別紙１（附属資料）

ＡＩ利活用原則の各論点に対する詳説

令　和　元　年　８　月　９　日
ＡＩネットワーク社会推進会議

凡例

「AIサービスプロバイダ、ビジネス利用者及びデータ提供者」の視点で記載した内容

（参考）
「消費者的利用者」の視点で記載した内容

①適正利用の原則

利用者は、人間とAIシステムとの間及び利用者間における適切な役割分担のもと、
適正な範囲及び方法でAIシステム又はAIサービスを利用するよう努める。

主な論点

ア）　適正な範囲・方法での利用

イ）　人間の判断の介在

ウ）　関係者間の協力

①ーア）　適正な範囲・方法での利用(1/2)

AIサービスプロバイダ及びビジネス利用者は、開発者等からの情報提供や説明を踏まえ、AIを利活用する際の社会的文脈に応じ、AIを利用する目的、用途とAIの性質、能力等を適切に認識した上で、AIを適正な範囲・方法で利用することが期待される。また、AIサービスプロバイダは、AIサービスの公平な条件による利用を確保するとともに、以下に関する情報を以下のとおり適時に提供することが期待される。

[提供すべき情報]

- 提供するAIの用途・方法
- AIの性質、利用の態様等に応じた便益及びリスク
- 提供するAIの利活用の範囲・方法に関する定期的な確認方法（特に、AIが自律的に更新される場合の観測、確認方法）、確認の重要性・頻度、未確認によるリスク等
- 利活用の過程を通じて、AIの機能を向上させ、リスクを抑制するために実施するAIソフトのアップデート[1]及びAIの点検・修理等

[情報を提供すべきタイミング]

- AIの利用前に当該情報を提供できることが望ましい。
- 事前に当該情報を提供できない場合には、AIの性質、利用の態様等に基づき想定されるリスクに応じ、消費者的利用者等からのフィードバックに対応する体制が整備されていることが望ましい。

また、利活用の過程を通じて、AIの機能を向上させ、リスクを抑制するため、AIソフトのアップデート及びAIの点検・修理等を提供することが期待される。特に、アップデートにより、連携する他のAIに影響を及ぼすこと[2]が想定される場合は、そのリスクに関する情報を提供することが期待される。

また、AIの利用により人の生命・身体・財産に危害を及ぼすことが想定される場合など、提供するAIシステム又はAIサービスの性質、利用の態様等によっては、提供を受ける利用者の信頼性等について事前に確認することが期待される場合も想定される。さらに提供をした後であっても、最終利用者がAIシステム又はAIサービスを誤って使用していないか、悪意をもって使用していないかについて、入出力等のログを記録・保存すること等により確認することが期待される場面も想定される。

1)問題が発見されてからアップデートが提供されるまでの間は、問題点を最終利用者に適時適切に情報提供するとともに注意喚起することが期待される。
2)アップデートを適用するAIの動作が周囲のAIに影響を及ぼすことが想定される。例えば、家庭内の家電に含まれるAIソフトがアップデートにより更新された場合、全体を統括する家庭内執事ロボットや周辺のAIを含む家電が当該アップデートに適応していないと、（家電同士、または家電とロボットの）相互の判断に齟齬が生じうる（「報告書2018」別紙３「AIが想定外の動作を行うなどのおそれ」の事例）。

①ーア）　適正な範囲・方法での利用（2/2）

＜参考＞

消費者的利用者は、開発者及びAIサービスプロバイダからの情報提供や説明を踏まえ、AIを利活用する際の社会的文脈にも配慮して、AIを適正な範囲・方法で利用することが望ましい。その際、以下の事項に留意することが望ましい。

[留意することが望ましい事項]

（利用前）

- AIの性質、利用の態様等に応じて、便益及びリスクを認識し、適正な用途を理解するとともに、必要な知識・技能を習得すること。

（利用中）

- 自らのAIの利活用が適正な範囲・方法で行われているか定期的に確認すること。
- 利活用の過程を通じて、AIの機能を向上させ、リスクを抑制するため、AIソフトのアップデート及びAIの点検・修理等を行うよう努めること。（ただし、アップデートにより連携する他のAIに影響を及ぼしうること[1]に留意することが望ましい。）
- 何らかの問題が発生した場合、問題が起こる予兆があった場合等には、開発者及びAIサービスプロバイダに対し、当該情報をフィードバックすること。

①ーイ）　人間の判断の介在

AIサービスプロバイダ及びビジネス利用者は、AIによりなされた判断について、必要かつ可能な場合には、その判断を用いるか否か、あるいは、どのように用いるか等に関し、人間の判断を介在させることが期待される。その場合、人間の判断の介在の要否については、例えば以下の基準を踏まえ、利用する分野やその用途等に応じて検討することが期待される。

［人間の判断の介在の要否について、基準として考えられる観点（例）］

- AIの判断に影響を受ける最終利用者等の権利・利益の性質及び最終利用者等の意向
- AIの判断の信頼性の程度（人間による判断の信頼性との優劣）
- 人間の判断に必要な時間的猶予
- 判断を行う利用者に期待される能力
- 判断対象の要保護性（例えば、人間による個別申請への対応か、AIによる大量申請への対応か等）

また、AIによりなされた判断について人間が最終判断をすることが適当とされている場合に、人間がAIと異なる判断をすることが期待できなくなることも想定されることから、説明可能性を有するAIから得られる説明を前提として、人間が判断すべき項目を事前に明確化しておくこと等により、人間の判断の実効性を確保することが期待される[1]。

また、アクチュエータ等を通じて稼働するAIの利活用において、一定の条件に該当することにより人間による稼働に移行することが予定されている場合には、移行前、移行中、移行後等の各状態における責任の所在があらかじめ明確化されている必要がある。また、AIサービスプロバイダは、移行条件、移行方法等を最終利用者に事前に説明し、必要な訓練を実施するなど、人間による稼働に移行した場合に問題が生じないための事前対策を講じることが期待される。

１)加えて、人間が確認するAIの判断の適正性を確保するため、他のAIを利用したダブルチェック、AIへの入力を摂動させることよるAI動作の確認などの措置を検討することが望ましい。

＜参考＞

消費者的利用者は、AIの判断に対し、消費者的利用者が最終判断をすることが適当とされている場合には、適切に判断ができるよう必要な能力及び知識を習得しておくことが望ましい。

また、開発者及びAIサービスプロバイダにより人間の判断の実効性を確保するための対応が整理されている場合は、それに基づき適切に対応することが望ましい。

また、アクチュエータ等を通じて稼働するAIの利活用において、一定の条件に該当することにより人間による稼働に移行することが予定されている場合には、消費者的利用者は、移行前、移行中、移行後等の各状態における責任の所在を予め認識しておくことが望ましい。また、AIサービスプロバイダから、移行条件、移行方法等についての説明を受け、必要な能力及び知識を習得しておくことが望ましい。

①ーウ）　関係者間の協力

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIサービスを提供し又はAIシステムを利用するに当たり、AIの利活用により生じ得る又は生じた事故、セキュリティ侵害・プライバシー侵害等によりもたらされる又はもたらされた被害の性質・態様等に応じて、関係者と協力して予防措置及び事後対応（情報共有、停止・復旧、原因解明、再発防止措置等）に取り組むことが期待される。

その際、以下に掲げる原則の記載内容に留意することが期待される。

[関係者間で協力して行う予防措置及び事後対応（例）]

- ①適正利用の原則　論点ア：適正な範囲・方法での利用（適正な範囲・方法による利用のための情報の提供等）
- ④安全の原則　論点ア：人の生命・身体・財産への配慮（AIがアクチュエータ等を通じて人の生命・身体・財産に危害を及ぼした場合に講ずるべき措置等）
- ⑤セキュリティの原則　論点ア：セキュリティ対策の実施（セキュリティが侵害された場合に講ずるべき措置等）
- ⑥プライバシーの原則　論点ア：他者のプライバシーの尊重（他者のプライバシーを侵害した場合に講ずるべき措置等）

＜参考＞

消費者的利用者は、AIを利用するに当たり、開発者又はAIサービスプロバイダからの情報提供を踏まえ、AIの利活用により生じ得る又は生じた事故、セキュリティ侵害・プライバシー侵害等によりもたらされる又はもたらされた被害の性質・態様等に応じて、関係者と協力して予防措置及び事後対応（情報共有、停止・復旧、原因解明、再発防止措置等）に取り組むことが望ましい。

②適正学習の原則

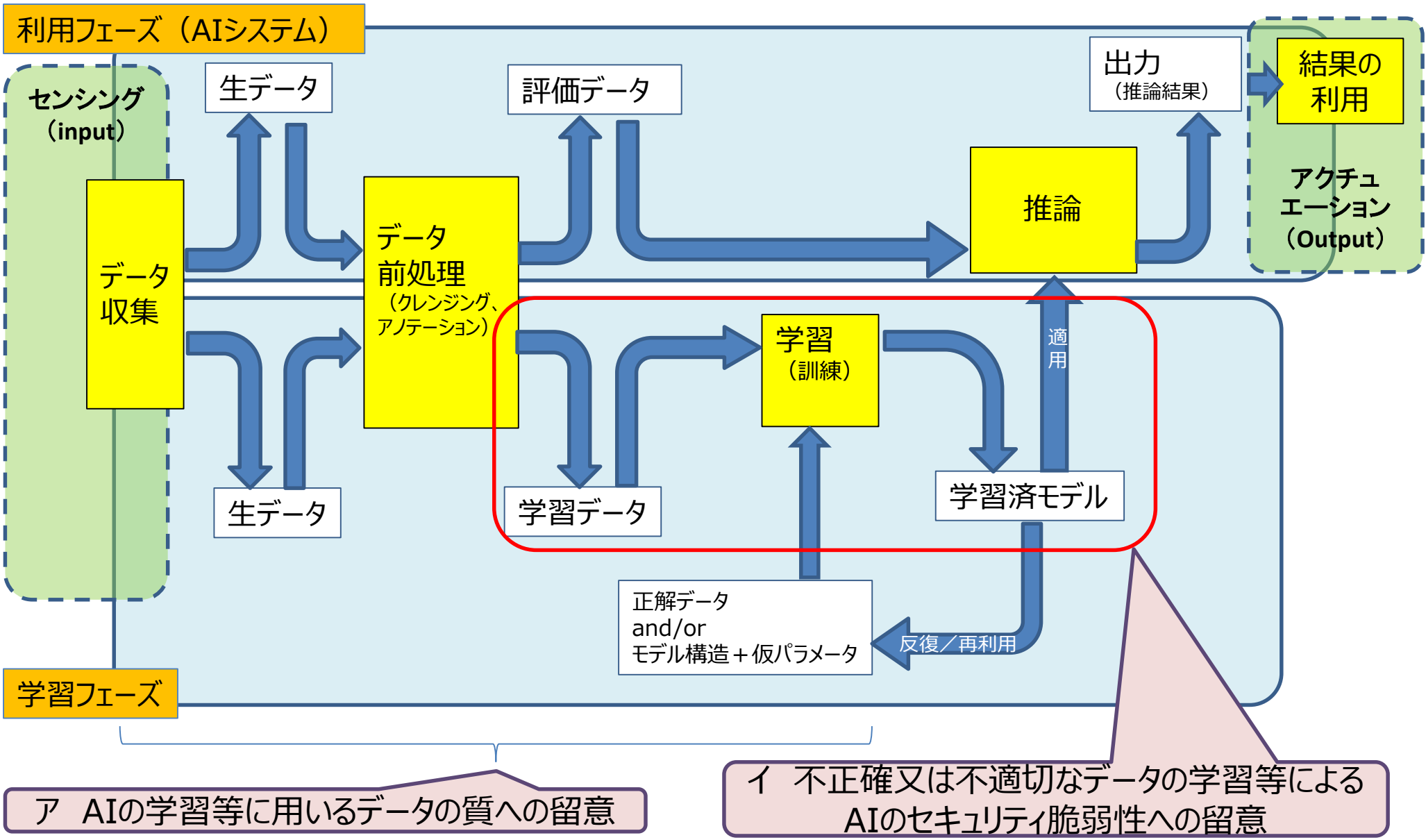
利用者及びデータ提供者は、AIシステムの学習等に用いるデータの質に留意する。

主な論点

ア）　AIの学習等に用いるデータの質への留意

イ）　不正確又は不適切なデータの学習等によるAIのセキュリティ脆弱性への留意

（機械学習を中心とした）学習と利用の流れにおける②適正学習の原則の各論点



②ーア） AIの学習等に用いるデータの質への留意（1/2）

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、利用するAIの特性及び用途を踏まえ、AIの学習等に用いるデータの質（正確性や完全性など）に留意することが期待される。特に機械学習においては、以下の方法によりデータの質を確保することが考えられる。

[データ収集時の対策（例）]
- 収集するデータがAIの利用目的に適ったものかを確認する。
- 社会的に信用の高い者が公開するデータを活用する。
- データの作成来歴を確認した上で収集する。
- 自らデータを収集する際には、データに付随する権利に留意する。

[データ前処理時の対策（例）]
- 人間でも判定が困難と考えられるデータは、学習等の対象から除外する[1]。
- 機械（学習器）が誤認識しやすいと考えられるデータは積極的に学習の対象とする[2]。
- （特に教師あり学習等で）アノテーション（ラベル付与）を行う際には、誤って行わないよう留意する。
- 利用時に利用（入力）されるデータの形式を意識してデータセットを作成する。
- 前処理をどのように行ったのか（データ前処理に関する来歴）について、ログを取得し保存する。

[学習時の対策（例）]
- 既存の学習モデルを利用して転移学習[3]等を行う。
- 学習の精度を上げるため、特定のデータを拡張[4]した上で学習を行う。
- 一過性のある時系列データを学習する場合などは、学習対象とすべきデータの範囲を適切に画定する。

1) 例えば、画像認識などで、対象となるオブジェクトが人間の目で見ても同定できない場合など。
2) 例えば、画像認識などで、対象となるオブジェクトが端にあるなど。
3) 転移学習(Transfer Learning)とは、深層学習を含む機械学習で用いられる技術の１つで、特定の領域（ドメイン）で学習させたモデルを別の領域に適用する技術である。少ないデータで精度の高い学習結果を得ることが出来る可能性がある点がメリットである。
4) 「データの拡張」(Data Augmentation)とは、特定の学習データが少ない際に、汎化性能（未知のデータに対する性能）を高めることにより、データの正確性を確保するためにとられる手段の１つである。学習に用いるデータを拡張し（例えば画像データであれば、反転、拡大、縮小を適用し）それぞれを別のソースに基づくデータとして用いることにより、汎化性能が改善されることがある。

②ーア） AIの学習等に用いるデータの質への留意（2/2）

また、AIによりなされる判断は、事後的に精度[1]が損われたり、低下することが想定されるため、想定される権利侵害の規模、権利侵害の生じる頻度、技術水準、精度を維持するためのコスト等[2]を踏まえ、あらかじめ精度に関する基準を定めておくことが期待される。精度が当該基準を下回った場合には、データの質に留意して改めて学習させることが期待される。

なお、消費者的利用者から提供されるデータを用いることが予定されている場合には、AIの特性及び用途を踏まえ、データ提供の手段、形式等について、あらかじめ消費者的利用者に情報を提供することが期待される。

1) ここで言う「精度」には、AIが正しい判断を行っているか、例えばAIが暴力的な表現を行っていないか、ヘイトスピーチなどを行っていないか等も含まれる。
2) 例えば、機械学習を中心としたAIは帰納的な処理を行うため、当該AI単体では、判断結果につき原理的に100%の精度を担保できないこと等が挙げられる。

＜参考＞

消費者的利用者は、自らデータを収集し、利用するAIの学習等を行うことが予定されている場合には、データの形式及び内容[3]について、開発者、AIサービスプロバイダ等から提供された情報を踏まえた上でデータの収集、保存を行うことが望ましい。

3) 誤ったデータでないか、悪意を持って入力されたデータでないかなど。

②ーイ） 不正確又は不適切なデータの学習等によるAIのセキュリティ脆弱性への留意

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIが不正確又は不適切なデータを学習することにより、AIのセキュリティに脆弱性が生じるリスクが存在することに留意することが期待される。また、消費者的利用者に対し、そのようなリスクが存在することを予め周知することが期待される。

[リスクの例]

- 学習が不十分であること等の結果、学習モデルが正確に判断することができるデータに、人間には判別できない程度の微少な変動を加え、そのデータをインプットすること等により、作為的に当該学習モデルの判断を誤らせることができるリスク（例：Adversarial example攻撃）
- 学習において不正確なラベリング等がなされたデータを混在させることで、誤った学習が行われるリスク

＜参考＞

消費者的利用者は、開発者、AIサービスプロバイダ及びデータ提供者からの情報を踏まえ、AIが不正確又は不適切なデータを学習することにより、AIのセキュリティに脆弱性が生じるリスクに留意することが望ましい。

また、AIを利用するに当たり、セキュリティ上の疑問を感じた場合は、開発者、AIサービスプロバイダ及びデータ提供者にその旨を報告することが望ましい。

③連携の原則

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIシステム又はAIサービス相互間の連携に留意する。
また、利用者は、AIシステムがネットワーク化することによってリスクが惹起・増幅される可能性があることに留意する。

主な論点

ア）　相互接続性と相互運用性への留意

イ）　データ形式やプロトコル等の標準化への対応

ウ）　AIネットワーク化により惹起・増幅される課題への留意

③ーア）　相互接続性と相互運用性への留意

AIサービスプロバイダは、利用するAIの特性及び用途を踏まえ、AIネットワーク化の健全な進展を通じて、AIの便益を増進するため、AIの相互接続性と相互運用性に留意することが期待される。

③ーイ）　データ形式やプロトコル等の標準化への対応

AIサービスプロバイダ及びビジネス利用者は、AI相互間及びAIと他のシステム等との連携を促進するため、AIの入出力等におけるデータ形式（構文（syntax）及び意味（semantics）[1]）、連携のための接続方式（特にネットワークを介す場合は各レイヤにおけるプロトコル）等の標準に準拠することが期待される。

また、データ提供者についても、AI相互間及びAIと他のシステム等との連携を促進するため、データ形式（構文（syntax）及び意味（semantics）[1]）の標準に準拠することが期待される。

1)データの構文だけが示されていても、意味が示されていないと連携は正しく動作しない。

＜参考＞

消費者的利用者は、自らデータを収集し、利用するAIの学習等を行うことが予定されている場合には、データの形式について、開発者、AIサービスプロバイダ等から提供された情報を踏まえた上でデータの収集、保存を行うことが望ましい。

③ーウ）　AIネットワーク化により惹起・増幅される課題への留意

AIが連携することによって便益が増進することが期待されるが、AIサービスプロバイダ及びビジネス利用者は、自ら利用するAIがインターネット等を通じて他のAI等と接続・連携することにより制御不能となる等、AIがネットワーク化することによってリスクが惹起・増幅される可能性があることに留意することが期待される。このため、開発者等からの情報を踏まえ、考えられるリスクを分析し、当該リスクを連携の相手方と共有するとともに、予防策や問題が生じた場合の対応策等を整理し、消費者的利用者等に対し、必要な情報提供を行うことが期待される。

［AIがネットワーク化することによってリスクが惹起・増幅される可能性の例］

- 個別の事業者のトラブル等がシステム全体に波及するおそれ
- AIシステム間の連携・調整が成立しないなどのおそれ
- AIの判断・意思決定を検証できないおそれ（システム間の相互作用が複雑となり解析が困難になるおそれ）
- 少数のAIの影響力が強くなりすぎるなどのおそれ（少数のAIの判断によって企業や個人が不利な立場になるなどのおそれ）
- 多数のAIが同一の判断をし、又は行動をとることにより、市場における競争及び制御が機能しなくなるおそれ
- 領域横断での情報の共有と特定の基盤的なAIへの情報の集中によるプライバシー侵害のおそれ
- AIが想定外の動作を行うなどのおそれ

＜参考＞

AIが連携することによって便益が増進することが期待されるが、消費者的利用者は、自ら利用するAIがインターネット等を通じて他のAI等と接続・連携することにより制御不能となる等、AIがネットワーク化することによってリスクが惹起・増幅される可能性があることに留意することが望ましい。また、事前の予防策や問題が生じた場合の対応策等について、開発者及びAIサービスプロバイダから情報提供があった場合には、利用にあたり留意することが望ましい。

④安全の原則

利用者は、AIシステム又はAIサービスの利活用により、アクチュエータ等を通じて、利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。

主な論点

ア）　人の生命・身体・財産への配慮

④ーア　人の生命・身体・財産への配慮

人の生命・身体・財産に危害を及ぼし得る分野でAIを利活用する場合には、AIサービスプロバイダ及びビジネス利用者は、想定される被害の性質・態様等を踏まえ、開発者等からの情報をもとに、必要に応じて以下の対応策を講ずることにより、AIがアクチュエータ等を通じて人の生命・身体・財産に危害を及ぼすことのないよう配慮することが期待される。

[対応策の例]

- AIの点検・修理及びAIソフトのアップデート[1]を行うこと、また消費者的利用者にこれらの実施を促すこと
- AIが想定外の動作を起こした場合でも、AIが組み込まれたシステム全体で安全を確保できる仕組み[2]を構築するなど、フェイルセーフ[3]の実現を図ること

また、AIサービスプロバイダ及びビジネス利用者は、AIがアクチュエータ等を通じて人の生命・身体・財産に危害を及ぼした場合に講ずるべき措置について、あらかじめ整理しておくことが期待される。加えて、当該措置について、消費者的利用者に対し、必要な情報提供を行うことが期待される。

[危害時の措置の例]

- 初動措置（当該AIを含むシステムの急用度等の文脈に応じ、必要な手順にて実施）
  - 当該システムのロールバック[4]、代替システムの利用などによる復旧
  - システムの停止（キルスイッチ）：可能な場合
  - ネットワークからの遮断：可能な場合
  - 危害の内容の確認
  - 関係者への報告
- 補償・賠償等（補償・賠償等を円滑に行うための保険の利用）
- 重大な損害が生じた場合等は、第三者機関の設置とその機関による原因調査・分析・提言など

1) 問題が発見されてからアップデートが提供されるまでの間は、問題点を最終利用者に適時適切に情報提供するとともに注意喚起することが期待される。
2) AI単体で技術的に安全性を保証することが困難な状況では、AI以外のシステムによりAI実装システムの安全確保を実施し、当該システムの運用経験によりAIの安全性を実証していくことも可能である。
3) 誤操作、誤動作などによる不具合が発生した場合に、損害が発生しないよう安全な方向に導くこと
4) 障害が起こった際等に、直前の（保存した）状態まで戻ること。

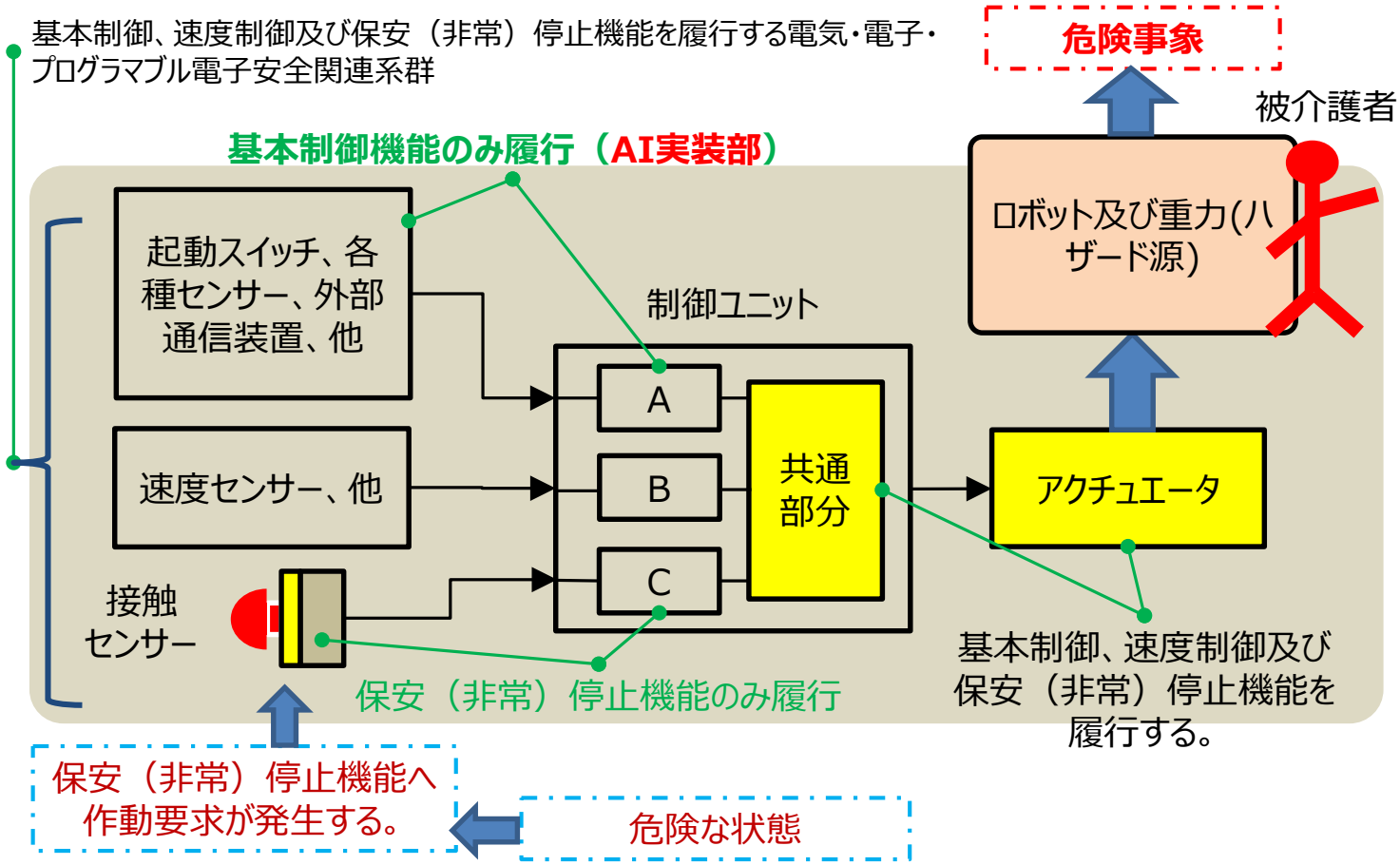
＜参考＞

人の生命・身体・財産に危害を及ぼし得る分野でAIを利活用する場合には、消費者的利用者は、想定される被害の性質・態様等を踏まえ、開発者及びAIサービスプロバイダからの情報をもとに、必要に応じてAIの点検及びAIソフトのアップデートを行うことなどにより、AIがアクチュエータ等を通じて人の生命・身体・財産に危害を及ぼすことのないよう配慮することが望ましい。

また、消費者的利用者は、AIがアクチュエータ等を通じて人の生命・身体・財産に危害を及ぼした場合に講ずるべき措置について、開発者及びAIサービスプロバイダから情報提供があった場合には、利用にあたり留意することが望ましい。

【参考】 AIの外部での安全確保の仕組みの例

衝突ハザード及び転倒ハザード等をもつ介護ロボットの電気・電子・プログラマブル電子安全関連系群（多重防護層）



引用：AIガバナンス検討会（第４回）ナブテスコ（株）佐藤吉信技術顧問　講演資料

→制御ユニットにおいて、AI実装部（A）のみでなく、他のシステム（B、C）を含めた安全確保

⑤セキュリティの原則

利用者及びデータ提供者は、AIシステム又はAIサービスのセキュリティに留意する。

主な論点

ア）　セキュリティ対策の実施

イ）　セキュリティ対策のためのサービス提供等

ウ）　AIの学習モデルに対するセキュリティ脆弱性への留意

⑤ーア）　セキュリティ対策の実施

AIサービスプロバイダ及びビジネス利用者は、AIのセキュリティに留意し、AIシステムの機密性・安全性・可用性を確保するため、その時点での技術水準に照らして合理的な対策を講ずることが期待される。

また、セキュリティが侵害された場合に講ずるべき措置について、当該AIの用途や特性、侵害の影響の大きさ等を踏まえ、あらかじめ整理しておくことが期待される。

[セキュリティ侵害時の措置の例]

- 初動措置（当該AIを含むシステムの急用度等の文脈に応じ、必要な手順にて実施）
  - 当該システムのロールバック[1]、代替システムの利用などによる復旧
  - システムの停止（キルスイッチ）：可能な場合
  - ネットワークからの遮断：可能な場合
  - セキュリティ侵害の内容確認
  - 関係者への報告
- 補償・賠償等（補償・賠償等を円滑に行うための保険の利用）
- 重大な損害が生じた場合等は、第三者機関の設置とその機関による原因調査・分析・提言など

1) 障害が起こった際等に、直前の（保存した）状態まで戻ること。

＜参考＞

消費者的利用者は、（消費者的利用者側で）セキュリティ対策を実施することが想定されている場合には、開発者及びAIサービスプロバイダからの情報提供を踏まえ、AIのセキュリティに留意し、必要な対策を講ずることが望ましい。

⑤ーイ）　セキュリティ対策のためのサービス提供等

AIサービスプロバイダは、自ら提供するAIサービスについて、最終利用者にセキュリティ対策のためのサービスを提供するとともに、過去のアクシデントやインシデント情報の共有を図ることが期待される。

また、AIサービスプロバイダ及びビジネス利用者はセキュリティが侵害された場合の措置について、消費者的利用者に対し必要な情報提供を行うことが期待される。

＜参考＞

消費者的利用者は、セキュリティが侵害された場合に講ずるべき措置について、開発者及びAIサービスプロバイダから情報提供があった場合には、利用にあたり留意することが望ましい。

また、AIを利用するに当たり、セキュリティ上の疑問を感じた場合は、開発者、AIサービスプロバイダ、データ提供者等にその旨を報告することが望ましい。

⑤ーウ） AIの学習モデルに対するセキュリティ脆弱性への留意

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、学習モデルの生成及びその管理において、セキュリティに脆弱性が存在するリスクに留意することが期待される。また、消費者的利用者に対し、そのようなリスクが存在することを予め周知することが期待される。

[リスクの例]

- 学習が不十分であること等の結果、学習モデルが正確に判断することができるデータに、人間には判別できない程度の微少な変動を加え、そのデータをインプットすること等により、作為的に当該学習モデルの判断を誤らせることができるリスク（例：Adversarial example攻撃）
- （教師あり学習において）学習において不正確なラベリング等がなされたデータを混在させることで、誤った学習が行われるリスク
- 学習モデルが容易に複製できるリスク
- 学習モデルから学習に用いられたデータをリバースエンジニアリングできるリスク

＜参考＞

消費者的利用者は、開発者、AIサービスプロバイダ及びデータ提供者からの情報を踏まえ、学習モデルの生成及びその管理において、セキュリティに脆弱性が存在するリスクに留意することが望ましい。

また、AIを利用するに当たり、セキュリティ上の疑問を感じた場合は、開発者、AIサービスプロバイダ、データ提供者等にその旨を報告することが望ましい。

⑥プライバシーの原則

利用者及びデータ提供者は、AIシステム又はAIサービスの利活用において、他者又は自己のプライバシーが侵害されないよう配慮する。

（注）日本においては、前提として、個人情報保護法を遵守することが必要である。

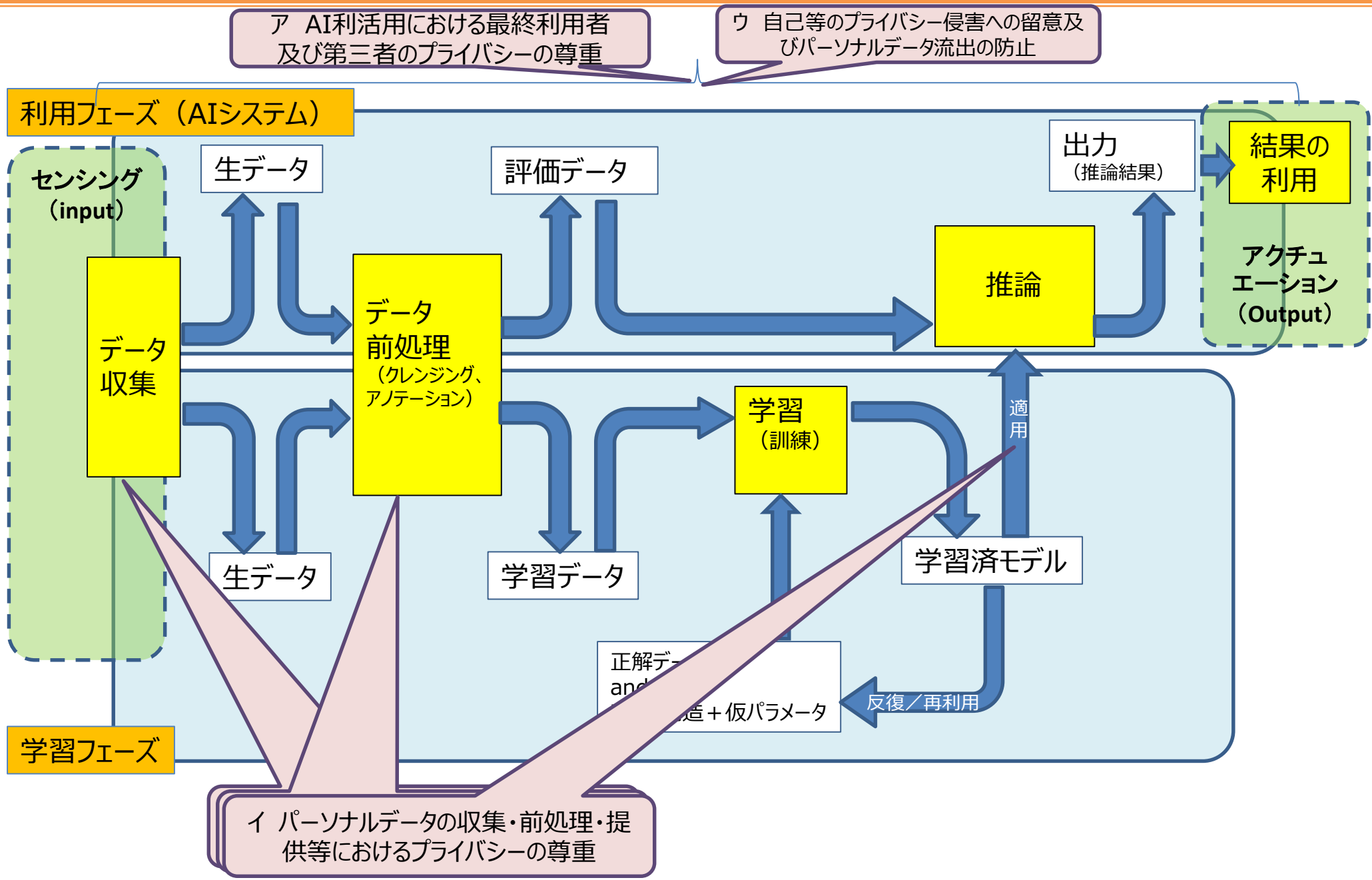
主な論点

ア）　AI利活用における最終利用者及び第三者のプライバシーの尊重

イ）　パーソナルデータの収集・前処理・提供等におけるプライバシーの尊重

ウ）　自己等のプライバシー侵害への留意及びパーソナルデータ流出の防止

（機械学習を中心とした）学習と利用の流れ（簡易版）と⑥プライバシーの原則の各論点



⑥ーア） AI利活用における最終利用者及び第三者のプライバシーの尊重

AIサービスプロバイダ及びビジネス利用者は、AIを利活用する際の社会的文脈や人々の合理的な期待を踏まえ、AIの利活用において最終利用者及び第三者のプライバシーを尊重する。

また、最終利用者及び第三者のプライバシーを侵害した場合に講ずるべき措置について、あらかじめ整理しておくことが期待される。

加えて、当該措置について、最終利用者及び第三者に対し、必要な情報提供を行うことが期待される。

[プライバシー侵害時に講ずるべき措置の例]

- 最終利用者及び第三者のプライバシーを侵害する情報を誤って取得した場合における、当該情報の消去、AIのアルゴリズムの更新等
- 最終利用者及び第三者のプライバシーを侵害する情報を拡散した場合における、保存先への消去の依頼、AIのアルゴリズムの更新等

＜参考＞

消費者的利用者は、AIを利活用する際の社会的文脈や人々の合理的な期待を踏まえ、AIの利活用において第三者のプライバシーを尊重する。

加えて、第三者のプライバシーを侵害した場合に講ずるべき措置について、開発者及びAIサービスプロバイダから情報提供があった場合には、利用にあたり留意することが望ましい。

⑥ーイ）　パーソナルデータの収集・前処理・提供等におけるプライバシーの尊重

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIの学習等に用いられるパーソナルデータの収集・前処理・提供等[1,2]において、また、それらを通じて生成された学習モデルの提供等において、最終利用者及び第三者のプライバシーを尊重する。

1) 他者に提供後のパーソナルデータの取り扱いについても例えば消去を行う等の留意が期待される。
2) AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、自ら提供したデータに個人情報が含まれる場合、当該データを誰がどのように利用しているのかを把握しておくことが求められる。

＜参考＞
消費者的利用者は、自らデータを収集し、利用するAIの学習等を行うことが予定されている場合には、収集等において第三者のプライバシーを尊重する。

⑥ーウ）　自己等のプライバシー侵害への留意及びパーソナルデータ流出の防止

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIの判断により本人同意なくパーソナルデータが第三者に提供されることがないよう、同意が得られていないデータはシステム上第三者に提供できないこととするなど、適切な措置を講ずることが期待される。

＜参考＞
消費者的利用者は、ペットロボットなどのAIに過度に感情移入すること等により、特に秘匿性の高い情報（自己の情報のみならず他者の情報を含む。）をむやみにAIに与えることのないよう留意することが望ましい。

⑦尊厳・自律の原則

利用者は、AIシステム又はAIサービスの利活用において、人間の尊厳と個人の自律を尊重する。

主な論点

ア）　人間の尊厳と個人の自律の尊重

イ）　AIによる意思決定・感情の操作等への留意

ウ）　AIと人間の脳・身体を連携する際の生命倫理等の議論の参照

エ）　AIを利用したプロファイリングを行う場合における不利益への配慮

⑦ーア）　人間の尊厳と個人の自律の尊重

AIサービスプロバイダ及びビジネス利用者は、AIを利活用する際の社会的文脈を踏まえ、人間の尊厳と個人の自律を尊重することが期待される[1]。

1)例えば、人間とAIの異質性を前提に、AIは人間の活動を支援するものであるとの認識を持つこと等が考えられる。なお、人間とAIの異質性とは、人間とAIが異なる性質を持つことを言い、この前提が成り立つことにより、AIを人間と同様に扱うべきではないと考える（すなわち、人間の尊厳と個人の自律を尊重する）ことが可能となる。

＜参考＞

消費者的利用者は、AIを利活用する際の社会的文脈を踏まえ、人間の尊厳と個人の自律を尊重することが望ましい[1]。

⑦ーイ）　AIによる意思決定・感情の操作等への留意

AIサービスプロバイダ及びビジネス利用者は、消費者的利用者にはAIにより意思決定や感情が操作される可能性[1]や、AIに過度に依存するリスクが存在することを踏まえ、例えば以下のような対策を講じることが期待される。

[意思決定・感情操作に対する対策例]

- サービス提供時における利用者への注意喚起
- AIシステムを含んだコンピュータシステムの開発側による対応
- 教育現場等において、上記リスクがあることの共有の支援

1) AIによりナッジ（合理的選択のための支援）を行う場合など、AIによる消費者的利用者の意思決定・感情の操作はすべてがリスクにつながるとは限らないことから、ここでは「可能性」という語を用いている。なお、AIでナッジを行う際には、AI開発ガイドライン案における「⑧利用者支援の原則」（利用者に選択の機会を適切に提供する）を参照することが期待される。

＜参考＞

消費者的利用者は、開発者及びAIサービスプロバイダからの情報等[2]を踏まえ、AIにより意思決定や感情が操作される可能性や、AIに過度に依存するリスクがあることを認識することが望ましい。

2)　開発者及びAIサービスプロバイダから直接入手する情報のみならず、教育現場等において得られる情報等を含む。

⑦ーウ） AIと人間の脳・身体を連携する際の生命倫理等の議論の参照

AIサービスプロバイダ及びビジネス利用者は、AIを人間の脳・身体と連携させる場合、特に、エンハンスメント（健康の維持や回復を超えた人間の能力の増進の追求）を行う場合には、その周辺技術に関する開発者等からの情報を踏まえつつ、生命倫理の議論等を参照し、人間の尊厳と自律が侵害されないよう特に慎重に配慮することが期待される。

また、提供するAIの機能及びその周辺技術に関する情報を消費者的利用者に提供することが期待される。

＜参考＞

消費者的利用者は、AIを人間の脳・身体と連携させたAIを用いる場合には、AIの機能及びその周辺技術に関する開発者及びAIサービスプロバイダからの情報を踏まえ、自律性に影響を及ぼす可能性が生じうることに留意して、利用することが望ましい。

⑦ーエ）　AIを利用したプロファイリングを行う場合における不利益への配慮

AIサービスプロバイダ及びビジネス利用者は、個人の権利・利益に重要な影響を及ぼす可能性のある分野においてAIを利用したプロファイリングを行う場合には、対象者に生じうる下記の不利益等に慎重に配慮[1,2]する。

[プロファイリングにおいて不利益を生じさせることとなる例]

- プロファイリング結果が事実と異なることにより誤った判断が下されること
- 対象者の特定の特徴のみがプロファイリングで用いられることにより、対象者が過小に評価されてしまうこと
- 対象者のプロファイリング結果の一部が特定の集団の特徴と共通である場合に、当該集団にネガティブな判断が下されると、対象者も同様にネガティブな判断が下されうること
- プロファイリングの結果、特定の個人又は集団に対する不当な差別を助長するなど人の権利・利益を損なう取扱いがなされること
- プロファイリング結果をもとに不確実な未来を予測（外挿）する過程で、ネガティブな判断が入り込むこと
- 匿名の個人に関する情報に基づくプロファイリング結果と、特定の個人に関する情報に基づくプロファイリング結果とを突合することにより、匿名の個人が特定されてしまうこと

1) GDPRにおいては、同22条において、人間が介在せずに最終的な決定を自動処理のみにより行われない権利が保障されている。
2) 「①適正利用の原則」の「論点イ　人間の判断の介在」参照。

＜参考＞

消費者的利用者は、AIによるプロファイリングが行われている可能性があることを踏まえ、自らの情報が正しく利用されているかを意識し、必要に応じ、AIサービスプロバイダ及びビジネス利用者に確認することが望ましい。

⑧公平性の原則（全体構成）

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIシステム又はAIサービスの判断にバイアス[1]が含まれる可能性があることに留意し、また、AIシステム又はAIサービスの判断によって個人及び集団が不当に差別されないよう配慮する。

（注）「公平性」には集団公平性・個人公平性など、複数の基準があることに留意する必要がある。
論点イに基準の例が記載されている。

<u>主な論点</u>

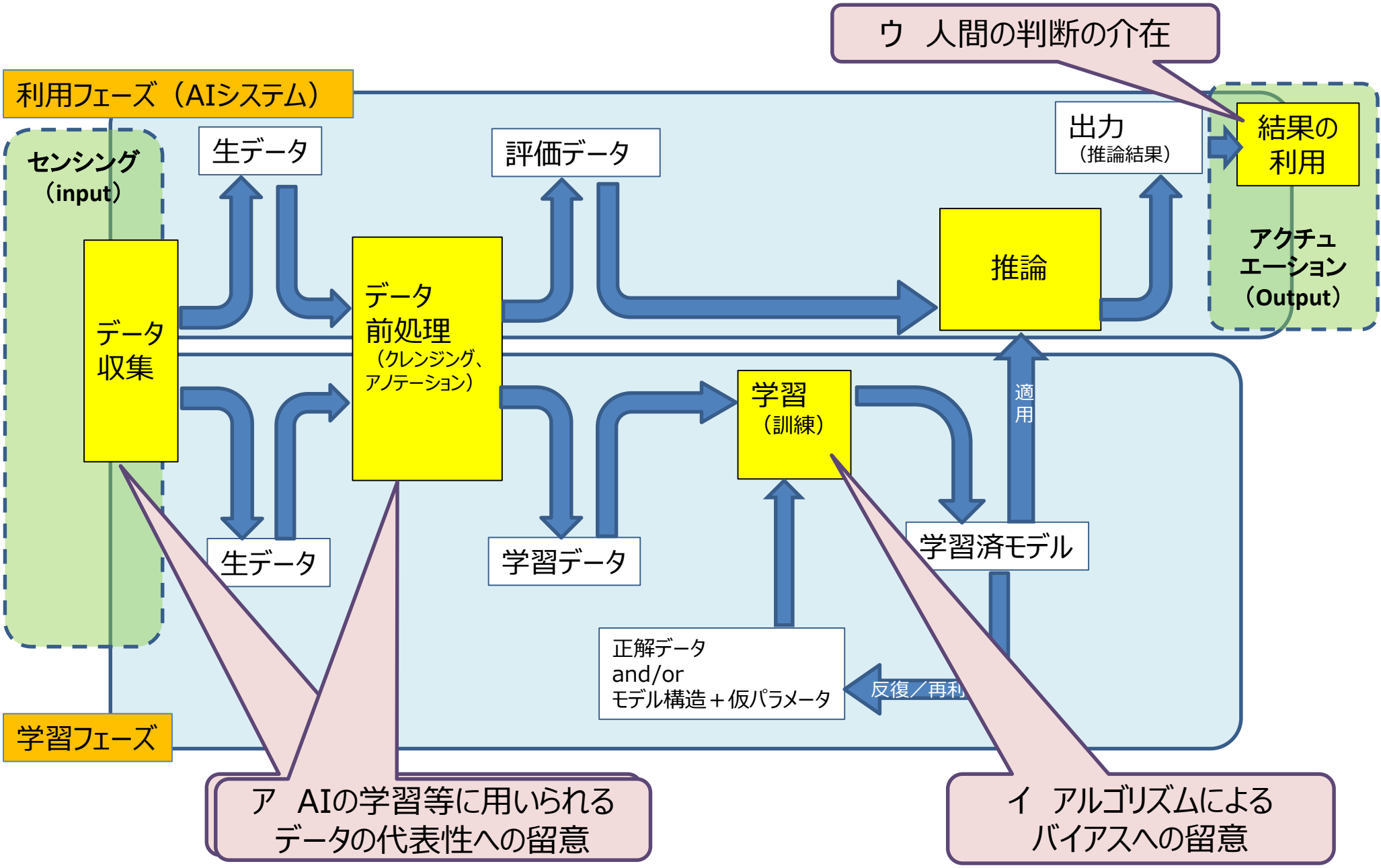
ア）　AIの学習等に用いられるデータの代表性への留意

イ）　学習アルゴリズムによるバイアスへの留意

ウ）　人間の判断の介在（公平性の確保）

1）「バイアス」という言葉には例えば以下の通り様々な解釈が考えられ、ここではそれらを総称したものとして用いている：
- 統計的な用語（サンプリングバイアス。偏り・偏差など。）
- 心理学的な用語（認知バイアス（思い込み等に起因。集団ごとの社会通念等による社会的バイアスも含む）、感情バイアス（人間の感情・都合等に起因）等）

（機械学習を中心とした）学習と利用の流れ（簡易版）と⑧公平性の原則の各論点



⑧ーア） AIの学習等に用いられるデータの代表性への留意

AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIの判断が学習時のデータによって決定づけられる可能性があることを踏まえ、AIを利活用する際の社会的文脈に応じて、以下のとおり、AIの学習等に用いられるデータの代表性[1]やデータに内在する社会的なバイアス等に留意することが期待される。

[公平性の観点からAIの学習等を行う際に留意すべき事項（例）]

- 不公平な判断が行われることのないようアルゴリズムが設計されていても、データの代表性が確保されないことによってバイアスが生じる可能性があることへの留意。
- センシティブ情報[2]を含む場合[3]に加え、センシティブ情報が含まれていない場合であっても、社会的バイアスを内在するデータを用いることによってバイアスが生じる可能性があること[4]への留意。
- （教師あり学習の場合）データの前処理において、学習データのラベルは多くの場合人間が作成・付与するため、（意図的にまたは意図せずに）ラベル付与を行う人のバイアスが入り込むことへの留意[5]。
- データの代表性を満足するためにパーソナルデータを含む大量のデータを集めようとする場合において、データに含まれる個人のプライバシーの尊重。

1) データの「代表性」とは、サンプルとして抽出され利活用に供されているデータが、その母集団の性質を正確に反映している度合いのことをいう。
2) 公平性の観点から排除すべき対象者の性別や人種等の個人の属性に関する情報。
3) センシティブ情報を含むデータを用いて学習等を行う場合に公平性を確保するための基準が検討されており、その一部を⑧ーイ[公平性の基準（例）]に掲載している。
4) 例えば、性別に依存しない採用試験を行おうとした場合に、仮に特定の項目に対する割合が男女間で相当程度異なるとすると、当該項目を一属性に加えて採用試験を行うアルゴリズムにより、結果として性別によるバイアスが生じる可能性がある。
また、ある集団ではセンシティブではないものが、別の集団ではセンシティブなものとして扱われるなど、国・地域などの集団ごとの社会通念の差異により、バイアスが生じる可能性があることにも留意が必要である。
5) 対策として、ラベリングの統一的な基準を作ることも考えられる。

＜参考＞
消費者的利用者は、AIの判断結果について疑義を感じた場合には、必要に応じて、開発者、AIサービスプロバイダ、ビジネス利用者等に問い合わせを行うことが望ましい。

⑧ーイ）　学習アルゴリズムによるバイアスへの留意

AIサービスプロバイダ及びビジネス利用者は、AIに用いられる学習アルゴリズムにより、AIの判断にバイアスが生じる可能性があることに留意することが期待される。特に、機械学習においては、一般的に、多数派がより尊重され、少数派が反映されにくい傾向にあり（バンドワゴン効果）、この課題を回避するため、例えば以下の方法が考えられる。

[機械学習アルゴリズムによるバイアスを生じさせないための方法（例）]

- AIを利活用する際の社会的文脈を踏まえ、センシティブ属性（公平性の観点から排除すべき対象者の性別や人種等の個人の属性）を明確化する[1]。
- センシティブ属性に関し確保すべき公平性の内容を、例えば以下の基準のとおり明確化する。
- 上記の公平性を満たす制約を機械学習アルゴリズムに付加する。
- ただし、（アルゴリズムにもよるが）公平性について上記の制約を課すことにより、機械学習の精度に影響を及ぼす可能性がある。

[公平性の基準（例）][2]

<集団公平性>

- センシティブ属性を取り除き、非センシティブ属性のみに基づき予測を行う（unawareness）。
- センシティブ属性の値が異なる複数のグループ間で、同じ予測結果を確保する（demographic parity）。
- 実際の結果に対する予測結果の誤差の比率を、センシティブ属性の値によらないように調整する（equalized odds）。

<個人公平性>

- センシティブ属性以外の属性値が等しい個人に対してはそれぞれ同じ予測結果を与える。
- 類似した属性値を持つ個人には類似した予測結果を与える（Fairness through awareness）。

1）例えば、入社試験で個人の属性情報から採否を判断する場合を考える。性別に依存することが問題視されている場合には、性別がセンシティブ属性となる。
2）1の例で、（機械学習）アルゴリズムにより採否を予測するとして、
-unawareness：性別に関する属性を取り除いてアルゴリズムを適用すること。
-demographic parity：アルゴリズムによる採否予測の分布（比率等）を男女間で同一となるよう調整すること。
-equalized odds：実データの男女間の採否の分布（比率等）がアルゴリズムによる男女間の採否予測の分布（比率等）と同一になるよう調整すること。

＜参考＞

消費者的利用者は、AIの判断結果について疑義を感じた場合には、必要に応じて、開発者、AIサービスプロバイダ、ビジネス利用者等に問い合わせを行うことが望ましい。

⑧ーウ）　人間の判断の介在（公平性の確保）

AIサービスプロバイダ及びビジネス利用者は、AIによりなされた判断結果の公平性[1]を保つため、AIを利活用する際の社会的文脈や人々の合理的な期待を踏まえ、その判断を用いるか否か、あるいは、どのように用いるか等に関し、人間の判断を介在させることが期待される。

人間の判断の介在の要否については、[①－イ] に掲げる内容を参照しつつ、公平性の観点から、例えば以下の基準も踏まえ、検討することが期待される。

[人間の判断の介在の要否について、基準として考えられる観点（例）]

- 統計的な将来予測が（不確定性が高く）難しい場合。[2]
- 意思決定（判断）に対し納得ある理由を必要とする場合。[3]
- 学習データにマイノリティなどに対する社会的バイアスが含まれていること等により、人種・信条・性別に基づく差別が想定される場合。

1) AIの学習に用いるデータに内在する社会的バイアス等がAIによる判断結果の公平性に影響を及ぼしうることを前提としている。
2) 例えば、人事においては、従業員の能力や生産性といった時間とともに変動する変数が用いられること、また、記録として残らない情報は用いることができないこと等から、統計的な将来予測が難しい。
3) 例えば、人事評価に当たっては、社員に対し評価の理由を説明できることが期待される。

⑨透明性の原則（全体構成）

AIサービスプロバイダ及びビジネス利用者は、AIシステム又はAIサービスの入出力等の検証可能性及び判断結果の説明可能性に留意する。

（注）本原則は、アルゴリズム、ソースコード、学習データの開示を想定するものではない。また、本原則の解釈に当たっては、プライバシーや営業秘密への配慮も求められる。

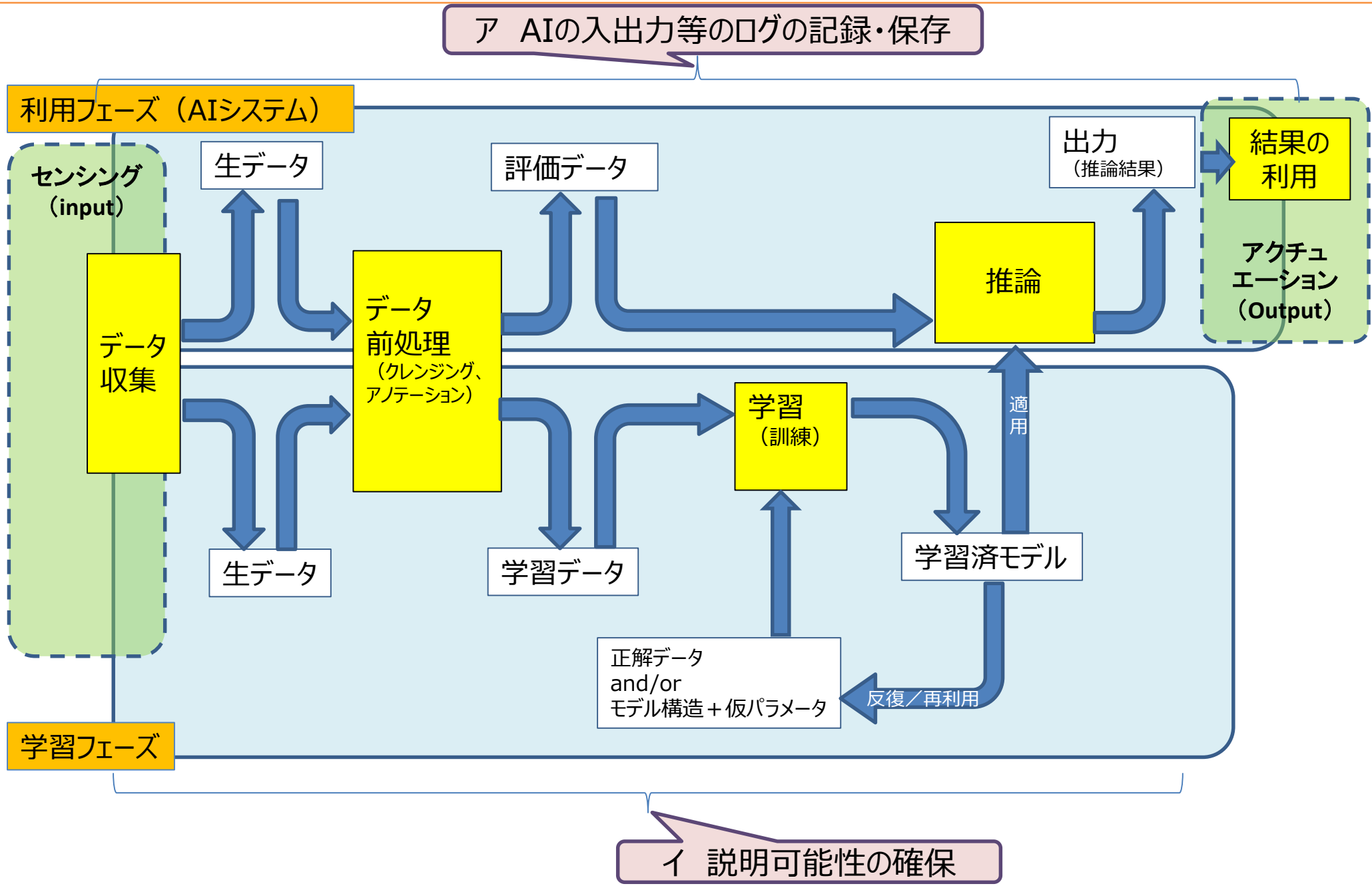
主な論点

ア）　AIの入出力等のログの記録・保存

イ）　説明可能性の確保

ウ）　行政機関が利用する際の透明性の確保

（機械学習を中心とした）学習と利用の流れ（簡易版）と⑨透明性の原則の各論点



⑨ーア） AIの入出力等のログの記録・保存

AIサービスプロバイダ及びビジネス利用者は、AIの入出力等の検証可能性を確保するため[1]、入出力等のログを記録・保存することが期待される。ログの記録・保存に当たっては、利用する技術の特性及び用途に照らして、例えば以下の事項について考慮することが期待される。

[ログの記録・保存に当たり、考慮すべき事項（例）][2]

- ログの記録・保存の目的（人の生命・身体・財産に危害を及ぼしうる分野における事故の原因究明や再発防止を目的としたものであるか等）[3]
- ログ取得・記録の頻度
- ログの精度
- ログの保存期間
- ログの保護（機密性、完全性等の確保）
- ロク保存場所の容量
- ログの時刻の記録
- 開示するログの範囲

1) 入出力等の検証可能性を確保することが期待される場面としては、事故発生時において原因解明を行う場合に加え、最終利用者がAIシステム又はAIサービスを誤って使用していないか、悪意をもって使用していないかについて確認を行う場合等も想定される。

2) ここで記載した考慮すべき事項は相互にトレードオフの関係にある場合があり、AIを利活用する際の社会的文脈やAIの用途に応じてバランスを考慮することが必要である。例えば、ログ取得頻度や保存期間等はログの機密性、完全性の確保とトレードオフの関係にある。

3) 人の生命・身体・財産に危害を及ぼし得る分野では、事故の原因究明や再発防止の必要性が高いことから、ログ取得・記録の頻度を高めたり、ログの精度を上げたり、保存期間を長期にする等の対応が必要となることも想定される。

⑨ーイ） 説明可能性の確保

AIサービスプロバイダ及びビジネス利用者は、個人の権利・利益に重大な影響を及ぼす可能性のある分野においてAIを利用する場合など、AIを利活用する際の社会的文脈を踏まえ、利用者の納得感や安心感の獲得、また、そのためのAIの動作に対する証拠の提示等を目的として、AIの判断結果の説明可能性を確保することが期待される。その際、AIを利活用する際の社会的文脈を踏まえ、どのような説明が求められるかを分析・把握し、以下の手段を参考に総合的な対応を講じることにより、AIの判断結果の説明可能性を確保することが期待される。

[説明可能性確保のための手段（例）]

（解釈可能なアルゴリズムを実装するAIソフトの採用）

- 利用するAIソフトとして、予め可読性の高い解釈可能なモデル[1]を採用する。

（アルゴリズムによる判断結果の説明を行うための技術的手法の採用）

- ブラックボックスモデルに対する説明を可能とする以下の技術的手法[2]を採用する
  - 「AIの予測・認識プロセスの可読化」など、解釈可能なモデルに置き換えて説明を行う大域的な説明法
  - 「重要な特徴の提示」、「重要な学習データの提示」およびその「自然言語による表現」など、特定の入力に対する予測の根拠を提示する局所的な説明法

（データの来歴の管理）

- AIの学習等に利用されたデータが、いつ、どこで、どういう目的で集められたデータなのかを管理（data provenance）する。

（学習モデルの入出力の傾向の分析）

- AIに対する複数の入力と出力の組合せをもとに、AIの判断の傾向を分析する（例えば、入力パターンを少しずつ変化させたときの出力の変化の観測など）。

（総合的な対策）

- 消費者的利用者等のニーズ、意見等も踏まえつつ、説明が不足している部分を明確にし、どのような説明が必要か、開発者とも連携して解決策を模索する[3]。

1) 一般に、学習モデルを解釈可能な形とすることは、AIの判断の精度の確保とトレードオフの関係にある。
2) 技術的な「説明可能性」の確保には実装や検証等が必要となるため、一般にコストとトレードオフの関係にある。
3) 「必要とする説明の明確化」と開発者による「その説明に関する技術開発」が相互に繰り返され、当該技術が広く共有されることにより、説明可能性に関する課題の本質的な解決へとつなげることが期待できる。

⑨ーウ）　行政機関が利用する際の透明性の確保

行政機関がAIを利用する場合には、法の支配、行政の透明性確保、適正手続等の要請を踏まえ、AIを利活用する際の社会的文脈に応じ、AIの判断結果の説明可能性を確保することが期待される。なお、説明可能性を向上させるため、例えば以下の方法などが考えられる。

[説明可能性を向上させるための方法の例]

- 行政機関が利用するAIのアルゴリズムの開発・設計プロセスに、様々な社会的少数派を包摂すること（コ・デザイン）
- 学習データの構成の考え方（学習データへの包摂・排除の考え方）、アルゴリズムの設計段階において行った政策的判断、AIを導入することによる社会的影響評価、AIに対する監査方法等を説明すること
- AIの判断を説明する諸要素について、開発者やAIサービスプロバイダが不開示とする範囲を限定した形で開発者やAIサービスプロバイダと契約を締結すること

⑩アカウンタビリティ※の原則

利用者は、ステークホルダに対しアカウンタビリティを果たすよう努める。

主な論点

ア）　アカウンタビリティを果たす努力

イ）　AIに関する利用方針の通知・公表

※アカウンタビリティ：　判断の結果についてその判断により影響を受ける者の理解を得るため、責任者を明確にした上で、判断に関する正当な意味・理由の説明、必要に応じた賠償・補償等の措置がとれること。

⑩ーア）　アカウンタビリティを果たす努力

AIサービスプロバイダ及びビジネス利用者は、人々と社会からAIへの信頼を獲得することができるよう、本ガイドラインの掲げる利活用原則①～⑨の趣旨に鑑み、消費者的利用者、AIの利活用により影響を受ける第三者等に対し、利用するAIの性質及び目的等に照らして、それぞれが有する知識や能力の多寡に応じ、AIシステムの特性について情報提供と説明を行うことや、多様なステークホルダとの対話を行うこと等により、相応のアカウンタビリティを果たすよう努めることが期待される。

＜参考＞

消費者的利用者は、それぞれが有する知識や能力の多寡に応じ、相応のアカウンタビリティを果たすよう努めることが望ましい。

また、AIの判断結果について疑義を感じた場合には、必要に応じて、開発者、AIサービスプロバイダ、ビジネス利用者等に問い合わせを行うことが望ましい。

⑩ーイ） AIに関する利用方針の通知・公表

AIサービスプロバイダ及びビジネス利用者は、消費者的利用者等がAIの利活用について適切に認識できるよう、以下のとおり、AIに関する利用方針を作成・公表し、通知を行うことが期待される。

(i) AIの判断が直接に消費者的利用者や第三者に対して影響を及ぼす態様によりAIを利活用する場合には、消費者的利用者や第三者がAIの利活用について適切に認識することができるよう、以下の事項を参考にAIに関する利用方針を作成・公表し、問い合わせがあった場合には通知を行う。

(ii) (i)について、消費者的利用者や第三者の権利・利益に重大な影響を及ぼす可能性のある場合には積極的に通知を行う[1]。

[AIに関する利用方針に記載する事項（例）]

- AIを利用している旨（具体的な機能・技術を特定できるのであれば、その名称と内容等[2]）
- 利活用の範囲及び方法
- 利活用に伴うリスク
- 相談窓口

また、通知又は公表は、利用開始前だけではなく、AIの動作に変更が生じたときや利用終了時も含め（特にAIの動作変更に伴い想定されるリスクに変更が生じる場合など）実施することが期待される。

1) AIサービスプロバイダ及びビジネス利用者が、AIに関する利用方針を公表することが求められるのは、利活用するAIの判断が、消費者的利用者や第三者に直接の影響を及ぼす場合であると考えられる。すなわち、人間の思考に供するための分析道具としてAIを利活用するにとどまる場合や、AIが原案を作成しつつも、最終的に人間が判断することが実質的に担保されている場合には、AIに関する利用方針の公表が必ずしも求められるわけではない。（もっとも、そうした場合であっても、自主的に公表されることが望ましい。）

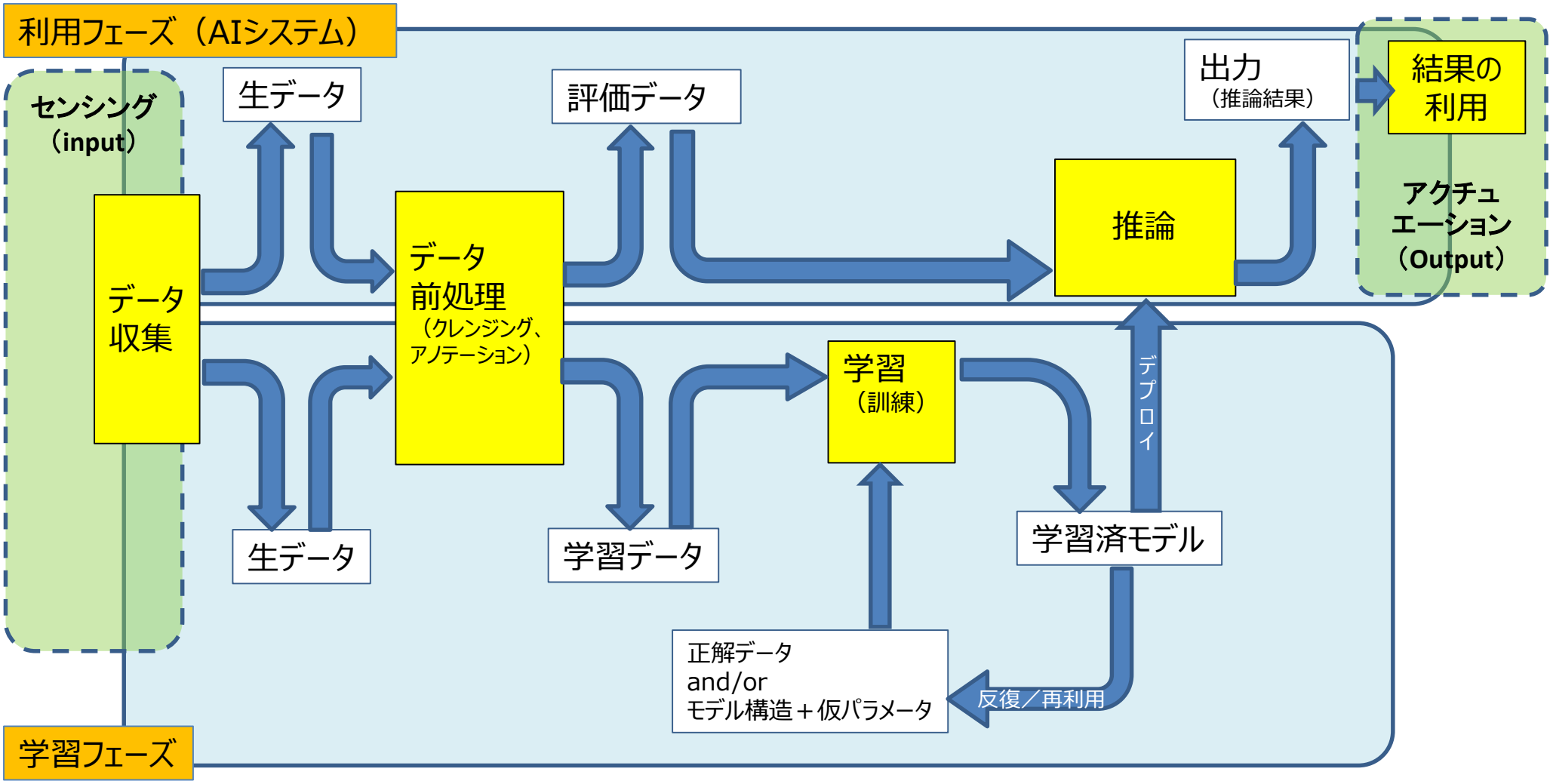
2) 例えば、「XXのサービスにおいて、深層学習モデルによる判定が行われており、想定どおりの動作を100%保証できない（ため、利用時はご注意いただきたい）」など、利用しているAIの機能・特徴とともに適正な利活用の方法や利活用に伴うリスクも含めて公表することが考えられる。

＜参考＞

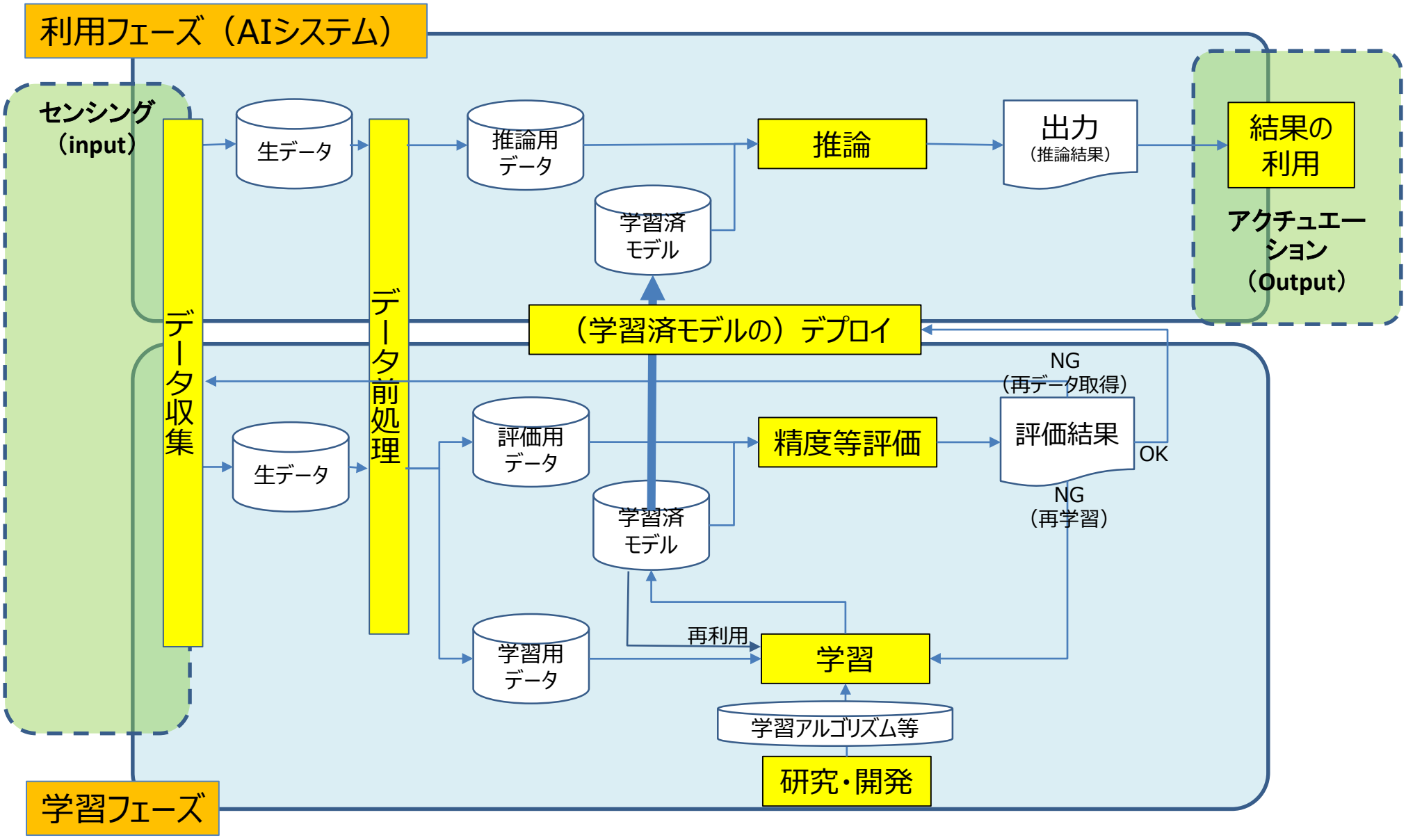
消費者的利用者は、AIの判断結果について疑義を感じた場合には、必要に応じて、開発者、AIサービスプロバイダ、ビジネス利用者等に問い合わせを行うことが望ましい。

（参考１）AIシステムの概要

（機械学習を中心とした）学習と利用の流れ（簡易版）



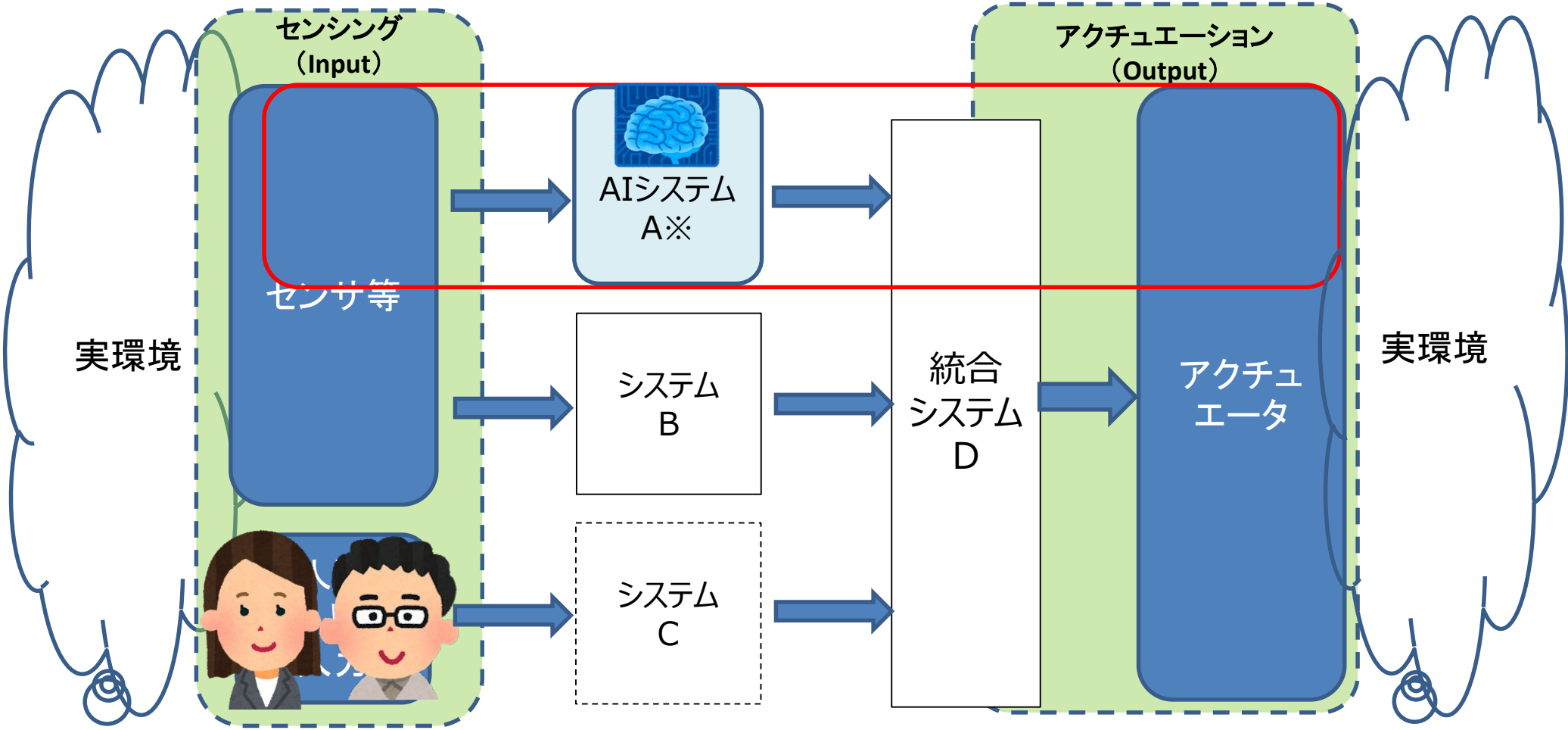
（機械学習を中心とした）学習と利用の流れ



（他システムと統合された）AIシステムの例

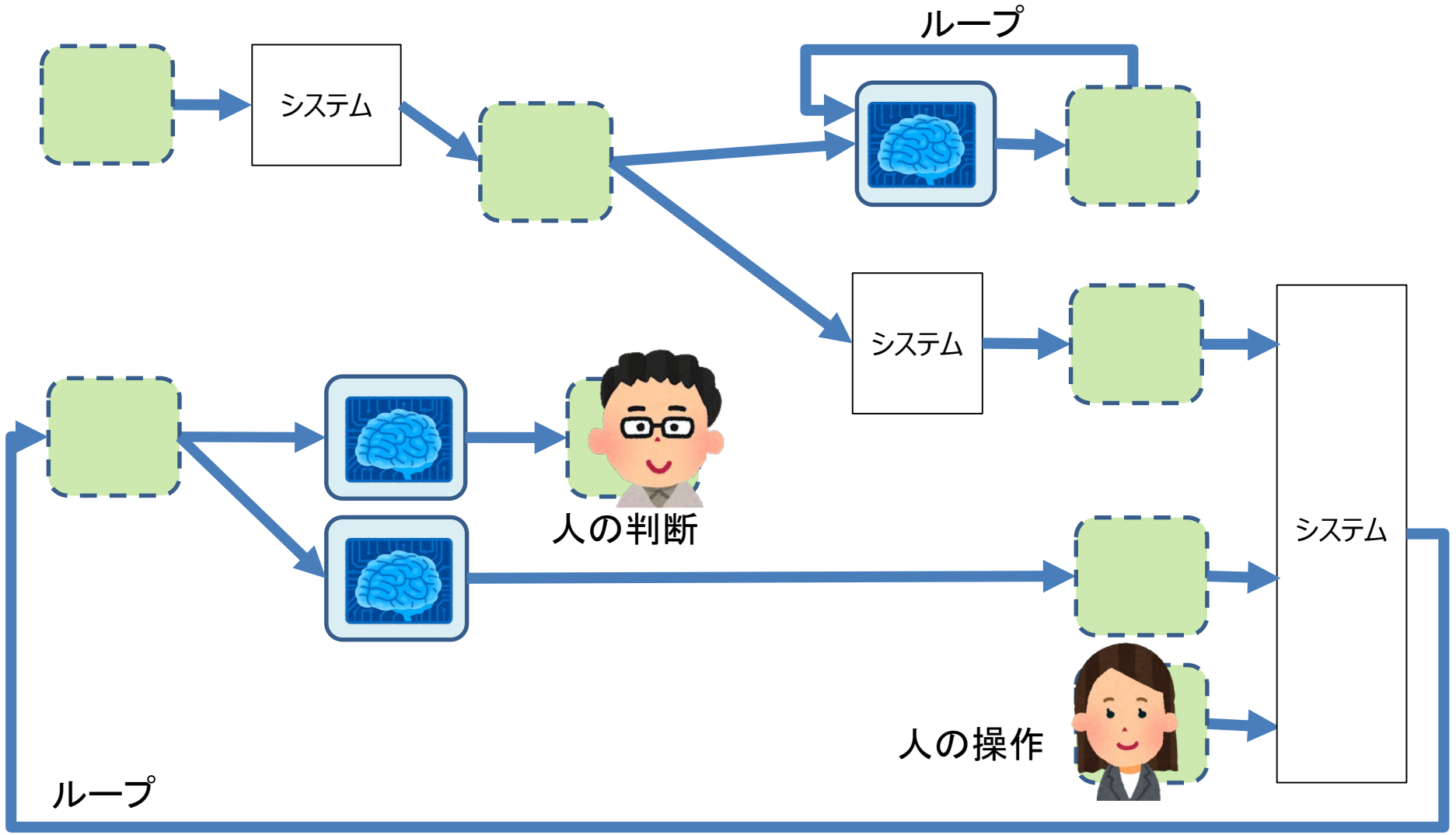
実環境からセンサ等を経て得られる入力を受け、AIシステムA及びシステムBにより処理された結果並びに人間による入力を受けシステムCにより処理された結果が、システムDで統合され、実環境に影響を及ぼすアクチュエータの制御に反映される。

（例）自動運転において、通常時は、センシングされた周辺画像情報等をもとに走行、停止等の推論を行うAIシステムAが統合システムDに指示を行い、それに基づき自動運転車の動作に係わるアクチュエータの制御が行われるが、人間が危険を察知した場合等の異常時への対応のためキルスイッチ（安全に停止するためのスイッチ）がシステムCに設けられており、人間からの強制停止の入力を受け、統合システムDがアクチュエータに対し指示を出し、自動運転車を安全に停止させるシステム。



※前スライドの利用フェーズ（AIシステム）に相当

AIシステムは他（AI）システムとも相互に連携

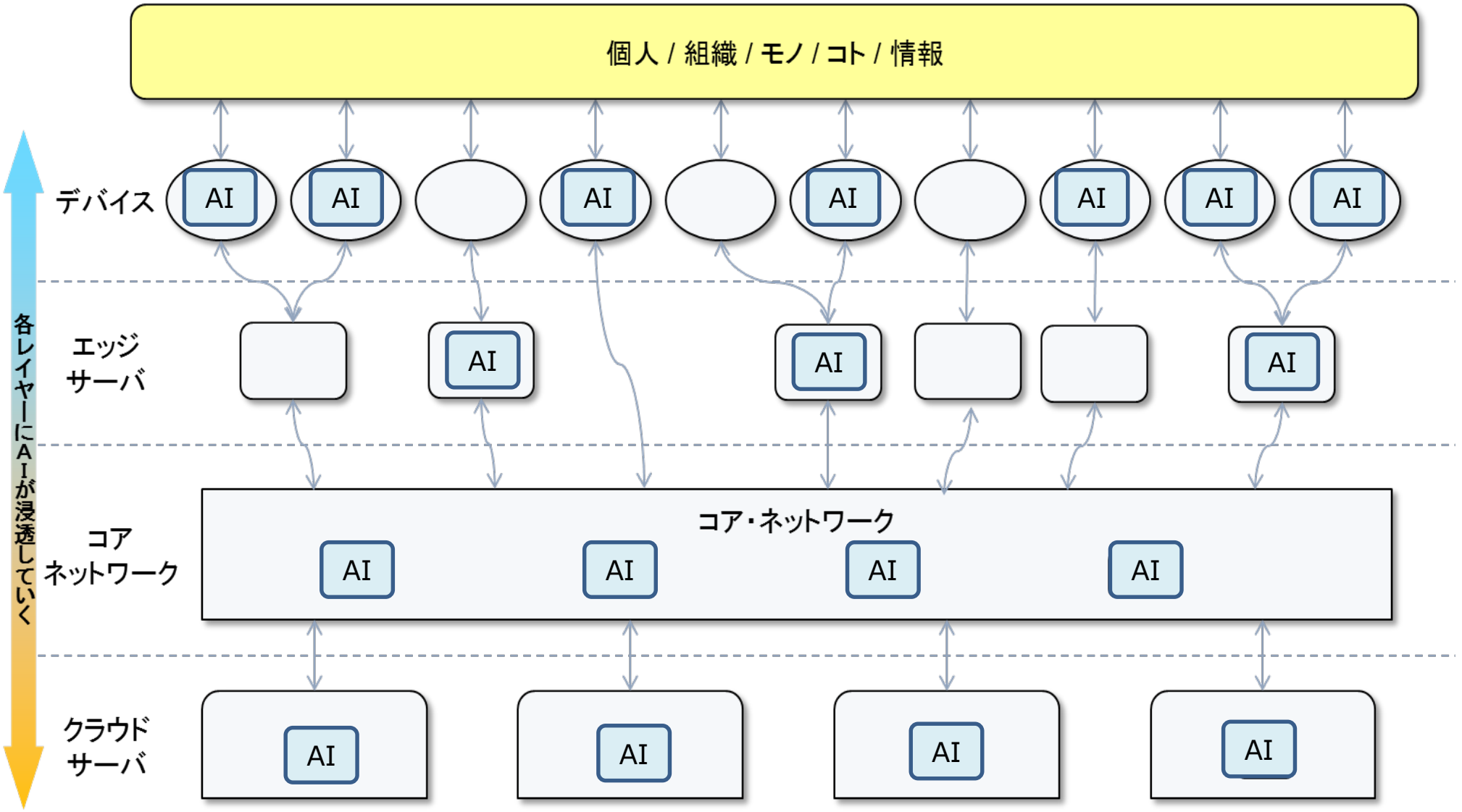


システム　一般のシステム

AIシステム

Input：センサ等の情報、人間による操作、判断
Output：判断情報、（それによる）アクチュエーション
に相当

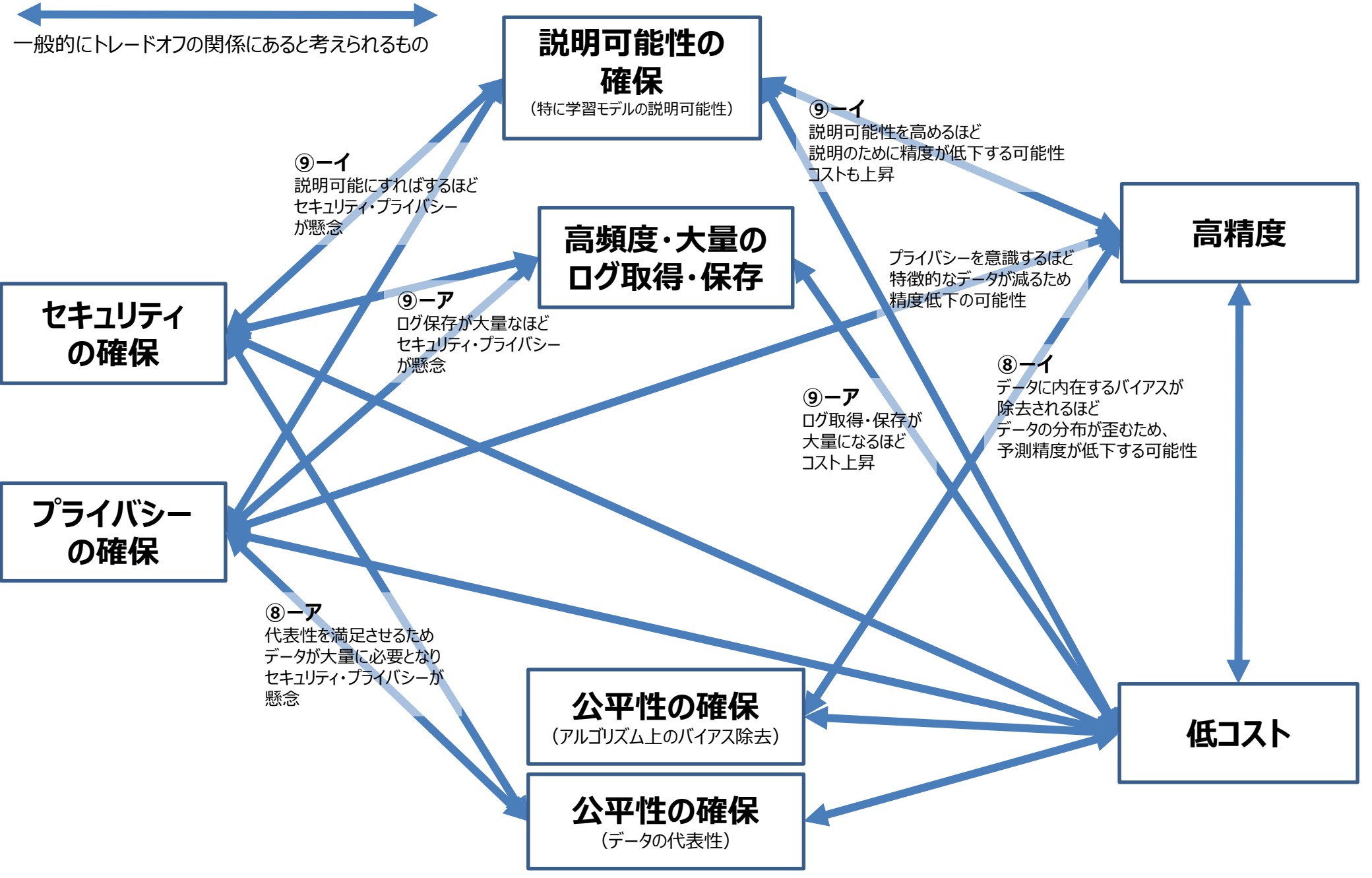
AI（システム）は様々なレイヤに浸透し、相互に連携・協調



引用：AIネットワーク化検討会議中間報告書「AIネットワーク化が拓く智連社会（ＷＩＮＳ(ウインズ)）―第四次産業革命を超えた社会に向けて―」

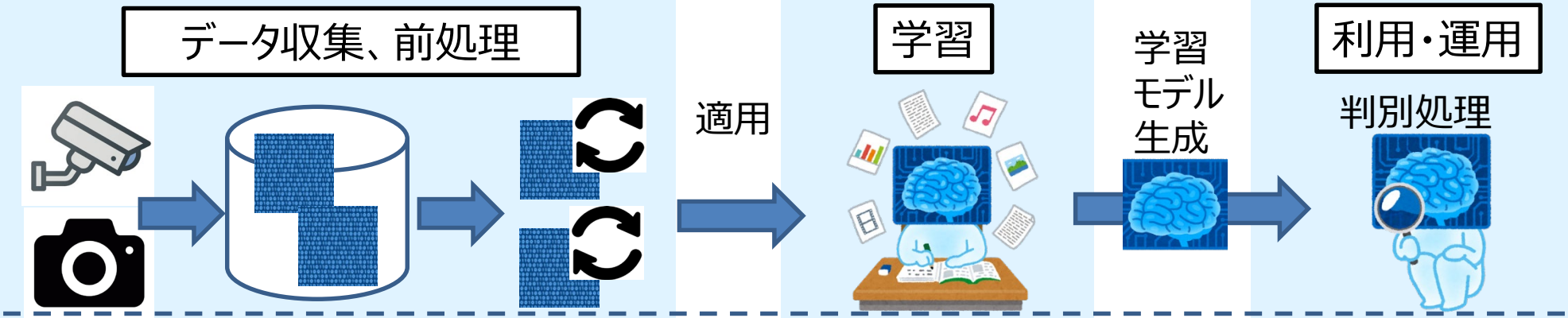
（参考２）各項目のトレードオフの例

トレードオフの例


一般的にトレードオフの関係にあると考えられるもの
説明可能性の確保
（特に学習モデルの説明可能性）
⑨ーイ
説明可能にすればするほど
セキュリティ・プライバシー
が懸念
⑨ーイ
説明可能性を高めるほど
説明のために精度が低下する可能性
コストも上昇
高頻度・大量の
ログ取得・保存
高精度
セキュリティ
の確保
⑨ーア
ログ保存が大量なほど
セキュリティ・プライバシー
が懸念
プライバシーを意識するほど
特徴的なデータが減るため
精度低下の可能性
⑧ーイ
データに内在するバイアスが
除去されるほど
データの分布が歪むため、
予測精度が低下する可能性
⑨ーア
ログ取得・保存が
大量になるほど
コスト上昇
プライバシー
の確保
⑧ーア
代表性を満足させるため
データが大量に必要となり
セキュリティ・プライバシーが
懸念
公平性の確保
（アルゴリズム上のバイアス除去）
低コスト
公平性の確保
（データの代表性）

（参考３）原則の適用例

例：人の映像を入力とし、罪を犯す可能性があるかどうかを判断する仕組み



[公平性の原則]
データの代表性確保
特定の地域の人のみを対象としない
[公平性の原則]
学習時のラベリングにおけるバイアス
怖そう＝犯人になりそうと意図的にラベリングしない
[透明性の原則]
データをどのように集めたのかを管理
[プライバシーの原則]
データ収集・前処理等における
プライバシー尊重

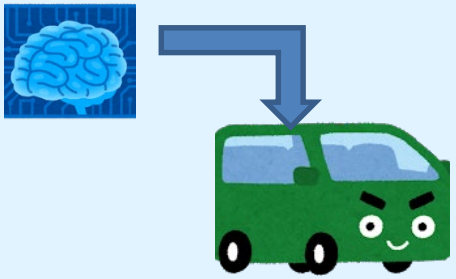
[公平性の原則]
アルゴリズム上のバイアス
人種差別などが起こらないような計算上の配慮

[適正利用の原則]
人間の判断の介在
AIの判断に影響を受ける最終利用者の権利を考慮

AI利活用ガイドライン（例）

例：自動運転

AIを含むシステムの構築



デプロイ
（利用可能な状態にすること）

利用・運用

[安全の原則]
AI単体ではなく、**システム全体での安全性の確保**（フェイルセーフ）

[セキュリティの原則]
システムハックがおこらないような**その時点での技術水準に照らした合理的な対策**

[安全・セキュリティの原則]
侵害された場合に講ずるべき措置の周知

[適正利用の原則]
人間の判断の介在
人間の判断に状態が移行する条件などの周知等

[安全・セキュリティの原則]
AIを含むシステムアップデート、そのための情報提供

[透明性・アカウンタビリティの原則]
事故時の**説明可能性の確保**、**アカウンタビリティを果たす努力**

[連携の原則]
- 自動運転車間の交渉・調整の必要性、そのための**データ形式／プロトコルへの対応**
- 個々のトラブルがシステム全体に波及するおそれへの対応