別紙１－１

AI利活用原則の各論点に対する詳説（案）

平成３１年４月
事務局

凡例

AI SP、ビジネス利用者等	「AIサービスプロバイダ、ビジネス利用者等」の視点で記載した内容
消費者的利用者	「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

AISP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、AIを消費者的利用者等に提供し、又は、自ら運用するに当たり、開発者等からの情報提供や説明を踏まえ、以下に関する情報を適時適切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の利用に関する適正な用途・方法。 • AIの性質、利用の態様等に応じた、便益及びリスクに関する情報。 • 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）、及び確認の重要性、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、AIソフトのアップデート及びAIの点検・修理等を行うための情報。 [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備え、AIの性質、利用の態様等に基づき考えられるリスクに応じ、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、アップデートのための機能を提供する際に、他のAIとの関係でリスク[1]が想定される場合は、当該リスク情報を提示した上で提供することが望ましい。 また、提供されるAIの性質、利用の態様等によっては、提供対象となる利用者が当該AIを提供するにふさわしい者であるか（信頼性）について事前に考慮することが期待される場合も想定される。 1)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しうる。例えば、家庭内の家電に含まれるAIソフトの動作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の内容を把握していないと、（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る場合がある（「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）。
消費者的利用者	消費者的利用者は、開発者、AIサービスプロバイダ、ビジネス利用者等からの情報提供や説明を踏まえ、社会的文脈や状況にも配慮して、AIを適正な範囲・方法で利用することが期待される。具体的には以下に留意することが期待される。 [実施が期待される内容] （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等。 （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、他のAIとの関係でのリスクが存在しうること[1]を理解した上で、アップデートを行うこと。 • 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、もしくは、問題に対するフィードバッグを要求された場合、開発者及びサービスプロバイダ等に対し、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 SP、ビジネス利用者等	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、例えば以下の基準の下、利用分野、用途等に応じて検討され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・利益の性質及び意向（例えば、人生を左右しかねない意思決定に係わるAIの利用時等） • 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 • 利用者に期待される能力 • 判断対象の要保護性（例えば、AIによる大量申請への対応等） また、AIの判断に対し、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。こうした場合に人間が行うべき判断についてその項目、手段などを明確化することにより、人間の判断の実効性を確保することが考えられる。 ［実効性を確保するための手段の例］ • 説明可能性を有するAIから得られる説明を前提として、最終判断に必要となるチェック項目の作成 • AIの判断の適正性を確認するためのチェック項目の作成（能動的な判断の実施（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）） 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、移行前、移行中、移行後等の各状態に伴い、予め責任の所在が明確になっている必要がある。また、前述の移行条件、移行方法等を利用者に事前に告知し、必要な訓練などを前もって実施するなど、人間による稼働に移行した場合に問題が起こらないよう、注意喚起をしておくことが期待される。
消費者的利用者	消費者的利用者は、AIの判断に対し、人間が最終判断をすることが適当とされている場合に、適切に判断ができるよう能力を習得しておくことが期待される。加えて、こうした場合に人間が行うべき判断について開発者、AIサービスプロバイダ、ビジネス利用者等により人間が行うべき判断についての項目、手段などが整理されている場合は、当該情報を入手し、これに基づき対応することが期待される。 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、移行前、移行中、移行後等の各状態に伴う責任の所在を予め認識しておく必要がある。また、前述の移行条件、移行方法等についての説明をサービスプロバイダ等から受け、必要な訓練などを受けておくことが期待される。

①ーウ）　関係者間の協力

AI SP、ビジネス利用者等	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を提供または利用するに当たり、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る被害の性質・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。 また、具体的には、例えば以下に記載の内容等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（例）] • ①適正利用の原則（本原則）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相互提供等） • ④安全の原則（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（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（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）等
消費者的利用者	消費者的利用者は、AI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る被害の性質・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。 また、その実効性を確保するため、開発者、AIサービスプロバイダ及びビジネス利用者等が提供する情報に基づき、関係者と協力して適切に対応することが期待される。

②適正学習の原則

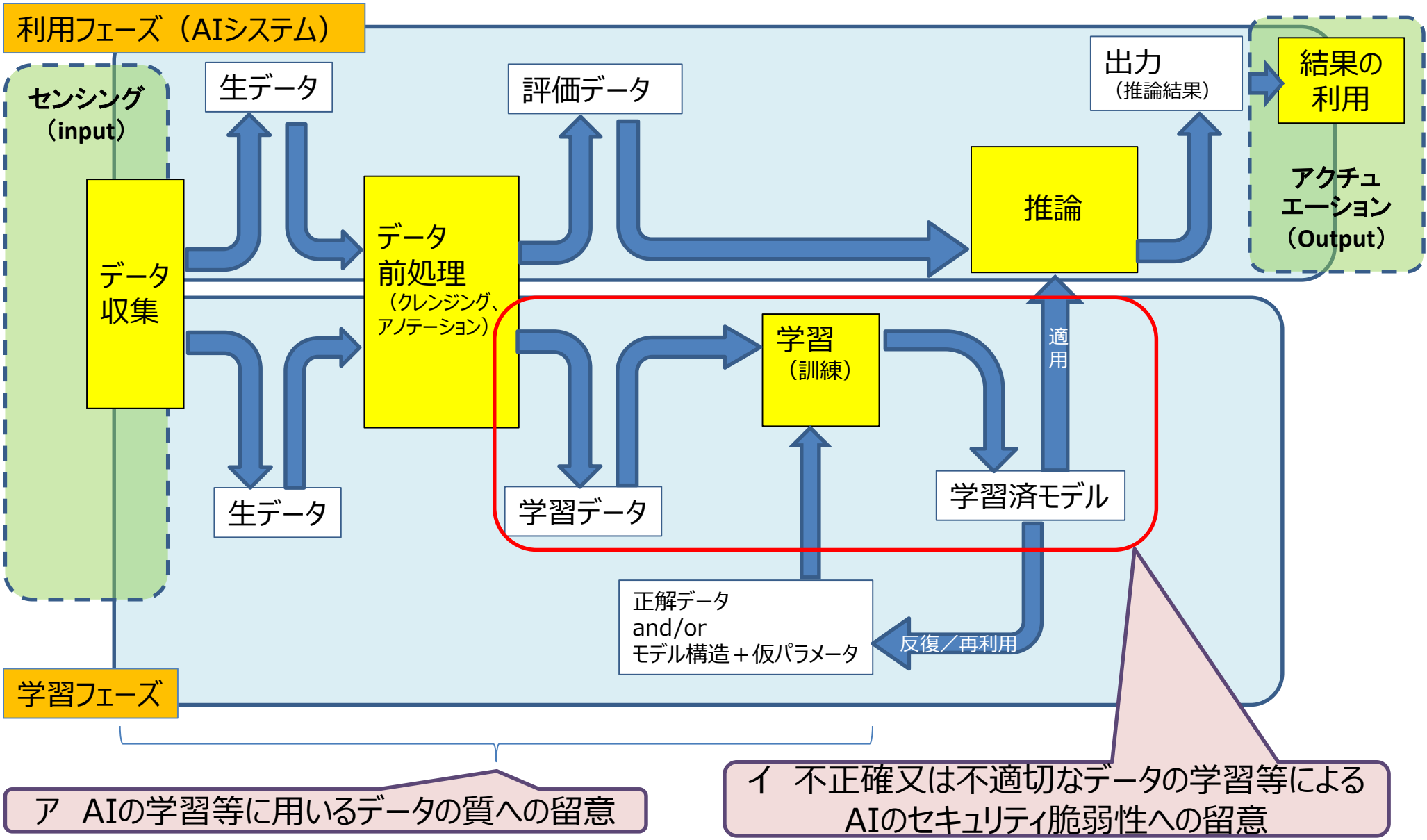
利用者及びデータ提供者は、AIシステムの学習等に用いるデータの質に留意する。

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（簡易版）



②ーア） AIの学習等に用いるデータの質への留意

AISP、ビジネス利用者等	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、提供もしくは利用する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。なお、特に機械学習においてデータの質を担保するための方法として、例えば以下の対策が考えられる。 また、AIによりなされた判断の精度が損われたり、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、実装コスト、及び、技術水準等[1]を踏まえ、精度に関する基準を予め定めておくことが期待される。他方、精度が当該基準を下回った場合には、データセットの質に留意して改めて学習させることが期待される。 加えて、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、提供もしくは利用する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、消費者的利用者に情報を提供することが期待される。 [データ収集時の対策（例）] • 過去のデータ収集時の経験を活かし、収集するデータが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などデータ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も注意する。特に個人情報については同意の取得などの留意が必要となる。 [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[2]。 • 一方、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[3]。 • （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、ログを取得するなど明確にしておく。 [学習時の対策（例）] • 既存の学習モデルを利用して転移学習[4]等を行う。 • 学習の精度を上げるため、特定のデータを拡張[5]した上で学習を行う。 • 特に一過性のある時系列データを学習する場合などは、どの範囲のデータを学習対象とすべきかを見極める。 1)例えば、機械学習を中心としたAIが帰納的であることから、当該AI単体で原理的に100%の精度を担保できないこと等が挙げられる。 2)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。 3)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。 4)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。 5)「データの拡張」(Data Augmentation)とは、データの正確性を高めるに当たって、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）を高めるためにとられる手段の１つである。学習に用いる当該データを拡張し（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。
消費者的利用者	消費者的利用者は、利用するAI等の学習に用いるデータを自ら収集することが予定されている場合には、（データの提供に関する）手段、形式等について、開発者、AIサービスプロバイダ、ビジネス利用者等からの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。

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 SP、ビジネス利用者等	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。 [リスクの可能性の例] • 学習モデルが、一部の学習データからの微少な変動により悪影響を受ける場合に、悪意あるインプット等により当該学習モデルが誤った結果を導出する可能性（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、間違った学習が行われる可能性
消費者的利用者	消費者的利用者は、サービスプロバイダ、ビジネス利用者及びデータ提供者等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が存在することを認識しておくことが期待される。 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、AIサービスプロバイダ、ビジネス利用者、データ提供者等にその旨を報告することが期待される。

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。

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 SP、ビジネス利用者等	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。
消費者的利用者	N/A

③ーイ）　データ形式やプロトコル等の標準化への対応

AI SP、ビジネス利用者等	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、以下AIの入出力等におけるデータ形式（構文（syntax）及び意味（semantics）[1]）や、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。
消費者的利用者	消費者的利用者は、利用するAI等の学習に用いるデータを自ら収集する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、ビジネス利用者等から提供された情報を踏まえた上で収集、保存を行うことが期待される。

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 SP、ビジネス利用者等	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、他者に提供し、または、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した上で、提供・利用するAIの設計を行うことが期待される。また、開発者等からの情報を基に考えられるリスクを分析し、当該情報を連携の相手方と共有するとともに、問題が生じた場合の対応策等を作成の上、消費者的利用者に情報提供することが期待される。 ［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
消費者的利用者	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期待される。また、問題が生じた場合の対応策等について、開発者、AIサービスプロバイダ及びビジネス利用者等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・増幅される可能性と対応策の例］ • 個別の事業者のトラブル等がシステム全体に波及する可能性がある場合には、開発者、AIサービスプロバイダ及びビジネス利用者等からの情報を踏まえ、当該事業者のシステムを停止させる、もしくは、システム全体を停止させるなどの処理を行う。 • AIが想定外の動作を行う場合は、開発者、AIサービスプロバイダ及びビジネス利用者等からの情報を踏まえ、他AIまたは他システムとの連携を解消する、もしくは当該AIシステム全体を停止させる等の処理を行う。 • 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があることに留意し、問題の予兆が見られる場合には、当該事業者にその旨報告する。

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、利用者等及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。

論点

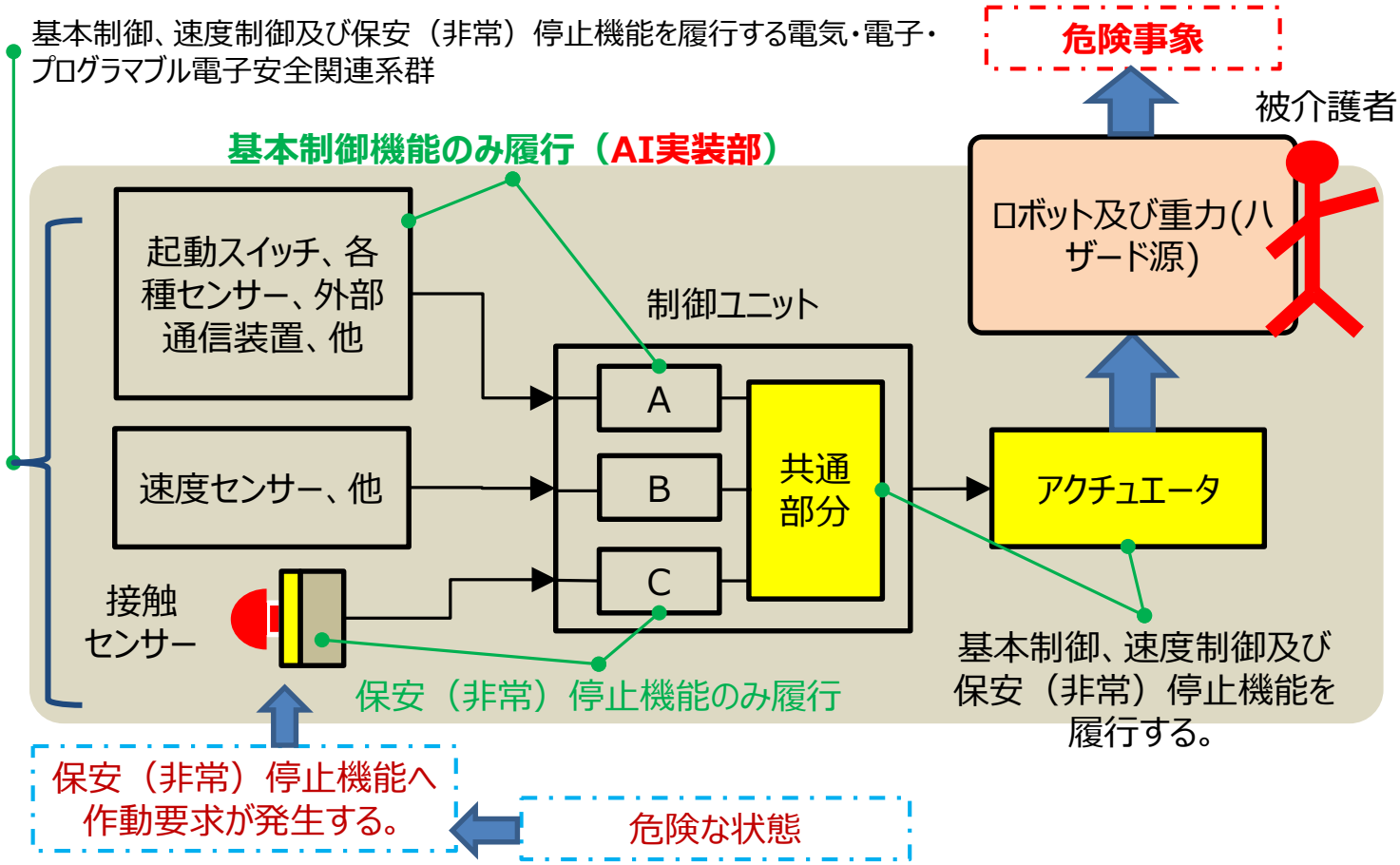
ア）　人の生命・身体・財産への配慮

④ーア　人の生命・身体・財産への配慮

AI SP、ビジネス利用者等	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基に、必要に応じて下記の対応策を講じることにより、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。 [対処方法の例] • AIの点検・修理及びAIソフトのアップデートを行う、また消費者的利用者にこれらの実施を促す • AIにはリスクがあるとの前提で、フェールセーフ[1]を意識し、AIが想定外の動作を起こした場合も、その外部で安全を確保できる仕組み[2]を構築する 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、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) 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 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。
消費者的利用者	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、AIサービスプロバイダ及びビジネス利用者からの情報を基に、必要に応じてAIの点検・修理及び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期待される。 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、AIサービスプロバイダ及びビジネス利用者から提供された情報に留意することが期待される。

【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。

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）　不正確又は不適切なデータの学習によるAIの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 SP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、AIはfragileであるという認識を持つ[1]と同時にAIのセキュリティに留意し、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、あらかじめ整理しておくことが期待される。また、その措置の内容について、開発者等からの情報を踏まえ、当該AIの用途や特性、侵害の影響の大きさ等社会的文脈に応じたものとすることが期待される。 [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[2]、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）例えば、学習モデルから学習データをリバースエンジニアリングできる可能性があることが報告されている。 2）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。
消費者的利用者	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、AIのセキュリティに留意し、必要な対策を講ずることが期待される。

⑤ーイ）　セキュリティ対策のためのサービス提供等

AISP、ビジネス利用者等	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。
消費者的利用者	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、AIサービスプロバイダ及びビジネス利用者から提供された情報に留意することが期待される。

⑤ーウ） 不正確又は不適切なデータの学習によるAIのセキュリティ脆弱性への留意

[②ーイ 再掲]

AI SP、ビジネス利用者等	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。 [リスクの可能性の例] • 学習モデルが、一部学習データからの微少な変動により悪影響を受ける場合に、悪意あるインプット等により当該学習モデルが誤った結果を導出する可能性（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、間違った学習が行われる可能性
消費者的利用者	消費者的利用者は、サービスプロバイダ、ビジネス利用者及びデータ提供者等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が存在することを認識しておくことが期待される。 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、AIサービスプロバイダ、ビジネス利用者、データ提供者等にその旨を報告することが期待される。

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。

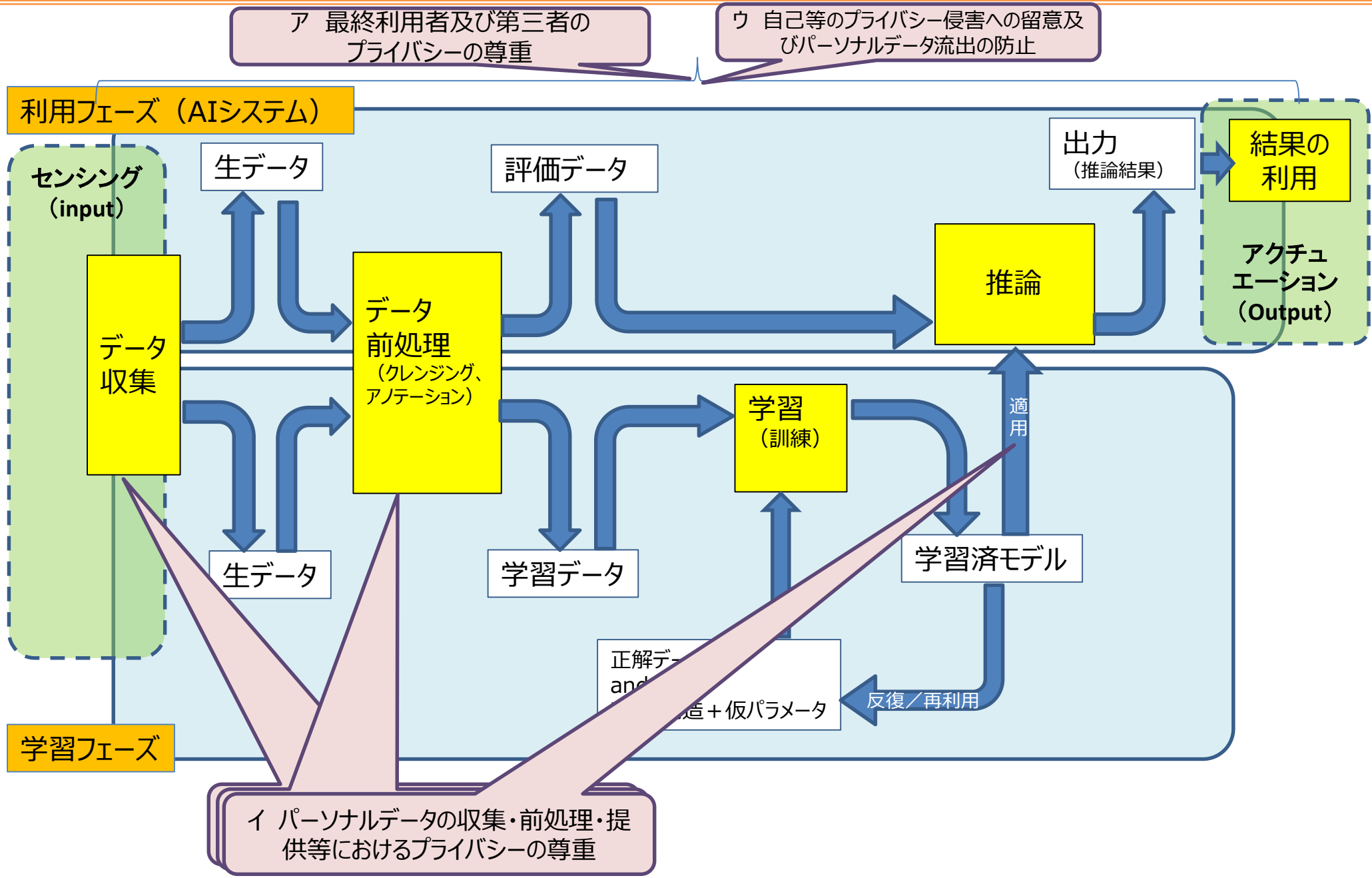
論点

ア）　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）



⑥ーア）　最終利用者及び第三者のプライバシーの尊重

AI SP、ビジネス利用者等	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] • 他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 • 他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
消費者的利用者	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、AIサービスプロバイダ及びビジネス利用者等から提供された情報に留意することが期待される。

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

AISP、ビジネス利用者等	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]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、第三者のプライバシーを尊重する。 また、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。 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必要である。
消費者的利用者	消費者的利用者は、利用するAI等の学習に用いるデータを自ら収集する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

AISP、ビジネス利用者等	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、同意がないデータはシステム上提供できないこととするなど適切な措置を講ずる。
消費者的利用者	消費者的利用者は、ペットロボットなど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期待される。

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。

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 SP、ビジネス利用者等	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。その際には、人間とAIの異質性[1]を前提とするとともに、人間の活動を支援するものであるとの認識が重要となる。 1)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、人間をAI同様に扱うべきではない（すなわち、人間の尊厳が必要）と考えることが可能となる。
消費者的利用者	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期待される。その際には、人間とAIの異質性[1]を前提とするとともに、人間の活動を支援するものであるとの認識が重要となる。 1)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、人間をAI同様に扱うべきではない（すなわち、人間の尊厳が必要）と考えることが可能となる。

⑦ーイ） AIによる意思決定・感情の操作等への留意

AISP、ビジネス利用者等	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必要がある。 [意思決定・感情操作に対する対策例] • 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 • 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 • 利用者の自覚を促すこと（注意喚起） 1)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。その例として、ナッジ（合理的選択のための支援）が考えられるが、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」を踏まえ、開発者からの情報をもとに行うことが期待される。
消費者的利用者	消費者的利用者は、AIサービスプロバイダ及びビジネス利用者からの情報等[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自覚することが期待される。 2)例えば、教育現場等において、上記リスクがあることの情報を取得するなど

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

AISP、ビジネス利用者等	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及）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な配慮が求められる。 また、上記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。
消費者的利用者	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、当該機能及びその周辺技術に関する開発者、AIサービスプロバイダ及びビジネス利用者からの情報を基に、自らの自律性が侵害されないよう留意して利用することが求められる。

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 SP、ビジネス利用者等	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]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 • 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 • 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 • 匿名データによるプロファイリング結果に（匿名でない）特定個人のデータを突合することにより、匿名データから個人が特定されてしまうこと • プロファイリングの結果、特定の個人又は集団に対する差別を助長するなど人の権利・利益を損なう取り扱いがなされること 1)GDPRにおいては、同22条において、自動処理に基づき重要な決定を下されない権利が保障されている。
消費者的利用者	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、確認することが期待される。

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が不当に差別されないよう配慮する。

（注）「公平性」には複数の定義・基準があることに留意する必要がある。

論点

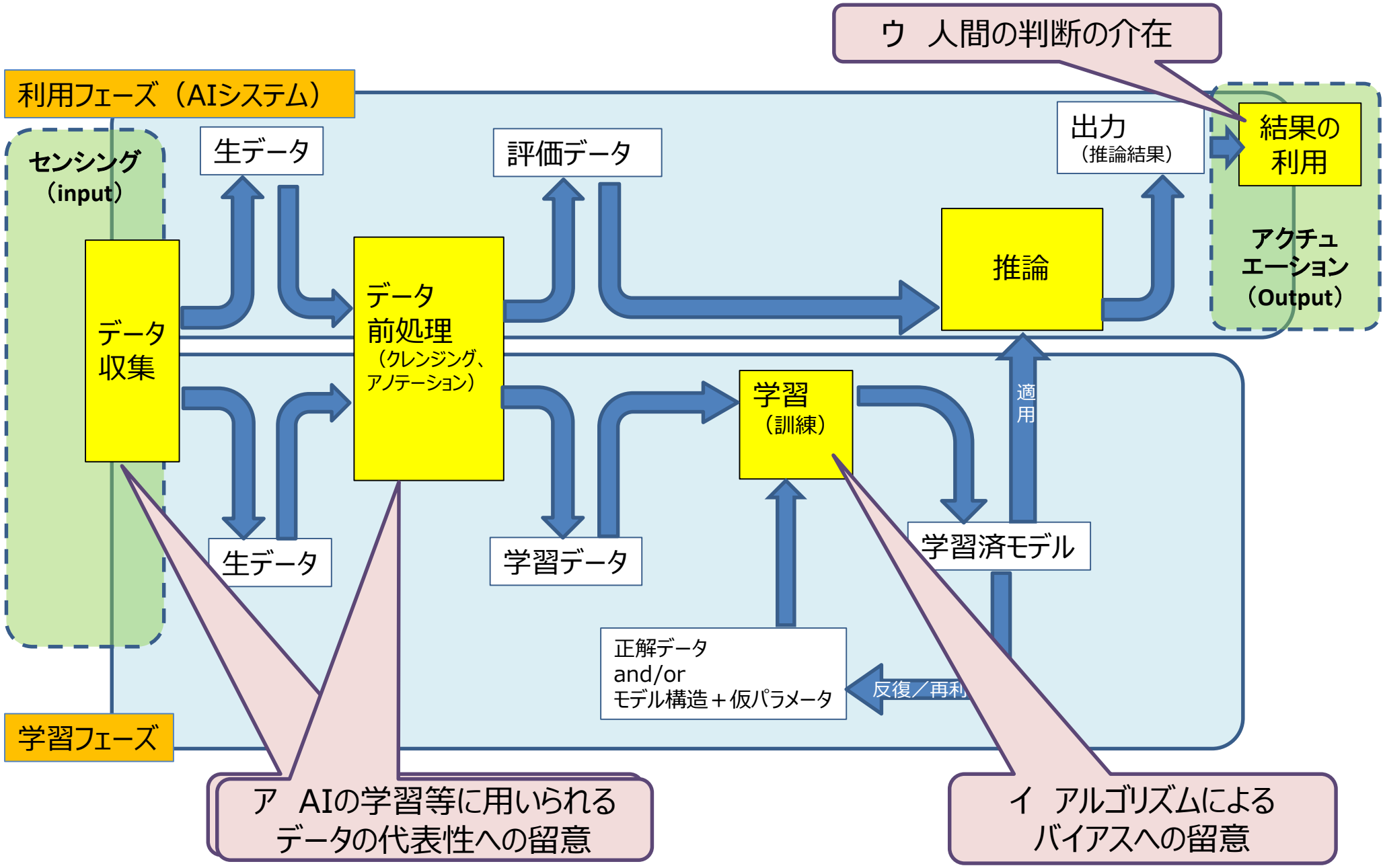
ア）　AIの学習等に用いられるデータの代表性への留意

イ）　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）論点イに公平性基準の例が記載されている。

（機械学習を中心とした）学習と利用の流れ（簡易版）



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 SP、ビジネス利用者等	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、利用時の社会的文脈及び用途に照らして、例えば以下のような観点で、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に留意することが期待される。 [データの代表性やデータに内在する社会的なバイアスに留意する際に考慮する点（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、不当な差別は起こりうる点に留意する必要がある。 • 差別に直結する属性（センシティブ属性）を持つデータを直接使用しなくても、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差別が起こりうる点[2]に注意が必要である。 • データの前処理においては、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に留意する必要がある[3]。 • データの代表性を満足するために大量のデータを集めようとする場合、特に個人データを含む場合には、プライバシーを尊重する必要がある。 1) データの「代表性」とは、サンプルされて利活用に供されているデータが母集団の性質を歪めないという性質のことをいう。 2) 例えば、性別に依存しない与信審査を行おうとした場合に、仮に収入の高い人の割合が男女間で数倍異なるとすると、収入の高低を一属性に加えて与信審査を行うアルゴリズムからは、結果として性別による差別が生ずることになる。 3) 対策として、対象や用途に合わせたラベリングの統一基準を作ることも考えられる。
消費者的利用者	N/A

⑧ーイ）　アルゴリズムによるバイアスへの留意

AISP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、AIに用いられるアルゴリズムにより、それによる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る（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り不当な差別を生み出さないための対策（例）] • （統計的）機械学習のアルゴリズムに、「公平性基準(definition of fairness)」を組み込む。具体的にはセンシティブ属性（前項参照）を定義し、それをアルゴリズムに組み込むことにより中立性を確保（バイアスを除去）する。 ➢ センシティブ属性は、用途に基づく社会の要請[1]に基づき決定される。 ➢ 公平性基準は、センシティブ属性に係わる特定方向のバイアスを除去するものであり、その方法として単純にセンシティブ属性を利用しない方法（unawareness）に加え、特定のグループ間の公平性を担保するための人口学的等価性（demographic parity）[2]を確保する等の方法がある。 ➢ 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ことに留意すべきである。 なお、公平性基準を定義し、それをアルゴリズムに組み込むことにより、公平性を満足する判断は可能となるため、公平性基準を選択・決定するための意思決定に関する制度設計（メカニズムデザイン）が重要となることに留意が必要である。メカニズムデザインにあたっては、どのようなメカニズムとすべきか（特定の者が不利益を被らないルールなど）、どのようにメカニズムを確保するのか（契約、不文律、世論など）等について検討を行う。 1)例えば、入社試験で性別が問題視されている場合、性別をセンシティブ属性とし、性別に対し中立に（独立に）学習を行うなどが考えられる。 2)上記1の例で、男女の平均評価点を同一にする方法のこと（アファーマティブ・アクションの考え方）。
消費者的利用者	N/A

⑧ーウ）　人間の判断の介在（公平性の確保）

AISP、ビジネス利用者等	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ーイ] に掲げる内容を基に、利用する技術の特性及び用途に照らして検討することが期待される。特に、公平性の観点からは以下の基準を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。[1] • 意思決定（判断）に対し納得ある理由を必要とする場合。[2] • マイノリティなどに対する人種・信条・性別に基づく差別が想定される場合。 1) 例えば、人事評価・採用に当たっては、能力や生産性といった変数が重視されるが、これらは時間と共に変わるため、将来予測が難しい。 2) 例えば、人事評価に当たっては、社員に対し判断の理由を説明できることが期待される。
消費者的利用者	N/A

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。

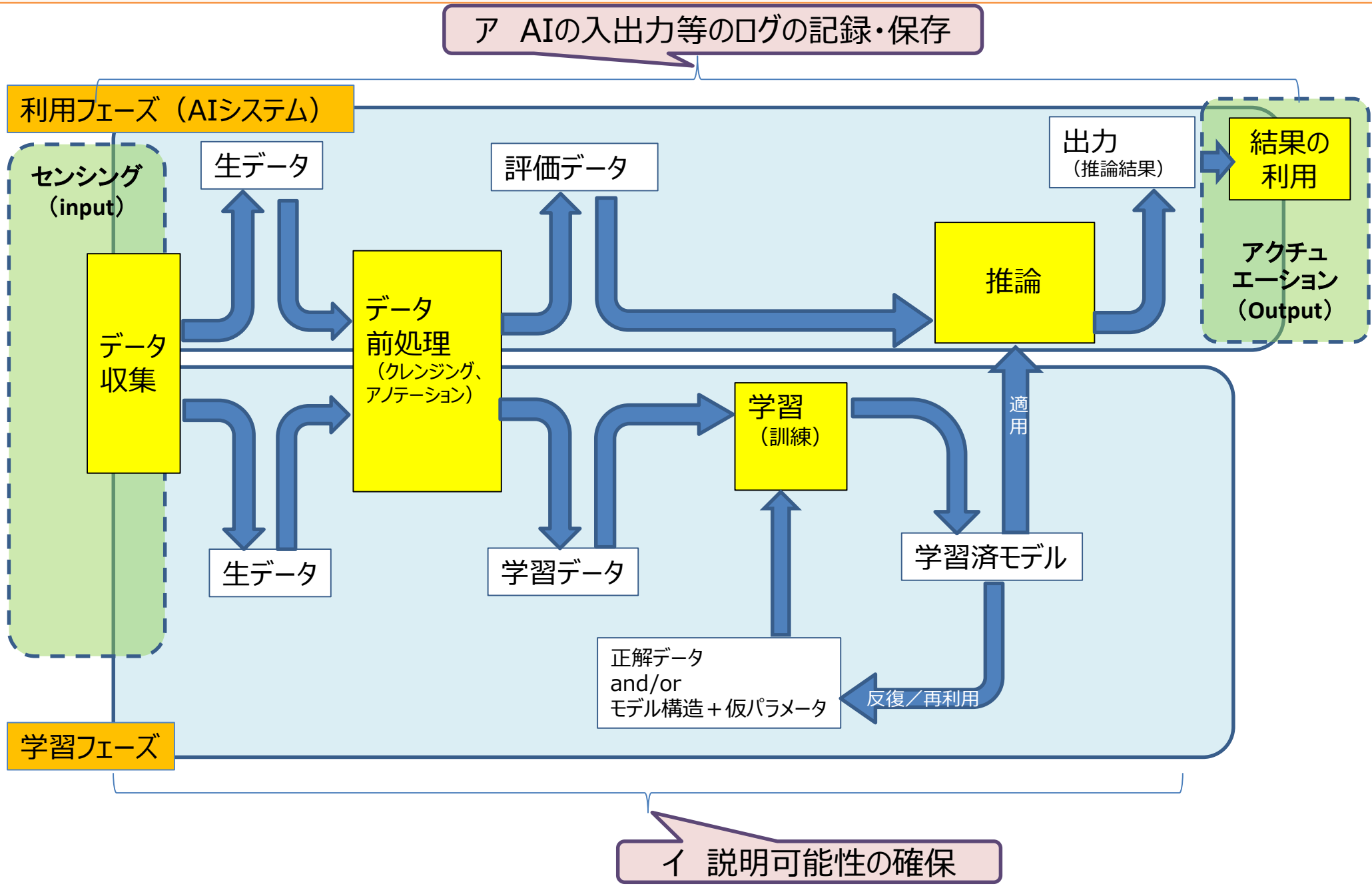
論点

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）



⑨ーア） AIの入出力等のログの記録・保存

AISP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、AIの提供にあたり、問題が生じた場合の原因究明（トレース）等を目的として、入出力等のログを記録・保存することが期待される。 ただし、ログの記録・保存に当たっては、利用するAIのアルゴリズムや用途に基づき、以下について考慮することが望ましい。 [ログの記録・保存に当たり、考慮すべき事項][1] [2] • 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） • ロク保存場所の容量 • ログの時刻の記録 1）掲げられた考慮すべき事項のいくつかは、それぞれが相互にトレードオフの関係にあるため、社会的文脈及び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、完全性を高めることとトレードオフの関係にある。 2）人の生命・身体・財産に危害を及ぼし得る分野では、原因究明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求められることが想定される。
消費者的利用者	N/A

⑨ーイ） 説明可能性の確保

AI SP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、利用者の納得感や安心感の獲得、及びそのための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（特に、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提供・利用する場合）。 ただし、説明可能性の確保に当たっては、その目的に鑑み、どのような説明が必要かを分析・把握し、それに対し、以下の手段を総合的に用いて対処することが期待される。 [説明可能性確保のための手段（例）] （データに関する対策） • 学習された学習済モデルが具備されたAIソフトならびにAIシステムを提供するに当たり、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。 （学習モデルに関する対策） • AIに対する複数の入力と出力の組合せをもとに、その傾向を分析する（例えば、入力パターンを少しずつ変化させたときに出力がどうなるかを観測するなど）。 • 下記に示す技術的な手法を用いる（ただし、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、一般的にコストとトレードオフの関係にあることにも十分留意する必要がある。） ➢ 予め可読性の高い解釈可能なモデルを作るための手法[1] ➢ ブラックボックスモデルに対する説明を可能とする手法 ✓「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 （総合的な対策） • 説明が不足している部分について、どのような説明が必要なのかを明確にし、開発者とも連携して解決策を模索する。 1）一般的に、学習モデルを解釈可能な形とすることは、（学習）精度の確保とトレードオフの関係にあることに注意が必要である。
消費者的利用者	N/A

⑨ーウ）　行政機関が利用する際の透明性の確保

AISP、ビジネス利用者等	行政機関がAIを利用する場合には、「法の支配」の原理に基づく行政の透明性確保の要請と、行政手続法に基づく適正手続の要請を踏まえ、AIの判断結果の説明可能性を確保するため、必要に応じ、以下の措置を講じることを検討すべきである。 [説明可能性を確保するための措置の例] ・ 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）、アルゴリズムの調整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 ・ 開発者やAIサービスプロバイダと契約を結ぶ際に、AIの判断を説明する諸要素について不開示の範囲を限定すること
消費者的利用者	N/A

⑩アカウンタビリティ※の原則

AIサービスプロバイダ及びビジネス利用者は、消費者的利用者を含む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。

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を納得させるため、（判断に関する）正当な意味・理由を説明したり、（必要に応じて）賠償・補償したりする等の措置をとること。

⑩ーア）　アカウンタビリティを果たす努力

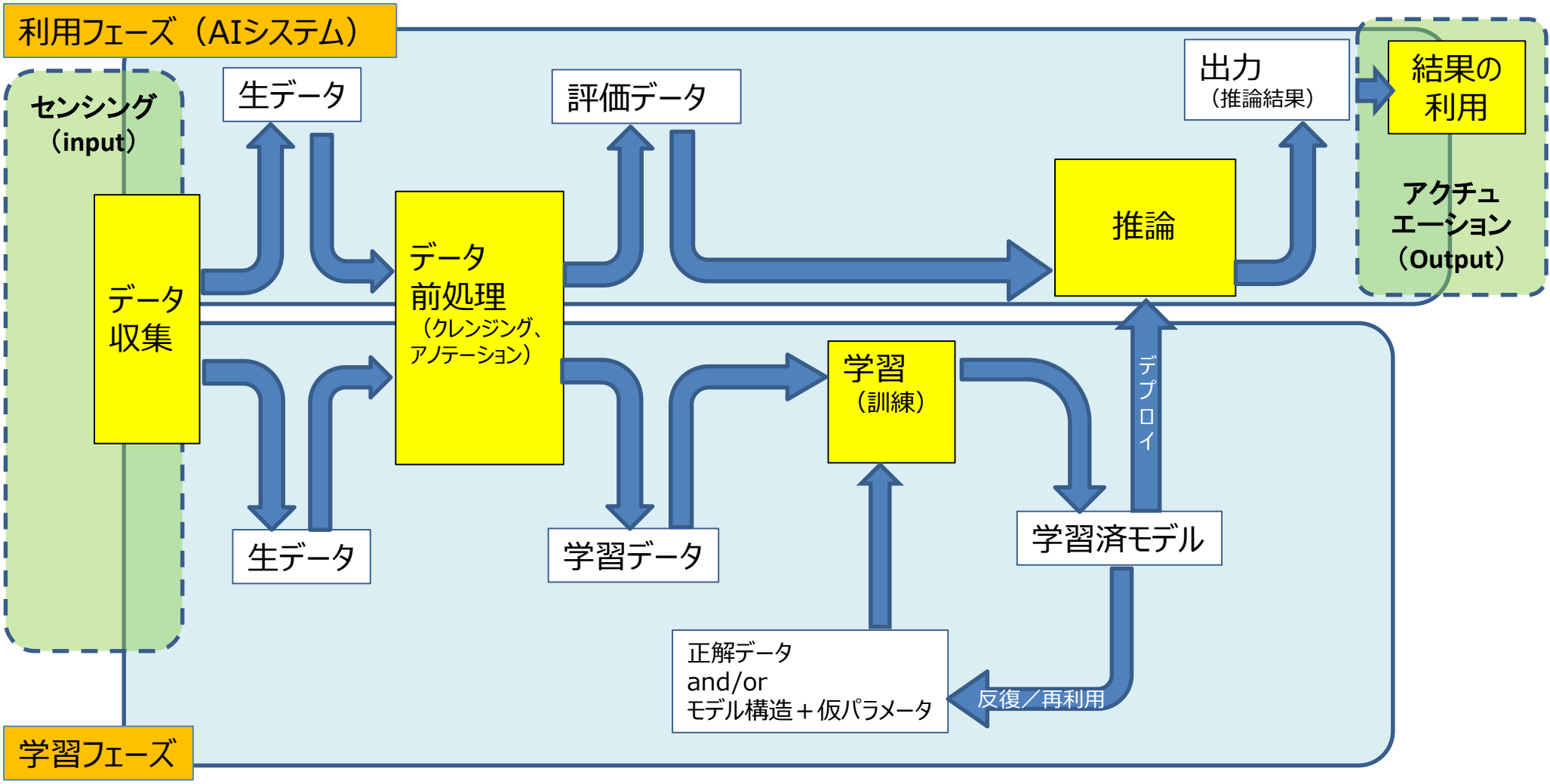
AI SP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、人々と社会からAIへの信頼を獲得することができるよう、消費者的利用者等AIの利活用により影響を受ける第三者等に対し、利用するAIの性質及び目的等に照らして、それぞれが有する知識や能力の多寡に応じ、AIシステムの特性について情報提供と説明を行う他、多様なステークホルダとの対話を通じて様々な意見を聴取する等、相応のアカウンタビリティを果たすよう努めることが期待される。
消費者的利用者	消費者的利用者は、AIの判断結果について疑義を感じた場合には、必要に応じて、同サービスを提供した開発者、AIサービスプロバイダ、及びビジネス利用者に問い合わせを行うことが期待される。

⑩ーイ） AIに関する利用方針の通知・公表

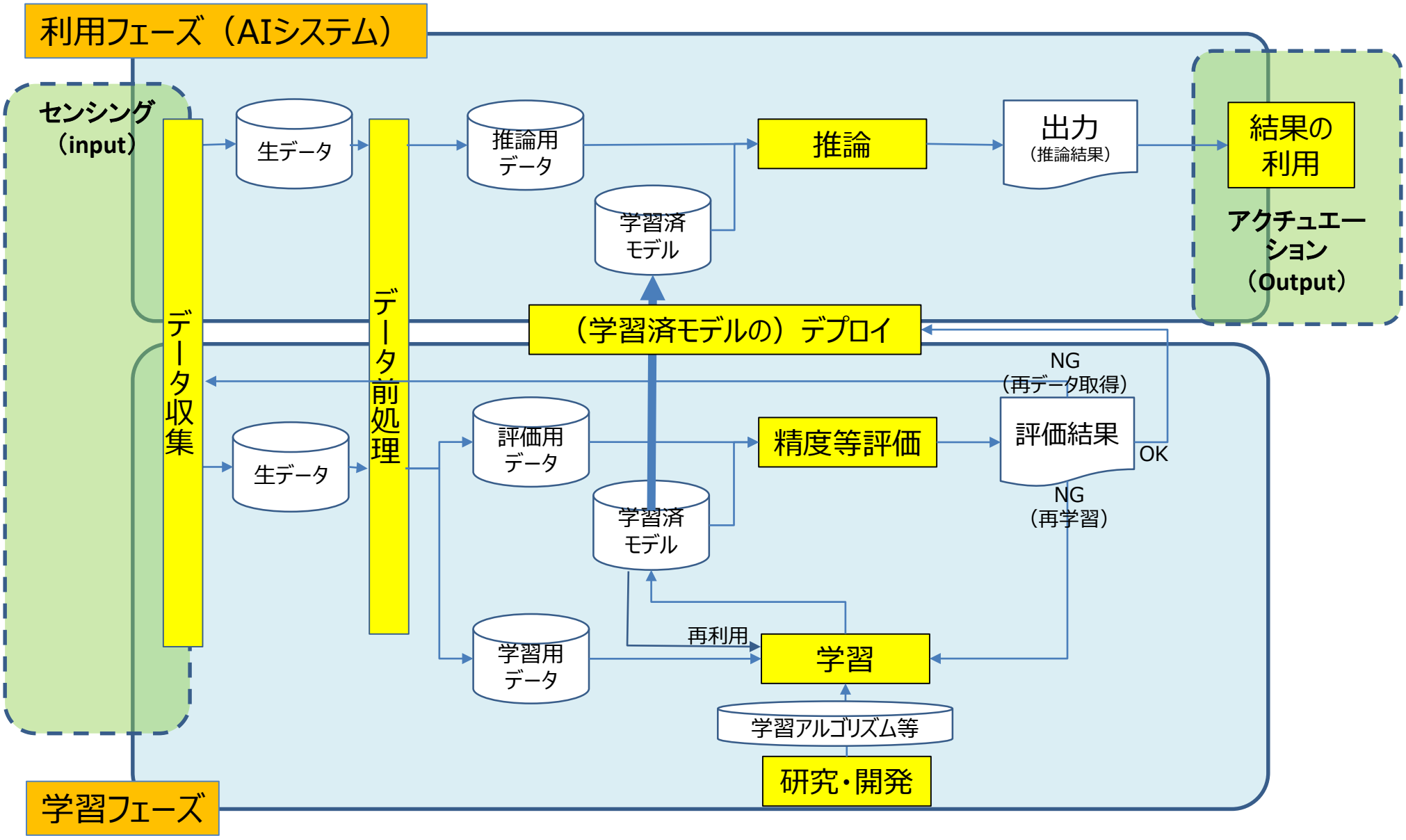
AI SP、ビジネス利用者等	AIサービスプロバイダ及びビジネス利用者は、AIの判断が直接に消費者的利用者や第三者に対して影響を及ぼす態様によりAIを利活用する場合には、消費者的利用者や第三者がAIの利活用について適切に認識することができるよう、以下の事項を含むAIに関する利用方針を作成・公表し、問い合わせがあった場合には通知を行うこと、加えて、個人の権利・利益に重大な影響を及ぼす可能性のある場合には積極的に通知することが期待される[1]。 [AIに関する利用方針に含まれる事項（例）] • AIを利用している旨（具体的な機能・技術を特定できるのであれば、その名称と内容等[2]） • 利活用の範囲及び方法 • 利活用に伴うリスク • 相談窓口 • 個人情報を取得する場合の目的など また、通知または公表は、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実施すること（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）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。
消費者的利用者	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、同サービスを提供した開発者、AIサービスプロバイダ、及びビジネス利用者に問い合わせを行うことが期待される。

（参考）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）



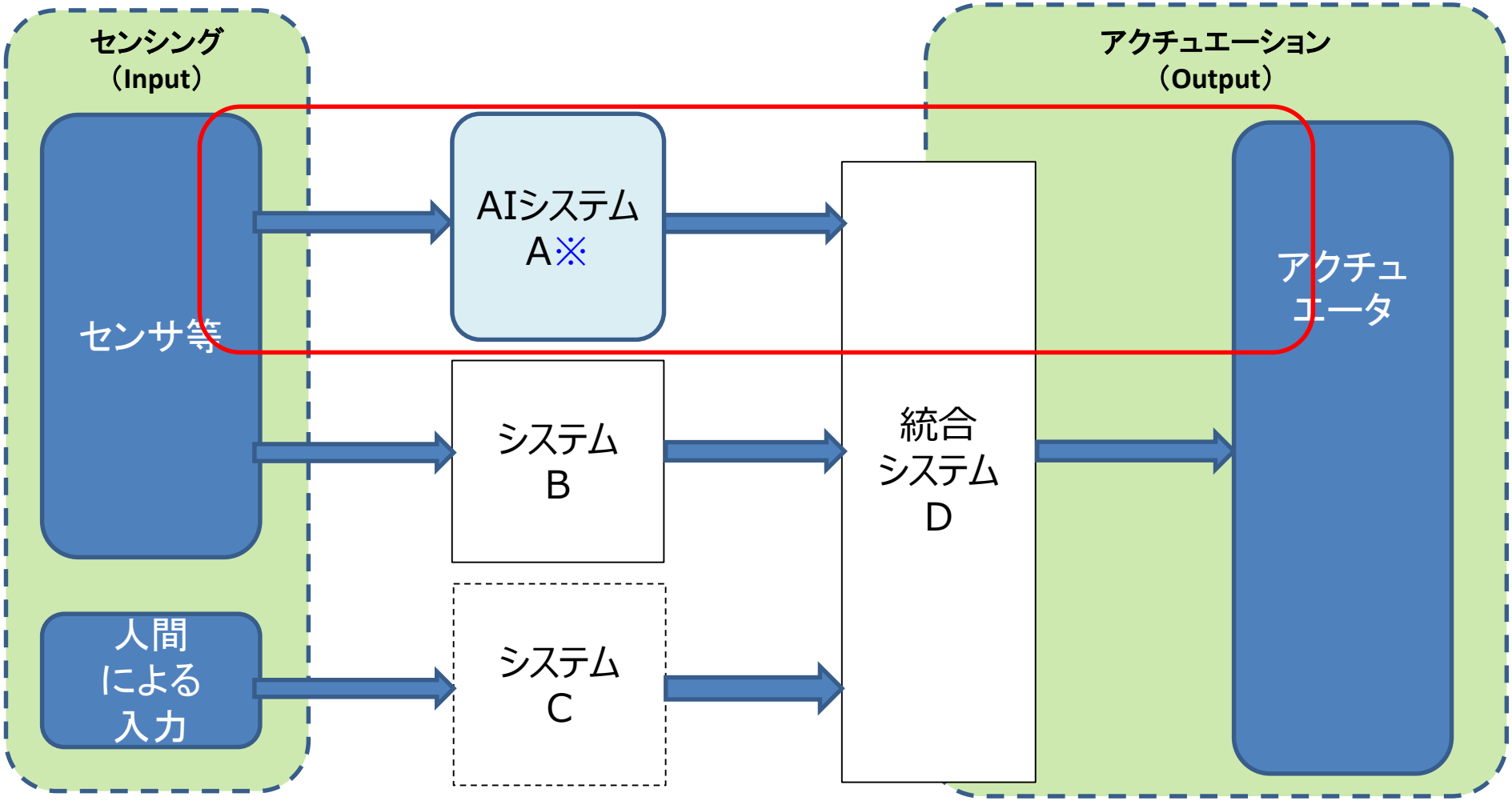
（機械学習を中心とした）学習と利用の流れ



（他システムと統合された）AIシステムの例

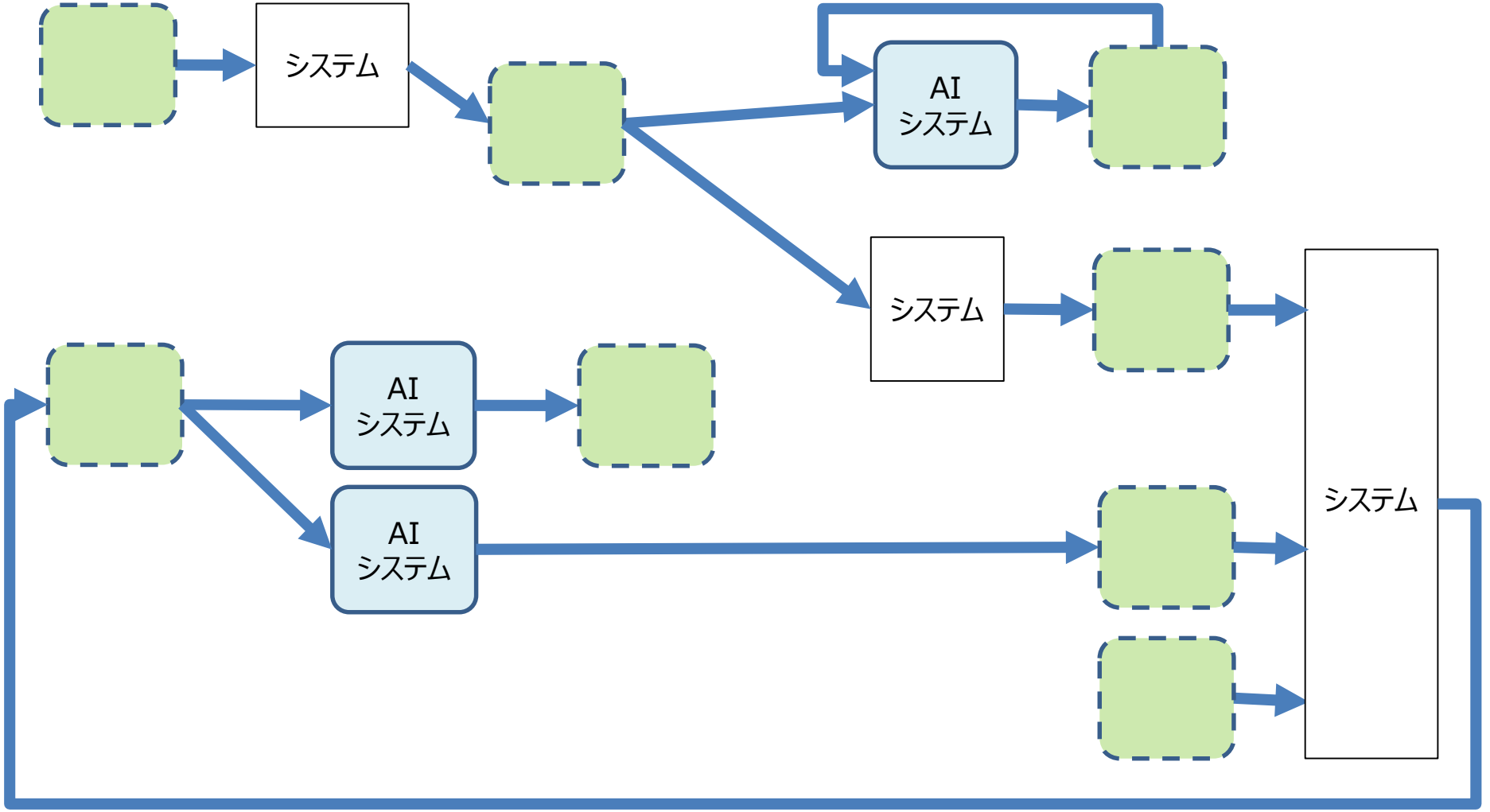
センサ等からの入力を受けAIシステムA、及びシステムBにより処理された結果、及び人間による入力を受けシステムCにより処理された結果がそれぞれシステムDで統合され、アクチュエータの制御に反映される。

（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基に走行、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を踏まえキルスイッチ（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

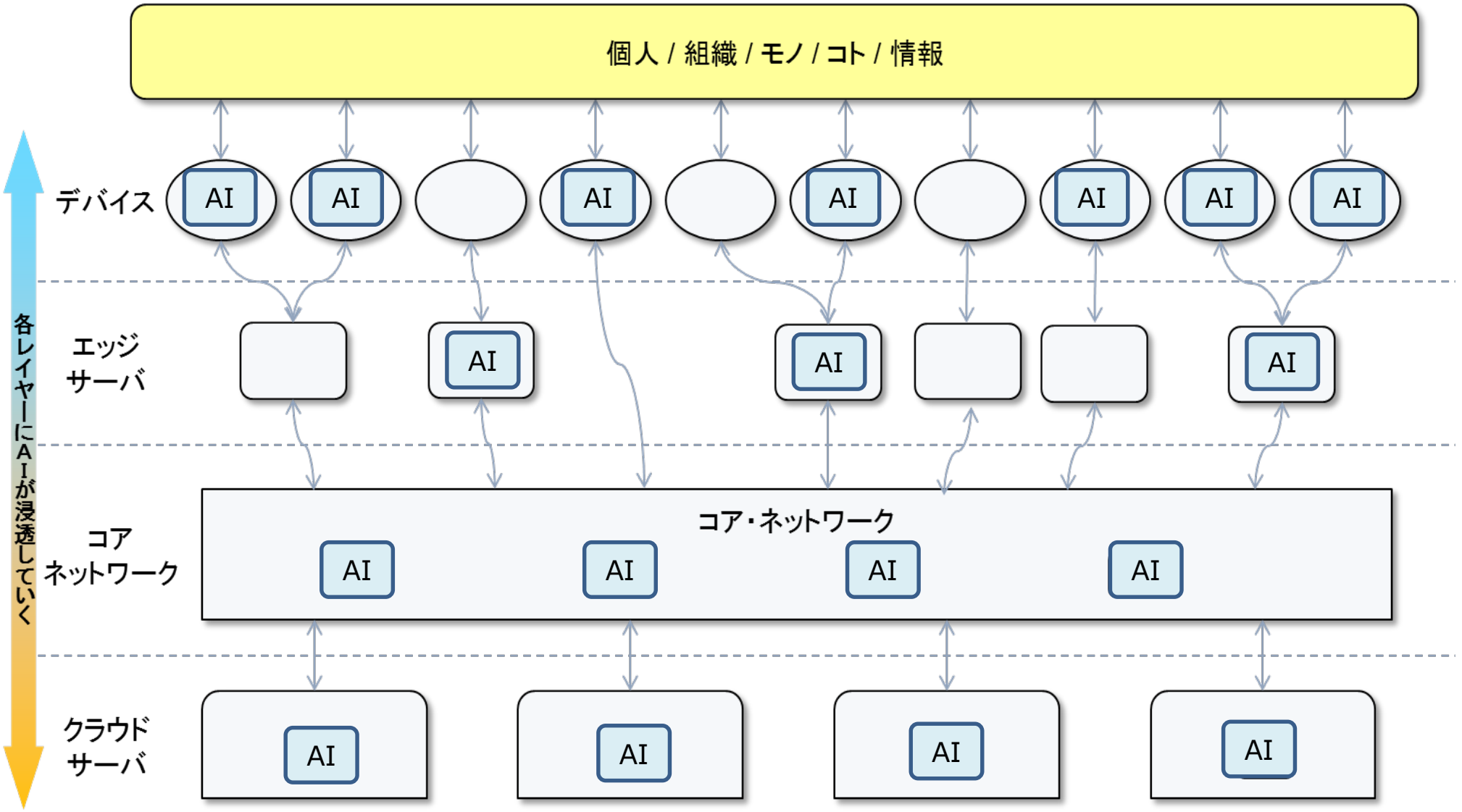


システム　一般のシステム

AI
システム　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―」