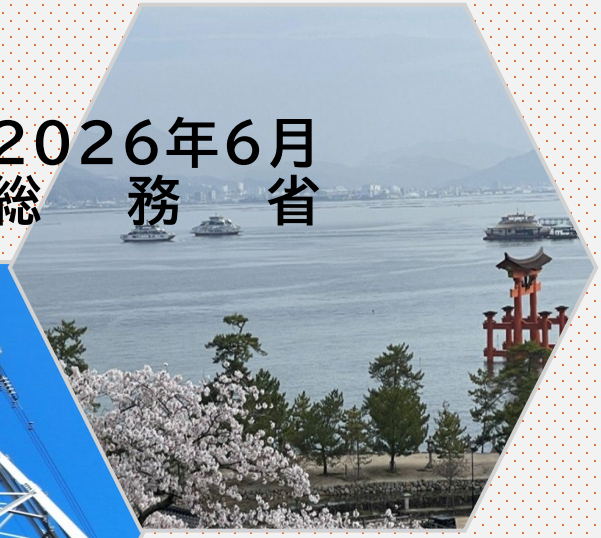
広島AIプロセス「報告枠組み」参加に向けた手引き

- 参加組織の報告分析 -





2026年6月
総　務　省

- 「報告枠組み」への参加はこちらから（OECD Webサイト）
  https://oecd.ai/en/transparency/overview

目次

1.はじめに
1-1.手引きの背景・目的

本手引きの背景・目的

- 我が国が議長国を務めた2023年のG7広島サミットを受け、生成AIに係る国際的なルール形成の枠組みである「広島AIプロセス」が立ち上げられ、「国際指針」及び「国際行動規範」が策定されました。

- 2024年のG7イタリア議長国の下では、生成AI開発における透明性及び説明責任を促進するため、**広島AIプロセスの「国際行動規範」の遵守状況をAI開発企業等が自ら確認・報告するための手法（「報告枠組み」）**を開発・導入するべく、G7で議論を重ねました。その結果、OECDの協力のもと、同年末に「報告枠組み」の基本的な運用方法及び質問票について合意し、2025年2月7日から、「報告枠組み」がOECDのWebサイト上で運用開始されました。

- 2025年6月までに20組織が自主的に参加し、OECDのWebサイト上で遵守状況を報告しました。同年9月には、OECDがこれらの報告を分析したペーパー（※）を公表しました。2026年5月1日時点では25組織が参加しています。

- 2026年5月28日には、中小企業を含むより広範な組織の参加を促すため、自由記述式の「報告枠組み（1.0）」から、回答選択式等を導入した**「報告枠組み2.0」**にアップデートされました。アップデート時点において、57組織が報告枠組み2.0に参加を表明しています。

- 本手引きは、**報告枠組みに新たに参加する組織がその報告を行うに当たっての参考**となるよう、**報告枠組み（1.0）に参加した組織からの報告をもとに、蓄積された共通項やベストプラクティスを分析・紹介**するものです。あわせて、報告枠組みで使用されている**用語の解説や、参考となるガイダンス、ガバナンスフレームワーク等**についても紹介しています。

- 報告枠組み2.0の運用が開始され、今後これに参加する組織にとって、1.0で蓄積された共通項やベストプラクティスを把握することは有意義であると考えられます。本手引きをぜひご活用ください。

- もっとも、報告枠組み2.0は、報告枠組み（1.0）とセクション構成は同一であるものの、各セクション内の設問には一部相違があります。また、回答様式が、選択式と記述式を組み合わせた様式へと変更されている点にもご留意ください。

※ OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025), https://www.oecd.org/en/publications/how-are-ai-developers-managing-risks_658c2ad6-en.html

1.はじめに
1-2.手引きの活用方法(1/2)

本手引きが想定する読者

本手引きは、これから広島AIプロセス報告枠組みに参加し、回答を作成する組織の担当者を主な読者として想定しています。

＜具体的な想定担当者＞

- 経営層、事業責任者（全社方針の決定、回答に関する最終判断）
- AI開発・運用部門（AIシステムやAIを活用した製品・サービスの設計、開発、提供、運用）
- 管理・統制部門（法務・コンプライアンス、リスク管理、情報セキュリティ、個人データ保護、品質保証、内部監査、委託先管理　等）

本手引きの活用方法

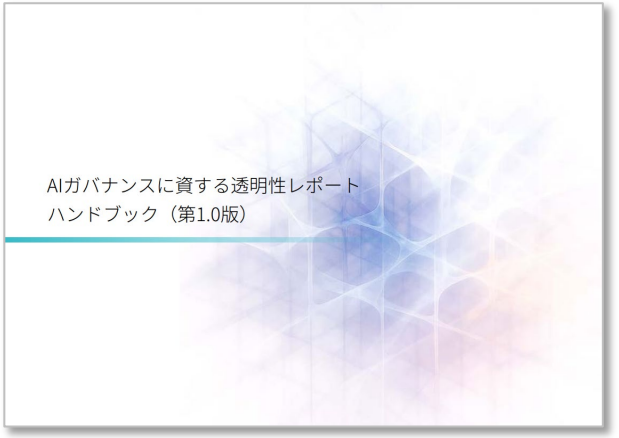
本手引きの構成は以下のとおりです。前提となる知識や取組状況に応じて必要とする部分を参照ください。

1.はじめに	本手引きの背景、活用方法を紹介
2.広島AIプロセス及び報告枠組みの概要	**「広島AIプロセス」「報告枠組み」についてよく知らないという方に** 広島AIプロセス全体の位置付けと、その中で報告枠組みが果たす役割を紹介します。回答に着手する前に、報告枠組みに参加する意義・メリットを理解できます。
3.設問別ガイダンス	**報告枠組みの各セクションの意図や考え方について知りたいという方に** 各セクションの設問ごとに、意図、回答の考え方、回答の要素を解説しています。

1.はじめに
1-2.手引きの活用方法(2/2)

「AIガバナンスに資する透明性レポート ハンドブック（第1.0版）」の概要

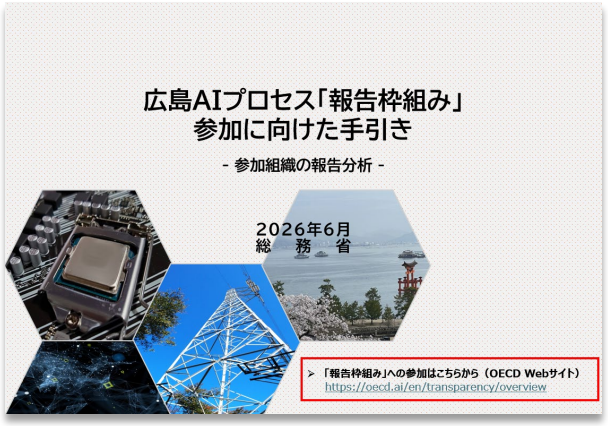
報告枠組みに関するハンドブックとして、江間有沙准教授（東京大学）・工藤郁子特任准教授（大阪大学）・実積寿也教授（中央大学）が中心となり作成された「AIガバナンスに資する透明性レポート ハンドブック（第1.0版）」があります。同ハンドブックは、報告枠組みの意義やレポート作成のための社内プロセス等について整理しています。（※1）



本手引きの位置付け

本手引きは、「報告枠組み」に新たに参加する組織がその報告を行うに当たっての参考となるよう、OECDペーパー（※2）と同様、報告枠組み（1.0）に参加した組織からの報告をもとに、蓄積された共通項やベストプラクティスを分析・紹介しています。報告枠組みの設問・回答で使用されている用語の解説や参考となるガイダンス、ガバナンスフレームワーク等についても紹介しています。

なお、報告枠組み2.0は、報告枠組み（1.0）とセクション構成は同一であるものの、各セクション内の設問には一部相違があり、また、回答様式が、選択式と記述式を組み合わせた様式へと変更されている点にはご留意ください。



（出典）
※1 江間有沙・工藤郁子・実積寿也（2025）『AIガバナンスに資する透明性レポート ハンドブック（第1.0版）』，東京大学東京カレッジ 江間研究室、ライセンス：CC BY 4.0
https://www.tc.u-tokyo.ac.jp/blog/wp-content/uploads/2025/12/HAIP_Handbook_JP.pdf
※2 OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025),
https://www.oecd.org/en/publications/how-are-ai-developers-managing-risks_658c2ad6-en.html

2.広島AIプロセス及び報告枠組みの概要
2-1.広島AIプロセスの概要(1/2)

広島AIプロセスの概要

2023年5月に開催されたG7広島サミット（議長国：日本）を受け、生成AIに係る国際的なルール形成の枠組みである「広島AIプロセス」が立ち上げられました。本プロセスは、マルチステークホルダー型の枠組みを採用し、生成AIを含む高度なAIシステムに関する国際的なガバナンスルールの構築に向けた議論を推進するものです。安全、安心で信頼できるAIの活用を通じて、世界中の人々がその恩恵を享受できる環境の整備を目的としています。

2023年12月には、広島AIプロセスの作業の集大成として、「国際指針」及び「国際行動規範」という二つの主要文書を含む包括的な政策枠組みがG7において承認されました。

さらに日本は、広島AIプロセスの精神に賛同する国々による「フレンズ・グループ」を立ち上げるとともに、企業、国際機関、市民社会などマルチステークホルダーによる議論を促進するため、「パートナーズ・コミュニティ」を設立しました。今後、フレンズ・グループ及びパートナーズ・コミュニティのさらなる拡大により、AIに関する包摂的な国際ガバナンスの形成が一層進展することが期待されます。

G7広島首脳宣言（2023年5月20日） 抜粋

我々は、関係閣僚に対し、**生成AIに関する議論を年内に行うために、包摂的な方法で、OECD及びGPAIと協力しつつ、G7の作業部会を通じた、広島AIプロセスを創設するよう指示する。**これらの議論は、ガバナンス、著作権を含む知的財産権の保護、透明性の促進、偽情報を含む外国からの情報操作への対応、これらの技術の責任ある活用といったテーマを含み得る。

（出典）
・総務省、「広島AIプロセスについて」（2023）https://www.soumu.go.jp/main_content/000919809.pdf
・外務省、「G7広島首脳コミュニケ」（2023） https://www.mofa.go.jp/mofaj/files/100507034.pdf

2.広島AIプロセス及び報告枠組みの概要
2-1.広島AIプロセスの概要（2/2）

「国際行動規範」の概要

「国際行動規範」は、高度なAIシステムを開発する組織による自主的な行動規範を提供するものです。

「国際指針」の11項目が指針を示すのに対して、「国際行動規範」はその具体的な行動を示すものです。

高度なAIシステムを開発する組織向けの広島プロセス国際行動規範の11項目

1. AI ライフサイクル全体にわたるリスクを特定、評価、軽減するために、高度な AI システムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる
2. 市場投入を含む導入後、脆弱性、及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する
3. 高度な AI システムの能力、限界、適切・不適切な使用領域を公表し、十分な透明性の確保を支援することで、アカウンタビリティの向上に貢献する
4. 産業界、政府、市民社会、学界を含む、高度な AI システムを開発する組織間での責任ある情報共有とインシデントの報告に向けて取り組む
5. 個人情報保護方針及び緩和策を含む、リスクベースのアプローチに基づく AI ガバナンス及びリスク管理方針を策定し、実施し、開示する
6. AI のライフサイクル全体にわたり、物理的セキュリティ、サイバーセキュリティ、内部脅威に対する安全対策を含む、強固なセキュリティ管理に投資し、実施する
7. 技術的に可能な場合は、電子透かしやその他の技術等、ユーザーが AI が生成したコンテンツを識別できるようにするための、信頼できるコンテンツ認証及び来歴のメカニズムを開発し、導入する
8. 社会的、安全、セキュリティ上のリスクを軽減するための研究を優先し、効果的な軽減策への投資を優先する
9. 世界の最大の課題、特に気候危機、世界保健、教育等（ただしこれらに限定されない）に対処するため、高度な AI システムの開発を優先する
10. 国際的な技術規格の開発を推進し、適切な場合にはその採用を推進する
11. 適切なデータインプット対策を実施し、個人データ及び知的財産を保護する

（出典）
・総務省、「高度なAIシステムを開発する組織向けの広島プロセス国際行動規範」（2023） https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05.pdf
・総務省、「全てのAI関係者向けの広島プロセス国際指針」（2023） https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03.pdf

2.広島AIプロセス及び報告枠組みの概要
2-2.広島AIプロセス報告枠組みの概要(1/2)

「報告枠組み」の概要

「報告枠組み」は、「国際行動規範」の遵守状況をAI開発企業等が自主的に確認・報告する枠組みとして、2025年2月7日から正式に運用開始されました。

2026年5月1日時点で、日本企業9社を含む25の組織が参加し、OECDのWebサイト上でレポートが公開されています。
（OECD Webサイト：https://oecd.ai/en/transparency/overview）

2026年5月28日にはOECDから「報告書枠組み2.0」がローンチされました。回答様式として選択式＋記述式が採用され、回答が容易となりました。

報告枠組み参加組織一覧（2026年5月1日時点）

○日本

1. KDDI Corporation
2. SoftBank Corp.
3. Preferred Networks
4. NEC Corporation
5. Fujitsu
6. Rakuten Group, Inc.
7. NTT
8. Hitachi, Ltd
9. ABEJA.inc

○アメリカ

1. West Lake research & education service (Palo Alto Research)
2. Microsoft
3. Salesforce
4. Anthropic
5. OpenAI
6. Google
7. Amazon

○その他の国

1. Data Privacy and AI(ドイツ)
2. KYP.ai GmbH(ドイツ)
3. TELUS(カナダ)
4. Fayston Preparatory School(韓国)
5. ai21(イスラエル)
6. MGOIT（ヨーロッパ）
7. TELUS Digital(カナダ)
8. Milestone(デンマーク)
9. Infosys Limited(インド)

2.広島AIプロセス及び報告枠組みの概要
2-2.広島AIプロセス報告枠組みの概要(2/2)

質問票の概要

広島AIプロセス報告枠組みの質問票はOECDのWebサイト上で公開されており、参加を希望する企業・団体はWebサイトから回答することができます。 ⇒ OECDのWebサイト https://oecd.ai/en/transparency/overview
質問票の内容は以下のとおり7セクション、39項目から構成されています。

質問票の内容

#	セクション名	主なテーマ	関連する国際行動規範の項目
1	リスクの特定及び評価	AIシステムのライフサイクル全体を通じてリスクを特定・分類・評価すること（リスク評価手法、テストの実施、技術標準の活用等を含む）	1,4,6,10
2	リスクの管理及び情報セキュリティ	リスクの緩和、データ品質の確保、プライバシー及び知的財産（IP）の保護、サイバーセキュリティ対策の実施、ならびにインシデントや脆弱性の管理に関すること	1,2,4,6,10,11
3	高度なAIシステムに関する透明性報告	AIシステムの能力及び限界について、明確な報告書や文書を提供すること（プライバシーポリシー、データソースの情報、ステークホルダーとの情報共有を含む）	3,4,5,10
4	組織のガバナンス、インシデント管理及び透明性	ガバナンスポリシーの策定・実施、インシデントの記録及び管理、職員研修の実施、ならびに法的枠組みへの遵守の確保に関すること	2,4,5,10,11
5	コンテンツの認証及び来歴確認の仕組	ユーザーがAIシステムやAI生成コンテンツとやり取りしていることを認識できるようにするための仕組みの開発・実装に関すること（コンテンツのラベリング、ウォーターマーキング、国際標準に沿った来歴管理を含む）	7,10
6	AIの安全性向上や社会リスクの軽減に向けた研究及び投資	AIの安全性に関する研究及び投資の実施、ならびにコンテンツ来歴、リスク緩和、社会的・経済的・環境的リスクの最小化に関するプロジェクトへの協力に関すること	1,8
7	人類と世界の利益の促進	社会経済的及び環境的利益を促進するプロジェクトへの投資、デジタル・リテラシーの支援、ならびに地球規模課題に対応するための市民社会との協働に関すること	9

（出典）OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025), P.9

2.広島AIプロセス及び報告枠組みの概要
2-3.カナダ政府「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」（1/2）

2025年G7議長国カナダ政府発行「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」

2025年のG7（議長国カナダ）の結果を踏まえ、カナダ政府から発行された「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」において、広島AIプロセスの国際行動規範のうち中小企業にとって特に適用が難しい規範が特定されています。これらの規範を中小企業が適用する上で参考となる国内外のリソースやツールが整理・提示されています。

2025年G7議長国カナダ政府発行「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」の目的・背景

本ツールキットは、2025年のG7サミット（議長国カナダ）において合意された広島AIプロセスの成果を踏まえ、カナダ政府が中小企業向けに作成したものであり、G7で策定された国際的な指導原則及び行動規範を中小企業の「AI導入者（ユーザー）」の立場から実務的に理解し、段階的に実践できるよう支援することを目的として作成されたものである。G7サミットにおける首脳レベルのコミットメントやOECD等との連携による議論を踏まえ、中小企業が経済全体の中で果たす重要な役割と、AIガバナンスにおけるリソース制約の双方を前提に設計されている。

中小企業にとって適用が難しい広島AIプロセスの規範

規範1：AIライフサイクル全体にわたるリスクを特定、評価、軽減するために、高度な AIシステムの開発全体を通じて、その導入前及び市場投入前も含め、適切な措置を講じる

➢AIシステムを業務に導入する前に、全ての機能において、業務遂行における潜在的なリスクを慎重に検討する必要がある。

規範2：市場投入を含む導入後、脆弱性及び必要に応じて悪用されたインシデントやパターンを特定し、緩和する

➢リスクの高い高度なAIシステムを導入する場合、意図しない影響やエラー、不適切な利用といったインシデントを継続的に監視し、問題が発生した際には関係者と連携して迅速に対応する必要。

規範10：国際的な技術規格の開発を推進し、適切な場合にはその採用を推進する

➢AI標準に準拠した製品を提供するベンダーを選ぶことで、AI導入に伴う潜在的なリスクを軽減できる。よって、中小企業は業界主導のAI標準策定に参加する機会を見つけることが推奨される。

中小企業の課題解決をサポート

適用が難しい規範に対応する上での参考リソース

①組織の認知と投資	②リスクの特定と優先度設定	③技術的専門知識
• 事前審査済みのプロバイダー一覧 • 信頼できるAIプロバイダーを特定するための評価ツール	• 信頼できるAIの基準 • 国際的な信頼できるAIの枠組みとガイドライン • OECDのAIインシデント及び危険監視	• HAIP報告枠組み • OECD政策ナビゲーター • AI事業者ガイドライン

（出典）TOOLKIT FOR SMALL AND MEDIUM-SIZED ENTERPRISES (SMEs) DEPLOYING ARTIFICIAL INTELLIGENCE (AI)
https://ised-isde.canada.ca/site/ised/sites/default/files/documents/g7-mtl-toolkit-en.pdf

2.広島AIプロセス及び報告枠組みの概要
2-3.カナダ政府「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」（2/2）

参考リソース

「中小企業における広島AIプロセスに基づくAI導入のためのツールキット」で整理・提示されている参考リソースのうち、特に日本企業にとって参考となるリソースには以下のようなものがあります。

参考リソース		策定主体	概要	URL
①組織の認知と投資	信頼できるAIのためのツール&指標カタログ	OECD	中小企業に対し、責任ある倫理的AIの原則に沿ったAIプロバイダーを特定するためのリソースや評価ツールの有用なディレクトリを提供している。	https://oecd.ai/en/catalogue/tools
②リスクの特定と優先度設定	AIシステム分類フレームワーク	OECD	利用状況、人の関与度、機能、リスクプロファイルに基づいて、様々な種類のAIを特定・評価するために設計されているフレームワーク。	https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html
	ISO/IEC 42001:2023	ISO・IEC	組織内でAIマネジメントシステム（AIMS）を確立、実施、維持、継続的に改善するための要件を規定した国際規格。	https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en
	AIインシデント&ハザードモニター（AIM）	OECD	業界特有のインシデントや類似導入での一般的な問題を提示し、注意すべき共通リスク領域に関する洞察を提供している。業界特有のインシデントや類似導入での一般的な問題の提示による、注意すべき共通リスク領域に関する洞察の提供や、AIのインシデントやハザードのカタログ化によりリスクや被害を明示し、時間の経過とともにパターンを明らかにし、AIに関する信頼できる共通理解の支援等を行う。	https://oecd.ai/en/incidents?search_terms=%5B%5D&and_condition=false&from_date=2020-12-18&to_date=2025-12-18&properties_config=%7B%22principles%22:%5B%5D,%22industries%22:%5B%5D,%22harm_types%22:%5B%5D,%22harm_levels%22:%5B%5D,%22harmed_entities%22:%5B%5D,%22business_functions%22:%5B%5D,%22ai_tasks%22:%5B%5D,%22autonomy_levels%22:%5B%5D,%22languages%22:%5B%5D%7D&order_by=date&num_results=20
	責任ある生成AIガバナンスガイド	Vector Institute	生成AIの責任あるガバナンスのための実践的なリソースを提供するガイド。	https://res-ai.ca/#home-panel
③技術的専門知識	報告枠組み	G7・OECD	組織が国際行動規範に準拠するための実践的なツールを提供することを目的とした取組。	https://oecd.ai/en/transparency/overview
	OECD. AIポリシーナビゲーター	OECD	世界各国のAIに関する政策や規制を追跡するための中央リポジトリとして機能している。	https://oecd.ai/en/dashboards/international/g7
	AI事業者ガイドライン	総務省・経済産業省	人間中心の設計、安全性、公平性、プライバシー、セキュリティ、透明性、説明責任といった主要原則を整理し、AIビジネスユーザー向けに具体的な助言を提供している。	https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

3.設問別ガイダンス

本手引きで使われる用語集（組織）

#	日本語通称	正式名	説明
1	OECD	Organisation for Economic Co-operation and Development	経済成長、貿易自由化、開発途上国支援を目的とする国際経済協力機構。AI原則の策定でも知られる。（※1）
2	ISO/IEC	International Organization for Standardization / International Electrotechnical Commission	国際標準化機構（ISO）と国際電気標準会議（IEC）による合同技術委員会で、情報技術分野（AI含む）の国際規格を策定。（※2）
3	CEN/CENELEC	European Committee for Standardization / European Committee for Electrotechnical Standardization	欧州の標準化機関で、EU域内における技術規格（AI法関連規格を含む）を策定。（※3）
4	MLCommons	MLCommons	機械学習のベンチマーク（MLPerf等）やデータセット、ベストプラクティスを開発するオープンな業界コンソーシアム。（※4）
5	IEEE	Institute of Electrical and Electronics Engineers	電気・電子・情報工学分野における世界最大の専門家組織で、技術標準策定（IEEE 802等）や学術活動を行う。（※5）
6	C2PA	Coalition for Content Provenance and Authenticity	デジタルコンテンツの出所・来歴を証明する技術標準を策定する業界連合。ディープフェイク対策等で注目されている。（※6）
7	フロンティア・モデル・フォーラム	Frontier Model Forum	Anthropic、Google、Microsoft、OpenAIが設立した、最先端AIモデルの安全性と責任ある開発を推進する業界団体。（※7）
8	パートナーシップ・オン・AI	Partnership on AI (PAI)	AIの責任ある開発・利用に関するベストプラクティスを研究・提言する非営利マルチステークホルダー組織。（※8）

（出典）
※1：OECD, https://www.oecd.org/
※2：ISO, https://www.iso.org/home.html / IEC, https://www.iec.ch/homepage
※3：CEN/CENELEC, https://www.cencenelec.eu/
※4：MLCommons, https://mlcommons.org/
※5：IEEE, https://www.ieee.org/
※6：C2PA, https://c2pa.org/
※7：フロンティア・モデル・フォーラム, https://www.frontiermodelforum.org/
※8：パートナーシップ・オン・AI, https://partnershiponai.org/

3.設問別ガイダンス
3-0.用語集

本手引きで使われる用語集（法令、ガイドライン、フレームワーク等）

#	英語	日本語	用語説明
1	AI Guidelines for Business	AI事業者ガイドライン	日本における従来の3ガイドライン（AI開発・AI利活用・AI原則実践）を統合し、AIの開発者・提供者・利用者の各主体が遵守すべき事項を示した指針。国際的な議論（OECD・広島AIプロセス等）との整合性を考慮している。（※1）
2	EU AI Act	EU AI 法	EUの包括的AI規制法。AIシステムをリスクに応じて4段階（許容できないリスク・高リスク・限定リスク・最小リスク）に分類し、規制を課す。（※2）
3	NIST AI Risk Management Framework（NIST AI RMF）	NIST AIリスクマネジメントフレームワーク	米国国立標準技術研究所（NIST）が策定した、AIシステムのリスクを管理するためのガバナンスフレームワーク。「統治（Govern）・マッピング（Map）・測定（Measure）・管理（Manage）」の4機能で構成されている。（※3）
4	OECD AI Principles	OECD AI原則	政策立案者及びAI関係者に向けた、5つの価値に基づく原則と5つの勧告で構成される、信頼できるAIの責任ある管理のための国際的な政策枠組み。（※4）
5	–	AIセーフティに関する評価観点ガイド	AIシステム、特に大規模言語モデル（LLM）を構成要素とするAIシステムの開発・提供に関わる事業者を対象に、安全性評価の観点や手法を整理したガイドライン。（※5）
6	–	AIのセキュリティ確保のための技術的対策に係るガイドライン	AI自体への脅威・攻撃（プロンプトインジェクション攻撃やDoS攻撃など）への具体的対策について整理したガイドライン。（※6）

（出典）
※1：総務省・経済産業省「AI事業者ガイドライン（第1.2版）」（2024年4月19日公表、2025年3月28日、2026年3月31日改訂），
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html
※2：欧州議会・理事会「Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence」（2024年6月13日採択、Official Journal of the EU, 2024年7月12日公布），https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
※3：NIST「Artificial Intelligence Risk Management Framework (AI RMF 1.0)」NIST AI 100-1（2023年1月26日公表），
https://www.nist.gov/itl/ai-risk-management-framework
※4：OECD「Recommendation of the Council on Artificial Intelligence」（OECD/LEGAL/0449、2019年5月22日採択、2024年5月改訂），
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
※5：Japan AI Safety Institute「AIセーフティに関する評価観点ガイド（第1.10版）」（2024年9月18日公表、2025年3月28日改訂），
https://aisi.go.jp/output/output_information/250328_1/
※6：総務省「AIのセキュリティ確保のための技術的対策に係るガイドライン」（2026年3月公表）
https://www.soumu.go.jp/menu_news/s-news/01cyber01_02000001_00282.html

3.設問別ガイダンス
3-0.用語集

本手引きで使われる用語集（技術的用語）(1/2)

#	英語	日本語	用語説明
1	Annotation	アノテーション	機械学習の教師データ作成のため、画像・テキスト・音声等の生データに意味情報（ラベル・タグ・領域指定等）を付与する作業を指す。AIモデルの精度・公平性を左右する重要工程となる。（※1）
2	Bias and Fairness Evaluation	バイアス・公平性評価	AIシステムの判断や出力に含まれる不当な偏り（性別・人種・年齢等の属性による差別的傾向）を、統計的指標（Demographic Parity、Equalized Odds等）で定量的に測定・評価する手法。（※2）
3	Drift Analysis	ドリフト検定	運用中のAI／機械学習モデルにおいて、入力データや予測分布が学習時と比べて統計的に有意に変化（ドリフト）しているかを、統計的仮説検定により判定する手法。（※1）
4	Edge Case	エッジケース	システムの学習データに十分に含まれていない、稀または非典型的な入力のことで、予期しない、または誤った出力を引き起こす可能性があるもの。エッジケースは、システムの弱点を特定するために、安全性評価やレッドチーミングの演習でしばしば対象とされる。（※3）
5	Guardrails	ガードレール	AIシステム（特にLLM・生成AI）の入出力を監視・制御し、有害コンテンツ生成、機密情報漏洩、規約違反、ハルシネーション等を防ぐための制約。入力検証・出力フィルタ・ポリシー適用等で実装する。（※4）
6	Penetration Testing	ペネトレーションテスト	システムに対して実際の攻撃者を模した手法で擬似的な攻撃を行い、セキュリティ上の脆弱性や防御機能の有効性を検証する試験。AI領域ではAIレッドチーミングとして、モデル固有の脅威（敵対的サンプル・データ汚染等）への評価に発展できる。
7	Purple-teaming	パープルチーミング	レッドチーム（攻撃側）とブルーチーム（防御側）が協力してシステムセキュリティを向上させる、協調的なサイバーセキュリティテスト手法。（※5）
8	Red-teaming	レッドチーム	企業のセキュリティ体制に対して潜在的な敵対者の攻撃能力や悪用能力を模倣するよう権限を与えられ、組織化されたグループ。潜在的なセキュリティ脆弱性を特定し、対処するために使用される。（※5、6）

（参考）
※1：ISO/IEC 22989:2022「Artificial intelligence concepts and terminology」, https://www.iso.org/standard/74296.html
※2：ISO/IEC TR 24027:2021「Bias in AI systems and AI aided decision making」,
https://cdn.standards.iteh.ai/samples/77607/c0664994eace4bd597db80bb10c18dec/ISO-IEC-TR-24027-2021.pdf
※3: OECD,「How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct」（2025), P.33
https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/09/how-are-ai-developers-managing-risks_fbaeb3ad/658c2ad6-en.pdf
※4：NIST AI 600-1「Artificial Intelligence Risk Management Framework: Generative AI Profile」（2024年7月）,
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
※5：NIST「Artificial Intelligence Risk Management Framework (AI RMF 1.0)」NIST AI 100-1（2023年1月26日公表）,
https://www.nist.gov/itl/ai-risk-management-framework
※6：AISI「 AIセーフティに関するレッドチーミング手法ガイド(第1.10版)」（2025年3月31日公表）,
https://aisi.go.jp/output/output_information/250331_1/

本手引きで使われる用語集（技術的用語）(2/2)

#	英語	日本語	用語説明
9	Regression Testing	回帰テスト	システムの変更（バグ修正・機能追加・モデル再学習等）により、既存機能に予期しない不具合（デグレード）が生じていないかを確認するためのテスト。AI分野ではモデル更新時の性能劣化検出に利用する。（※7）
10	Stress Testing	ストレステスト	システムに対して想定を超える負荷・異常条件・極端な入力を与え、限界性能や破綻時の挙動、堅牢性（robustness）を評価する試験手法。AIモデルでは敵対的入力や分布外データへの耐性評価にも用いられる。
11	Systemic risk	システミックリスク	バリューチェーン・社会全体に重大な影響を及ぼし得るリスク。（※8）
12	Threat Modeling	脅威モデリング	システムに対する潜在的な脅威・攻撃ベクトルを体系的に特定・列挙し、リスクを分析・優先順位付けして対策を検討するセキュリティ設計手法。（※9）

（参考）
※7：ISO/IEC/IEEE 29119-1:2022「Software and systems engineering — Software testing — Part 1: General concepts」, https://www.iso.org/standard/81291.html
※8：EU AI Act, Article 3:Definitions (65), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
※9：NIST SP 800-154「Guide to Data-Centric System Threat Modeling」（2016年）, https://csrc.nist.gov/pubs/sp/800/154/ipd

3.設問別ガイダンス

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～

• Section1-Risk identification and evaluation／セクション1～リスクの特定及び評価～（1/2）

#	設問内容（英語）	日本語訳
a	How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?	AIに関連する様々な種類のリスク（不合理なリスク等）について、どのように定義及び／または分類していますか。
b	What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?	AIのライフサイクル全体を通じて、脆弱性・インシデント・新たなリスク・悪用等のリスクの特定と評価をどのような方法で実践していますか。
c	Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?	モデル／システムが開発段階から次に移行する際の適合性を評価するために、どのようにテスト（レッドチーミング等）を行っているかを説明してください。
d	Does your organization use incident reports, including reports shared by other organizations, to help identify risks?	リスクを特定するために、他組織から共有された報告書を含め、インシデントレポートを利用していますか。
e	Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?	定量的及び／または定性的なリスク評価指標を使用していますか。使用している場合、どのようなことに配慮していますか。 脆弱性やインシデントの報告の仕組みを、社内外の多様な関係者が利用しやすいよう整備していますか。 リスク・インシデント・脆弱性の責任ある開示に対するインセンティブ制度を持っていますか。
f	Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?	リスクの特定、アセスメント、評価のために、外部の独立した専門知識を活用していますか。活用している場合、どのようなものを活用していますか。また、第三者からリスク・インシデント・脆弱性の報告を受ける仕組みを持っていますか。

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～

• Section1-Risk identification and evaluation／セクション1～リスクの特定及び評価～（2/2）

#	設問内容（英語）	日本語訳
g	Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?	リスクの特定、アセスメント、評価のための国際的な技術標準やベストプラクティスの開発に貢献及び／または採用していますか。
h	How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?	特にシステミックリスクを含むリスクへの対応にあたり、分野横断的に関係するステークホルダーとどのように連携し、リスク緩和策の評価及び導入を行っていますか。

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問a

設問a

AIに関連する様々な種類のリスク（不合理なリスク等）について、どのように定義及び／または分類していますか。
How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

設問の意図

- AIに関連する様々な種類のリスクについて、自組織においてどのような基準・方法で定義し、具体的に定義・分類しているかの結果を示すことが求められています。

回答にあたっての考え方

- リスク定義・分類の方法としては、リスクの程度（影響度、発生確率）に基づいてリスクを定義・分類する方法のほか、プライバシー保護、偽誤情報といったAIに特有のリスクに基づいて定義・分類する方法が見受けられます。

- なお、多くの組織で、リスクの程度を定義・分類するための閾値を設定している例が見受けられます。閾値の設定にあたっては、独自のスコアカードの運用や、参照先となる法令等を示す例が見受けられます。

回答の要素

1　リスクの程度による定義・分類の例

- リスクの程度に基づいてリスクを定義・分類する場合、以下のような例があります。
  - 禁止、高、中、低リスク
  - 不合理なリスクと予測可能なリスク　等

2　法令、ガイドライン、フレームワーク等を参照する例

- EU AI法、NIST AI RMF、AI事業者ガイドライン、AIビジネスガイドライン、広島AIプロセス国際行動規範、OECD AI原則、 AIセーフティに関する評価観点ガイド (第1.10版)等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.12

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問b

設問b

AIのライフサイクル全体を通じて、脆弱性・インシデント・新たなリスク・悪用等のリスクの特定と評価をどのような方法で実践していますか。
What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

設問の意図

- AIライフサイクル全体を通じて、どの段階でリスク特定・評価を行うか、リスクに関する特定手法、評価体制等を具体的に記載することが重要です。AIライフサイクルは、設計・開発段階から、リリース前・テスト段階、運用段階におけるリスクの特定と評価を行うことが想定されます。

回答にあたっての考え方

- AIに関連するリスクの特定と評価は、多段階アプローチが採用され、設計・開発段階から導入後のモニタリングに至るまで、AIシステムのライフサイクル全体にわたって行われているケースが多いと見受けられます。
- リスクの特定・評価手法として技術的手法（レッドチーム演習、脆弱性スキャナ、監視システム等）を採用している組織も見受けられます。
- 多くの組織が敵対的テストの活用を強調しており、また、AIシステムをテストまたは評価するためにAIツールを使用する例も見受けられます。

回答の要素

1 設計・開発段階の例

- チェックリストを用いて潜在的なリスクについて議論・特定していること、脅威モデリングやベンチマーク評価を行っていること等を記載することが想定されます。

2 リリース前・テスト段階の例

- 敵対的テスト（レッドチーム・パープルチーム演習等）、ストレステスト、エッジケーステスト、バイアス・公平性評価、回帰テスト、脅威モデリングの他、AIツールを活用した評価等が想定されます。
- ガードレール設計やコンテンツフィルタリング、関連部門を含むレビュー体制の構築といった技術的・組織的対策を記載することが想定されます。

3 リリース後・運用段階の例

- ユーザーフィードバックやインシデント報告、リアルタイムの異常検知ツールによる監視等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.12,13

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問c

設問C

モデル／システムが開発段階から次に移行する際の適合性を評価するために、どのようにテスト（レッドチーミング等）を行っているかを説明してください。
Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

設問の意図

- モデル／システムを開発段階から次の段階（リリース等）へ進める際に実施するテストの方法・内容について説明が求められています。テストの手法、実施主体等について整理することが想定されます。

回答にあたっての考え方

- 多様なテスト手法を採用し、様々な側面（性能、品質、安全性、脆弱性等）からテストを実施する例が見受けられます。
- 多くの組織がレッドチームやパープルチーム演習等の敵対的テストを導入しており、AIシステムに厳しい課題を与え、現実的な攻撃状況下でどのように反応するかを確認していると見受けられます。
- 人間によるテスト・評価を補完する方法として、AIツールを使用する例も見受けられます。

回答の要素

1 テスト手法の例

- 敵対的テスト（レッドチーム・パープルチーム演習等）、ストレステスト、エッジケーステスト、バイアス・公平性評価、回帰テスト、脅威モデリングの他、AIツールを活用した評価等が想定されます。
- なお、テストの実施主体としては、自組織で実施するケースと、外部のレッドチームや研究所、評価会社等の外部機関を活用するケースがあります。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.13

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問d

設問d

リスクを特定するために、他組織から共有された報告書を含め、インシデントレポートを利用していますか。Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

設問の意図

- 自組織内で発生したインシデントだけでなく、他社、業界全体で共有されているインシデント情報も活用して、AIシステムに潜むリスクを把握・予防できているかを確認するための設問です。

回答にあたっての考え方

- 他組織から共有されたインシデントレポートを利用しているかどうかをYes/No形式で回答し、多くの組織はYesと回答していることが見受けられます。

回答の要素

（Yes/No形式のため、省略）

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問e

設問e

定量的及び／または定性的なリスク評価指標を使用していますか。使用している場合、どのようなことに配慮していますか。
脆弱性やインシデントの報告の仕組みを、社内外の多様な関係者が利用しやすいよう整備していますか。
リスク・インシデント・脆弱性の責任ある開示に対するインセンティブ制度を持っていますか。
Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats?
Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders?
Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

設問の意図

- リスクの評価指標、脆弱性やインシデントの報告を受ける仕組み、報告を促す制度をどのように設計し、運用しているかについての説明が求められています。

回答にあたっての考え方

- 多くの組織が、リスク評価指標について、バイアススコアや誤検知率等の定量的指標と、専門家による判断等の定性的指標の両方を組み合わせて実施していると見受けられます。
- 脆弱性やインシデント報告を受ける仕組みとして、社内外からの報告窓口等を設置する例が見受けられます。
- 外部のユーザーや研究者からの情報を収集する仕組みとして、バグバウンティ制度を設けることで、継続的なリスク監視を推進する事例が見受けられます。

回答の要素

1 定量・定性的指標の例

- 定量的指標は、バイアススコア、誤検知率、ドリフト分析、脆弱性の件数等が想定されます。
- 定性的指標は、専門家の判断、倫理審査、ステークホルダーからのフィードバック等が想定されます。

2 報告の仕組みの例

- 社内外の報告窓口（フォーム、メール、製品内）の設置等が想定されます。

3 インセンティブ制度の例

- バグバウンティ制度、表彰制度等を設けることが想定されます。（バグバウンティとは、脆弱性の発見・報告に対して報奨金を支払う制度を指します。）

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.12, 13

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問f

設問f

リスクの特定、アセスメント、評価のために、外部の独立した専門知識を活用していますか。活用している場合、どのようなものを活用していますか。また、第三者からリスク・インシデント・脆弱性の報告を受ける仕組みを持っていますか。
Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

設問の意図

- 独立性を確保した形で外部の視点を取り入れ、社内判断の偏りや見落としを減らす設計になっているか、また社外からの指摘を受け止めて改善につなげられる窓口の有無についての説明が求められています。

回答にあたっての考え方

- 多くの組織が外部の研究者や専門家の意見を取り入れることを重視し、テストの外部委託や、内部テストと外部テストの結果比較等を実施していると見受けられます。一方で、外部機関を活用していない例も見受けられます。

- セクション1 設問eでは、社内外の多様な関係者からインシデント等に関する報告を受ける仕組みの有無について問われていましたが、本設問では特に第三者から報告を受ける仕組みの有無について問われています。

回答の要素

1 外部機関の例

- 外部機関の例としては、学術・研究機関、政府機関、認証機関、業界団体等が挙げられます。
- 外部機関の具体的な連携方法には次のようなものがあります。
  - 外部機関との協働によるリスク評価手法・閾値の検討等
  - 外部レッドチームによるペネトレーションテスト、脆弱性評価等
  - 外部専門家の登用、諮問委員会の設置等

2 第三者報告の仕組みの例

- 第三者による監査の実施、専用窓口の設置等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.13

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問g

設問g

リスクの特定、アセスメント、評価のための国際的な技術標準やベストプラクティスの開発に貢献及び／または採用していますか。
Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

設問の意図

- リスクの特定、アセスメント、評価という目的に対して、標準をどのように使っているか、どのように貢献しているかを説明することが求められています。

回答にあたっての考え方

- 複数の組織が国際的な技術標準やベストプラクティスを採用するとともに、業界コンソーシアムへの参加を通じて策定に貢献していると見受けられます。
- また、多くの組織が外部基準を採用していると見受けられます。

回答の要素

1 基準策定への貢献の例

- OECD、ISO/IEC、CEN/CENELEC、MLCommons、 IEEE、C2PA等の標準化・評価コミュニティ、また、Frontier Model Forum、Partnership on AI等の業界コンソーシアムに参画し、ドラフト起草、WGリード、資金拠出、共同ホワイトペーパー等の成果物に貢献すること等が想定されます。

2 外部基準の採用の例

- 国際標準・フレームワーク・法令等（例：ISO 42001 /27001、NIST AI RMF、EU AI法等）を組織内で実装すること等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.13

3.設問別ガイダンス
3-1.セクション1　～リスクの特定及び評価～　設問h

設問h

特にシステミックリスクを含むリスクへの対応にあたり、セクターを超えた関係者とどのように協力し、リスク緩和策の評価及び採用を行っていますか。
How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

設問の意図

- システミックリスクを含むリスクへ対応するための緩和策を評価・採用するために、自組織以外の関係者（ステークホルダー）とどのように協力しているかについて説明することが求められています。

回答にあたっての考え方

- 「システミックリスク」とは、バリューチェーン・社会全体に重大な影響を及ぼし得るリスクを指します。
- 外部ワーキンググループ等に参画し、システミックリスクの緩和策を共同で検証している例や、部門を横断した社内ステークホルダーとの協働によりリスク評価を実施している例が見受けられます。

回答の要素

1 協働の例

- AIガバナンス協会、C2PA、NIST、UNESCO AI倫理グローバルビジネス評議会、AI Allianceといった産官学・業界におけるワーキンググループやフォーラム等に参画し、共同で知見・対策等を形成している例等が想定されます。

3.設問別ガイダンス

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～

- Section2-Risk management and information security／セクション2～リスクの管理及び情報セキュリティ～（1/2）

#	設問内容（英語）	日本語訳
a	What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?	AIのライフサイクル全体にわたるリスク・脆弱性に対処するために、どのような措置を講じていますか。
b	How do testing measures inform actions to address identified risks?	テスト結果は、特定されたリスクへの対策の実施にどのように反映されていますか。
c	When does testing take place in secure environments, if at all, and how?	安全な環境でテストを実施している場合、いつ、どのように実施していますか。
d	How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?	訓練及びデータ収集のプロセスを含み、どのようにデータの質を確保し、また、有害なバイアスのリスクを緩和していますか。
e	How does your organization protect intellectual property, including copyright-protected content?	知的財産（著作物等）をどのように保護していますか。
f	How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?	プライバシーをどのように保護していますか。機密／センシティブ情報の漏洩をどのように防止していますか。

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～

- Section2-Risk management and information security／セクション2～リスクの管理及び情報セキュリティ～（2/2）

#	設問内容（英語）	日本語訳
g	How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?	運用セキュリティ及びサイバー／物理セキュリティに関するAI特有の情報セキュリティ対策をどのように実施していますか。
g-i	How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?	高度なAIシステムのサイバーセキュリティ強化に向け、サイバーセキュリティのリスク評価、ポリシーの導入をどのように行っていますか。
g-ii	How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?	セキュリティリスクから、最も重要な知的財産及び企業秘密をどのように保護していますか（専有及び未公開のモデルウェイトへのアクセスを制限する等）。また、モデルウェイト・アルゴリズム・サーバー・データセット、またはその他の関連する要素について、安全に保管・運用するためにどのような仕組みを整えていますか。これらが適切に保護された環境で、限られたアクセス権のもと、管理するための対策を教えてください。
g-iii	What is your organization’s vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?	脆弱性管理プロセスはどのようなものですか。他の利害関係者との協働を含め、特定されたリスクや脆弱性に対処するための行動をとっていますか。
g-iv	How often are security measures reviewed?	セキュリティ対策はどれくらいの頻度で見直していますか。
g-v	Does your organization have an insider threat detection program?	内部脅威検知プログラムを持っていますか。
h	How does your organization address vulnerabilities, incidents, emerging risks and misuse across the AI lifecycle, including after the deployment of advanced AI systems?	AIライフサイクルにおける脆弱性・インシデント・新たなリスクや悪用について、AIシステムの導入後も含めてどのように対処していますか。

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～　設問a

設問a

AIのライフサイクル全体にわたるリスク・脆弱性に対処するために、どのような措置を講じていますか。
What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

設問の意図

- AIのライフサイクル全体にわたるリスク・脆弱性に対処するために、どのような措置を講じているかを具体的に説明することが求められています。

回答にあたっての考え方

- セクション1設問bと近しい設問内容となっていますが、本設問では、リスク・脆弱性に対処するための体制・プロセス・技術を含むリスク管理の取組を示すことが想定されます。
- 多くの組織では、AIのライフサイクル全体を通じて、体制・プロセス・技術を組み合わせたリスク管理対策例が見受けられます。
- 一部の組織では、段階的にリリース対象範囲を広げているアプローチ例も見受けられます。

回答の要素

1 体制を示す例

- NIST AI RMF等のドキュメントを参照したフレームワークの導入や自社のガイドラインを定めていることなどを示すことが想定されます。

2 設計・開発段階の取組の例

- リスクアセスメント、ユースケース審査・倫理レビュー、脅威モデリング、リスク予防策の組み込み、設計レビュー、ベンチマーク評価、モデルのファインチューニングなどを実施していることを示すことが想定されます。

3 リリース前・テスト段階の取組の例

- 複数のテストを含むリスク評価、リスク評価に応じた対策の実施やリリース判定を行っていること示すことなどが想定されます。

4 リリース後・運用段階の取組の例

- 利用者への情報の提供、継続的モニタリング、インシデント対応を行っていることなどを記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct' (2025), P.14

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～　設問b

設問b

テスト結果は、特定されたリスクへの対策の実施にどのように反映されていますか。
How do testing measures inform actions to address identified risks?

設問の意図

- テスト結果が特定されたリスクへの対策に反映されているかを説明することが求められています。

回答にあたっての考え方

- テスト結果をシステム改善や運用手順の変更に反映している例が見受けられます。
- 社内規程で、定期的なシステムや運用手順の更新について定めている例が見受けられます。

回答の要素

1 システム改善を行っている例

- テスト結果は関連チーム内に共有され、新たなガードレール作成やモデルの改善等、技術的な改善に反映されていることが想定されます。

2 運用手順を変更している例

- テスト結果に応じて、インシデント対応手順等の運用プロセスを更新することが想定されます。
- また、テスト結果は、リスク登録記録や倫理ガイドライン、運用ポリシー等の規程・ガバナンスの見直しに反映されることが想定されます。

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～　設問c

設問c

安全な環境でテストを実施している場合、いつ、どのように実施していますか。
When does testing take place in secure environments, if at all, and how?

設問の意図

- 安全な環境でのテスト実施がどのようなタイミング、方法で実施されているかについて説明することが求められます。

回答にあたっての考え方

- 多くの組織が、安全でコントロールされた環境（しばしば本番環境から隔離された環境）でテスト実施していると見受けられます。
- 一部の組織は、外部テスターに対してテスト環境へのアクセスを許可している例も見受けられます。

回答の要素

1 実施タイミングの例

- プレリリース時、大幅な更新前、事故後の再現テスト等のタイミングを記載することが想定されます。

2 安全な環境でのテスト実施方法の例

- 開発・テスト環境と本番環境を分離し、バリデーションチェック、データの暗号化、監視、アクセス権限の管理等が整備された専用テスト環境で実施していること等を示すことが想定されます。
- 外部テスター運用、テストデータの匿名化といった内容についても記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.15

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～　設問d

設問d

訓練及びデータ収集のプロセスを含み、どのようにデータの質を確保し、また、有害なバイアスのリスクを緩和していますか。
How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

設問の意図

- データの質を確保し、また、有害なバイアスのリスクを緩和させるために、データ収集・訓練プロセスのほか運用時等において、どのような対策を講じているかを具体的に説明することが求められています。

回答にあたっての考え方

- 具体的な対策としては、データ収集時におけるフィルタリング、訓練段階でのモデルの微調整、運用段階のモデレーションツールを使用したシステム出力の管理等が見受けられます。

回答の要素

1　データ収集・訓練時におけるリスク緩和の例

- 有害な偏見、差別、プライバシー・著作権侵害となり得るデータを収集・訓練データからフィルタリングすることが想定されます。
- 人間のフィードバックによる強化学習により、有害なリスクを緩和することも想定されます。

2　運用段階におけるリスク緩和の例

- 危険なプロンプトや応答、出力バイアスを検出するためのガードレールを設置
- 公平性ダッシュボードで属性別指標を常時監視し、逸脱時は一時停止・是正（再学習、ルール更新）

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.15

3.設問別ガイダンス
3-2.セクション2 ～リスク管理と情報セキュリティ～ 設問e

設問e

知的財産（著作物等）をどのように保護していますか。
How does your organization protect intellectual property, including copyright-protected content?

設問の意図

- 知的財産（著作物等）に関するリスクについて、どのような方法で低減させているかを説明することを求められています。

回答にあたっての考え方

- 多くの組織が、データの収集段階、訓練段階、出力段階のそれぞれで知的財産権を侵害するリスクの緩和策を講じていると見受けられます。

回答にあたって盛り込む要素

1 技術的な取組を行っている例

- 以下のような取組を記載することが想定されます。
  - 違法コンテンツやポリシーに反するドメインをフィルタリングしている。
  - robots.txtでモデル訓練のためのデータ利用を権利者がオプトアウトしている場合、これを尊重している。
  - 学習データの重複除去や正則化、出力時の類似度チェック等により、逐語的再現（メモリゼーション）を抑止する。
  - オンラインで入手可能な歌詞や書籍を抜粋して提供しない仕組みを設ける。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.15

3.設問別ガイダンス
3-2.セクション2　～リスク管理と情報セキュリティ～　設問f

設問f

プライバシーをどのように保護していますか。機密／センシティブ情報の漏洩をどのように防止していますか。
How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

設問の意図

- プライバシー保護の観点から、個人情報・機密情報の漏えい防止として、どのような対策を講じているかを説明することを求められています。

回答にあたっての考え方

- 多くの組織においては、プライバシー保護に優先して取り組むことを述べており、個人データに対するユーザーの管理権限を強調しているケースが多く見受けられます。
- 技術的保護措置を設けている例が多く見受けられます。
- プライバシー保護について契約条項で規定している例も多く見受けられます。

回答の要素

1　技術的保護措置を設けている例

- データの暗号化や匿名化に関する取組を記載することが想定されます。
- データ保持ゼロの原則を採用している企業では、ユーザーの入力や出力が大規模言語モデルや関連ログに一切保存されないことを記載する例も見受けられます。

2　契約上の合意による取組を示している例

- 契約上、個人データをAIモデルの訓練に利用しない旨定めることが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.15

3.設問別ガイダンス
3-2.セクション2 ～リスク管理と情報セキュリティ～ 設問g(1/2)

設問g（小分類i～vを含む）

運用セキュリティ及びサイバー／物理セキュリティに関する、AI特有の情報セキュリティ対策をどのように実施していますか。

i. 高度なAIシステムのサイバーセキュリティ強化に向け、サイバーセキュリティのリスク評価、ポリシーの導入をどのように行っていますか。

ii. セキュリティリスクから、最も重要な知的財産及び企業秘密をどのように保護していますか（専有及び未公開のモデルウェイトへのアクセスを制限する等）。また、モデルウェイト・アルゴリズム・サーバー・データセット、またはその他の関連する要素について、安全に保管・運用するためにどのような仕組みを整えていますか。これらが適切に保護された環境で、限られたアクセス権のもと、管理するための対策を教えてください。

iii. 脆弱性管理プロセスはどのようなものですか。他の利害関係者との協働を含め、特定されたリスクや脆弱性に対処するための行動をとっていますか。

iv. セキュリティ対策はどれくらいの頻度で見直していますか。

v. 内部脅威検知プログラムを持っていますか。

How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?

ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?

iii. What is your organization’s vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

iv. How often are security measures reviewed?

v. Does your organization have an insider threat detection program?

3.設問別ガイダンス
3-2.セクション2 ～リスク管理と情報セキュリティ～ 設問g(2/2)

設問g（ i～vの詳細は前頁に記載）

運用セキュリティ及びサイバー／物理セキュリティに関する、AI特有の情報セキュリティ対策をどのように実施していますか。
How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

設問の意図

- AI特有の脅威を踏まえ、情報セキュリティ対策として、どのような施策を取っているかを説明することが求められています。回答にあたっては、サイバーセキュリティに関するリスク評価・ポリシーの導入や、データ保護方法、脆弱性管理プロセス、情報セキュリティ対策の見直し頻度、内部脅威検知プログラム等の観点から回答することが求められています。

回答にあたっての考え方

- 情報セキュリティ対策について、g(ｉ)~(ｖ)の5つの観点から具体的な対応策について記載することが求められています。
- 回答にあたっては、自組織の情報セキュリティ対策の一般的な施策に加えて、AI固有のリスクに対する施策の例も見受けられます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.15

回答の要素

1 g-i：サイバーセキュリティのリスク評価、ポリシーの導入の例

- 業界標準、ガイドライン、フレームワーク等に基づきリスク評価を行い、生成AI固有の攻撃に備え、入出力の検証、権限・コンテキストの分離、データの健全化をしている例が想定されます。

2 g-ii：データ保護の例

- 重要資産は隔離環境で暗号化保管し、最小権限・多要素認証でアクセスを限定することなどが想定されます。その他関連する要素は、暗号保護やリアルタイム監視等により管理している例を記載することが想定されます。

3 g-iii：脆弱性管理プロセスの例

- 継続的な監視、自動セキュリティスキャン、ペネトレーションテスト、外部からの情報収集等の取組が想定されます。

4 g-iv：セキュリティ対策の見直し頻度の例

- 定期的な見直し、脆弱性検出時の見直し等が想定されます。

5 g-v：内部脅威検知プログラムの例

- UEBA等の内部脅威検知と監査ログを用いて、早期に内部脅威を検知・是正することが考えられます。

3.設問別ガイダンス
3-2.セクション2 ～リスク管理と情報セキュリティ～ 設問h

設問h

AIライフサイクルにおける脆弱性・インシデント・新たなリスクや悪用について、AIシステムの導入後も含めてどのように対処していますか。
How does your organization address vulnerabilities, incidents, emerging risks and misuse across the AI lifecycle, including after the deployment of advanced AI systems?

設問の意図

- AI導入後を含むライフサイクル全体を通して発生し得るリスク等に対し、どのように対処しているかを具体的に説明することが求められています。

回答にあたっての考え方

- セクション1設問bはAIライフサイクル全体でのリスクの特定と評価方法に関して回答するのに対し、本設問では、特定・評価されたリスクへどのように対処しているかについて説明することが求められています。
- 多くの組織において、AIに特化した多層的なリスク対策を採用していることが見受けられます。第三者機関による監査を利用している企業も見受けられます。

回答の要素

1 リスク対策の実施の例

- リスク対策の実施例として、ゼロトラストアーキテクチャ、アクセス制御、レッドチーム演習、ペネトレーションテスト、セキュリティアラート、リアルタイム監視、内部脅威検知等を含む複数のアプローチを実施していることが想定されます。
- ISMS（ISO/IEC 27001）等を準拠として記載することも想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.15

3.設問別ガイダンス

3.設問別ガイダンス
3-3.セクション3　～高度なAIシステムに関する透明性報告～

- Section3-Transparency reporting on advanced AI systems／セクション3～高度なAIシステムに関する透明性報告～

#	設問内容（英語）	日本語訳
a	Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?	高度なAIシステムの能力、限界、適切・不適切な使用領域に関する明確で理解しやすい報告書及び／または技術文書を公表していますか。
a-i	How often are such reports usually updated?	通常、このような報告書はどのくらいの頻度で更新していますか。
a-ii	How are new significant releases reflected in such reports?	このような報告書において、重要な新規公表はどのように反映されているか。
a-iii	Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.	以下の情報のうち、公表している文書に含まれているものはどれですか： ・潜在的な安全性、セキュリティ、社会的リスク及び人権の享受に対するリスクについて実施された評価の詳細及び結果 ・モデルやシステムの安全性や社会に対する影響やリスク（有害な偏見、差別、プライバシーや個人情報保護への脅威、公平性に関連するもの等）についての評価 ・開発段階以降のモデル／システムの適合性を評価するために実施されたレッドチーミング又はその他のテストの結果。 ・適切な使用領域に影響を及ぼすモデル／システムの能力及び性能上の重大な限界 ・ 関連するその他の技術文書や使用説明書

3.設問別ガイダンス
3-3.セクション3　～高度なAIシステムに関する透明性報告～

- Section3-Transparency reporting on advanced AI systems／セクション3～高度なAIシステムに関する透明性報告～（2/2）

#	設問内容（英語）	日本語訳
b	How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?	高度なAIシステムに関連するリスクや影響の評価の結果について、多様なステークホルダー（他の組織・政府・市民社会・学術界等）とどのように情報共有を行っていますか。
c	Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?	個人データの使用・ユーザープロンプト、及び／または高度なAIシステムのアウトプットに対応するプライバシーポリシー（個人情報保護方針）を開示していますか。
d	Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?	高度なAIシステムの訓練に使用されるデータのソースに関する情報（データのアノテーションやエンリッチメントのソースに関する情報を含む）を提供していますか。
e	Does your organization demonstrate transparency related to advanced AI systems through any other methods?	高度なAIシステムに関連する透明性について、他の方法でも示していますか。

3.設問別ガイダンス
3-3.セクション3 ～高度なAIシステムに関する透明性報告～ 設問a(1/2)

設問a（小部類i～iii含む）

高度なAIシステムの能力、限界、適切・不適切な使用領域に関する明確で理解しやすい報告書及び／または技術文書を公表していますか。

i. 通常、このような報告書はどのくらいの頻度で更新していますか。

ii. このような報告書において、重要な新規公表はどのように反映されているか。

iii. 以下の情報のうち、公表している文書に含まれているものはどれですか：
・潜在的な安全性、セキュリティ、社会的リスク及び人権の享受に対するリスクについて実施された評価の詳細及び結果
・モデルやシステムの安全性や社会に対する影響やリスク（有害な偏見、差別、プライバシーや個人情報保護への脅威、公平性に関連するもの等）についての評価
・開発段階以降のモデル／システムの適合性を評価するために実施されたレッドチーミング又はその他のテストの結果。
・適切な使用領域に影響を及ぼすモデル／システムの能力及び性能上の重大な限界
・ 関連するその他の技術文書や使用説明書

Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

i. How often are such reports usually updated?

ii. How are new significant releases reflected in such reports?

iii. Which of the following information is included in your organization’s publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model’s or system’s effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model’s/system’s fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.

3.設問別ガイダンス
3-3.セクション3 ～高度なAIシステムに関する透明性報告～ 設問a(2/2)

設問a（a i～iiiの詳細は前頁に記載）

高度なAIシステムの能力、限界、適切・不適切な使用領域に関する明確で理解しやすい報告書及び／または技術文書を公表していますか。
Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

設問の意図

- 本設問は、高度なAIシステムについて、「何ができるか（能力）」「何ができないか（限界）」「どのような利用が適切/不適切か（利用領域）」を、ユーザーにとって明確で理解しやすい形で公表しているかを確認するものです。（a-i）更新頻度の例、（a-ii）重要リリースの反映の例、（a-iii）公表している情報の例を記載することが想定されます。

回答にあたっての考え方

- モデル・システムカード、透明性報告書等を公表する例が見受けられる一方、BtoB契約書の中で情報提供する例も見受けられます。

回答の要素

1 a-i：更新頻度の例

- 四半期・年次等の定期更新、AIシステムの重大リリース等における随時更新が想定されます。
- また、変更履歴ついて記載されている例が見受けられます。

2 a-ii：重要リリースの反映の例

- 重要リリースが行われる場合、反映される報告書等に加え、Web更新通知、リリースノート等の周知方法の記載も想定されます。

3 a-iii：公表している情報

- a-iiiの設問で列挙された情報を含む報告書等（モデル・システムカード、透明性報告書、技術文書、使用説明書、FAQ等）、概要及び参照先(URL)を記載することが想定されます。
- また、公表していない場合は、その理由を記載する例が見受けられます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.16,17

3.設問別ガイダンス
3-3.セクション3 ～高度なAIシステムに関する透明性報告～ 設問b

設問b

高度なAIシステムに関連するリスクや影響の評価の結果について、多様なステークホルダー（他の組織・政府・市民社会・学術界等）とどのように情報共有を行っていますか。
How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

設問の意図

- 高度なAIシステムに関するリスクや影響の評価結果について、社内や利用者にとどめず、他の組織・政府・市民社会・学術界等の多様なステークホルダーを対象に、どのような相手に、どのような方法で情報共有しているかを具体的に説明することを求めています。

回答にあたっての考え方

- 情報共有は、研究成果の公表から専用の透明性プラットフォームに至るまで、幅広い方法が用いられています。
- 多様なステークホルダーとの情報共有は一般的に行われており、多くの組織が、他の組織、政府、学術界との知見を共有しています。一方で、NGO等の市民社会組織と情報共有を報告している組織は少ない状況です。

回答の要素

1 情報共有している例

- 情報共有について、以下のような内容を記載することが想定されます。
  - 情報共有しているステークホルダー
  - 共有方法（公表文書、学術論文、ブログ、説明会・イベント、フォーラム参加、個別説明等）
  - 共有内容（対象リスク、リスク・影響の評価結果、緩和策・対応方針等）
  - 公表している場合には参照先（URL）

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.17

3.設問別ガイダンス
3-3.セクション3　～高度なAIシステムに関する透明性報告～　設問c

設問c

個人データの使用・ユーザープロンプト、及び／または高度なAIシステムのアウトプットに対応するプライバシーポリシー（個人情報保護方針）を開示していますか。
Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

設問の意図

- 個人データ、ユーザープロンプト及び高度なAIシステムのアウトプットの取扱いについて、プライバシーポリシー（個人情報保護方針）として利用者へ開示しているかを説明を求めるものです。

回答にあたっての考え方

- 本設問におけるプライバシーポリシー（個人情報保護方針）の対象は個人データに加え、生成AI特有の論点であるユーザープロンプト及び高度なAIシステムのアウトプットの取扱いにも及びます。
- BtoB契約書の中で開示する例も見受けられます。

回答の要素

1　開示している例

- プライバシーポリシー（個人情報保護方針）の参照先（URL）及び概要を記載することが想定されます。
- 概要として、個人データ、ユーザープロンプト、アウトプット等に関わる規律を抜粋する例が見受けられます。

3.設問別ガイダンス
3-3.セクション3 ～高度なAIシステムに関する透明性報告～ 設問d

設問d

高度なAIシステムの訓練に使用されるデータのソースに関する情報（データのアノテーションやエンリッチメントのソースに関する情報を含む）を提供していますか。

Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

設問の意図

- 高度なAIシステムの訓練に用いたデータのソースついて、データのアノテーション（注釈付け）やエンリッチメント（拡充）に関する情報も含め、適切な範囲で情報を提供しているかの説明を求めています。

回答にあたっての考え方

- 訓練データのソースについて、モデル・システムカードや技術文書等で公表している例が見受けられますが、公表の粒度には組織差があります。
- 要約は公表しつつ、より詳細な情報（アノテーション、エンリッチメントを含む）については、契約を締結した顧客にのみ提供する例も見受けられます。

回答の要素

1 提供している例

- 公表している文書名（モデルカード、技術レポート、ブログ、arxiv論文）、概要及び参照先（URL）を記載することが想定されます。
- なお、限定して提供している場合には、提供先、要約、限定している理由を記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.18

3.設問別ガイダンス
3-3.セクション3　～高度なAIシステムに関する透明性報告～　設問e

設問e

高度なAIシステムに関連する透明性について、他の方法でも示していますか。
Does your organization demonstrate transparency related to advanced AI systems through any other methods?

設問の意図

- 設問3a～3d以外に透明性を高める取組を行っているかについて説明を求めるものです。

回答にあたっての考え方

- 3a～3d以外の方法で透明性を高める取組を行っている場合に、本設問へ回答することが想定されます。

回答の要素

3.設問別ガイダンス

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～

- Section4-Organizational governance, incident management and transparency／セクション4～組織のガバナンス、インシデント管理及び透明性～

#	設問内容（英語）	日本語訳
a	How has AI risk management been embedded in your organization's governance framework? When and under what circumstances are policies updated?	組織のガバナンスフレームワークにどのようにAIリスク管理を組み込んでいますか。方針はいつ、どのような状況で更新していますか。
b	Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?	関連するスタッフは、組織のガバナンス方針とリスク管理慣行について研修を受けていますか。もしそうなら、どのように研修を行っていますか。
c	Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?	リスク管理方針と実践を、ユーザー及び／または公に伝えていますか。伝えている場合、どのように伝えていますか。
d	Are steps taken to address reported incidents documented and maintained internally? If so, how?	報告されたインシデント対応の内容は、社内で文書化・管理していますか。もしそうなら、どのように文書化・管理していますか。
e	How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?	脆弱性・インシデント・新たなリスク・悪用に関する情報をどのように他者と共有していますか。
f	Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?	高度なAIシステムのインシデントについて、必要に応じて、関連する他のステークホルダーと情報共有を行っていますか。行っている場合、どのような方法で共有していますか。また、インシデントに関連する情報を公に共有し、報告していますか。
g	How does your organization share research and best practices on addressing or managing risk?	リスクへの対応または管理に関する研究やベストプラクティスをどのように共有していますか。
h	Does your organization use international technical standards or best practices for AI risk management and governance policies?	AIのリスク管理やガバナンスポリシーについて、国際的な技術標準やベストプラクティスを採用していますか。

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問a

設問a

組織のガバナンスフレームワークにどのようにAIリスク管理を組み込んでいますか。方針はいつ、どのような状況で更新していますか。
How has AI risk management been embedded in your organization's governance framework? When and under what circumstances are policies updated?

設問の意図

- 組織のガバナンスフレームワークへのAIリスク管理の組込みについて具体的な説明を求めています。あわせて、方針がいつ、どのような状況で更新されるのかについて説明が求められています。

回答にあたっての考え方

- AIリスク管理について、多くの組織が既存のリスク管理や品質管理の枠組みに統合している一方で、新たにAIに特化した枠組みを構築した例も一部見受けられます。
- 多くの組織で経営陣がAIリスクに関する意思決定に関与し、取締役会や委員会等によって監督している例が見受けられます。
- また、内部通報窓口設置、高リスク領域に係る審査プロセスの導入、サービス提供時における部門横断的なレビューを実施している例も見受けられます。
- 加えて、外部専門家も加わった委員会、弁護士・外部監査への相談等、外部専門知識を活用する例も見受けられます。

回答の要素

1　ガバナンスフレームワークと具体的対応の例

- 以下の記載が想定されます。
  - ガバナンス構造（取締役会・各種委員会等の監督主体）と意思決定に関するプロセス（リスク評価・重要リリース判断等にかかる報告・承認ルート）
  - 具体的なフレームワーク（NIST AI RMF、ISO/IEC 42001、EU AI法、独自で策定等）と対応内容（ライフサイクルを踏まえたレビューの実施、外部専門知識の活用等）

2　方針の更新の例

- 定期的な見直しに加えて、環境変化に応じた更新が想定されます。
  - 技術進歩や新たなリスクの顕在化、法令等の改正等
  - 規模の拡大、重大インシデントの発生、内部監査指摘等

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.18, 19

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問b

設問b

関連する職員は、組織のガバナンス方針とリスク管理慣行について研修を受けていますか。もしそうなら、どのように研修を行っていますか。
Are relevant staff trained on your organization’s governance policies and risk management practices? If so, how?

設問の意図

- ガバナンス方針とリスク管理慣行を熟知していることを確保するため、関連するスタッフに対して研修を実施しているか、また実施している場合にはどのように行っているかについて、具体的な説明を求めています。

回答にあたっての考え方

- ガバナンス及びリスク管理の実効性を担保するには、関連する職員におけるガバナンス方針の理解・浸透やリスク対応能力の向上が重要です。
- 研修プログラムとしては、年1回必須のセキュリティ及びプライバシーに関する研修のほか、専門的な認定資格取得のための研修や、シナリオベースの研修、ハッカソン、パープルチーム演習等を行ってる例が見受けられます。

回答の要素

1 研修プログラムの構成の例

- 全社研修、役割別研修、個別研修等の研修プラクラムを有している場合には、その点を記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.19

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問c

設問c

リスク管理方針と慣行を、利用者及び／または公に伝えていますか。伝えている場合、どのように伝えていますか。
Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

設問の意図

- AIリスク管理方針と慣行について、利用者や公に伝えているか、伝えている場合には、どのように伝えているかについて、具体的に説明を求めています。

回答にあたっての考え方

- AIモデル・システムのリスク及びその緩和策に関する透明性については、セクション３で取り扱っているため、本設問の回答では、AIリスク管理方針及び慣行（組織体制、プロセス）を中心に記載することが想定されます。
- 基本方針は公表しつつ、より詳細な管理方針と慣行については、契約書の中で顧客にのみ提供する例も見受けられます。

回答の要素

1 一般向けに公表している例

- 公表内容（リスク管理方針、リスク管理体制・手続きの整備状況）及び参照先（URL）を記載することが想定されます。

2 伝えている相手先を限定している例

- 伝えている相手先と方法、限定している理由を記載することが想定されます。

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問d

設問d

報告されたインシデント対応の内容は、社内で文書化・管理していますか。もしそうなら、どのように文書化・管理していますか。
Are steps taken to address reported incidents documented and maintained internally? If so, how?

設問の意図

- 報告されたインシデントへの対応が、属人的に終わらず、対応内容が文書化され（電子的記録を含む）、継続的に管理されているかについて説明を求めています。

回答にあたっての考え方

- 多くの組織で、インシデント対応手順が整備されているほか、監視・記録・対応する専門チームを設置している例も一部見受けられます。

回答の要素

1　システム等でインシデント対応を管理している例

- 専用システム等でインシデントを一元的・構造的に記録・追跡し、トリアージやエスカレーション、原因分析、クローズまでの運用プロセスを記載することが想定されます。
- 社内規程や手順書・様式を整備し、インシデントの基本情報を記録・保存していることを記載することが想定されます。
- 後から検証でき、再発防止に活用される形で残し、維持していることを記載することが期待されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.19

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問e

設問e

脆弱性・インシデント・新たなリスク・悪用に関する情報をどのように他者と共有していますか。
How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

設問の意図

- 脆弱性、インシデント、新たなリスク、悪用に関する情報を社外へ共有する方法についての説明が求められています。具体的な共有方法や共有先について回答することが想定されます。

回答にあたっての考え方

- 複数の組織で、ブログ記事や透明性レポートを通じてインシデント関連の最新情報を公表している例が見受けられます。その内容はサイバーセキュリティ上の脅威だけでなく、より広範なAIの安全性リスクやシステムの脆弱性についてもカバーしています。

- 米国では、Frontier Model Forumの参加企業が、フロンティアAIに関連・懸念される脆弱性、脅威等に関する情報共有を促進することに合意する等、競合他社と協働する例も見受けられます。

回答の要素

1　共有方法の例

- ブログ記事、透明性レポート、Webページ、システム画面等を通じてユーザや一般に広く共有していることを記載することが想定されます。
- 共有先としては、ユーザー、競合他社、業界団体、政府機関、研究機関等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.19, 20

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問f

設問f

高度なAIシステムのインシデントについて、必要に応じて、関連する他の利害関係者と情報共有を行っていますか。行っている場合、どのような方法で共有していますか。また、インシデントに関連する情報を公に共有し、報告していますか。
Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

設問の意図

- 高度なAIシステムでインシデントが発生し際に、ユーザー等のインシデントの影響を受ける関係者に対してどのように情報共有を行っているかについての説明が求められています。また、インシデント関連情報の公表方針についても説明する必要があります。

回答にあたっての考え方

- インシデントが発生した場合、情報共有を行う利害関係者の範囲や共有方法について記載することが想定されます。また、公表を行う際の基準、公表方法についても説明する必要があります。
- 広範な利害関係者へ情報共有するケースがある一方で、BtoB企業においては顧客へ個別に情報共有する例も見受けられます。
- 自組織において定められているインシデント対応フローに沿って記載することが想定されます。

回答の要素

1 広範な利害関係者へ情報共有する例

- 情報共有を行う利害関係者としては、顧客、ユーザー、行政機関、業界団体等が想定されます。
- 自組織のWebサイト等で公表している例を示すことが想定されます。

2 限定された範囲に情報共有する例

- 契約関係にある顧客やパートナー等、限られた利害関係者に情報共有している例を示すことが想定されます。
- 公への情報公表の基準や公表範囲についても説明することが求められます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.20

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問g

設問g

リスクへの対応または管理に関する研究やベストプラクティスをどのように共有していますか。
How does your organization share research and best practices on addressing or managing risk?

設問の意図

- リスクへの対応やリスク管理に関するベストプラクティスを社内外の関係者と共有する具体的な方法についての説明が求められています。

回答にあたっての考え方

- 多くの組織が自組織のWebサイト等を通じてリスク対応や管理に関するベストプラクティスを社外に公表していると見受けられます。団体等での活動を通して、業界全体でベストプラクティスを共有するケースも多く見受けられます。

回答の要素

1　共有方法の例

- Webサイト、ブログ、透明性レポート、論文、学会発表等による共有が想定されます。
- Frontier Model Forum、Partnership on AIといった団体等への参加を通じた情報交換についても記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.20

3.設問別ガイダンス
3-4.セクション4　～組織のガバナンス、インシデント管理及び透明性～　設問h

設問h

AIのリスク管理やガバナンスポリシーについて、国際的な技術標準やベストプラクティスを採用していますか。
Does your organization use international technical standards or best practices for AI risk management and governance policies?

設問の意図

- AIのリスク管理やガバナンスポリシーについて、国際的な技術標準やベストプラクティスの採用有無について説明することが求められています。

回答にあたっての考え方

- 大多数の組織が国際的な技術標準やベストプラクティス等を採用しており、内部規定やリスクアセスメント、ガバナンスポリシーへ反映していると見受けられます。

回答の要素

1　国際標準・ベストプラクティスの例

- 採用する国際標準・ベストプラクティスの例として、ISO/IEC 42001、NIST AI RMF、OWASP Top 10等が想定されます。

3.設問別ガイダンス

3.設問別ガイダンス
3-5.セクション5　～コンテンツの認証及び来歴確認の仕組～

- Section5-Content authentication & provenance mechanisms／セクション5～コンテンツの認証及び来歴確認の仕組～

#	設問内容（英語）	日本語訳
a	What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?	可能かつ適切な場合、ユーザーが高度なAIシステムと相互作用していることを知ることができるよう、どのような仕組みを設けていますか。
b	Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?	高度なAIシステムによって生成されたコンテンツをユーザーが識別できるように、コンテンツの来歴検出・ラベリング・透かしの仕組みを設けていますか。設けている場合、どのような仕組みですか。 また、コンテンツの来歴検出を開発または導入する際に、国際的な技術標準またはベストプラクティスを採用していますか。

3.設問別ガイダンス
3-5.セクション5　～コンテンツの認証及び来歴確認の仕組～　設問a

設問a

可能かつ適切な場合、ユーザーが高度なAIシステムと相互作用していることを知ることができるよう、どのような仕組みを設けていますか。

What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

設問の意図

- ユーザーがAIとやり取りしていることを認識できるようにするための仕組みについての説明が求められています。

回答にあたっての考え方

- 多くの組織が、製品内のメッセージ、免責事項等の目に見える手がかりを用いて、ユーザーがAIシステムとやり取りしていることを通知していると見受けられます。

回答の要素

1 通知方法の例

- 製品内のメッセージ、免責事項での通知が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD (2025) “How Are AI Developers Managing Risks?”, P.21

3.設問別ガイダンス
3-5.セクション5　～コンテンツの認証及び来歴確認の仕組～　設問b

設問b

高度なAIシステムによって生成されたコンテンツをユーザーが識別できるように、コンテンツの来歴検出・ラベリング・透かしの仕組みを設けていますか。設けている場合、どのような仕組みですか。
また、コンテンツの来歴検出を開発または導入する際に、国際的な技術標準またはベストプラクティスを採用していますか。
Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how?
Does your organization use international technical standards or best practices when developing or implementing content provenance?

設問の意図

- AI生成コンテンツをユーザーが識別できるようにする仕組みの有無と具体的な内容についての説明が求められています。また、仕組みの開発・実装における国際的標準・ベストプラクティスの採用有無についても回答することが求められています。

回答にあたっての考え方

- AI生成コンテンツを検証・追跡するための技術的ツールの開発を行っている例もあり、一部組織では既に実装されています。
- コンテンツ認証の取組は、国際的なフレームワークに準拠して行われるケースが多く見受けられます。
- BtoB組織においては、顧客がコンテンツ来歴を管理できるシステム機能を提供する例も見受けられます。

回答の要素

1 来歴検出・透かしを用いている例

- 透かし（コンテンツに埋め込まれた微細なデジタル信号）、メタデータタグ付け（コンテンツの由来を記述する情報を付加すること）、及びコンテンツクレデンシャル等が想定されます。

2 国際的なフレームワークに準拠している例

- C2PA、NIST等を参照していること等が想定されます。

3 製品機能として提供する例

- コンテンツ来歴を顧客が管理できるダッシュボードの提供等。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.22

3.設問別ガイダンス

3.設問別ガイダンス
3-6.セクション6　～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～

- Section6-Reserch & investment to advanced AI safety & mitigate societal risks／セクション6～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～

#	設問内容（英語）	日本語訳
a	How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?	高度なAIシステムのセキュリティ・安全性・バイアスと偽情報・公平性・説明可能性と解釈可能性・透明性・堅牢性及び／または信頼性についての研究・投資をどのように推進していますか。
b	How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?	コンテンツ認証及び来歴の技術を前進させるために、どのように研究に対して協力・投資していますか。
c	Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?	AIの安全性・セキュリティ・信頼性、及びリスク評価と緩和ツールの推進を支援する研究のプロジェクト・協力・投資に参加していますか。
d	What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?	AIがもたらす社会経済及び/または環境面でのリスクを最小化するために、どのような研究や投資を進めていますか。

3.設問別ガイダンス
3-6.セクション6　～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～　設問a

設問a

高度なAIシステムのセキュリティ・安全性・バイアスと偽情報・公平性・説明可能性と解釈可能性・透明性・堅牢性、及び／または信頼性についての研究・投資をどのように推進していますか。

How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

設問の意図

- 高度なAIシステムの信頼性を支える主要論点（セキュリティ、安全性、バイアス・偽情報、公平性、説明可能性・解釈可能性、透明性、堅牢性等）について、研究と投資をどのように進めているかについて具体的に説明を求めています。

回答にあたっての考え方

- 本設問は、安全性・公平性・堅牢性等の高度AIシステムの信頼性を支える横断テーマへの研究・投資を広く問う総論的な位置づけです。特に、研究対象のテーマを中心に記載することが想定されます。
- 大規模テクノロジー組織は、堅牢性・公平性・レッドチーミング・コンテンツ認証及び来歴・解釈可能性に焦点を当てた幅広い分野の取組を行っていると見受けられます。一方で小規模組織は、特定ユースケース及びセクター固有リスクに焦点を当てている傾向が見受けられます。
- 複数の組織では、サイバーセキュリティ、偽誤情報、バイアス検出の分野において、より具体的な研究・投資内容が記載されています。

回答の要素

1 研究・投資分野の例

- 研究・投資を行っている分野（列挙された分野、より具体的な重点分野）を記載することが想定されます。

2 推進手段の例

- 研究：論文公表、共同研究ネットワーク・社内研究組織の構築、評価手法開発等の例が見受けられます。
- 投資：助成金、データセンターの設置、ツール開発、オープンソース化等の例が見受けられます。

3 実践への反映や成果物の例

- 研究成果が開発・評価・運用にどう組み込まれるか（例：レッドチーム手法、評価手法・ベンチマーク、規程・基準書の更新）を可能な範囲で記載するほか、論文・オープンソースを公表している場合は参照先を記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025), P.23, 24

3.設問別ガイダンス
3-6.セクション6　～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～　設問b

設問b

コンテンツ認証及び来歴の技術を前進させるために、どのように研究に対して協力・投資していますか。
How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

設問の意図

- 生成AIの普及により、AI生成コンテンツが「本物か」「どこで作られ、どう改変されたか」を示す仕組み（認証及び来歴）の重要性が高まっています。認証及び来歴の技術を推進させるために、研究への協力・投資をどのように行っているかを具体的に説明することを求めています。

回答にあたっての考え方

- 前設問aが総論であるのに対し、本設問はコンテンツ認証及び来歴技術に関する協力・投資にテーマを絞っています。

回答の要素

1 標準化団体による貢献の例

- 参画している標準化団体（C2PA等）及びその主な活動内容を記載することが想定されます。

2 協働による貢献の例

- 協働先（大学、研究機関、産業パートナー等）、協働内容（共同研究、技術開発、ベストプラクティスの共有、フォーラム参加）、重点テーマ（偽情報分析技術、暗号署名、メタデータ追跡）、成果物（論文、オープンソース化等）等を記載することが想定されます。

3 自組織の取組による貢献の例

- 自組織の取組内容（調査、研究、開発）を記載することが想定されます。

3.設問別ガイダンス
3-6.セクション6　～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～　設問c

設問c

AIの安全性・セキュリティ・信頼性、及びリスク評価と緩和ツールの推進を支援する研究のプロジェクト・協力・投資に参加していますか。
Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

設問の意図

- 自組織内の取組にとどまらず、外部のプロジェクト・共同研究・研究投資を通じて、①AI安全性・セキュリティ・信頼性を向上させ、②リスク評価・緩和ツールの進歩を支えているかについて、説明を求めています。

回答にあたっての考え方

- 多くの組織が学術界・市民社会・政府の各セクターと協力していると見受けられます。

- 組織は国際的イニシアチブへの参画を通じて、多国間の安全性枠組みの形成に寄与しています。以下のような代表的な国際的イニシアチブの例が見受けられます。

- G7 広島AIプロセス
- NIST AI Safety Consortium
- C2PA
- Frontier Model Forum
- AI Verify Foundation

回答の要素

1 国際的イニシアチブへの参画の例

- 参画イニシアチブ先
- 対象テーマ（AI倫理・ガバナンス全般、セキュリティ、信頼性、リスク評価枠組み、軽減ツール等）
- 貢献内容（枠組み・ベストプラクティスの形成・共有、技術標準・仕様の策定・改善、イベント・会議・ワークショップを通じた議論参加）
- 取組と成果を公表している場合には参照先（URL）

2 研究プロジェクト・共同研究・投資による推進の例

- 協働先（学術界、政府、他社・業界内）
- 推進形態（共同研究、共同プロジェクト、資金・リソース提供、センター・ラボの設置・運営、コンテスト・ネットワークの運営等）
- 対象テーマ（サイバーセキュリティ・脅威対応、バイアス・公平性、プライバシー・知的財産権の保護、リスク評価・リスク軽減ツール）
- 取組と成果成果を公表している場合には参照先（URL）

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.24

3.設問別ガイダンス
3-6.セクション6　～AIの安全性向上や社会リスクの軽減に向けた研究及び投資～　設問d

設問d

AIがもたらす社会経済及び／または環境面でのリスクを最小化するために、どのような研究や投資を進めていますか。
What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

設問の意図

- AIの影響は社会経済・環境面にまで及ぶことが想定され、これらの負の影響を最小化する研究や投資を進めているかについて、具体的に説明することを求めています。

回答にあたっての考え方

- 環境面に配慮した研究・投資例が多く見受けられます。

- なお、国際行動規範第8項では優先的に対処すべき重要なリスクとして以下が例示されています。
  - 民主的価値の維持
  - 人権の尊重
  - 子どもや社会的弱者の保護
  - 知的財産権とプライバシーの保護
  - 有害な偏見
  - 偽・誤情報
  - 情報操作の回避　等

回答の要素

1 リスク最小化の取組の例

- リスク最小化の具体的な取組の例として以下の記載が想定されます。

  研究
  - AIの社会経済的影響の測定・分析
  - バイアスや不利益の検出・軽減に関する研究
  - 偽誤情報への対策に関する研究
  - 計算効率や資源効率を高めるアルゴリズムの研究

  投資
  - 省エネルギーなモデル設計や推論手法への投資
  - 再生可能エネルギーの利用
  - 効率的で持続可能なAIインフラへの投資

3.設問別ガイダンス

3.設問別ガイダンス
3-7.セクション7　～人類と世界の利益の促進～

• Section7-Advancing human and global interests／セクション7～人類と世界の利益の促進～

#	設問内容（英語）	日本語訳
a	What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.	AIがもたらす社会経済及び／または環境面での利益を最大化するために、どのような研究や投資を進めていますか。例を示してください。
b	Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.	ユーザーの意識を向上させる、及び／または高度なAIシステムの性質・能力・限界・影響を人々が理解できるようにするために、デジタルリテラシー・教育またはトレーニングの取組を支援していますか。例を示してください。
c	Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.	国連の持続可能な開発目標（SDGs）を支援するために、信頼できる人間中心のAIのスチュワードシップのAIプロジェクトを優先していますか。例を示してください。
d	Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.	国連の持続可能な開発目標（SDGs）を支援し、世界が直面する最も重要な課題に取り組むために、市民社会やコミュニティグループと協力してAIソリューションを特定・開発していますか。例を示してください。

3.設問別ガイダンス
3-7.セクション7　～人類と世界の利益の促進～　設問a

設問a

AIがもたらす社会経済及び／または環境面での利益を最大化するために、どのような研究や投資を進めていますか。例を示してください。
What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

設問の意図

- AIを活用して社会経済や環境面でのプラスの効果を最大化するために実施している研究や投資について、具体的に説明することが求められています。

回答にあたっての考え方

- 多くの組織が、気候変動、医療、教育等における多様な社会的課題を解決するためにAIを活用していると見受けられます。

回答の要素

1 研究・投資に関する取組の例

- 以下のような具体的な取り組み内容を記載することが想定されます。
  - エネルギー効率の高いAIモデルや環境に配慮したインフラへの投資
  - 創薬及び疾病予測用AIツールの開発
  - AIリテラシー向上ための教材・ツールの提供

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.25, 26

3.設問別ガイダンス
3-7.セクション7　～人類と世界の利益の促進～　設問b

設問b

ユーザーの意識を向上させる、及び／または高度なAIシステムの性質・能力・限界・影響を人々が理解できるようにするために、デジタルリテラシー・教育またはトレーニングの取組を支援していますか。例を示してください。
Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

設問の意図

- 高度なAIシステムの性質、能力、限界、影響についての理解促進、ユーザーの認識向上のために、組織としてどのような教育・トレーニングを実施しているかの説明が求められています。

回答にあたっての考え方

- 多くの企業が、特に十分な支援を受けていない層や技術に精通していないユーザーを対象に、AI教育へのアクセスを拡大し、AIへの理解を深めるためのトレーニングやツールを提供していると見受けられます。

回答の要素

1　提供する教育・トレーニングの例

- 教育の対象者としては、学生、教育者、企業ののほか、開発途上国や農村部等のデジタル資源へのアクセスが十分でない人々、非技術系ユーザーなどが想定されます。
- トレーニングとしては、デジタルリテラシーやデジタルスキル向上のための研修プログラム、オンラインコンテンツ、ワークショップ、ウェビナー・セミナー等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.26

3.設問別ガイダンス
3-7.セクション7 ～人類と世界の利益の促進～ 設問c

設問c

国連の持続可能な開発目標（SDGs）を支援するために、信頼できる人間中心のAIのスチュワードシップのAIプロジェクトを優先していますか。例を示してください。
Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

設問の意図

- 国連の持続可能な開発目標（SDGs）を支援するために、自組織で取り組んでいるAI関連のプロジェクトに説明することが求められています。

回答にあたっての考え方

- 多くの企業がSDGsの達成を重要視し、以下のような取組を行っていることが見受けられます。
  - 医療の行き届かない地域に対するAI搭載の遠隔医療ソリューションの提供
  - AIによる自然災害の予測と安全な非難経路の提案
  - AIを活用した学習支援による学習格差の縮小化
  - 情報へのアクセスと包摂性の向上を目的とした多言語AIモデルの開発
  - AIによる物流とサプライチェーンマネジメントの最適化

回答の要素

1 取組の例

- SDGsを支援するAI関連のプロジェクトの具体的な内容について記載することが想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, ‘How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct’（2025）, P.26

3.設問別ガイダンス
3-7.セクション7　～人類と世界の利益の促進～　設問d

設問d

国連の持続可能な開発目標（SDGs）を支援し、世界が直面する最も重要な課題に取り組むために、市民社会やコミュニティ・グループと協力してAIソリューションを特定・開発していますか。例を示してください。
Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

設問の意図

- SDG s の進捗を支援し、世界が直面する最も重要な課題を解決を目的とした、市民社会やコミュニティ・グループとの連携に関する設問です。市民社会やコミュニティ・グループとどのように優先課題を特定し、AIソリューションの開発を実施しているか、具体的に説明することが求められています。

回答にあたっての考え方

- 世界最大の課題とは、特に気候危機、世界保健、教育等を指します。（国際行動規範第９項）
- NGOやNPO等の市民団体や学術パートナーといった外部ステークホルダーと協力し、責任あるAIソリューションを共同設計するケースが拡大傾向にあります。
- 社会課題の解決に取り組む団体等に対して投資や資金援助することで、活動を支援する例も見られます。

回答の要素

1　取組の例

- 連携の対象としては、NGO、NPO、地域コミュニティ、研究機関、教育機関、学術団体等が想定されます。
- 取組としては、地域ニーズに即した医療・教育ツールのカスタマイズ、地域の洪水予測等の災害対策支援、投資による活動支援等の各社様々な例が見受けられます。
- 投資の例としては、AIを活用してSDGsの達成を目指す組織向けの資金提供、アクセラレータプログラムを通じたSDGsに貢献するAI搭載ソリューションへの投資等が想定されます。

※「回答にあたっての考え方」は、OECDペーパーも参考としています。
OECD, 'How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct'（2025）, P.26

参考文献一覧

- 江間有沙・工藤郁子・実積寿也、「AIガバナンスに資する透明性レポートハンドブック（第1.0版）」（2025）
  東京大学東京カレッジ江間研究室、ライセンス：CC BY 4.0.
  https://www.tc.u-tokyo.ac.jp/blog/wp-content/uploads/2025/12/HAIP_Handbook_JP.pdf
- 総務省、「広島AIプロセスについて」（2023）
  https://www.soumu.go.jp/main_content/000919809.pdf
- 外務省、「G7広島首脳コミュニケ」（2023）
  https://www.mofa.go.jp/mofaj/files/100507034.pdf
- 総務省、「全てのAI関係者向けの広島プロセス国際指針」（2023）
  https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03.pdf
- 総務省、「高度なAIシステムを開発する組織向けの広島プロセス国際行動規範」（2023）
  https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05.pdf
- 総務省、「広島AIプロセス」Webサイト
  https://www.soumu.go.jp/hiroshimaaiprocess/index.html
- Canada、「Toolkit for Small and Medium-Sized Enterprises (SMEs) Deploying Artificial Intelligence (AI)」（2025）
  https://ised-isde.canada.ca/site/ised/sites/default/files/documents/g7-mtl-toolkit-en.pdf
- OECD、「How are AI developers managing risks? –Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct」（2025）
  https://www.oecd.org/en/publications/how-are-ai-developers-managing-risks_658c2ad6-en.html
- OECD、「Hiroshima AI Reporting Framework」Webサイト
  https://oecd.ai/en/transparency/overview