Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

AI システムの透明性

ホワイトペーパー

変更履歴

バージョン	日付	名称	説明
1.0	01.07.2024	オリバー・ミュラー博士とヴェロニカ・ラザール	

表 *1*：変更履歴

連邦情報セキュリティ局

P.O. Box 20 03 63
53133 ボン referat-
t26@bsi.bund.de

インターネット：https://www.bsi.bund.de

補足

内容

補足

言葉の使い方に関する注意：読みやすさを考慮し、女性形と男性形は同時に使用していない。すべての情報は、すべての性別のメンバーについて言及している。

1　はじめに

本ホワイトペーパーでは、人工知能（AI）を統合した情報技術システムの透明性の定義を示す。本書の目的は、透明性という用語の共通理解を深め、様々な利害関係者と BSI にとっての透明性の関連性を強調することである。したがって、本稿は AI システムのすべての利害関係者を対象としており、とりわけ、利害関係者によっても透明性に関する要求が異なる可能性があることを示すことを意図している。

1.1　動機

AI は今や、プライベートとプロフェッショナルの両分野でデジタルツールとしての地位を確立している。スマートウォッチを使って個人の必要カロリーを測定したり、カスタマー・エクスペリエンスを向上させるために電話を自動的に転送したり、コンピューター・ネットワークにおける不審な活動を検知したりと、AI はあらゆる分野に浸透しており、応用可能な分野の例は常に増え続けている。これは、AI モデルのトレーニング、テスト、検証に利用できるデータ（ビッグデータ）の量と、それに対応するコンピューティングパワーを提供できるハードウェアリソースが利用可能になったことで可能になった。技術的な可能性が高まるにつれ、AI ベースのソリューション、特に効率性と生産性の向上に対する需要も高まっている。この需要は現在、特に増え続ける AI 新興企業によって満たされている。その結果、利用可能な AI システムの数は絶えず増加しており、その技術も急速に発展し、複雑さを増している。

抽象的な言い方をすれば、これらのシステムのほとんどはブラックボックスと呼ばれるもので、システムに対する入力とシステムからの出力だけが外界から見える（例えば、（Ribeiro, 2016）を参照）。システムがどのようにして出力に到達するのかは通常不明確なままであり、理解できないことが多い。加えて、出力の真実性は検証不可能であることが多く、システムの出力がどのようにしてもたらされたかを理解することを困難にしている。システムの複雑さと、システムに関する情報の欠如や不十分さにより、視覚的に評価することも、アウトプットの信頼性を判断することも難しくなっている。

AI システムの開発で使用される技術により、学習データに関する情報などの追加情報も関連してくる。例えば、機械学習モデルが使用する学習データセットを攻撃者が操作するデータポイズニング攻撃のリスクをシステム上で評価するためには、AI システムが使用される前に、学習データの出所と品質を評価できなければならない。

まとめると、これらの要素は、適切なトレーサビリティと説明可能性を可能にする AI システムの開発と使用を必要とする。両者はしばしば、透明性という関連基準と密接に関連している（第 2 節参照）。透明性は、AI システムの信頼性という広い分野にテーマとして組み込まれている。様々な基準は互いに明確に区別することはできないが、それぞれが異なるトピック領域に焦点を当てている。この重複を図 1 に示す。本ホワイトペーパーは、AI システムの透明性というトピックに焦点を当てている。

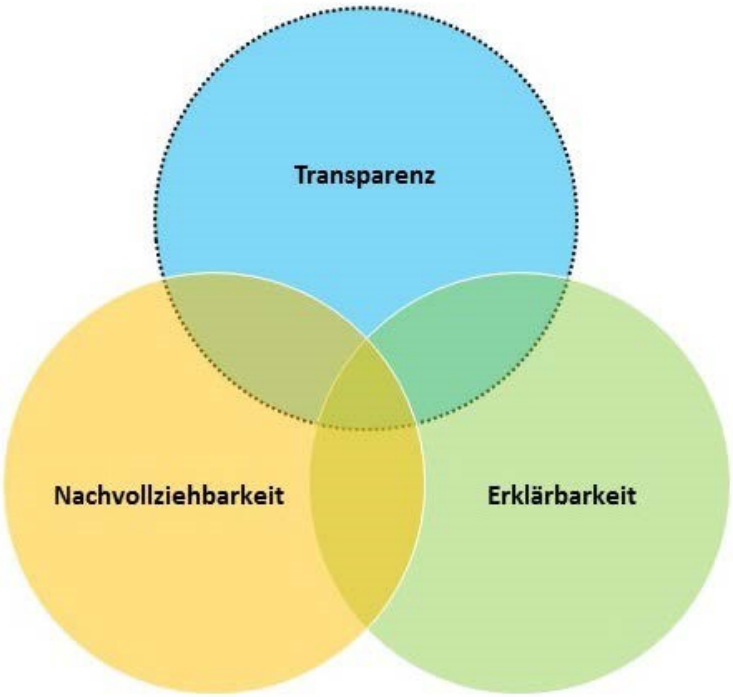

図 *1*：*AI* システムの信頼性の文脈における透明性、説明可能性、トレーサビリティの関係を示すベン図。異なる領域は重なり合っているが、それぞれに焦点が当てられている。

以下では、AI システムの文脈における透明性の概念を定義し、定義の各要素について議論する。次に、我々のアプローチと手順を説明し、欧州議会および理事会の人工知能に関する規則（AI 規則）における透明性要件との関連性を確立し、透明な AI システムの機会とリスクを強調する。

2　定義

AI システムの透明性とは、AI システムとそのエコシステムのライフサイクル全体に関する情報を提供することである。透明性は、すべての利害関係者のさまざまなニーズや目的に対するシステムの評価を可能にする情報へのアクセスを促進する。

2.1　エレメント

上記の定義は、（OECD, 2019）および（BSI, 2021a）の透明性の概念に基づいており、AI 規則の透明性要件に準拠し（詳細は 3.3 項参照）、BSI の立場を表している。定義の各要素については以下のサブセクションで詳述し、それらの関係を図 2 に図示した。

2.1.1　AI システム

人工知能の規制に関する AI 規則では、AI システムを「様々な程度の自律性で動作するように設計され、運用開始後は適応可能であり、明示的または暗黙的な目標のために受け取った入力から、物理的または仮想的環境に影響を与えることができる予測、コンテンツ、推奨、または決定などの出力を生成する方法を導き出す機械ベースのシステム」と定義している（AI 規則第 3 条参照）。BSI は、その定義においてハードウェアの要素を明確に定式化しており、AI システムを、物理的またはデジタル的な世界で「合理的」に行動するために人工知能を使用するソフトウェアおよびハードウェアシステムと定義している（BSI, 2021b）。AI と呼ばれるこれらのシステムに統合される技術には、機械学習、機械推論、ロボット工学など、さまざまなアプローチや技術が含まれる。AI システムで使われるエキスパート・システムやニューラルネットワークもこれに含まれる。これは、使用されているテクニックの網羅的なリストではないが、さまざまなテクニックの豊富さを示すことを意図している。単純なタスクから非常に複雑なタスクまで、AI システムのさまざまな機能性を見れば、同じ範囲が明らかになる。AI システムは、パターン認識、分類、予測、レコメンデーション、自然言語処理、コンピューター・ビジョンなどのタスクを実行することができ、さまざまな方法で互いに組み合わせることもできる。AI はまた、アプリケーションの要件や目的に応じて、さまざまな方法でシステムに実装することができる。一方では、例えばチャットボット・アプリケーションのように、独立したアプリケーションとして開発・使用し、システムの主要機能を表すこともできる。一方では、バックグラウンド・プロセスで機能を拡張したり、パフォーマンスを向上させたりするために、既存のシステムに統合することもできる。AI システムの実装タイプを見ると、自動化の程度も大きく異なる。AI の出力を推奨事項としてのみ使用し、最終的な意思決定権限として人間を必要とするシステムもあれば、人間が介入することなく AI の決定や分類を自律的に使用または実行する（サブ）システムもある。全体として、単一の AI システムというものは存在せず、さまざまな技術、機能、実装形態が豊富にあるということができる。

2.1.2　エコシステム

本稿では、エコシステムという用語は、AI システムが開発され、提供され、運用される文脈を指す。AI システムのエコシステムに関連する情報は、実際の AI システムにとどまらず、例えば、提供者の詳細（所在地、連絡先など）やシステムの開発プロセスも含むべきである。また、AI システムのサプライチェーン全体も含めるべきである。AI システムのエコシステムに関する情報も透明性の定義に含めるという決定は、実際の AI システムとそのエコシステムには（条件付きの）依存関係があるという事実に基づいている。例えば、AI システムが EU 域外の第三国で開発・運用されている場合、基盤となる IT セキュリティやデータ保護レベルに関して、対応する疑問や課題が生じる。このようなメタ情報は、十分な根拠に基づいた状況の評価と、利害関係者による情報に基づいた決定をサポートすることができる。

2.1.3　情報

情報は、利害関係者が AI システムとそのエコシステムの評価に到達するために必要な知識の基礎である。情報は、この目的のために利用できるように開示され、提供されなければならない。さらに、情報は知識を得るために適切でなければならない。3.3 項では、AI 規則における透明性要件について説明し、特定の AI システムの提供者及び運営者が開示しなければならない最低限の情報を示す。

2.1.4　ライフサイクル

ISO/IEC 22989:2022）によると、AI システムのライフサイクルは様々な段階から構成されており、透明性の概念との関連で以下に簡単に説明する。

2.1.4.1　企画・構想段階

AI システムの計画段階から透明性を考慮することが望ましい。こうすることで、ライフサイクルの後の段階での時間のかかる手戻りを最初から避けることができる。同時に、関係者は最初から関与することになる。計画段階は、AI システムとその使用に関する具体的な計画ができた時点で終了する。

2.1.4.2　設計、開発、検証段階

AI システムは、計画に基づいて開発、実施、テスト、検証される。計画された透明性と対応策が当初から考慮され、実施され、テストされる場合、これは設計による透明性と呼ばれる。対応策によって、透明性が望ましい程度に達成されない場合、AI システムおよび／または情報状況を再調整することができる。

2.1.4.3　試用と申請段階

開発／検証段階が完了すると、AI システムはロールアウトされ、生産的な運用に移される。運用中、利害関係者に関連するすべての情報が利用可能でなければならず、反復ループのための適切なオプションも利用可能でなければならない（2.1.4.4 節と 2.1.4.5 節を併せて参照）。

2.1.4.4　継続的評価段階

AI システムは動的であり、要件や環境は絶えず変化する可能性があるため、評価フェーズはアプリケーションフェーズの開始からシームレスに続かなければならない。自己変化するシステムの場合、永続的なモニタリングという意味で、適用段階と並行して行われる。関係者からのフィードバックは、損傷、緊急事態、故障の発生時などの評価に不可欠である。この目的のために、適切な対応とフィードバックのオプションが利用できなければならない（2.1.4.2 項と 2.1.4.3 項参照）。

2.1.4.5　システム更新

評価段階（2.1.4.4 節参照）で得られた結果に応じて、必要に応じて計画段階（2.1.4.1 節参照）を再開することができる。バグフィックスや性能向上に加え、更新はしばしば新機能をもたらす。既存の機能が削除されたり、変更されたり、他のモジュールに移されたりすることもある。この場合、既存の透明性と対応措置が、これらの適応プロセスによって、意図した機能が制限されたり、使用できなくなったりしないようにしなければならない。新たな機能に対しては、新たな手段を提供しなければならない。

AI システムを再教育する場合、透明性を確保するためにさらなる対策が必要となる。この措置は、例えば、新たに使用される訓練データセットに関連するものであり、差別／偏見と適性というトピックに特に関連するものである。

2.1.4.6　廃棄措置

レガシーシステムの廃棄には、(i)レガシーシステムを継続せずにシャットダウンする、(ii)新システムに移行する、という 2 つの選択肢がある。いずれの場合も、不要になったデータや移行されたデータがどのように扱われるのか、また、廃棄によって利害関係者にどのような変化や影響がもたらされるのかを透明化しなければならない。さらに、利害関係者は、データ主体としての権利を主張するなど、ここで対応するための選択肢を持たなければならない。

2.1.5　ニーズと目標

これらは個々のものであり、様々なユースケースにおいて大きく異なる可能性がある。この用語は、透明性が特定の情報へのアクセスを可能にすることを意図しているのではなく、それぞれの利害関係者が評価を行うことを可能にする情報が提供されることを説明することを意図している。各利害関係者が具体的に何を評価するかは、個別かつ文脈に依存する。利害関係者のニーズと目的は、適用シナリオによって異なる。目的は、利害関係者の特定のニーズに関してシステムの評価を可能にする、望ましいシステム情報を幅広くカバーすることである。

2.1.6　ステークホルダー

ステークホルダーという用語は、AI システムによって間接的に（すなわち、直接ではないが、例えば効果を通じて）、あるいは直接的に（すなわち、直接的に、例えばアプリケーションを通じて）影響を受けるか、あるいはシステムに影響を与える（例えば開発者）すべての関係者を指す。これらは、個々の人または人のグループであることができる。利害関係者は「積極的な」役割を果たす必要はない。考えられるさまざまな利害関係者と AI システムとの関係の概要を表 2 に示す。一般に、消費者や利用者は AI システムを利用するだけであるが、専門家、開発者、企業／組織も AI システムを提供する可能性がある。間接的に影響を受ける当事者／第三者は、AI システムを提供することも利用することもない。それにもかかわらず、影響を受ける可能性があるため、（受動的な）利害関係者となる。ここに示したさまざまな利害関係者のリストは、網羅的であると主張するものではなく、必要に応じて改良することができる。しかし、AI システムに関して異なる利害が存在しうること、そしてそれはとりわけ、AI システムの透明性（提供される情報の種類や詳細度など）に対する異なる要件に反映されうることを示すには、選んだ表現で十分である。したがって、透明性という言葉を定義する際には、さまざまな利害関係者を考慮に入れることが不可欠である。

ステークホルダー	使用	提供する	透明性に関する要求事項の例
消費者	+	-	データ保護に関するサーバーの場所
ユーザー	+	-	施工ミスを防ぐための使用説明書
専門家	+/-	+/-	基礎となるモデルの機能性
開発者	+	+	インターフェイスを特定するための技術文書
企業／団体	+	+	刑法上の結果を避けるためのライセンス条件
間接的な影響を受ける第三者	-	-	破損時の連絡先

表 *2*：*AI* システムのステークホルダーの例と、利害の違いによる透明性への要求の違い。使用した記号：″+″（該当する）、″-″（該当しない）。

3 ディスカッション

3.1 アプローチと手順

透明性という用語の定義がすでに公表されていることから、さらなる定義の必要性が疑問視されている。しかし、定義市場における透明性というトピックの幅の広さは、結局のところ、利害関係者や適用分野によって透明性に対する要求が異なることを反映している。幅広い利害関係者グループと一般的な適用分野に焦点を当てた BSI のさらなる作業の基礎を作るため、既存の定義は使用しなかった。

加えて、AI 分野の技術開発のスピードは非常に速い。そのため、一度確立された定義が、特に具体的すぎる場合、その有効性を失うリスクがある。技術の進歩に歩調を合わせ、常に定義を更新し、現在の技術状況に適合させることを避けるため、本白書で提示する透明性の定義は、可能な限り技術に中立的で、将来性を考慮したものである。一方では、理解しやすく、透明性に関連するすべての側面をカバーすると同時に、それぞれの利害関係者や使用する AI 技術に応じて、個別の解釈が可能なほどオープンでなければならない。第二に、BSI が将来この分野で取り組む際の一般的な基礎となるものでなければならない。

さらに、定義づけには全体的なアプローチがとられた。これを図 2 に示す：透明性には、AI システムそのものに関する情報と、AI システムのサプライチェーンやプロバイダーに関する詳細など、そのエコシステムに関する情報の提供が含まれる。提供される情報の重み付けは、それぞれのステークホルダーの責任である。

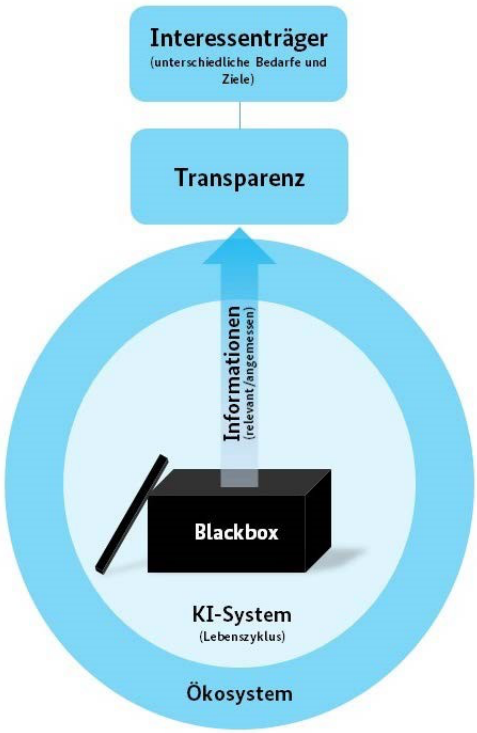

図 *2*：定義にある各要素を用いた全体的透明性アプローチの模式図：ステークホルダーにとって適切かつ適切な *AI* システム自体に関する情報に加え、そのエコシステムに関する情報も提供／開示される。これにより、利害関係者による *AI* システムの適合性と適切性の正当な評価が促進される。

3.2 透明性の目標

AI システムの透明性を促進することで、利害関係者の自主性を強化し、AI システムの使用、変更、提供が彼らにとって適切で正当なものかどうかを、彼ら自身が判断できるようにすることが目的である。AI システムの能力を説明するだけでは十分ではない。システムの限界も分析し、透明化しなければならない。このようにして初めて、利害関係者は、AI システムが特定の目的に適しているかどうかなど、総合的な評価を下すことができる。

デジタル消費者保護の分野では、デジタル化の進展にもかかわらず、消費者が安全で信頼できる AI システムを認識し、利用できるようにすべきである。また、企業や組織は、独自の AI システムを透明性をもって開発・運用できるようにすべきである。また、AI システムのエコシステムから得られる透明性の高い情報によって、影響を受ける第三者が、損害が発生した場合にどのように権利を主張できるかを認識できるようにすべきである。このように、AI システムの透明性は、利害関係者に力を与える役割を果たす。

3.3 AI 規則における透明性要件

AI 規則は、AI に関する世界初の包括的な法的枠組みである。2024 年 5 月 21 日に欧州連合（EU）理事会で採択され、EU における AI の利用を規制している。AI 規則は、データの倫理的かつ責任ある取り扱いを確保するための中核的要件として透明性を挙げている。したがって、一定のリスクレベルに該当する AI システムは、適切なレベルの透明性とシステムの十分な透明性のある運用に関する要件の対象となる。

とりわけ AI 規則は、特定の AI システムに関する調和された透明性規則を定めている（AI 規則第 1 条 2 項 d 参照）。AI 規則第 13 条は、リスクの高い AI システム（バイオメトリクスや重要インフラ（AI 規則附属書 III 参照）の分野など）は、プロバイダーやオペレーターが AI 規則に定められた義務を果たせるよう、透明性を確保しなければならないと定めている。このため、AI 規則は運用説明書に記載すべき最低限の情報も規定している。AI 規則はまた、自然人との直接対話を目的とする AI システム（チャットボット・アプリケーションなど）の提供者および運営者は、当該システムが AI システムであること、またはシステムの出力が AI によって生成されたものであることを、関係する自然人にも通知しなければならないと規定している（AI 規則第 50 条参照）。このような様々な利害関係者への情報開示・提供は、それぞれの AI システムに関する透明性を高めることを意図している。この具体的なシステムの種類も、AI システムの使用に関する単なる表示も、本白書で示した透明性の定義で理解されるもののサブセットに過ぎない。

欧州委員会は、AI 規則の実施および第 50 条に基づく透明性保持義務の実質的な実施に関するガイドラインを作成する（第 96 条 1 項(d)参照）。AI 規則に違反した場合に発効する制裁措置の手続きは、第 99 条に定められている。第 50 条に規定されたプロバイダーおよびオペレーターの透明性義務違反については、再

度明確に言及されており（AI 規則第 99 条 4 項(g)参照）、AI 規則におけるこのテーマの重要性が強調されている。

AI 規則の評価および見直しの一環として、欧州委員会は、追加的な透明性措置が必要な AI システムのリストの変更を 4 年ごとに評価する（AI 規則第 50 条による）。さらに、このリストに「各条項に定められた基準に基づくリスクレベルの評価と、新たなシステムを含めるための参加型手法」を作成することになっている（AI 規則第 112 条(11)(c)）。

AI 規則の附属書 XII は、「汎用 AI モデルのプロバイダーが、そのモデルを AI システムに統合する川下プロバイダーのために（...）技術文書を作成するための、（...）第 53 条(1)(b)に基づく透明性情報」を取り扱っている。第 1 項は、文書に最低限含まれなければならないモデルに関する情報に言及している。第 1 項が AI システムそのものに言及しているのに対し、第 2 項は実際の AI システムにとどまらず、例えばモデルの開発プロセスの構成要素に関する情報も要求している。ここでも、エコシステムに関する情報が含まれるため、AI 規則が定める透明性の義務は、本白書で示した透明性の定義と一致している。

具体的な透明性要件とは別に、AI 規則の目的は第 1 条 1 項に定められている。AI 規則は、域内市場の機能を向上させることに加え、「人間中心で信頼できる」AI の導入を促進することを目的としている。同時に、EU における AI システムの有害な影響に対する高水準の保護を確保することも目的としている。この白書で示された透明性の定義は、こうした目的の妨げになってはならない。AI システムとそのエコシステムに関する情報の提供は、AI システムの信頼性を高めるのに役立つはずである。さらに、透明性は、サプライチェーンに沿った様々な関係者／利害関係者間の相互作用を簡素化することができ、ひいては域内市場の機能にとっても有益である。同時に、透明性は、特定の（例えばセキュリティに関連する）情報の開示を通じて、AI システムが攻撃者に悪用されることにつながってはならない（これについては 3.5 節で詳述する）。また、AI 規則では、AI に関連する事故や誤作動によって、AI システムによって人や財産の健康や安全が脅かされてはならないと定めている。AI 規則はまた、「民主主義、法の支配、環境保護など、憲章に謳われている基本的権利」を促進することも目的としている。透明性は、利害関係者が AI システムの適合性を評価するのに役立つ。このエンパワーメントにより、利害関係者は前述の点について、AI システムを使用できるか／使用すべきかどうかを自ら判断できるようになる。責任を開示することで、透明性は損害を抑制し、損害が発生した場合に起こりうる結果を防止または最小化するのに役立つ。また、透明性はイノベーションの原動力にもなる。AI システムの限界について知ることで、もはやそのような限界を持たないアプリケーション／製品の開発につながる可能性がある。例えば、最初のチャットボットとの対話は、当初はテキスト言語を通じてのみ可能だった。キーボードを使って入力し、それに対してチャットボットが画面上でテキスト言語で応答するというものだった。現在では、チャットボットは様々な分野（例えば、保険分野での電話による顧客サポートなど）で使用されており、音声入力や音声出力による対話が可能となっている。

まとめると、AI 規則では透明性が考慮され、初期要件が策定されているといえる。同時に、AI 規則では透明性の概念が非常に広く定義されている。本白書で提示する透明性の定義は、AI 規則が定めるこれらの透

明性要件と矛盾するものではなく、透明性という用語のより包括的な定式化と定義を提供することを意図したものである。

3.4　透明性がもたらす機会

透明性の高い AI システムの使用は、意思決定のトレーサビリティとシステムの適切性の評価を促進することができる。また、潜在的なリスクや望ましくない影響を早い段階で認識できるようになるため、透明性は悪用からの保護にも役立つ。問題に適切に対応するためには、例えば、AI システムの出力に差別がないか、ライセンス条件に違反していないかを知ることが重要である。透明性は、消費者保護に関しても支援ツールとして機能する。透明性は、使用されるシステムの適切性を正しく評価するための基礎となる。このような評価を行うためには、システムに関する情報にアクセスできなければならない。AI システムの妥当性の正当な評価は、積極的な信頼と受容プロセスの基礎を形成する。初期の発表によると、例えば、透明性の高い AI モデルのダウンロード数が高いことが示されており、これは開発者の間でこれらのシステムがよりよく受け入れられていることを示している可能性がある（Liang, 2024）。透明性の欠如は、システムの適切性、ひいてはその信頼性を正当に評価することを困難にする。後者は、システムと関連する支出との良好な信頼関係を確立し、維持するための基本的な前提条件である。透明性の要求は、法的責任の明確な定義や、AI システムの使用に責任を負う者の特定と密接に関係しているため、透明性を確保することで、利用者が権利を主張しやすくなることもある。トレーサビリティ、誤用からの保護、受容性、信頼性、そして法的な説明責任といった項目は、AI システムの利用における透明性の重要性を示している。この関連性は、規制や法的要件にも反映されている（3.3 項参照）。

透明性は、一方では AI システムのセキュリティに貢献し、他方では AI アプリケーションの安全な利用を促進する。透明性は、潜在的な問題や脆弱性の特定を可能にし、望ましくないシステムの挙動を可視化し、問題の発見と悪用の防止に貢献する。IT セキュリティの文脈では、透明性は、システムの導入と使用に関連するリスクの開示と評価の基礎にもなる。透明性要件の一環として、損害や望ましくない事象が発生した場合の役割と責任を明確にすることは、利害関係者がシステムの欠陥動作を検出し、対応時間を短縮し、その結果、潜在的な損害を抑えることにも役立つ。

AI システムのライフサイクルの初期段階で透明性が実践されれば、開発チーム内の矛盾を最初から回避し、エラーの原因を最小限に抑え、トレーニング段階を短縮することができる。AI システムは開発されるとともに、新しいシステムの開発ツール（例えば、プログラム・コードの自動生成のための AI サポート・プログラミング）として使用されることも増えている。初期の開発段階では、開発者にとって、トレーニング／テスト／検証データがどこから来たのか、どのように整理されているのか、偏りがないのか（例えば差別を避けるため）を知ることも重要である。この情報は、トレーニング／テスト／検証の前にデータを正しく前処理するために重要である。現在の開発動向は、既存のモデルがしばしば使用されていることを示しており、既存のモデルに関するすべての安全上重要な情報が利用可能であり、アクセスしやすいことが特に重要である。この情報が欠落していると、基本モデルの安全リスクを自社製品に実装してしまうリスクがある。その場合、両システムは互いに依存し合うことになり、基本システムの内部情報が実装されたシステム

に転送されることになる。上記のようなセキュリティ・リスクが異なる領域から継承されることで、前述の例では全体的なリスクが増大することになる。このことは、AI システムを使用する際、セキュリティに関連する側面の透明性が爆発的に重要であることを改めて強調している。

さらに、前節ですでに述べたように、透明性を確保することは、ユーザーが AI システムをより適切に評価することにつながる。アプリケーション、適切なアプリケーション・シナリオ、起こりうる問題や安全性のリスクを正しく評価することで、ユーザーによる安全な利用を促進することができる。

3.5　透明性によるリスク

これまでのところ、透明性のポジティブな側面が主に紹介されてきた。しかし、AI システムの透明性を高める／向上させることは、意図しない悪影響をもたらす可能性もある。例えば、AI システムの機能やアーキテクチャに関する情報を提供することで、攻撃者がシステムを悪用・侵害するために利用できる新たな攻撃ベクトルが露呈する可能性がある。

AI システムの限界や適用除外領域に関する情報は、攻撃者によって意図的に悪用される可能性もある。例えば、誤った動作や破壊的な出力を意図的に生成するためなどだ。

逆に、攻撃者は、意図的に誤った情報を透明な方法で提供するために、透明性が生み出すはずの信頼を悪用することもできる。例えば、セキュリティ上重要なアプリケーションを、実際には重要でないと見せかけることができる。さらに、透明でないシステムを透明であるかのように見せかけることもできる。このような擬似的な透明性は、自社製品のマーケティング目的で使用される可能性があり、透明性ラベルがチェックされなければ、消費者側がシステムに対して誤った評価を下す結果となる。したがって、開示／提供される情報の信頼性についての疑問にも、今後答えていかなければならない。公的な透明性ラベルと検証可能な透明性基準は、ここでの救済策となりうる。

したがって、AI システムの透明性は諸刃の剣であり、慎重に使用されるべきである。目標と問題は時に矛盾し、同時に解決することはできない。利害関係者が意思決定するために必要な情報は何か」「関連性のない情報は何か」という重要な問いに答えることは有益である。EU 一般データ保護規則と同様、ここでもデータ最小化の原則の適用が推奨される。これは、個々のユースケースごとに個別に回答されるもので、必要なだけの情報は開示されるべきだが、絶対に必要な情報以上は開示されるべきではない。この「知る必要性」の原則は、特にセキュリティ上重要な情報に適用される。目指すべきは、十分であると同時に、セキュリティーなどの面で過度なペナルティーを与えない、適切なレベルの透明性である。

4 結論

多くの AI システムはブラックボックス化されているため、データや情報はユーザーにとって透明性のない方法で処理され、検証不可能な決定が下される。AI システムに関する知識の欠如は、システム出力のトレーサビリティと検証可能性の欠如と密接に関係している。システムの出力が正しく適切かどうかを評価するのは難しい。同様に、システムとそのエコシステムに関する情報が不足していれば、責任、法的責任、公平性に関する疑問にも答えることができない。結局のところ、透明性のない AI システムは、信頼を失い、システムを拒絶することにつながりかねない。また、既存のシステムに AI コンポーネントを実装したり、異なるシステムを組み合わせたりすることで、さらに複雑さが増し、関連情報へのアクセスがさらに困難になる可能性もある。システムの洞察力の欠如とシステムに関する情報の欠如によって引き起こされる問題は多岐にわたり、大きな課題となっている。透明性は、この情報不足の問題に対処し、システム情報へのアクセス性を高め、システムの正当な評価を可能にすることで、AI システムをより理解しやすくすることを目的としている。こうした理由から、透明性は AI システムのすべての利害関係者にとって決定的な役割を果たす。課題は、個々に異なる透明性の要求を持つすべての利害関係者に等しく対応することである。

AI システムの透明性分野におけるプロジェクト全体と今後の活動は、BSI のすべての利害関係者を対象としている。システムのユーザーである社会にとってのこのテーマの関連性は、期待されるトレーサビリティの向上、誤用に対するより良い保護、より妥当な受け入れと信頼性のプロセス、より拘束力のある法的説明責任に反映されている。透明性対策は、AI システムの選択と使用に関するエンドユーザーの信頼と自律性を高めることで、エンドユーザーのエンパワーメントに直接貢献するはずである。全体として、エンドユーザーのエンパワーメントは、AI システムの利用を民主化することを目的としている。さらに、透明性の向上や具体的な基準・措置の策定は、AI システムの信頼ある利用という包括的な目標に貢献するはずである。AI システムの開発に携わる企業に対しては、AI システムの開発・運用におけるテーマの関連性と対策の遵守を強調すべきである。サードパーティの AI システムを組織で利用したり、自社のシステムや製品に導入しようとするビジネス環境の利害関係者のためのオリエンテーションの補助として、ガイドラインやポジションが利用できるようにする。これらのガイドラインは、企業が適切で安全かつ高性能なシステムを特定しやすくすることを意図している。

この作業は、AI システムの利用を希望する公的機関にとっても指針となるはずである。自らの利用だけでなく、AI システムに関する安全保障に関連する新たな知見に日々対処する必要があるため、公共部門や行政の関係者は、職員の専門的な資格と適切な装備を確保するという課題に直面することになる。今回の透明性確保と今後の透明性確保は、職員の恒久的かつ適切な（再）訓練を促進・加速させるために活用できる。さらに、この作業と今後の作業を通じて期待される透明性基準の確立は、公的機関による有意義で信頼できる品質シールの開発を促進することができる。生活の様々な分野で AI システムのさらなる普及と広範な展開が予想されることから、社会全体にとっての関連性は常に高まっている。

今後、これらのシステムに対して適切かつ有効な評価を行うためには、透明性基準の確立が不可欠である。汎用の AI システムや感情認識システムなど、特定の AI システムのプロバイダーやオペレーターに対して

は、すでに AI 規則で透明性の義務が定められている（AI 規則第 50 条参照）。これらは、これらのシステムが EU で販売され、使用されるための前提条件のひとつである。透明性の基準は、十分な情報に基づいた意思決定を可能にすることで、AI システムの利害関係者の自主性を強化することができる。したがって、透明性は当初から考慮されうるし、また考慮されるべきである（設計による透明性）。

参考文献

BSI、連邦情報セキュリティ局。2021a.AI クラウドサービスコンプライアンス基準カタログ（AIC4）。2021.

BSI、連邦情報セキュリティ局。2021b.連邦情報セキュリティ局、AI の安全で堅牢かつ追跡可能な利用-問題点、対策、行動の必要性。2021.

ISO/IEC 22989:2022 情報技術-人工知能-人工知能の概念と用語。

Liang,W.、Rajani,N.、Yang,X.、Ozoani,E.、Wu,E.、Chen,Y.、Smith,D. S.、 & Zou,J. 2024. AI に何が記録されているのか？32K AI モデルカードの系統的分析。2024. http://arxiv.org/abs/2402.05160.

OECD.人工知能に関する理事会勧告**.**人工知能理事会の勧告。2019.

Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. "Why Should I Trust You?" **:** s.l. : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.11, 2016.