安心・安全で信頼性のあるＡＩの社会実装に向けて
～ＡＩネットワーク社会推進会議での議論及び関連動向より～

２０２０年８月２７日

高木　幸一

自己紹介

（株）KDDI総合研究所　フューチャーデザイン２部門　シニアアナリスト

2018年４月より総務省　情報通信政策研究所調査研究部　主任研究官
総務省AIネットワーク社会推進会議（有識者会議）事務局を担当

2020年７月より現職

Contents

- AIネットワーク社会推進会議とは
- AIに関する日本政府全体の取組と総務省の取組
- AI利活用ガイドラインとは
- 国際的な議論の動向
- 原則は具体化へ
- AI×セキュリティについての議論

Contents

AIに対し・・・

○ AIの研究開発が進み、様々な分野においてAIの利活用が進展することが想定される。

○ その際、AIネットワーク化※の進展が想定される。

※ 「AIネットワーク化」とは、AIシステムがインターネットその他の情報通信ネットワークと接続され、AIシステム相互間又はAIシステムと他の種類のシステムとの間のネットワークが形成されるようになること。

多大な便益への期待 ＋ リスクへの懸念

便益の例	リスクの例
高齢者や過疎地などで公共交通機関の利用が困難な者が、自動運転車を使うと、病院への通院や買い物などに出掛け易くなる。	自動運転車がユーザの意図に反し特定の場所を通過するようルートを設定するなどのおそれがある。
過去の症例を参考に、AIが患者の病名を推定するとともに、適切な処置方法を提示する。	AIが不適切な推定を行ったり、不適切な処置方法を提示したりすることより、医療過誤が生ずるおそれがある。
道路や橋などに設置されたセンサーや衛星写真から得られる情報等から、AIが異常検知や故障予測を行う。異常を検知した場合には、ロボットが自動的に点検・修理を行う。	AIが異常を見逃すこと等により、道路の陥没や橋の崩落など事故が発生するおそれがある。

○ AIは様々な分野で利活用され、また、そのサービスはネットワークを通じて（場合によっては国境を越えて）提供されることが想定される。

- AIは人間・社会に多大な便益を広範にもたらすことが期待される一方、リスクの抑制も今から考えておくことが必要ではないか？
- イノベーション促進を図りつつ、AIを安全・安心に社会実装していくためにはどのような手法が有効か？

総務省 「AIネットワーク社会推進会議」概要

背景

○ AIの研究開発・利活用の進展、AIの相互連携・ネットワークの形成（AIネットワーク化）

○ 様々な分野におけるAI利活用、ネットワークを通じた（国境を越えた）サービス提供

○ 多大な便益を広範にもたらすことが期待されるとともに、リスクの抑制も図ることが重要

⇒
- AIの便益の増進、リスクの抑制のための取組について**中長期的な視点**で検討が必要
- **産学民官の幅広い関係者の参画を得て**、国際的にも議論することが重要

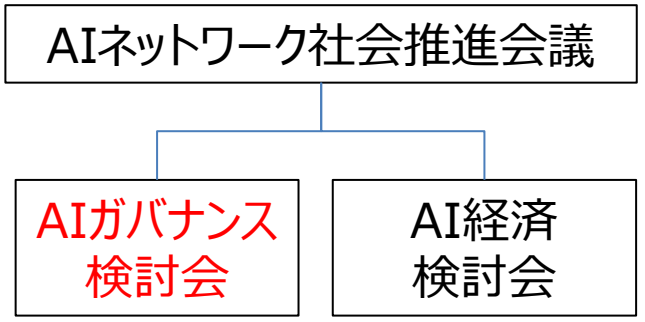
AIネットワーク社会推進会議

目的・検討事項

AIネットワーク化に関して、社会的・経済的・倫理的・法的課題に関する事項を検討。具体的には、以下について検討。

○ AI開発ガイドライン・AI利活用ガイドライン

○ AIに関する経済政策　　等



検討体制

【議　長】　須藤修（中央大学国際情報学部教授・東京大学大学院情報学環特任教授）

【構成員】　産学民の有識者（関係学会の会長経験者、関係企業の会長又は社長等）

【オブザーバ】 関係行政機関、関係国立研究開発法人 等

<引用> https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/index.html

「AIネットワーク社会推進会議」（親会）構成員

（2020年6月５日現在）

議　長　須藤 修　（中央大学国際情報学部教授、東京大学大学院情報学環特任教授）

副議長　三友 仁志　（早稲田大学国際学術院大学院アジア太平洋研究科教授）

構成員

【研究者（社会・人文系）】

大橋 弘　（東京大学大学院公共政策大学院・経済学研究科教授）
大屋 雄裕　（慶應義塾大学法学部教授）
小塚 荘一郎　（学習院大学法学部法学科教授）
宍戸 常寿　（東京大学大学院法学政治学研究科教授）
実積 寿也　（中央大学総合政策学部教授）
城山 英明　（東京大学大学院法学政治学研究科教授）
新保 史生　（慶應義塾大学総合政策学部教授）
鈴木 晶子　（京都大学大学院教育学研究科教授）
橋元 良明　（東京女子大学現代教養学部心理・ｺﾐｭﾆｹｰｼｮﾝ学科ｺﾐｭﾆｹｰｼｮﾝ専攻教授）
林 秀弥　（名古屋大学大学院法学研究科教授）
平野 晋　（中央大学国際情報学部教授・学部長）
福田 雅樹　（大阪大学社会技術共創研究センター教授、(兼任)大学院法学研究科教授）
柳川 範之　（東京大学大学院経済学研究科教授）
山本 勲　（慶應義塾大学商学部教授）

【産業界】

岩本 敏男　（株式会社エヌ・ティ・ティ・データ相談役）
遠藤 信博　（日本電気株式会社取締役会長）
金井 良太　（株式会社アラヤ代表取締役ＣＥＯ）
エリー キーナン　（日本アイ・ビー・エム株式会社取締役会長）
谷崎 勝教　（株式会社三井住友銀行取締役専務執行役員グループＣＤＩＯ）

【消費者団体】

木村 たま代　（主婦連合会事務局長）
近藤 則子　（老テク研究会事務局長）

【研究者（技術系）】

大田 佳宏　（東京大学大学院数理科学研究科特任教授、Arithmer代表取締役社長兼CEO）
喜連川 優　（国立情報学研究所所長、東京大学生産技術研究所教授）
杉山 将　（理化学研究所革新知能統合研究センター長、東京大学新領域創成科学研究科教授）
高橋 恒一　（理化学研究所生命機能科学研究センターチームリーダー）
中川 裕志　（理化学研究所革新知能統合研究センターチームリーダー）
中西 崇文　（武蔵野大学データサイエンス学部データサイエンス学科長・准教授）
西田 豊明　（福知山公立大学情報学部学部長・教授）
萩田 紀博　（大阪芸術大学アートサイエンス学科長・教授、株式会社国際電気通信基礎技術研究所萩田紀博特別研究所長）
堀 浩一　（東京大学大学院工学系研究科教授）
松尾 豊　（東京大学大学院工学系研究科教授）
村井 純　（慶應義塾大学教授）
森川 博之　（東京大学大学院工学系研究科教授）
山川 宏　（全脳アーキテクチャ・イニシアティブ代表）

時田 隆仁　（富士通株式会社代表取締役社長）
東原 敏昭　（株式会社日立製作所代表執行役 執行役社長兼ＣＥＯ）
田丸 健三郎　（日本マイクロソフト株式会社業務執行役員ナショナルテクノロジーオフィサー）
藤田 雅博　（ソニー株式会社VP、シニア・チーフ・リサーチャー、AIコラボレーションオフィス）
マシューズ 真里　（グーグル合同会社執行役員 公共政策担当）
村上 憲郎　（株式会社村上憲郎事務所代表取締役）

長田 三紀　（情報通信消費者ネットワーク）

オブザーバー　内閣府、内閣官房情報通信技術（ＩＴ）総合戦略室、個人情報保護委員会事務局、消費者庁、文部科学省、経済産業省、情報通信研究機構、科学技術振興機構、理化学研究所、産業技術総合研究所

顧　問　安西 祐一郎　（慶應義塾大学名誉教授）
長尾 真　（京都大学名誉教授）
西尾 章治郎　（大阪大学総長）
濱田 純一　（東京大学名誉教授）　（敬称略。五十音順）

AIネットワーク社会推進会議「AIガバナンス検討会」構成員

（2020年6月５日現在）

座長	平野 晋	中央大学国際政策学部教授・学部長
構成員	江間 有沙	東京大学未来ビジョン研究センター特任講師
	江村 克己	日本電気株式会社NECフェロー
	大屋 雄裕	慶應義塾大学法学部教授
	金井 良太	株式会社アラヤ代表取締役CEO
	河島 茂生	青山学院女子短期大学現代教養学科准教授／ 理化学研究所革新知能統合研究センター客員研究員
	木谷 強	株式会社エヌ・ティ・ティ・データ取締役常務執行役員
	木村 たま代	主婦連合会消費者相談室長
	久世 和資	日本アイ・ビー・エム株式会社執行役員 最高技術責任者
	小塚 荘一郎	学習院大学法学部法学科教授
	三部 裕幸	弁護士
	城山 英明	東京大学大学院法学政治学研究科教授
	鈴木 教洋	株式会社日立製作所執行役常務CTO兼研究開発グループ長
	高橋 恒一	理化学研究所生命機能科学研究センターチームリーダー／ 慶應義塾大学大学院政策・メディア研究科特任教授
	武田 英明	国立情報学研究所情報学プリンシプル研究系教授
	田丸　健三郎	日本マイクロソフト株式会社業務執行役員ナショナルテクノロジーオフィサー
	中川 裕志	理化学研究所革新知能統合研究センターグループディレクター
	長田 三紀	情報通信消費者ネットワーク
	原 裕貴	株式会社富士通研究所代表取締役社長
	西田 豊明	福知山公立大学情報学部学部長・教授
	堀 浩一	東京大学大学院工学系研究科教授
	ﾏｼｭｰｽﾞ 真理	グーグル合同会社執行役員 公共政策担当
	山本 龍彦	慶應義塾大学法科大学院教授
	湯淺 墾道	情報セキュリティ大学院大学学長補佐・情報セキュリティ研究科教授

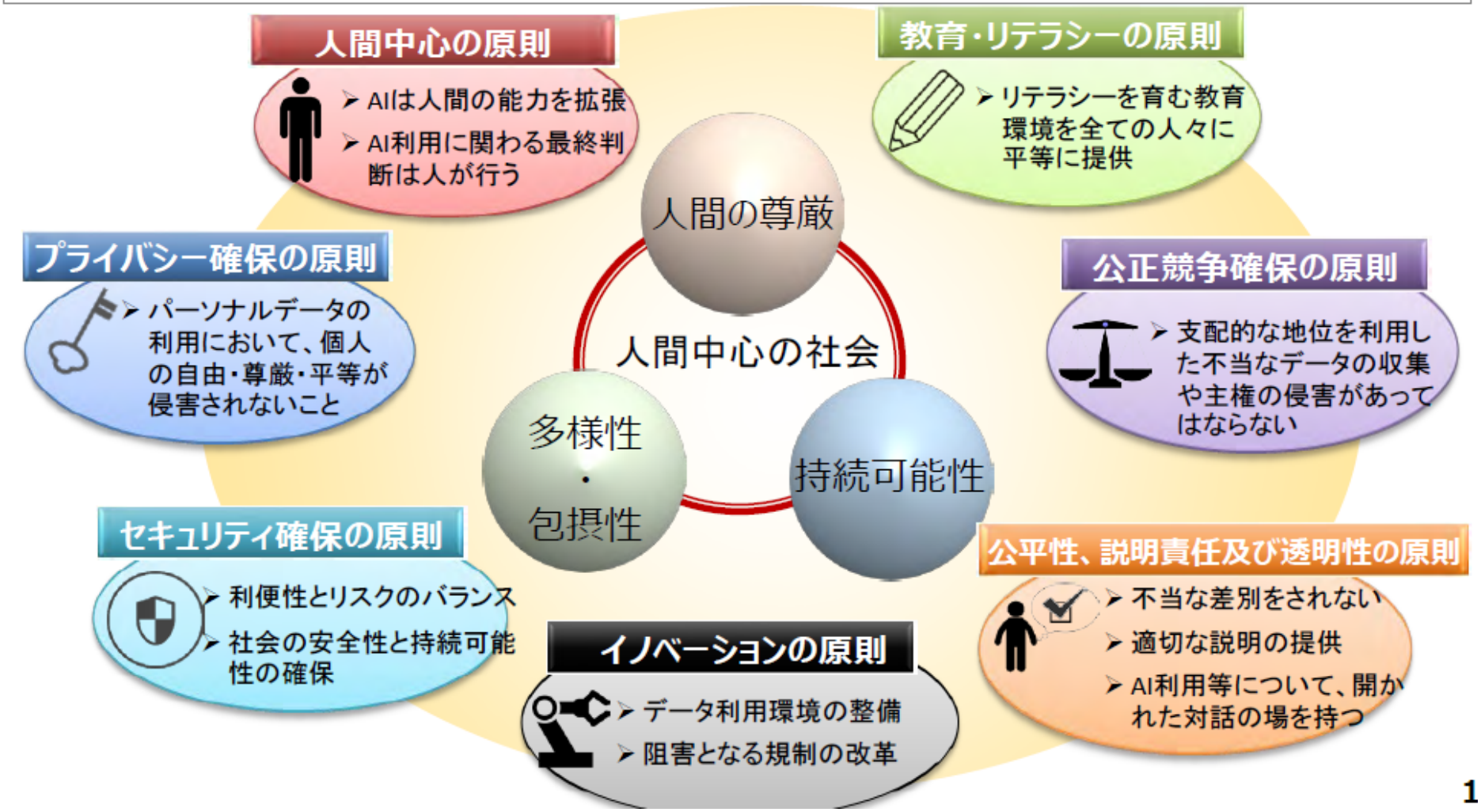
（敬称略、五十音順）

※須藤 修　　AIネットワーク社会推進会議議長
　三友 仁志　　同副議長
　実積 寿也　　同構成員
　がオブザーバとして参加。

Contents

【人間中心のAI社会原則】

- 世界でAIの倫理的側面に関する議論が進展
- AIに関する人々の不安を払拭し、積極的な社会実装を推進するため、我が国としての原則を3月に策定
- 今後、AI社会原則に関する多国間の枠組みを構築



<引用> https://www5.cao.go.jp/keizai-shimon/kaigi/special/reform/wg7/20191101/shiryou1.pdf

ＡＩ戦略 【基本的考え方】

- 「**人間尊重**」、「**多様性**」、「**持続可能**」の３つの理念を掲げ、Society 5.0を実現し、SDGsに貢献
- ３つの理念を実装する、**４つの戦略目標**（人材、産業競争力、技術体系、国際）を設定
- 目標の達成に向けて、「**未来への基盤作り**」、「**産業・社会の基盤作り**」、「**倫理**」に関する取組を特定

戦略目標Ⅰ：**人材**
人口比において最もAI時代に対応した人材を育成・吸引する国となり、持続的に実現する仕組みを構築

戦略目標Ⅱ：**産業競争力**
実世界産業においてAI化を促進し、世界のトップランナーの地位を確保

理念（実現する社会）
- 人間の尊厳の尊重（Dignity）
- 多様な人々が多様な幸せを追求（Diversity & Inclusion）
- 持続可能（Sustainability）

戦略目標Ⅲ：**技術体系**
理念を実現するための一連の技術体系を確立し、運用するための仕組みを実現

戦略目標Ⅳ：**国際**
国際的AI研究・教育・社会基盤ネットワークの構築

具体目標・取組

未来への基盤作り		産業・社会の基盤作り			倫理
教育改革	研究開発	社会実装	データ関連基盤	デジタル・ガバメント中小・新興企業支援	AI社会原則

<引用> https://www5.cao.go.jp/keizai-shimon/kaigi/special/reform/wg7/20191101/shiryou1.pdf

AI戦略2019 フォローアップ

- ✓ **昨年6月に策定した「AI戦略2019」の実施初年度として、各府省庁等が関連する取組を鋭意実施。**
- ✓ **取組の8割強は、計画通りに進捗。**
- ✓ **新型コロナウイルス感染症拡大に直面し、よりデジタル社会の深化が不可欠。**
  →AIの研究開発・社会実装、それらを支える情報通信環境の整備等の強化・充実が必要。

２０１９年度の進捗（進捗のあった主な取組）

- ＧＩＧＡスクール構想の前倒し実施、数理・データサイエンス・ＡＩ教育プログラム認定制度（リテラシーレベル）検討
- 「人工知能研究開発ネットワーク」を設立、3月末時点で104の機関が参画
- 医療画像診断支援やスマート農業、インフラデータプラットフォーム構築
- スマートシティ共通アーキテクチャ構築
- G20 AI原則や各省庁のAIガイドラインの策定

（参考）2019年度内を期限とした取組の進捗状況

	取組数	計画通り	未了/一部未了	進捗率
教育改革	31	27	4	87%
研究開発	16	11	5	69%
社会実装	26	24	2	92%
データ関連基盤	9	8	1	89%
デジガバ・中小	3	3	0	100%
倫理・その他	4	4	0	100%
Total	89	77	12	87%

進捗遅れの挽回、新たな課題や新型コロナウイルス感染拡大への対応、等

２０２０年度に実施する主な取組

- ＧＩＧＡスクール構想の加速、認定制度（応用基礎）の検討、社会人リカレント教育の拡充
- AIの信頼性確保や、人文・社会科学と数理・情報科学との融合に関する研究開発
- ものづくり現場の暗黙知（経験や勘）の伝承・効率的活用を支え、生産性を向上させるAI技術の開発
- ５Gや光ファイバ等のAI利活用に向けたネットワーク基盤の高度化、計算資源の増強
- 自治体でのAIサービスの標準化、自治体行政へのAI・RPA※1の実装
- 責任あるAIやイノベーション等の推進に向け、GPAI※2等における国際連携の強化

※1 RPA：Robotic Process Automation、 ※2 GPAI：Global Partnership on AI

<引用> https://www.kantei.go.jp/jp/singi/tougou-innovation/dai7/siryo3-1.pdf

総務省における主な取組と統合イノベーション戦略推進会議における位置づけ

2016

2017

2018

2019

G7情報通信大臣会合（高松、2016.4）
AIの開発に着目したルール作りに向け、国際的な議論を進めるよう提案（我が国から原則のたたき台を紹介）

「AI開発ガイドライン」策定
（2017年7月公表）

OECD、G7等における検討の場に日本からも有識者を派遣
（例：OECD/総務省共催AIカンファレンス（パリ、2017.10）、
OECD理事会勧告に向けたAI専門家会合（2018.9～2019.2）
G7マルチステークホルダ会合（モントリオール、2018.12） 等）

統合イノベーション戦略推進会議
＜議長：官房長官＞

AI技術：強化すべき主要分野

人間中心のAI社会原則（2019.3）

AI戦略 2019 ～人・産業・地域・政府全てにAI～（2019.6）

OECD閣僚理事会（2019.5）
AIに関する理事会勧告を採択

G20貿易・デジタル経済大臣会合
（つくば、2019.6）
「G20 AI原則」を採択

「AI利活用ガイドライン」策定
（2019年8月公表）

AI中核センター改革・
次世代AI基盤技術推進 等

国際的な議論への貢献（OECD等）

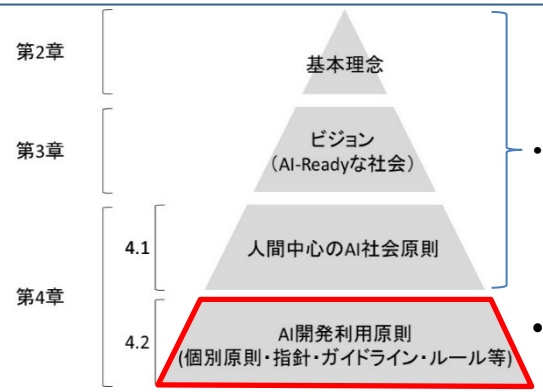
日本型モデル構築・創発研究の推進

今後は、目指すべき社会モデルを検討し、社会実装を加速

「AI開発ガイドライン」・「AI利活用ガイドライン」の位置づけ

「人間中心のAI社会原則」（2019年３月統合イノベーション戦略推進会議決定）より抜粋

人間中心のAI社会原則会議
（2018年５月～）



・・・「社会（特に、国などの立法・行政機関）が留意すべき『AI社会原則』」

・・・「開発者及び事業者において、基本理念及びAI社会原則を踏まえた
AI開発利用原則を定め、遵守するべき」

開発者・事業者それぞれにおいて、AI開発利用原則を策定することを期待

そのための参考となるガイドラインが必要

AIネットワーク社会推進会議
（2016年２月～）

総務省の取組

AI開発ガイドライン
開発者が留意すべき事項と解説

2017年７月とりまとめ

AI利活用ガイドライン
事業者が留意すべき事項と解説

2019年８月とりまとめ

関係省庁に共有の上、開発者・事業者に提供。自主的対応を支援。

Contents

AI利活用ガイドラインとは

Q	どのようなものか？
A	AIの利用者（AIを利用してサービスを提供する者を含む）が利活用段階において留意することが期待される事項を「原則」（全10原則）という形式でまとめ、その解説を記載したもの。
Q	なぜ作ったのか？
A	AIによる便益の増進とリスクの抑制を図り、AIに対する信頼を醸成することにより、AIの利活用や社会実装を促進するため。 政府全体で検討・決定した「人間中心のAI社会原則」において、「開発者・事業者それぞれにおいて、AI開発利用原則を策定することを期待」とある中で、事業者向けの解説書として。 我が国の考えを世界に共有するため（国際的な議論に資するため）。
Q	誰に対して作ったのか？
A	（事業者を中心に）AIを利用して事業を行う者。 消費者的利用者等の最終利用者に対しては「参考」として。
Q	どのような体制で作ったのか？
A	産業界、学術界、市民団体等によるマルチステークホルダによる会議で議論 （関係省庁もオブザーバ）。

関係する主体の整理

開発者

AIシステムの研究開発を行う者

利用者

AIシステム、AIサービス又はAI付随サービスを利用する者

データ提供者

他者が利用するAIシステムの学習等のためにデータを提供する者

第三者

他者の利用するAIにより自らの権利・利益に影響を受ける者

AIサービスプロバイダ

利用者のうち業としてAIサービス又はAI付随サービスを他者に提供する者

最終利用者

利用者のうち業としてAIサービス又はAI付随サービスを他者に提供することなくAIシステム又はAIサービスを利用する者

ビジネス利用者（非営利の専門職・行政機関を含む）

最終利用者のうち業としてAIシステム又はAIサービスを利用する者

（注）ビジネス利用者であっても、AIシステム又はAIサービスについて自ら運用等を行うことなく利用するのみの者については、他のビジネス利用者と同等の留意を期待することが困難であることも想定されるが、その場合でも、開発者やAIサービスプロバイダに対し、適切な措置を依頼する等の対応が期待される。

消費者的利用者

最終利用者のうちAIシステム又はAIサービスを利用する者（ビジネス利用者を除く）

（注）消費者的利用者であっても、AIシステム又はAIサービスについて、自ら運用等を行う場合には、開発者やAIサービスプロバイダ等と同等の留意が求められる場合がある。

・AIシステム　：　AIソフトを構成要素として含むシステム
・AIサービス　：　AIシステムの機能を提供するサービス
・AI付随サービス：　AIシステムのアップデート又は追加的な学習等に係るサービス

（注）同一の個人・事業者が複数の主体に該当する場合がある。

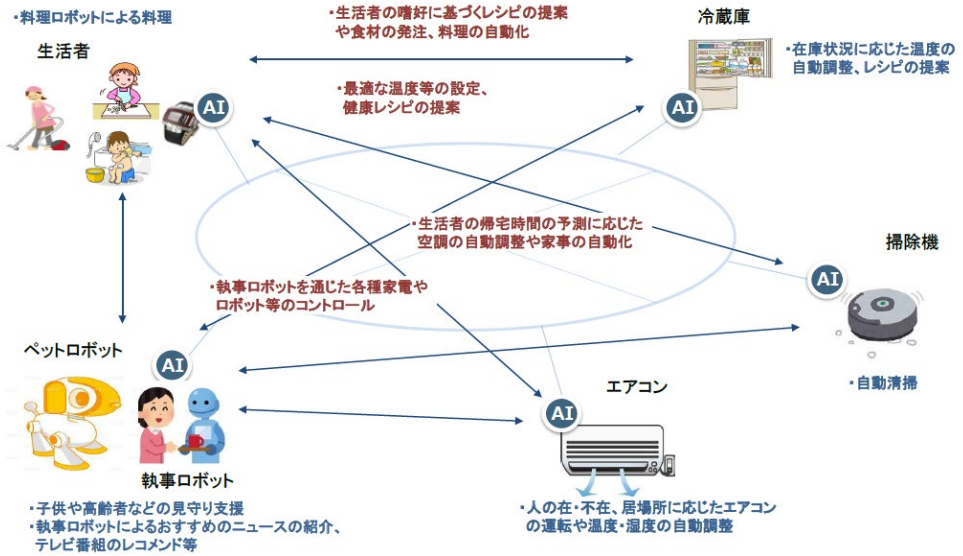
AI利活用ガイドラインとりまとめの経緯

課題の抽出
- ケーススタディ等
- 課題の整理

原則化
- 原則化（AI利活用原則案）
- 原則ごとの論点の抽出

具体化
- 論点ごとのヒアリング（構成員以外を中心に）
- 論点ごとの詳説作成

ケーススタディ例（家庭内での生活）



＜引用＞AIネットワーク社会推進会議 報告書2018 別紙１　を元に加工

想定される利活用	想定される便益	想定されるリスク
人の在・不在、居場所に応じたエアコンの運転や温度・湿度の自動調整	・快適な空間で生活可能に。 ・電力消費の削減やピークカット/ピークシフトが実現。	・在宅・不在や生活習慣等に関する情報が明らかになりプライバシーが侵害されるおそれ。**[プライバシー]**
子供や高齢者などの見守り支援	・安心して外出可能に。 ・就労・地域活動等への参加が容易に。	・ハッキング等により、ペットロボットの制御が失われ（暴走し）、子供等が怪我をしたり、家具が壊れるおそれ。**[安全]**
執事ロボットによるおすすめのニュースの紹介、テレビ番組のレコメンド等	・手が離せない状況などにおいてニュース等の検索が容易に。 ・見たいテレビ番組を視聴可能。	・生活者の趣味・趣向を誤って判断し、必要な情報・関心の高い情報を提供できないおそれ。　**[正当性・公平性、セキュリティ]**
生活者の帰宅時間の予測に応じた空調の自動調整や家事の自動化	・より快適な空間で生活可能に。 ・家事の負担を軽減可能に。	・料理ロボットや掃除ロボットやペットロボットなど異なるロボット間の連携・調整が十分でなく、ロボット同士が衝突したり、破損するおそれ。**[連携、安全]** ・家事をAIに依存しすぎると、災害発生等でAIが利用できなくなった場合、生活が困難になるおそれ。**[役割分担、連携]**
生活者の嗜好に基づくレシピの提案や食材の発注、料理の自動化	・家事の負担を軽減。 ・嗜好に合わせた料理や健康に良い料理を（容易に）選択可能に。	
執事ロボットを通じた各種家電やロボット等のコントロール	・執事ロボットだけで家庭内の複数のロボットや家電等をコントロール可能に。	

課題の整理→原則化

【主として生命・身体の安全、権利・利益等を守るための課題】

○ 生命・身体・財産の安全に関する課題（事故の防止など） ⇒ ①適正利用の原則、④安全の原則

→ どのように事故が発生しないようにするか、また、事故が生じた場合にどのように対応すべきか（責任の在り方を含む。）について検討が必要ではないか。

○ AIによる判断の正当性や公平性に関する課題（差別、生命倫理との関係など）⇒ ②適正学習の原則、⑦尊厳・自律の原則、⑧公平性の原則

→ どのようにAIによる判断の正当性や公平性を確保し差別的な取扱いがなされないようにするか、データの適正性・正確性や人間の介在の在り方を含めて検討が必要ではないか。

○ プライバシーに関する課題（プライバシーの尊重、プロファイリングなど）⇒ ⑥プライバシーの原則

→ どのようにプライバシーを尊重するのか、本人同意の在り方やプロファイリングの在り方などを含めて検討が必要ではないか。

【主として人間とAIとの関係等に関する課題】

○ 人間とAIとの役割分担等に関する課題（人間の判断の介在、関係者間の協力など）⇒ ①適正利用の原則、③連携の原則

→ どのような場合に人間の判断を介在させるべきか、その介在の要否の基準を含めて検討が必要ではないか。また、
安心して安全にAIを利活用するために、どのように関係者が協力すべきかについて検討が必要ではないか。

○ AIに対する受容性に関する課題（利用者に対する説明責任など） ⇒ ⑩アカウンタビリティの原則

→ どのように利用者・社会のAIの信頼性を醸成すべきかについて検討が必要ではないか。

【主として技術的な観点からの解決が求められる課題】

○ AIの判断のブラックボックス化に関する課題（事故が発生した場合の原因究明など）⇒ ⑨透明性の原則

→ どのような場合に、どの程度AIの判断の根拠・理由を明らかにすべきかについて検討が必要ではないか。

○ セキュリティに関する課題（ハッキング対策など） ⇒ ⑤セキュリティの原則

→ どのようにセキュリティを確保すべきかについて検討が必要ではないか。

○ AI間の連携に関する課題（AI間の交渉・調整など） ⇒ ③連携の原則

→ どのようにAI間の円滑な交渉・調整を実現するか、データ形式やプロトコル等の観点も含めて検討が必要ではないか。

【主としてデータに関する課題】

○ AIが学習するデータに関する課題（データの正確性など）⇒ ②適正学習の原則

→ どのようにデータの適正性・正確性を担保するか、また、どのように適切なデータを確保するかについて検討が必要ではないか。

＜引用＞AIネットワーク社会推進会議報告書2018 別紙２－１　AIの利活用の展望　を元に加工

AI利活用原則

AIの利用者（AIを利用してサービスを提供する者を含む）が**利活用段階において**留意することが期待される事項を「原則」という形式で記載

原　則	
適正利用	適正な範囲及び方法でAIを利用
適正学習	AIの学習等に用いるデータの質に留意
連携	AI相互間の連携に留意 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
安全	生命・身体・財産に危害を及ぼすことがないよう配慮
セキュリティ	AIのセキュリティに留意
プライバシー	他者又は自己のプライバシーが侵害されないよう配慮
尊厳・自律	人間の尊厳と個人の自律を尊重
公平性	AIの判断にバイアスが含まれる可能性があることに留意 個人及び集団が不当に差別されないよう配慮
透明性	AI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
アカウンタビリティ	アカウンタビリティを果たすよう努める

※　AIサービスプロバイダやビジネス利用者等が自主的に参照するものとして、また国際的な認識の共有を図るものとして取りまとめ

原則ごとの論点の整理

原則	原則に対する論点
①適正利用	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
②適正学習	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のセキュリティの留意
③連携	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
④安全	ア　人の生命・身体・財産への配慮
⑤セキュリティ	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
⑥プライバシー	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
⑦尊厳・自律	ア　他者の尊厳と自律の尊重
	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
⑧公平性	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）
⑨透明性	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
⑩アカウンタビリティ	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

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・完全性・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。

＜引用＞ https://www.soumu.go.jp/main_content/000637098.pdf

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。

＜引用＞ https://www.soumu.go.jp/main_content/000637098.pdf

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

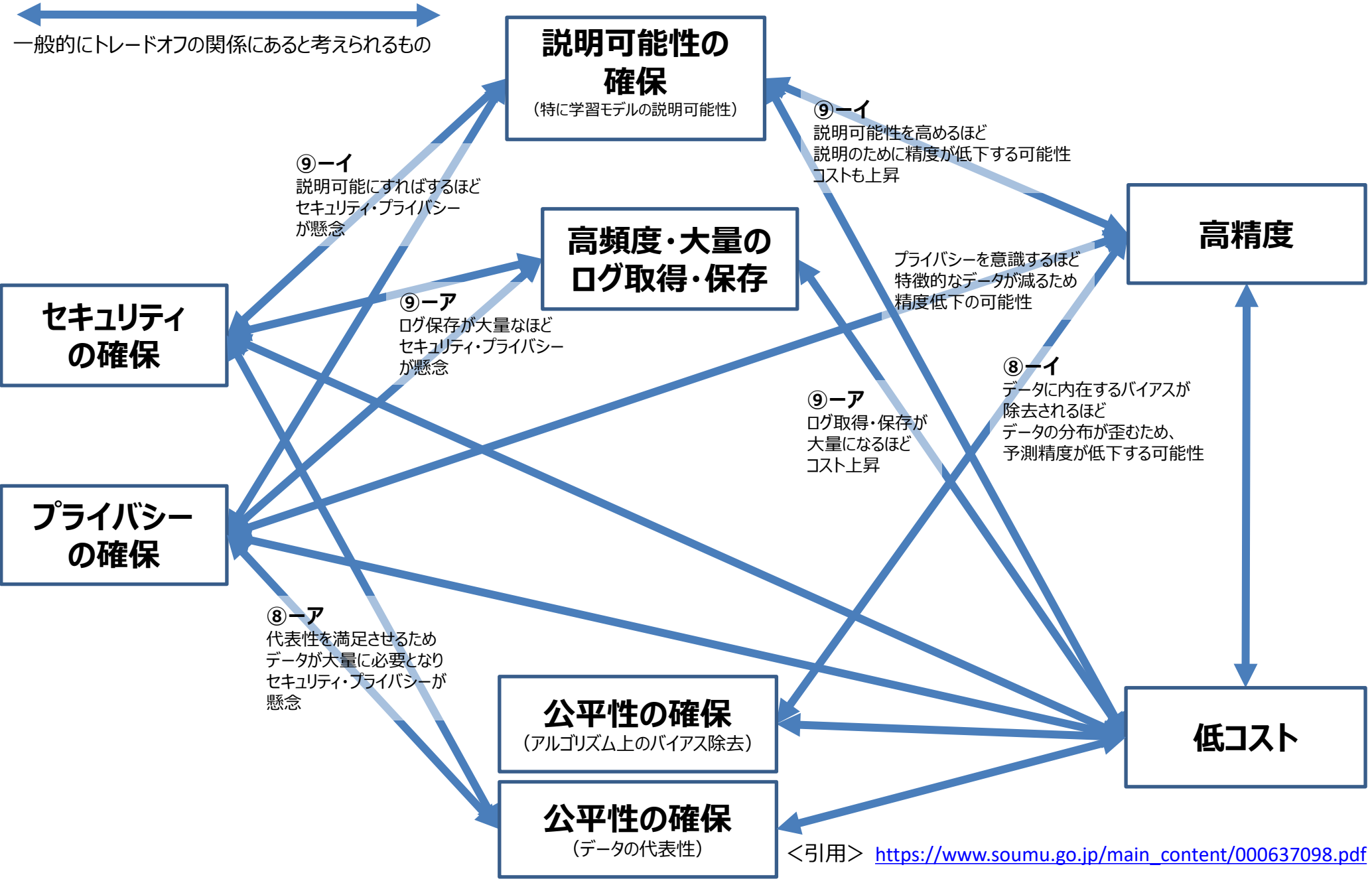
＜参考＞

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。

＜引用＞ https://www.soumu.go.jp/main_content/000637098.pdf

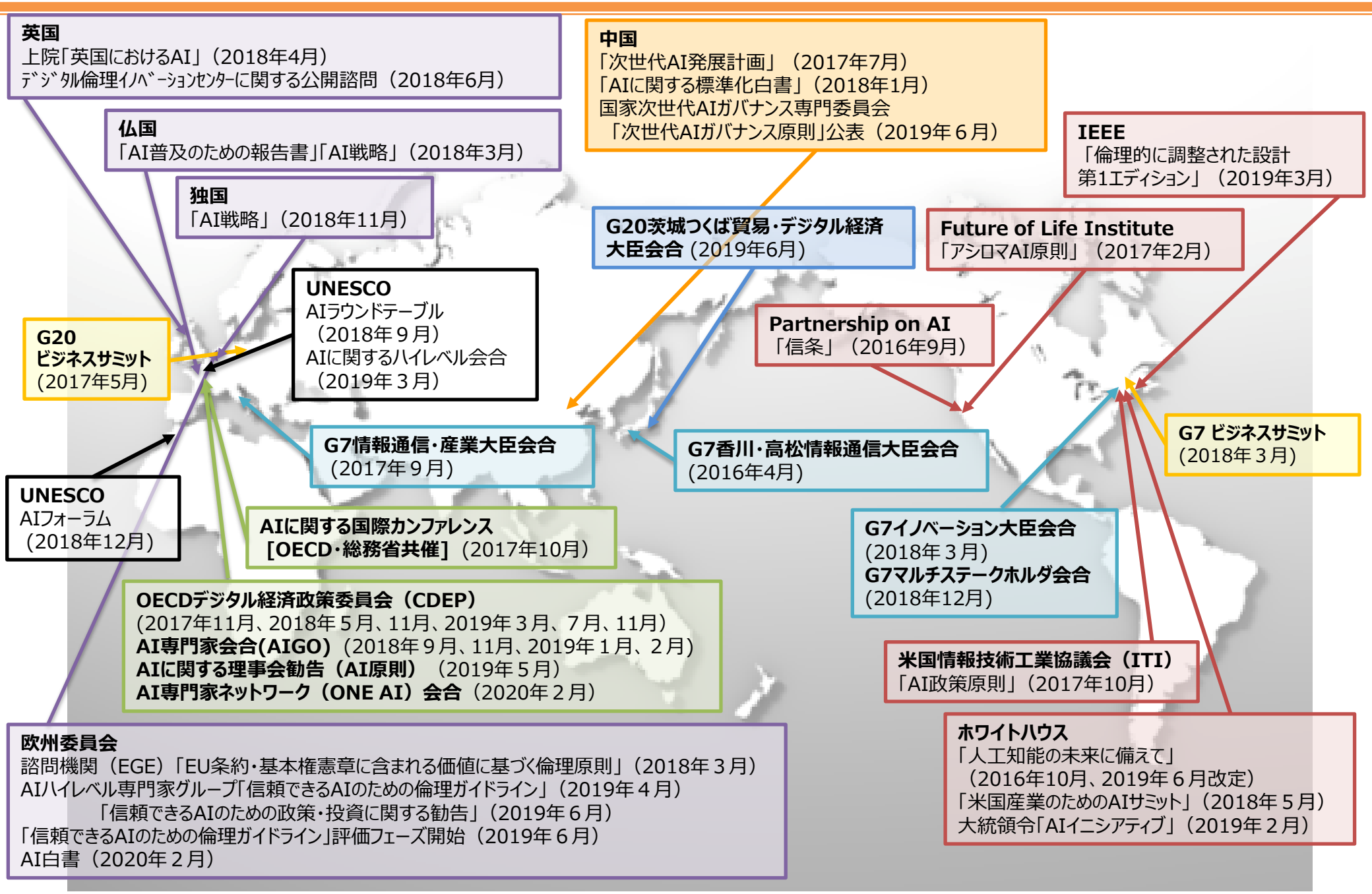
トレードオフの例



<引用> https://www.soumu.go.jp/main_content/000637098.pdf

Contents

国際的な議論


英国
上院「英国におけるAI」（2018年4月）
ﾃﾞｼﾞﾀﾙ倫理ｲﾉﾍﾞｰｼｮﾝｾﾝﾀｰに関する公開諮問（2018年6月）
仏国
「AI普及のための報告書」「AI戦略」（2018年3月）
独国
「AI戦略」（2018年11月）
中国
「次世代AI発展計画」（2017年7月）
「AIに関する標準化白書」（2018年1月）
国家次世代AIガバナンス専門委員会
「次世代AIガバナンス原則」公表（2019年６月）
IEEE
「倫理的に調整された設計
第1エディション」（2019年3月）
G20茨城つくば貿易・デジタル経済
大臣会合（2019年6月）
Future of Life Institute
「アシロマAI原則」（2017年2月）
UNESCO
AIラウンドテーブル
（2018年９月）
AIに関するハイレベル会合
（2019年３月）
G20
ビジネスサミット
（2017年5月）
Partnership on AI
「信条」（2016年9月）
G7情報通信・産業大臣会合
（2017年９月）
G7香川・高松情報通信大臣会合
（2016年4月）
G7 ビジネスサミット
（2018年３月）
UNESCO
AIフォーラム
（2018年12月）
AIに関する国際カンファレンス
[OECD・総務省共催]（2017年10月）
G7イノベーション大臣会合
（2018年３月）
G7マルチステークホルダ会合
（2018年12月）
OECDデジタル経済政策委員会（CDEP）
（2017年11月、2018年５月、11月、2019年３月、７月、11月）
AI専門家会合(AIGO)（2018年９月、11月、2019年１月、２月）
AIに関する理事会勧告（AI原則）（2019年５月）
AI専門家ネットワーク（ONE AI）会合（2020年２月）
米国情報技術工業協議会（ITI）
「AI政策原則」（2017年10月）
欧州委員会
諮問機関（EGE）「EU条約・基本権憲章に含まれる価値に基づく倫理原則」（2018年３月）
AIハイレベル専門家グループ「信頼できるAIのための倫理ガイドライン」（2019年４月）
「信頼できるAIのための政策・投資に関する勧告」（2019年６月）
「信頼できるAIのための倫理ガイドライン」評価フェーズ開始（2019年６月）
AI白書（2020年２月）
ホワイトハウス
「人工知能の未来に備えて」
（2016年10月、2019年６月改定）
「米国産業のためのAIサミット」（2018年５月）
大統領令「AIイニシアティブ」（2019年２月）

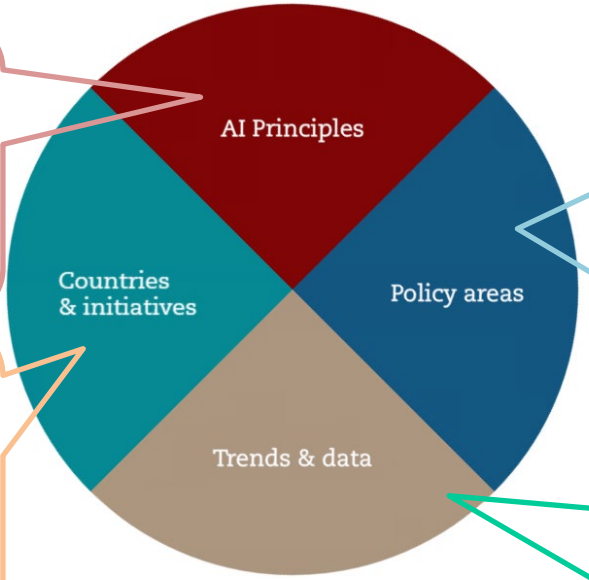
プラクティカルガイダンスの策定、AI政策に関するオブザーバトリ及び専門家会合の設置

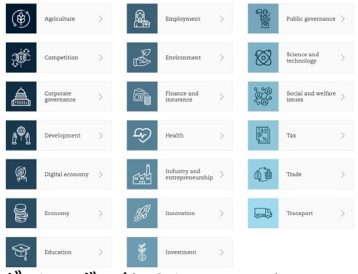
- プラクティカルガイダンス：上記理事会勧告の履行に係る実務者向けのガイダンスで、各原則等の具体的な解説や、AIに関する開発・運用者、政府関係者等に求められる対応、各国の取組事例を記載。AI利活用ガイドラインの内容も反映。本年２月に初版公表。その後も継続的に更新される予定。
- AI政策に関するオブザーバトリ：AIに関する取組の情報共有を進めるためのプラットフォーム（ライブ型のデータベース https://oecd.ai ）であり、本年２月から運用開始。各国のAI戦略や政策の共有、政策の比較分析等が可能。また、本オブザバートリーに関して、政策的、技術的、商業的な助言を行う専門家のネットワーク（ONE AI：OECD Network of Experts on AI）が設置（日本からは須藤教授が参加）。

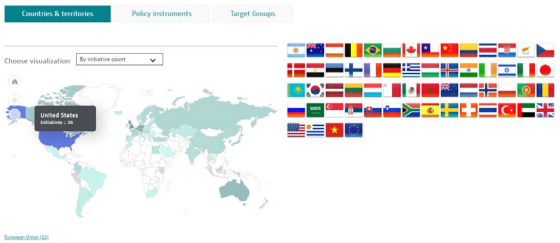
オブザーバトリー４つの柱

①**ＡＩ原則**：ＯＥＣＤのＡＩ原則及び実務者向けのガイダンス（プラクティガル・ガイダンス）を掲載。
②**政策分野**：各公共政策分野毎に、ＡＩ政策ニュースやＡＩ調査に関する公表内容等の様々なコンテンツにアクセス可能。
③**トレンドとデータ**：ＡＩに関する調査データを掲載。データの地域比較や時間的変化を観ることが可能。
④**国々と取組**：ＡＩに関する国家戦略や政策、取組に関するデータベースであり、各国のＡＩ政策を共有・比較することが可能。

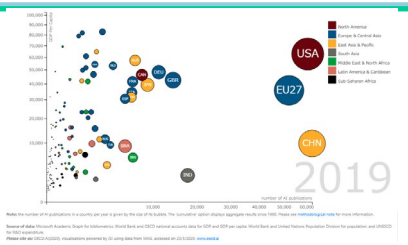
イメージ：OECDのＡＩ原則




イメージ：ダッシュボード（政策分野ごとにコンテンツを整理）



イメージ：国毎のAI政策に関するイニシアチブ数比較



イメージ：ＡＩに関する出版物の数とＧＤＰ

AIに関する国際的な議論の動向　欧州委員会

欧州委員会 「信頼できるAI（Trustworthy AI）のための倫理ガイドライン」公表 【2019年（平成31年） 4月8日】

○ ４月８日、欧州委員会は、選定した52名の専門家グループ（HLEG）により作成された信頼できるAIのための倫理ガイドラインを公表。
○ 同ガイドラインでは、信頼できるAIのためには合法的、倫理的、及び、頑健であるべきとし、その上で基本的人権に基づき尊重すべき４つの倫理原則（人間の自律性の尊重、危害の防止、公平性、説明可能性）、および７つの要求条件（人間の営みと監視、技術的な頑健性と安全性、プライバシーとデータガバナンス、透明性、多様性・無差別・公平性、環境及び社会の幸福、アカウンタビリティ）を掲げ、さらにそれらを評価するためのチェックリスト（Assessment list）を列挙している。
○ 掲げられた内容は非拘束的なものとしてAIを開発・利用する全ての関連するステークホルダーを対象としており、企業・団体等の各ステークホルダは、本ガイドラインの内容を咀嚼し、自らの憲章・行動規範等に適用することにより、信頼できるAIへの関与の表明が可能。
○ 今後、上記チェックリストについては、広く継続的にレビューを行い、2020年に取りまとめる予定

欧州委員会　AI白書公表【2020年（令和２年） ２月19日、ブラッセル】

○ 欧州委員会は2020年2月19日、欧州のデジタル未来形成のため、「AI白書」を公表
● 卓越性（excellence）と信頼性（trust）に基づくAIに向けた枠組案を提示。同案について６月14日までパブコメ実施。
● 卓越性の視点では、バリューチェーン全体にリソースを動員し、中小企業等のAIの展開を加速するためのインセンティブの作成に言及（テストセンター構築など）。
● 信頼性の視点では、リスクの低いシステムに過度の負担をかけることなく、リスクの高いAIシステムに対処できるよう、リスクの高い領域・用途等のスコープを明確にした上で、リスクの高低に応じた将来の規制のあり方を提示。また、消費者保護、不公正な商慣行に対処し、個人データとプライバシーを保護するための厳格なEU規則の適用を継続することに言及。
  - リスクが高いことが想定されるヘルスケア、輸送、公共部門等の特定の用途においては、AIシステムは透明で追跡可能であり、人間の監視を保証すること等、EUの信頼性のあるAI倫理ガイドラインに記載の条件等を踏まえることが必要（他、適切に機能するためのトレーニング、偏りのないデータの利用など）。
  - 今日、遠隔生体認証のための顔認識の使用は特定の条件を除き一般的に禁止されており、EUまたは国内法に基づいて例外として正当化された場合にのみ使用できるが、これについて幅広い議論を開始。
  - リスクが低いAIシステムにおいても、信頼を醸成すべく、EUにおける客観的なベンチマークによる任意のラベル付けスキーム（＝認定の仕組）を検討。

現在の議論の流れ：原則は内容から具体的な実行へ

原則レベルではコンセンサスが得られつつあり、今後はそれをどのように実行していくかが議論の焦点に→**各国の議論に貢献し、認識を共有**

欧州委員会

信頼できるAIのための必要条件

- 人間の営みと監視
- 技術的な頑健性と安全性
- プライバシーとデータガバナンス
- 透明性
- 多様性・無差別・公平性
- 環境及び社会の幸福
- アカウンタビリティ

（上記を評価する）
Assessment listを列挙
→　深掘り中

OECD

信頼できるAIのための責任あるスチュワードシップに関する原則

- 包摂的な成長・持続可能な開発及び幸福
- 人間中心の価値及び公平性
- 透明性及び説明可能性
- 頑健性・セキュリティ及び安全性
- アカウンタビリティ

プラクティカルガイダンス作成
→オブザーバトリ（共有のためのWebプラットフォームに）

日本

AI開発原則
AI利活用原則

- 適正利用、適正学習
- 連携
- 安全、セキュリティ
- プライバシー
- 公平性
- 透明性
- アカウンタビリティ

人間中心のAI社会原則

認識の共有

AI利活用ガイドライン

- 原則を実現する具体的な措置を列記
- 利活用フェーズとの関連性を明確化

Contents

AIネットワーク社会推進会議「報告書2020」

総務省 | AIネットワーク社会推進会

soumu.go.jp/menu_news/s-news/02iicp01_04000235.html

総務省 MIC Ministry of Internal Affairs and Communications

ご意見・ご提案 ENGLISH(TOP) ENGLISH(ICT POLICY) アクセシビリティ閲覧支援ツール 文字サイズの変更 小 標準 大

ENHANCED BY Google

総務省の紹介 | 広報・報道 | 政策 | 組織案内 | 所管法令 | 予算・決算 | 申請・手続 | 政策評価

総務省トップ > 広報・報道 > 報道資料一覧 > AIネットワーク社会推進会議 AI経済検討会 報告書2020の公表

報道資料

令和2年7月21日

AIネットワーク社会推進会議 AI経済検討会 報告書2020の公表

総務省情報通信政策研究所は、平成31年1月から、「AIネットワーク社会推進会議」の下に、「AI経済検討会」を置き、AIに関して経済的な見地から検討を進めてきました。
今般、同検討会において、「AI経済検討会 報告書2020」が取りまとめられましたので、公表します。

1 経緯等

総務省情報通信政策研究所は、平成28年10月から、社会全体におけるAIネットワーク化の推進に向けた社会的・経済的・倫理的・法的課題を総合的に検討することを目的として、産学民の有識者の参加を得て、「AIネットワーク社会推進会議」（議長：須藤 修 中央大学国際情報学部教授、東京大学大学院情報学環特任教授）を開催してきました。
平成31年1月より、同推進会議の下に、AI社会実装の推進により、どのような社会経済を目指すべきか、基本的な政策や中長期的な戦略のあり方について検討するため、「AI経済検討会」（座長：岩田 一政 公益社団法人日本経済研究センター理事長）を置き、令和元年5月には「AI経済検討会報告書」が取りまとめられました。
同報告書を踏まえ、更なる検討を行い、今般、「AI経済検討会 報告書2020」が取りまとめられましたので、公表します。

2 主な内容

(1) 背景（「AI経済検討会」の設置から現在に至るまでの経緯）
(2) AIの社会実装に向けて求められるデータ活用のあり方
(3) AI時代のデータ経済政策
(4) 将来像（「インクルーシブなAI経済社会」のイメージ）
(5) 提言

3 公表資料

AI経済検討会報告書2020
・AI経済検討会 報告書2020（本体）
・別紙1 AIネットワーク社会推進会議AI経済検討会／データ専門分科会 構成員
・別紙2 開催経緯
・別紙3 AIネットワークの進展に伴い形成されたエコシステムの展望に関する分析
・AI経済検討会 報告書2020（概要）

<引用> https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000091.html

AIネットワーク社会推進会議「報告書2020」構成

はじめに

第１章 AIネットワーク化をめぐる最近の動向

　１．AIとCOVID-19対策

　２．国内・海外及び国際的な議論の動向

第２章 AIネットワーク化の進展に伴い形成されるエコシステムの展望

第３章 開発者・AIサービスプロバイダーにおける取組

第４章 ビジネス利用者における取組

第５章 消費者的利用者に関する取組

第６章 セキュリティに関する取組

第７章 保険に関する取組

結びに代えて

はじめに ～「安心・安全で信頼性のあるAIの社会実装」に向けて～

データ提供者

データ提供

データ提供

※個人情報保護法等の規律／AIガイドライン（プライバシー原則等）との関係

開発者／AIサービスプロバイダー

AI開発・提供部門

経営判断部門（意思決定）

自己点検・自己評価の仕組み

AI原則等

第３章

※AI原則等は安心・安全で信頼性のあるAIの開発・利活用にどのように活かすことができるか

※安心・安全で信頼性のあるAIの開発・利活用に必要なガバナンス体制にはどのようなものが考えられるか（自己点検・自己評価の仕組み／外部による評価の仕組）

最終利用者

ビジネス利用者

ビジネス利用部門

経営判断部門（意思決定）

自己点検・自己評価の仕組み

AI原則等

例：【移動】【健康】【金融】【危機管理】【まちづくり】【ものづくり】【居住】【エンタメ】

第４章

※ビジネス利用者にとってＡＩの利活用を進めるにあたり課題となっていることは何か。またその課題解決に必要な取組はどのようなものが考えられるか

消費者的利用者

第５章

※消費者的利用者に関する取組をどのように進めていくことが考えられるか

※高齢者・障害者にとっての安心・安全で信頼性のあるAIの社会実装に必要な取組みとしてどのようなものが考えられるか

第三者

安心・安全で信頼性のあるAIの社会実装のための環境整備等

第２章

将来像→AIの利活用の進展等を踏まえ、近い将来（2025年頃）、中期的な将来（2035年頃）を視野にAIの利活用の将来像としてどのようなことが考えられるか

技術面→AIと情報セキュリティ：開発・利活用等の各フェーズで何が課題でどのような取組みが必要か
他、Certification、品質の確保等

第６章

セーフティネット→保険の仕組みにはどのようなものが考えられるか

第７章

責任のあり方：契約責任（私的自治の原則）、不法行為責任（過失責任主義）

AIネットワーク社会推進会議におけるガイドラインの策定

AI開発ガイドライン（開発者向け） ＜ 連携参照 ＞ AI利活用ガイドライン（主としてAIサービスプロバイダー、ビジネス利用者向け）

消費者的利用者部分

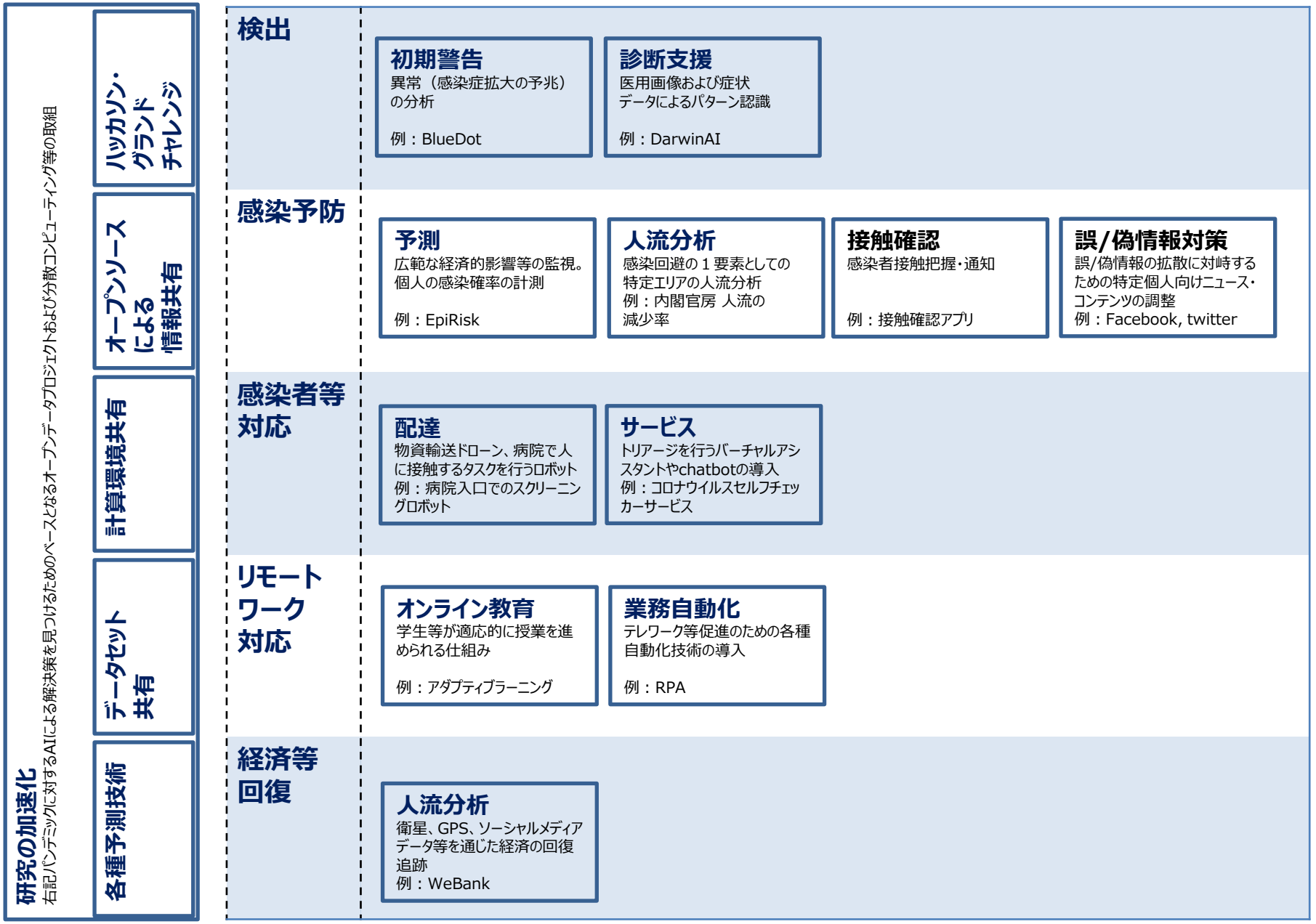
海外における安心・安全で信頼性のある社会実装促進の環境整備としてのガイドラインの策定（例）

第１章２節

OECD：プラクティカルガイダンス（オブザーバトリ（oecd.ai）に掲載・継続的に更新）

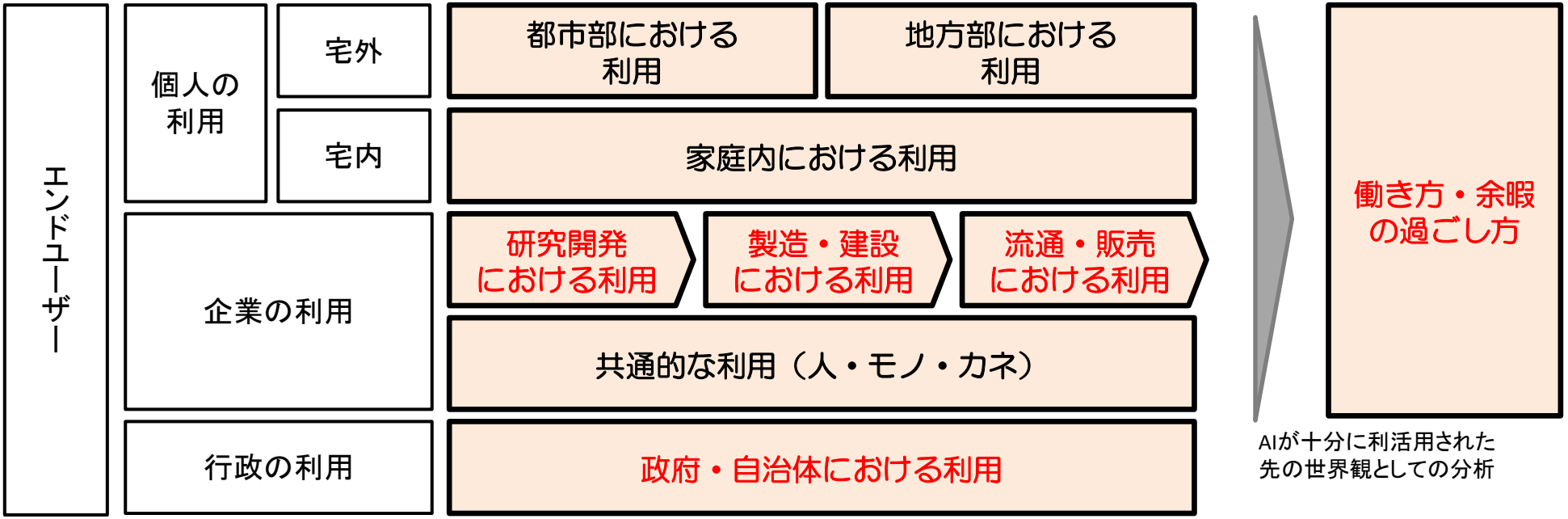
EU： Trustworthy AI Assessment List（2019年12月までPilot Phaseを試行、2020年内を目処にとりまとめ予定）、AI白書

米国：Guidance for Regulation of Artificial Intelligence Applications（民間AI向け規制検討のベース、2020年3月まで意見募集）


研究の加速化
右記パンデミックに対するAIによる解決策を見つけるためのベースとなるオープンデータプロジェクトおよび分散コンピューティング等の取組
各種予測技術
データセット共有
計算環境共有
オープンソースによる情報共有
ハッカソン・グランドチャレンジ
検出
初期警告
異常（感染症拡大の予兆）の分析
例：BlueDot
診断支援
医用画像および症状データによるパターン認識
例：DarwinAI
感染予防
予測
広範な経済的影響等の監視。個人の感染確率の計算
例：EpiRisk
人流分析
感染回避の１要素としての特定エリアの人流分析
例：内閣官房 人流の減少率
接触確認
感染者接触把握・通知
例：接触確認アプリ
誤/偽情報対策
誤/偽情報の拡散に対峙するための特定個人向けニュース・コンテンツの調整
例：Facebook, twitter
感染者等対応
配達
物資輸送ドローン、病院で人に接触するタスクを行うロボット
例：病院入口でのスクリーニングロボット
サービス
トリアージを行うバーチャルアシスタントやchatbotの導入
例：コロナウイルスセルフチェッカーサービス
リモートワーク対応
オンライン教育
学生等が適応的に授業を進められる仕組み
例：アダプティブラーニング
業務自動化
テレワーク等促進のための各種自動化技術の導入
例：RPA
経済等回復
人流分析
衛星、GPS、ソーシャルメディアデータ等を通じた経済の回復追跡
例：WeBank

① AIの利活用に着目し、生活者と事業者の両面から、**AIの利活用シーンを展望**。
AIの利活用シーンの展望に当たっては、次のように利活用シーンを分類。

＜AIの利活用シーンの分類＞



（AIネットワーク社会推進会議　報告書2018 別紙２「AIネットワーク化の進展に伴い形成されるエコシステムの展望について」をもとに作成、
赤字は報告書2018に対し加えた領域）

② 上記①の利活用シーンをもとに、いくつかの事例の**AIの社会実装に関するケーススタディ**を行い、
AIの利活用による便益及び課題を整理。

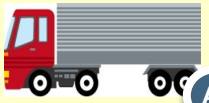
ケース：移動（完全自動運転）
ケース：健康（医療・介護）
ケース：金融
ケース：危機管理（防犯・公共インフラ・防災）
ケース：ものづくり
ケース：居住
ケース：エネルギー

個人（都市部）におけるＡＩの利活用シーン（例）

移動

MaaS：Mobility as a Service

● MaaSの普及により移動の自由度や利便性が飛躍的に向上するほか、完全自動運転の実現により、移動時間の有効活用を図ることなどができる。



AI 長距離トラック・バスの自動運転の一部実現（隊列走行の実現）により労働負荷が緩和

AI 高速道路への合流判断支援

AI 完全自動運転の実現による移動時間の有効利用・渋滞等の人為的ミス低減

AI AI

移動履歴やその際の健康状態の自主的な収集・提示による、人との接触抑制を意識した移動ルートや時間帯の提案

介護

● 生活状況や健康状態に応じて自立に向けた支援がなされるほか、自動運転車での送迎や、介護ロボットの活用等により人手不足を補うことができる。

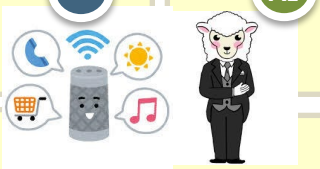
AI AI 自動運転車による介護施設への送迎

AI 完全自律ロボットによる介護支援

AI ウェアラブルロボット（装着型ロボット、ロボットスーツ）による身体能力拡張で要介護者が自立した生活を実施

AI

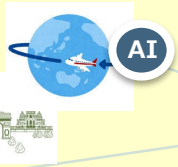
生活状況や健康状態に応じた最適なケアプランの提案・作成（自立支援含む）



AI AI AIスピーカーとの対話（生活・自立支援）執事ロボットの活用

観光・旅行

過去の旅行歴、趣味・嗜好、予算等に応じた最適な旅行計画の提案、手続きの実施



AI

AI 趣味・嗜好に応じたVR／ARと連動した歴史の疑似体験

AI 外出から帰宅まで全ての旅程を手ぶらで完結（顔認証等で全て完結）

多言語翻訳による外国人とのコミュニケーションの円滑化

AI

● 最適な旅行を計画すると同時にチケットが自動手配され、顔認証や翻訳により荷物や言語に悩むことなく快適に観光を楽しむことができる。

VR：Virtual Reality（仮想現実）
AR：Augmented Reality（拡張現実）

教育・人材育成

通信帯域が限定される中でAIでのデータダイエットにより効率的な遠隔教育を実施

AI

学校、塾、家庭での学習の連携により、自身の能力、適性等に応じた学習コンテンツの提供（場所や学年を問わず必要な学習を実施）

AI

AI

リタイア後の再就職支援、生涯学習支援

AI

技能習得のための最適なコンテンツの提供（熟練者のノウハウと知見の伝承など含む）

● 学校、塾、家庭などが連携し、最適な学習コンテンツが提供されるほか、必要な技能を習得するための最適なコンテンツが提供される。

AI：既に実用化されているもの
近い将来実現しそうなもの
（～2025年目処）

AI：中期的なもの
（～2035年目処）

（注）想定される利活用のうち、いくつかの例を記載
現行制度等を前提とせずに利活用の可能性を展望して記載

Contents

はじめに ～「安心・安全で信頼性のあるAIの社会実装」に向けて～

データ提供者

データ提供

データ提供

※個人情報保護法等の規律／AIガイドライン（プライバシー原則等）との関係

開発者／AIサービスプロバイダー

AI開発・提供部門

経営判断部門（意思決定）

自己点検・自己評価の仕組み

AI原則等

第３章

※AI原則等は安心・安全で信頼性のあるAIの開発・利活用にどのように活かすことができるか

※安心・安全で信頼性のあるAIの開発・利活用に必要なガバナンス体制にはどのようなものが考えられるか（自己点検・自己評価の仕組み／外部による評価の仕組）

最終利用者

ビジネス利用者

ビジネス利用部門

経営判断部門（意思決定）

自己点検・自己評価の仕組み

AI原則等

例：【移動】【健康】【金融】【危機管理】【まちづくり】【ものづくり】【居住】【エンタメ】

第４章

※ビジネス利用者にとってＡＩの利活用を進めるにあたり課題となっていることは何か。またその課題解決に必要な取組はどのようなものが考えられるか

消費者的利用者

※消費者的利用者に関する取組をどのように進めていくことが考えられるか

第５章

※高齢者・障害者にとっての安心・安全で信頼性のあるAIの社会実装に必要な取組みとしてどのようなものが考えられるか

第三者

安心・安全で信頼性のあるAIの社会実装のための環境整備等

第２章

将来像→AIの利活用の進展等を踏まえ、近い将来（2025年頃）、中期的な将来（2035年頃）を視野にAIの利活用の将来像としてどのようなことが考えられるか

技術面→AIと情報セキュリティ：開発・利活用等の各フェーズで何が課題でどのような取組みが必要か

他、Certification、品質の確保等

第６章

セーフティネット→保険の仕組みにはどのようなものが考えられるか

第７章

責任のあり方：契約責任（私的自治の原則）、不法行為責任（過失責任主義）

AIネットワーク社会推進会議におけるガイドラインの策定

AI開発ガイドライン（開発者向け）

連携参照

AI利活用ガイドライン（主としてAIサービスプロバイダー、ビジネス利用者向け）

消費者的利用者部分

海外における安心・安全で信頼性のある社会実装促進の環境整備としてのガイドラインの策定（例）

第１章２節

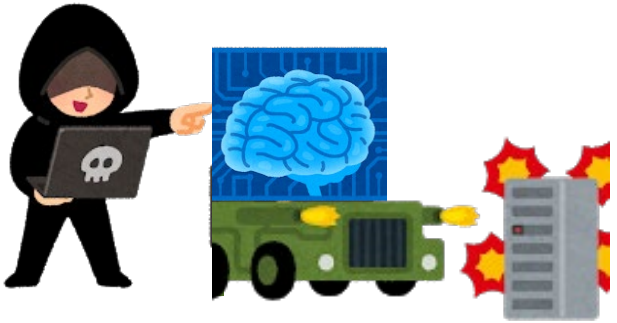
OECD：プラクティカルガイダンス（オブザーバトリ（oecd.ai）に掲載・継続的に更新）

EU： Trustworthy AI Assessment List（2019年12月までPilot Phaseを試行、2020年内を目処にとりまとめ予定）、AI白書

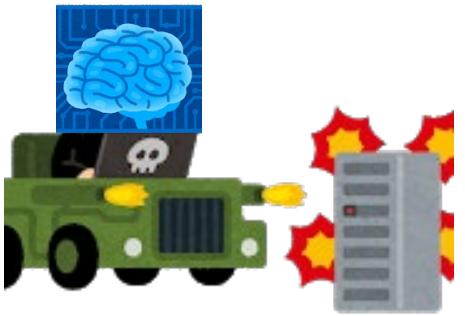
米国：Guidance for Regulation of Artificial Intelligence Applications（民間AI向け規制検討のベース、2020年3月まで意見募集）

AIとセキュリティの関係

東京電機大・佐々木顧問によると、AIとセキュリティの関係は以下の４つに大別。

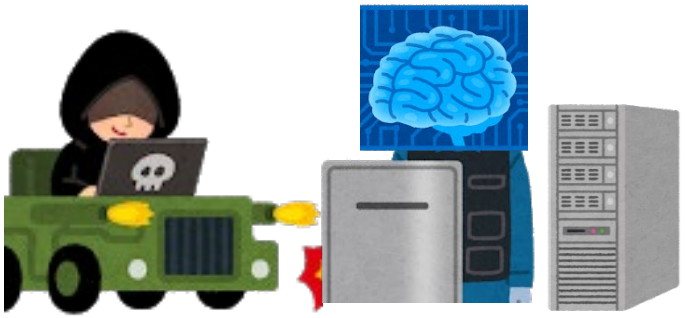


(a) Attack using AI
（AIを利用した攻撃）



(b) Attack by AI
（AI自身による攻撃）

(c) Attack to AI
（AIへの攻撃）



(d) Measure using AI
（AIを利用したセキュリティ対策）

AI利活用ガイドラインでは「セキュリティの原則」を打ち出し、AIを使う人を守るための視点(c)にフォーカス。
⇒視点(c)を個別に深掘りすることも重要だが、「それぞれ深めるのも大事であるが、組み合わせることで研究が深まる。」との意見もあるため、その他の視点も考慮することが重要。

＜引用＞佐々木良一（東京電機大学研究推進社会連携センター顧問　客員教授）「AIとセキュリティ」
https://digitalforensic.jp/2018/09/18/column531/　を元に図示化

(a) Attack using AI
（AIを利用した攻撃）
とその対策の例

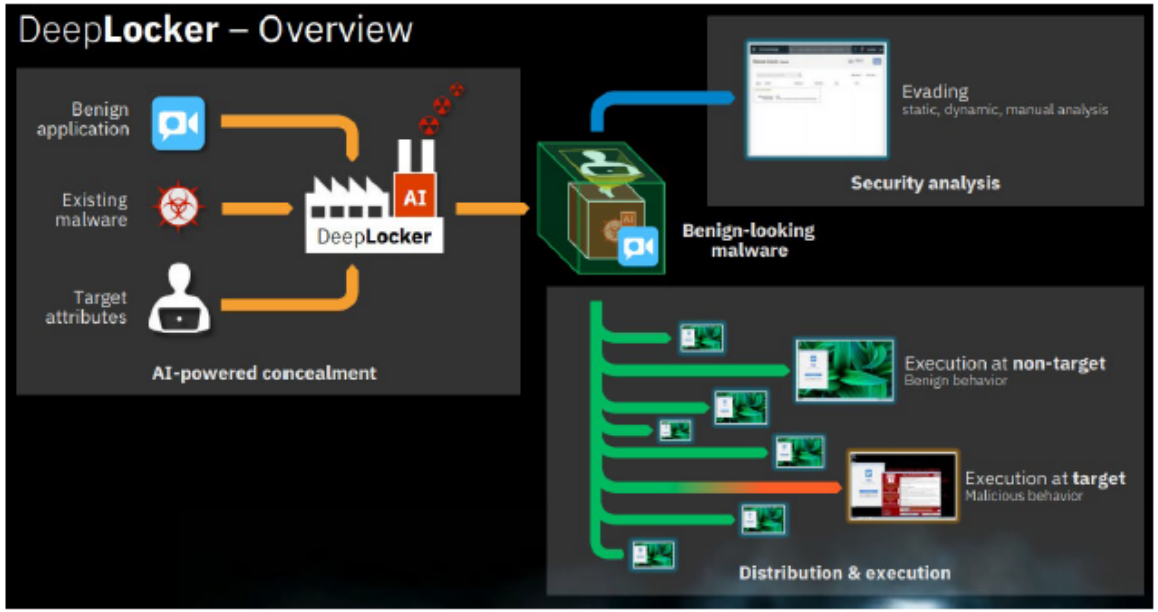
AIを利用した攻撃

攻撃名称	AIの用途	内容
DeepExploit	最適な攻撃手法の判断	深層強化学習を侵入テストに利用したPoC。 侵入対象のシステムから収集した情報（OS、製品名/バージョン等）を基に、システム侵入に成功する確率が最も高い攻撃手法を判断して侵入行為を実行。侵入に成功後、侵入したシステムを踏み台にし、内部のシステムに侵入を繰り返す。
DeepLocker	・標的の識別 ・マルウェアの秘匿	深層学習を標的型マルウェア攻撃に利用したPoC。 暗号化したマルウェアを内蔵し、平時は顔認証アプリやビデオアプリ等として振る舞いながら、Webカメラ/マイク経由で標的人物の情報を収集。標的人物を識別した場合、内蔵マルウェアを復号して攻撃を行う。アンチウィルスソフトに検知されずに標的のPC/スマートフォン等に入り込むことが可能。
tAIchi	マルウェアの自動生成	GANと強化学習をマルウェア生成に利用したPoC。 既知のマルウェアをGANと強化学習で変形させ、アンチウィルスソフトによる検知を回避するマルウェア（亜種）を自動生成。
Deepfake	・顔の入れ替え ・表情の再現	Autoencoder/decoderをフェイク動画の作成に利用した技術群。 オリジナル動画の顔部分を標的人物の顔に入れ替えることで、標的人物のフェイク動画を作成。

JNSA

＜引用＞JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

DeepLocker：概要



出典[1] "Black Hat USA 2018: DeepLocker – Concealing Targeted Attacks with AI Locksmithing"

- 深層学習モデルにマルウェアを埋め込んだ世界初の**標的型攻撃手法**
- **暗号化マルウェア**を良性アプリに内蔵し、**平時は良性アプリとして動作。**
- 良性アプリ経由で標的の情報を収集し、**標的の有無を深層学習モデルで判定。**
- **標的を認識した場合、マルウェアを復号**して攻撃を実行。

JNSA

<引用>JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

Deepfake：概要

標的人物の顔

出典[3] “MBSD Blog: DeepFake –動画編-”

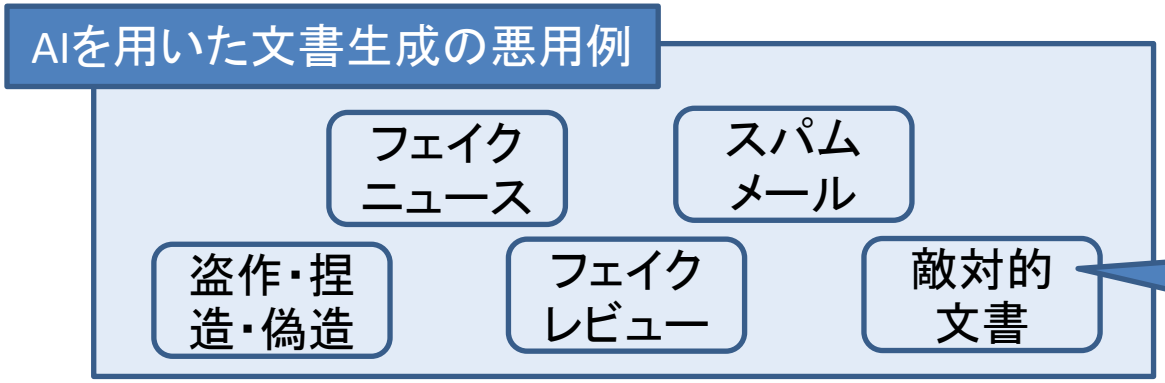
- オリジナル動画に**標的人物の顔をマッピング**する技術群。
- 顔の角度、唇/目の動き、表情等を**自然に再現可能**。
- 技術的に**音声の再現も可能**（オーディオフェイク）。
- 虚偽報道、詐欺、プロパガンダ、ポルノ等への悪用が懸念されている。

JNSA

<引用>JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

KDDI総合研究所での研究例：機械生成文書の検知

逆翻訳により機械生成文書の高精度な検知を実現



AIの高度化により、様々な文書が自動生成可能となり、機械生成文書を悪用した攻撃が多様化

【対策】文書から特徴抽出を行うための3種類の特緒抽出モジュールを開発

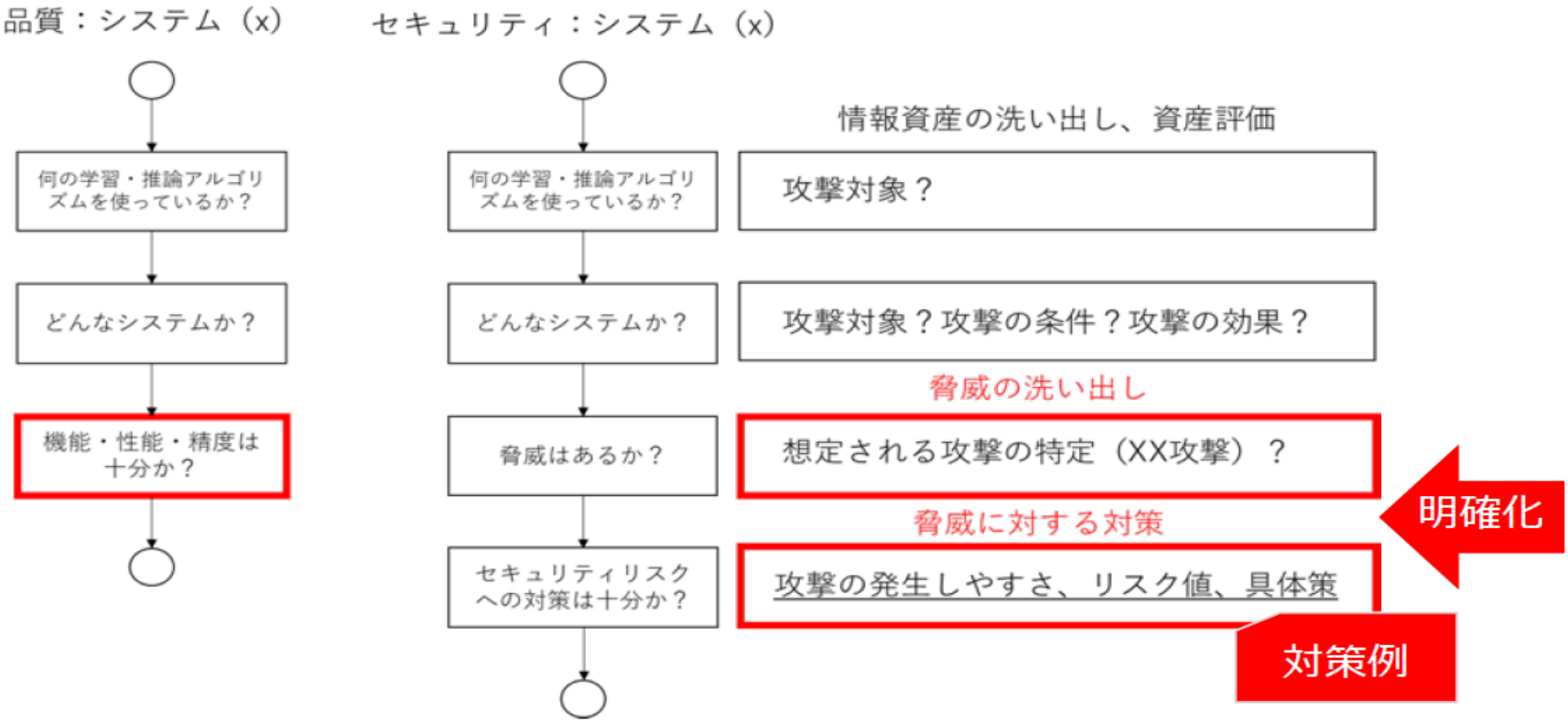
- 機械翻訳による盗作文書用
- 逆翻訳による偽造文書用
- 敵対的文書用

⇒人がある言語で書いた文章を、AIで別の言語に翻訳した後に元の言語に逆翻訳し、元の文章との差異を学習することで、約9割の精度で機械生成文書の検知に成功

<引用> https://www.kddi-research.jp/topics/2019/082801.html 別紙１
2020/1/31 NHKニュース おはよう日本 <おはBiZ> で報道

(c) Attack to AI
（AIへの攻撃）
とその対策の例

JNSA IoTセキュリティWG
～AIセキュリティ調査の動機～



機械学習の脅威と対策を明確化

JNSA

＜引用＞JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

AI特有の脅威

機械学習への攻撃

攻撃（脅威）	サブ分類	内容
回避攻撃 (Evasion Attacks)	・透明化攻撃 (Stealth) ・なりすまし攻撃 (Impersonate)	人間には認識できない「摂動を含んだデータ入力」により、人間と機械学習の推論エンジンとで異なる認識を起こす攻撃（画像、音声、文字等）
中毒攻撃 (Poisoning Attacks)	・可用性攻撃 (Availability) ・バックドア攻撃 (Backdoor)	学習データへの「不正データの入力（注入）」により、学習モデルの境界を何らかの方法によりシフトする攻撃 （機械学習のモデル境界を大量の不良データの注入により使用不能とする可用性攻撃と、少量の洗練されたデータ注入によりバックドアを生成するバックドア攻撃がある）
移転攻撃 (Inversion Attacks)	・プライバシー攻撃 (Privacy) ・メンバーシップ推論攻撃 （Membership）	機械学習の「推論エンジンへのデータ入出力または反応」によって、元データなどの機密情報（またはモデル自体）を抽出する攻撃 （メンバーシップ推論攻撃では、敵対者が手元データが相手のデータセットに含まれているかを探る攻撃）

※違う名称、分類もあるが、ここでは３つの攻撃まとめた

JNSA

＜引用＞JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

【再掲】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）

回避攻撃

- （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

中毒攻撃

- 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

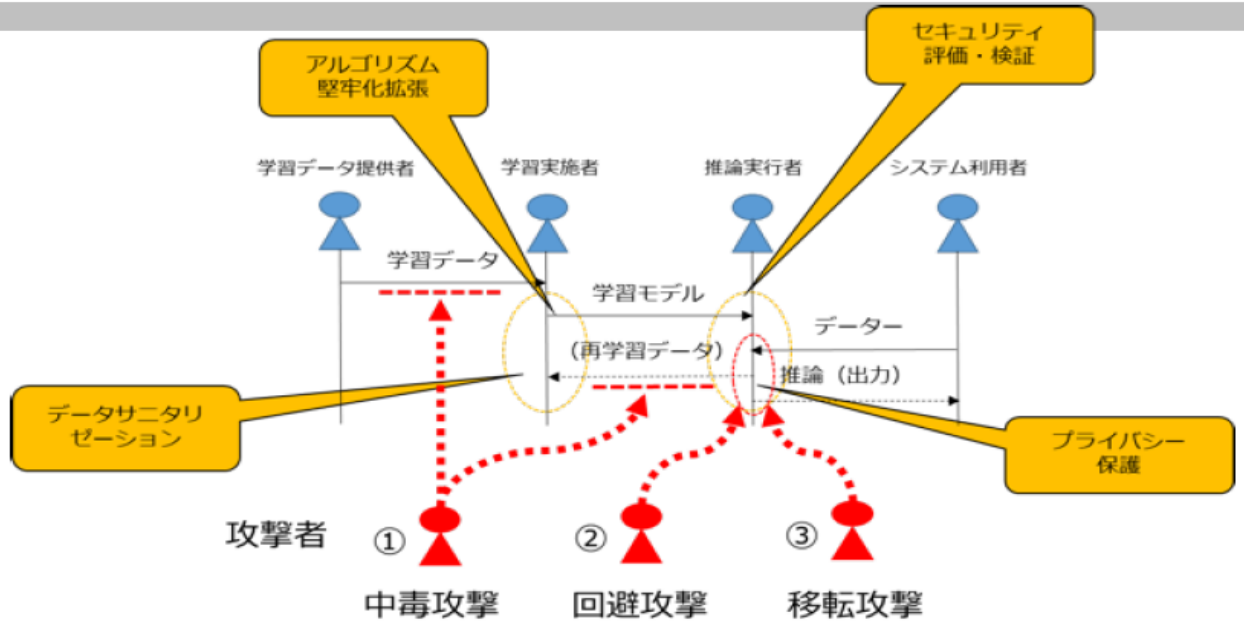
移転攻撃

＜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。

機械学習での脅威と対策（詳細）



対中毒攻撃

異常値検出／モデル精度に与える影響分析

- 入力を意図的に摂動（STRIP）
- 境界シフトの警告（LOOP）
- 回帰（TRIM）

対回避攻撃

全網羅／経験的防御

- 摂動攻撃パターンの網羅（困難）
- 敵対的訓練（中毒攻撃の副作用）
- 勾配マスキング
- 入力変更（ノイズ除去）
- 検出技術
- NULLクラス

対移転攻撃

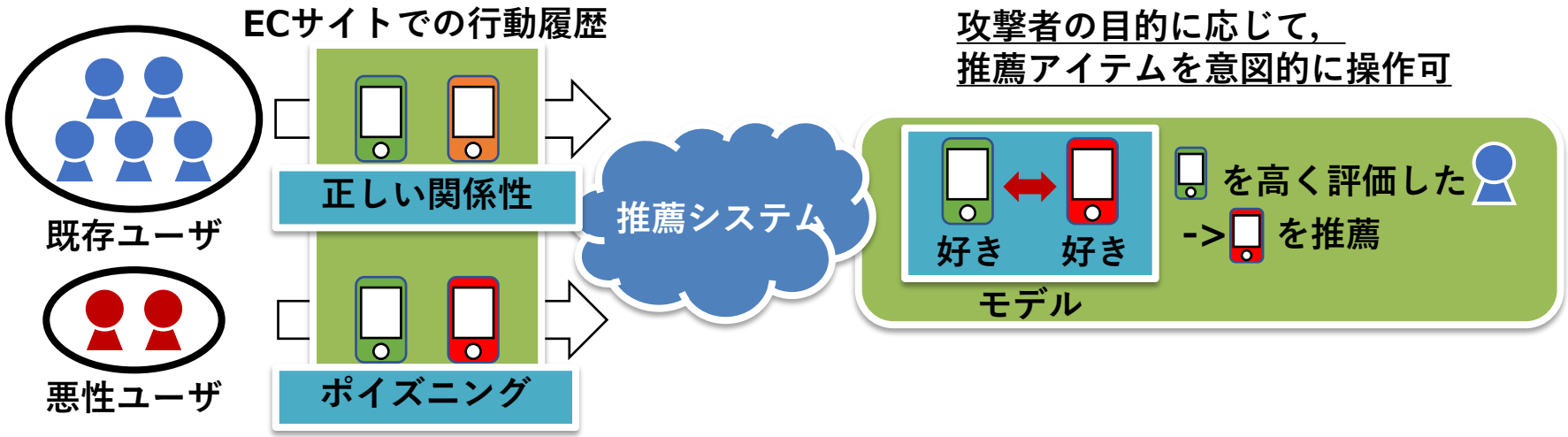
API強化／サニタイズ／モデル強化／検出／差分プライバシー

- API非公開
- 漏らしたくないデータの除去
- モデル選択／フィット制御／知識管理
- 攻撃パターンの異常性判定
- ユーザー入力のランダム化

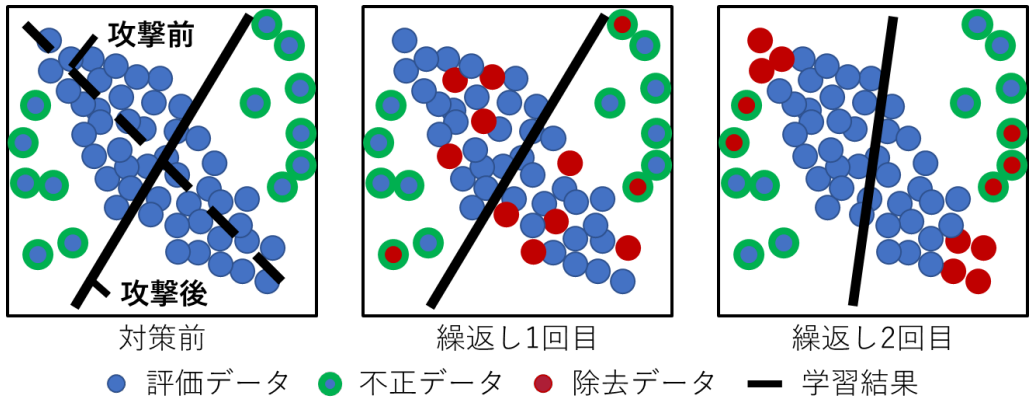
JNSA

＜引用＞JNSA（日本ネットワークセキュリティ協会）「AI×セキュリティ」
https://www.soumu.go.jp/main_content/000696633.pdf

安心してECサイトを利用できる商品推薦サービス向け攻撃対策



【対策】一般的な利用者の評価データと学習結果の違いが小さくなるように、データを選び直しながら学習を繰り返すことで異常なデータを除去



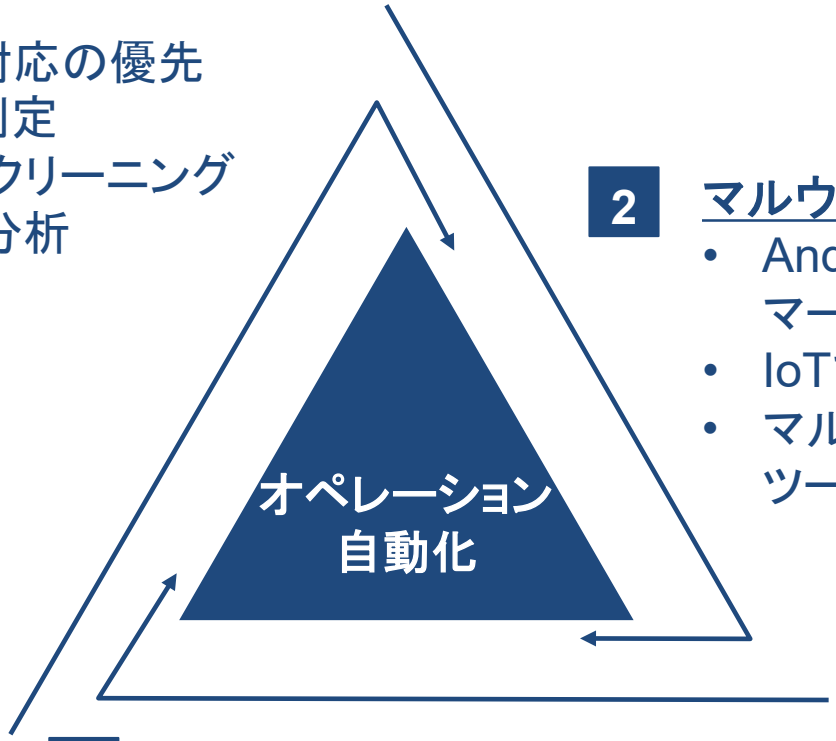
⇒既存の技術では約7割にとどまる
不正データの除去率を98%まで
高められることを実験的に確認

<引用> https://www.kddi-research.jp/newsrelease/2019/100801.html
2019/10/08 日経・日経産業掲載

(d) Measure using AI
（AIを利用したセキュリティ対策）
の例

機械学習を用いたサイバーセキュリティ技術の発展のためNICTで注力しているドメイン

1 インシデント対応の優先順位の自動判定
- アラートスクリーニング
- 脆弱性の分析



2 **マルウェア機能分析自動化**
- Androidアプリおよびマーケット分析
- IoTマルウェア分析
- マルウェア自動分析ツール開発

3 **攻撃の検知・脅威予測**
- ダークネット分析
- ユーザトラフィックの異常検出
- 脅威予測

引用：NICT で行われている機械学習を用いたサイバーセキュリティ技術の発展のための検討

アラートのスクリーニングと優先順位付け: 概要

重要なセキュリティアラートを特定する現在のプロセス

セキュリティアプライアンス

固定ルール

手動検証

CISCO, lastline, JUNIPER NETWORKS, AlaxalA, PFU a Fujitsu company, FFRI, A10

アラート/ログ
>= 800M/day

精査が必要な
アラート
~ 100/day

重要なアラート
~ 5 /day

重要なアラートが
漏れる可能性有

専門家が1日4時間もの時
間を費やしているのが現状

上記の2段階プロセスを効率化すべく、2段階のフィルタリング(固定ルール+手動検証)を機械学習技術に置き換える

引用：NICT で行われている機械学習を用いたサイバーセキュリティ技術の発展のための検討

Androidアプリの分析: 概要

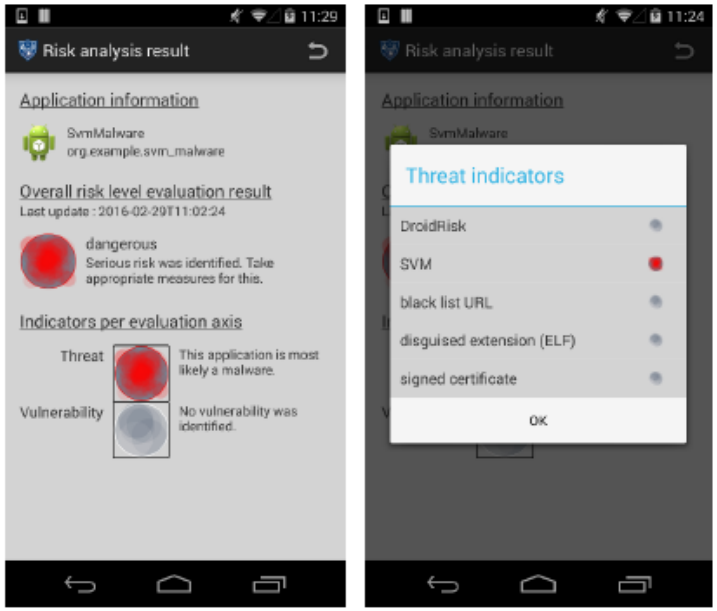
- 機械学習及びニューラルネットワークを用いてAndroidマルウェアを検知 (recall ≒99.52%)
  - 特徴量: パーミッション要求、APIコール、アプリカテゴリ、アプリクラスタ(アプリ説明文からアルゴリズムにより生成)
  - Step 2が計算コストを劇的に削減

- 幾つかの分析を実施中
  - Step2抜きでのパフォーマンスはSVM-RFEを活用することで94-95%
  - Step2抜きで、Step3に各種深層学習アルゴリズムを活用しても、精度はやはり94-95%
  - SVM-RFEに基づき分析した結果、影響度の高い特徴量は、APIコール、いくつかのパーミッションとカテゴリ情報

Step 1: データを収集し、特徴情報を生成

Step 2: ニューラルネットワークを用いて特徴情報の次元を削減

Step 3: 機械学習を用い、悪性・良性のアプリを識別



Sources: B.Sun et al., "A Scalable and Accurate Feature Representation Method for Identifying Malicious Mobile Applications," ACM SAC, 2019.
T.Takahashi et al., "Android Application Analysis using Machine Learning Techniques," Intelligent Systems Reference Library, 181 - 205, 2019.

引用：NICT で行われている機械学習を用いたサイバーセキュリティ技術の発展のための検討

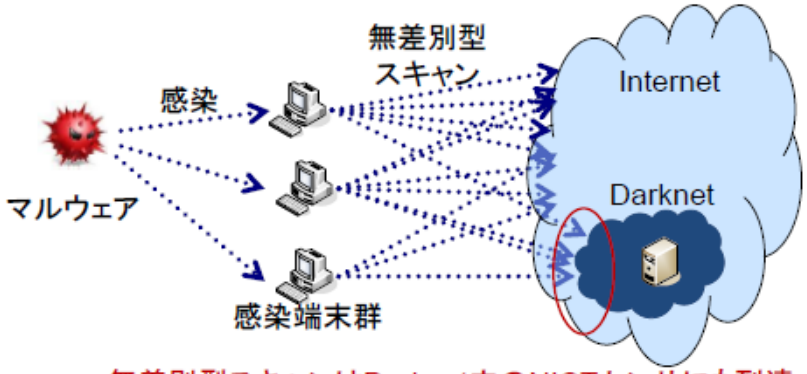
同期性検知: 概要

目的

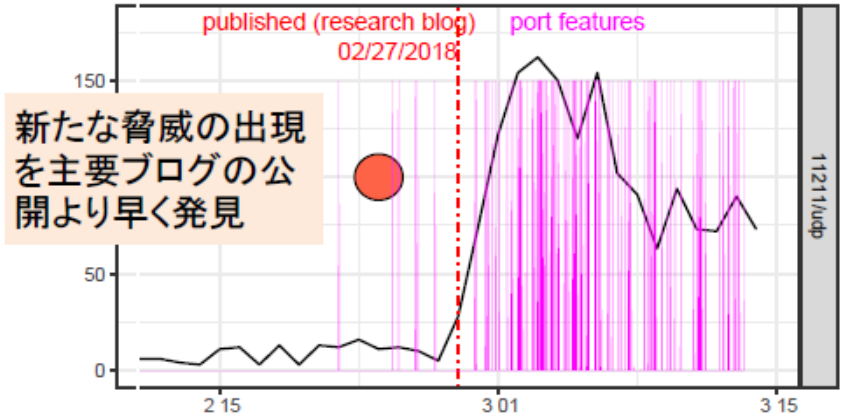
- ホストの同期性を検知
- その際には検知のリアルタイム性を実現し、かつ偽陽性/偽陰性を低減

アプローチ

- 同期性を特定することによりマルウェアの感染活動状況を把握
  - 多数のマルウェアは感染拡大に向けてスキャン活動を実施
  - 新たなマルウェアの登場・感染拡大は同期性という形で ダークネット空間で観測される
- ダークネットトラフィックを教師なし学習にて分析し、同期性を持つホストを特定
- 特定はリアルタイムに実施できるようにアルゴリズムを発展



無差別型スキャンはDarknet内のNICTセンサにも到達

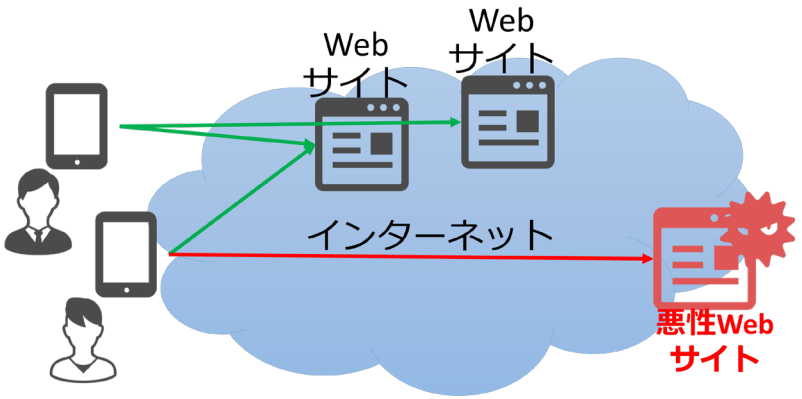


A sample case of a coordinated scan detection
(x: date/time, y: number of sources)

Source: H.Kanehara et al., "Real-Time Botnet Detection Using Nonnegative Tucker Decomposition," ACM SAC, 2019.

引用：NICT で行われている機械学習を用いたサイバーセキュリティ技術の発展のための検討

ユーザの振る舞い分析により発生前に被害を高精度で予測

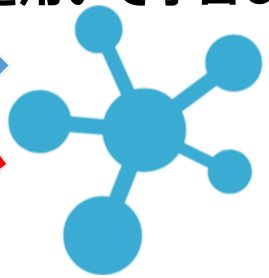


ユーザのWebアクセスに関する**振る舞いの変化**を捉えて、危険サイト**アクセスに至る前に**警告を発したい

Webアクセスの遷移

http://www.google.co.jp/
http://reddit.com/r/football
(現在) **http://football-stream.example.com/**
http://ad1.example.net/
.
.
.
http://188.255.xx.yy//?NDA1MDIw&oHW
hfaJsIUlnP&ZPtcnBf …

ニューラルネットワークを用いて学習＆予測



URL
リクエスト数
セッション継続時間
閲覧サイトカテゴリ比率
曜日／時間
・・・

同一セッション内での
悪性サイトアクセスを予測

⇒誤検知1%で数秒～数十秒後の悪性サイトアクセスを96%の精度で予測

<引用>国際会議CCS2018にて成果発表
https://dl.acm.org/doi/pdf/10.1145/3243734.3243779

（１）AIへの攻撃に対する対策の深化とそれ以外の論点への対応

- セキュリティの原則等を実用に資するものとするため、技術的には(c)AIに対する攻撃の分類等を踏まえ、可能な限り攻撃を制限すること、攻撃者が誰かを見極めて対策を強化することが重要。
- また、AIに対する攻撃だけでなく、AIに関連するそれ以外の論点も考慮しながら議論を進めることが重要。特に(b)AIによる攻撃は詳細な検討が進んでいないが、AI開発・利活用ガイドライン検討時に自律的に動作するAIやAGIについても考慮したのと同様に、リスクの１つとして捉えておく必要。

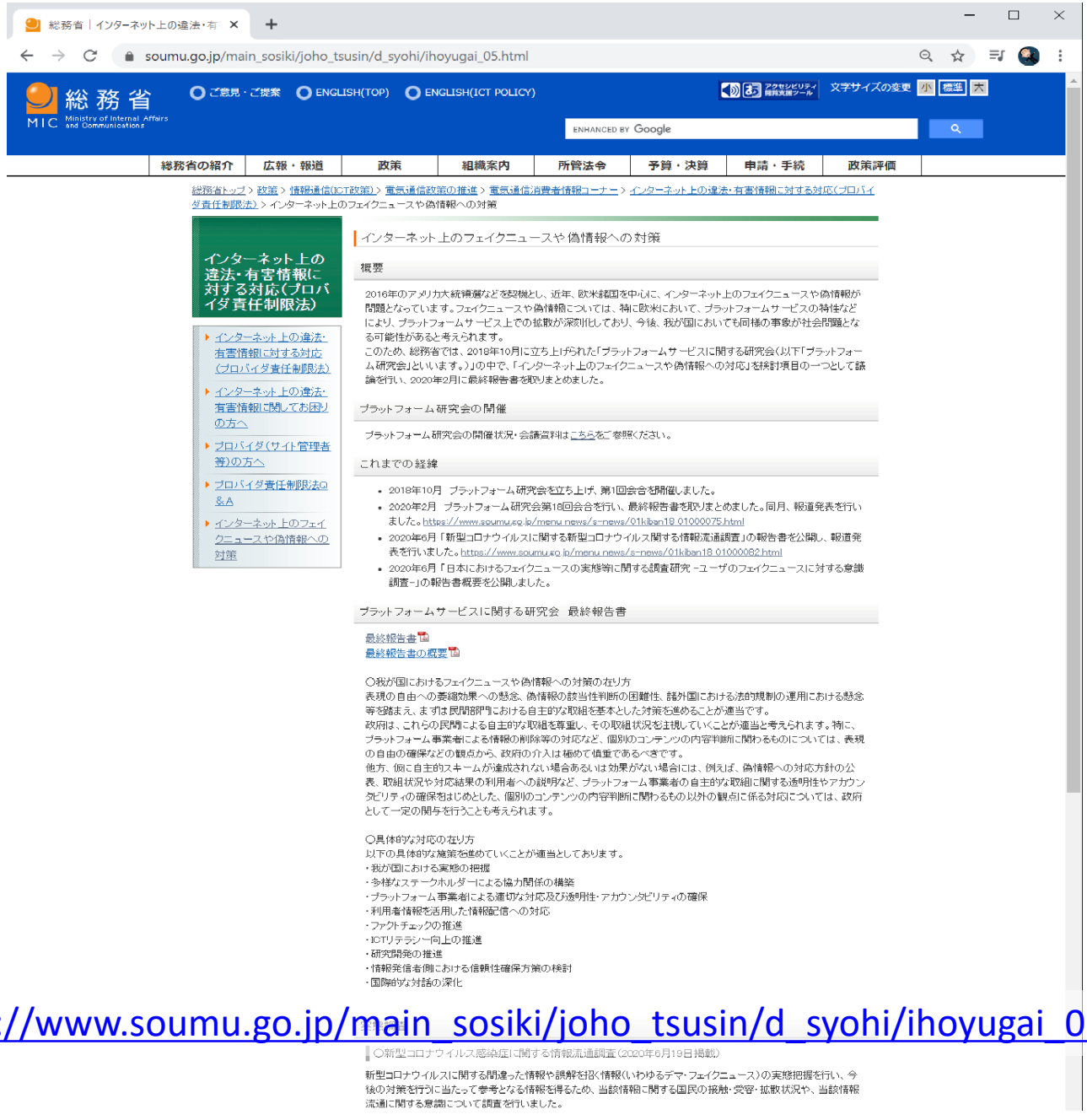
（２）攻撃における意図の見極めの必要性

- (c)AIに対する攻撃、(a)AIを使った攻撃の双方について、何をもって不正・悪であるかを見極めること（攻撃していると判断すること）が困難であるため、その意図をどのように見抜いていくかが重要。
- 上記参考として、総務省・プラットフォームサービスに関する研究会では同最終報告書の中で、偽情報（何らかの意図性を持った虚偽の情報）への対応の在り方として様々なレベル感があるとして、ファクトチェックの推進 やICTリテラシー向上の推進など複数の論点について紹介。

（３）マルチステークホルダによる学際的な議論の必要性

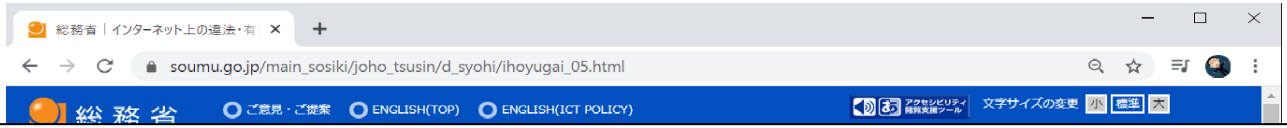
- 本分野については技術面だけでは解決できない問題が含まれるため、（セキュリティ）技術者だけの議論にとどめず、心理学・社会学等の知見も交えながら、学際的な議論を継続的に行っていくことが重要。

総務省・インターネット上のフェイクニュースや偽情報への対策



https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/ihoyugai_05.html

総務省・インターネット上のフェイクニュースや偽情報への対策



○新型コロナウイルス感染症に関する情報流通調査（2020年6月19日掲載）

新型コロナウイルスに関する間違った情報や誤解を招く情報（いわゆるデマ・フェイクニュース）の実態把握を行い、今後の対策を行うに当たって参考となる情報を得るため、当該情報に関する国民の接触・受容・拡散状況や、当該情報流通に関する意識について調査を行いました。

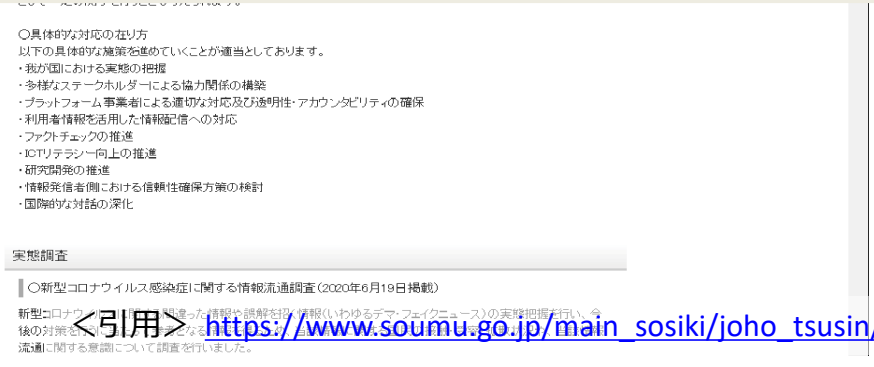
【報告書の要約】

＜調査結果＞

・新型コロナウイルスに関する間違った情報や誤解を招く情報について、一つでも見たり聞いたりしたと答えた人の割合は72%。

・新型コロナウイルスに関する間違った情報や誤解を招く情報を見聞きした人のうち、その場合に共有・拡散したことがあると答えた人の割合は35.5%（すべての人を母数とした場合の共有・拡散経験の割合は19.5%）。

・新型コロナウイルス感染症に関する間違った情報や誤解を招く情報があたかも真実又は真偽不明の情報として書かれているのを見かけたことがあると答えた人は、サービス・メディア別にみると、「Twitter」（57.0%）、「ブログやまとめサイト」（36.5%）で見かけたことがある人の割合が高かった。

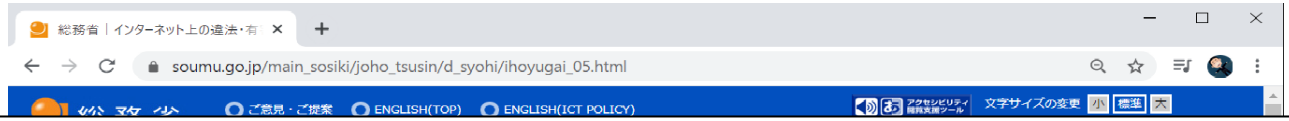

○具体的な対応の在り方
以下の具体的な施策を進めていくことが適当としております。
・我が国における実態の把握
・多様なステークホルダーによる協力関係の構築
・プラットフォーム事業者による適切な対応及び透明性・アカウンタビリティの確保
・利用者情報を活用した情報配信への対応
・ファクトチェックの推進
・ICTリテラシー向上の推進
・研究開発の推進
・情報発信者側における信頼性確保方策の検討
・国際的な対話の深化

実態調査

○新型コロナウイルス感染症に関する情報流通調査（2020年6月19日掲載）

<引用> https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/ihoyugai_05.html

総務省・インターネット上のフェイクニュースや偽情報への対策



○日本におけるフェイクニュースの実態等に関する調査研究
－ユーザのフェイクニュースに対する意識調査－（2020年6月19日掲載）

日本におけるフェイクニュースの実態について把握することを目的として、インターネット利用者が、フェイクニュースにどの程度接触しており、拡散経験があるか等の調査を行いました。

【報告書の要約】

・SNSやブログなどでフェイクニュースをみかけた頻度が「週1回以上」となったのは全体の3割。若い年代ほどフェイクニュースへの接触度は高かった。

・フェイクニュースの拡散経験について、全体では、「拡散したことはない」が最も高く（約7割）、年代が上がるほど「拡散したことがない」と回答した人の割合が高い傾向が見られた。一方、全体の約15%が「拡散した経験がある」と回答。若い年代ほど「拡散した経験がある」と回答した割合が高い傾向が見られた。

・フェイクニュース対策に取組むべき組織として「報道機関、放送局、ジャーナリスト」への期待が高く、個人のリテラシーの向上が必要であるという意見も強かった。リテラシー向上のために参加したい取組としては、「学校職場での授業や研修の実施」や「フェイクニュース対策を学べるテレビ番組の視聴」の希望が高かった。

・その他、フェイクニュース拡散経験者の特徴としては、
　・インターネットの利用に関する教育を受けたことがある人
　・フェイクニュースを見分ける自信がある人
　・メディアで見た情報が「怪しい」と思った場合に調べる頻度が高い人
　といった属性の人はフェイクニュースの拡散経験が多い傾向が見られた。

<引用> https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/ihoyugai_05.html