討論会

人工知能（AI）の普及と
安全・安心なシステムの構築

講演の内容

01. AI（人工知能）と安全

02. 課題と解決に向けて

03. 視点と論点（開発者・利用者）

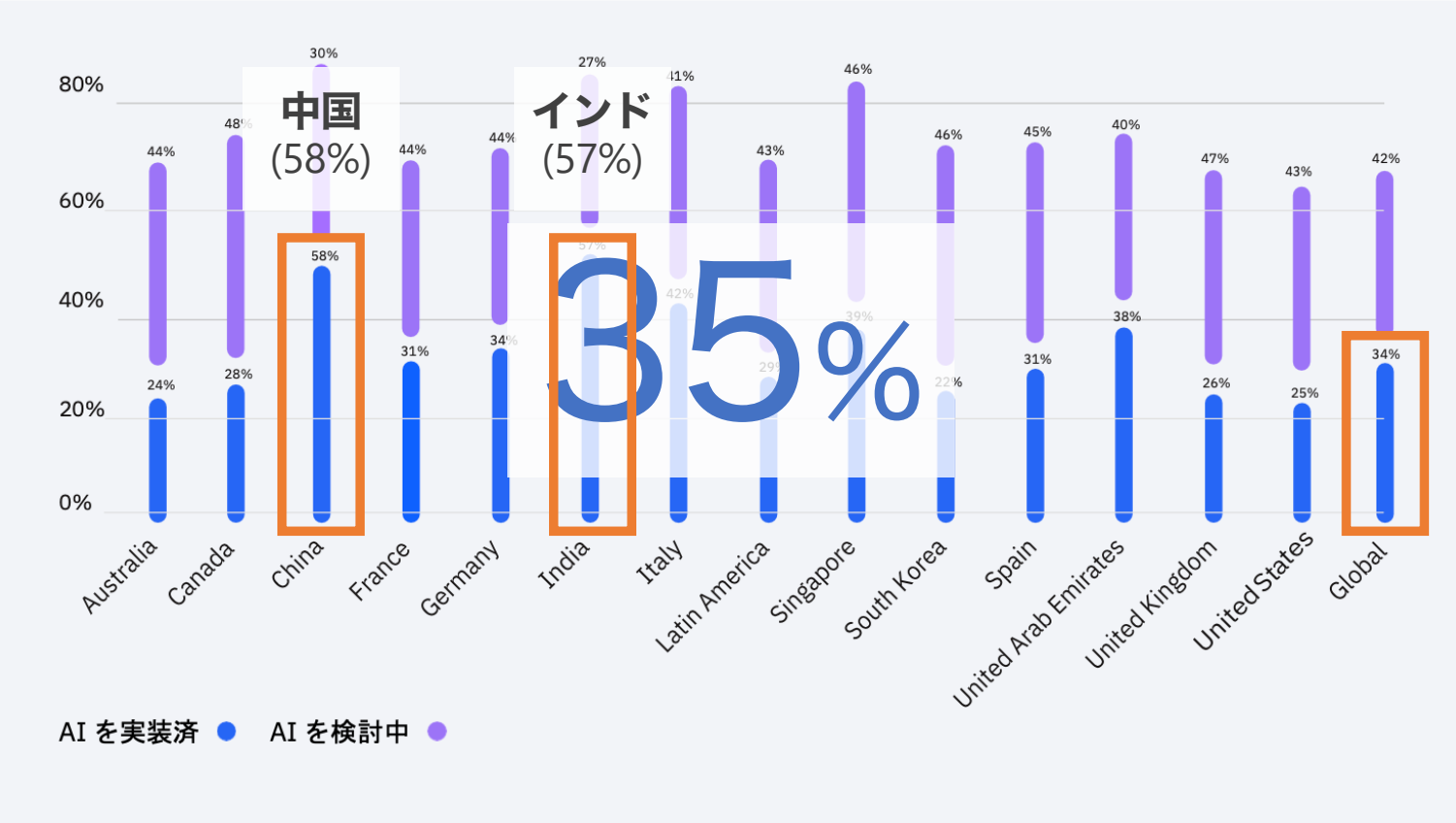
講演の内容

01. AI（人工知能）と安全

02. 課題と解決に向けて

03. 視点と論点（開発者・利用者）

世界のAI導入率（2022年）



出典｜ https://Global AI Adoption Index 2022

AIとは？

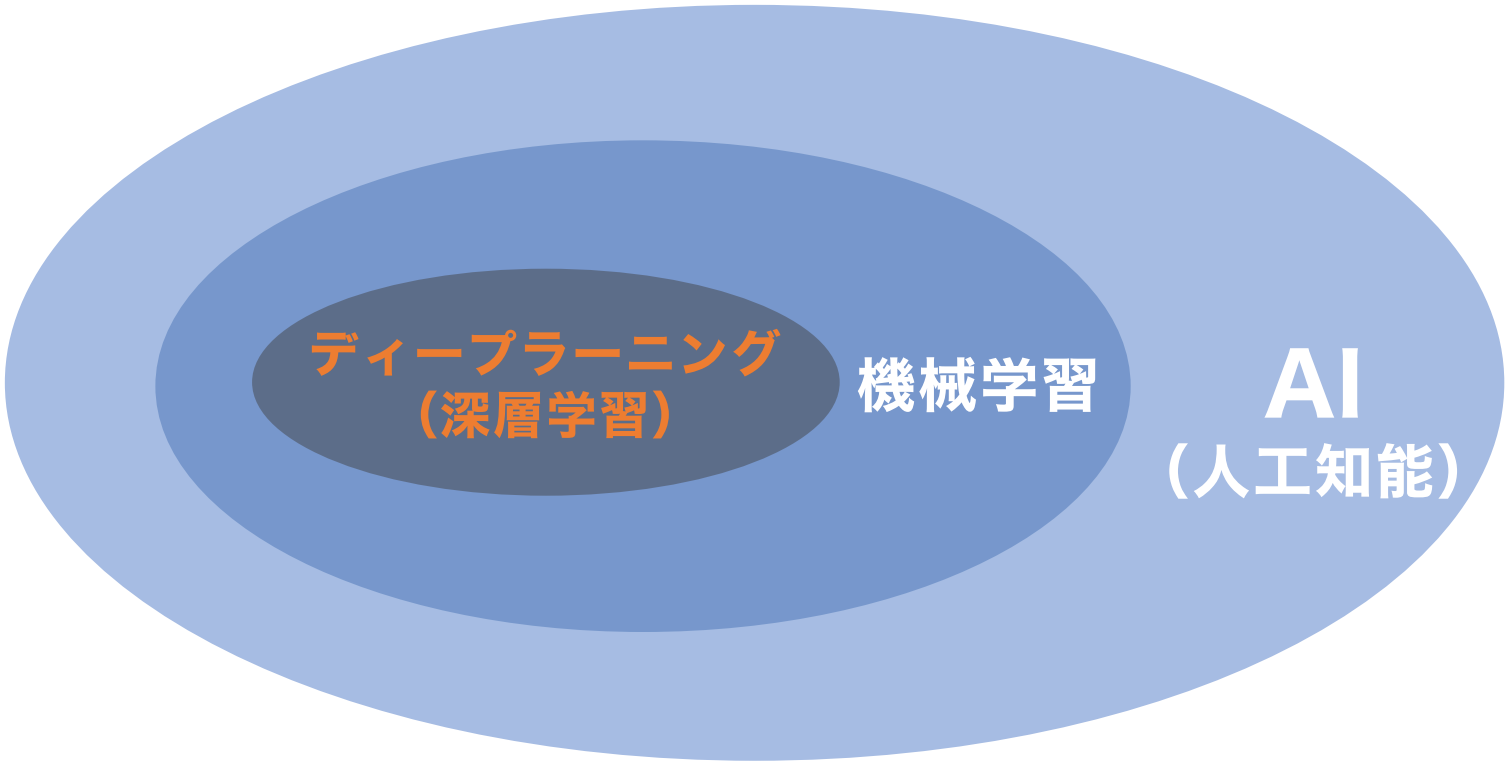
“ 人工的につくられた人間のような知能,
ないしはそれをつくる技術 ”

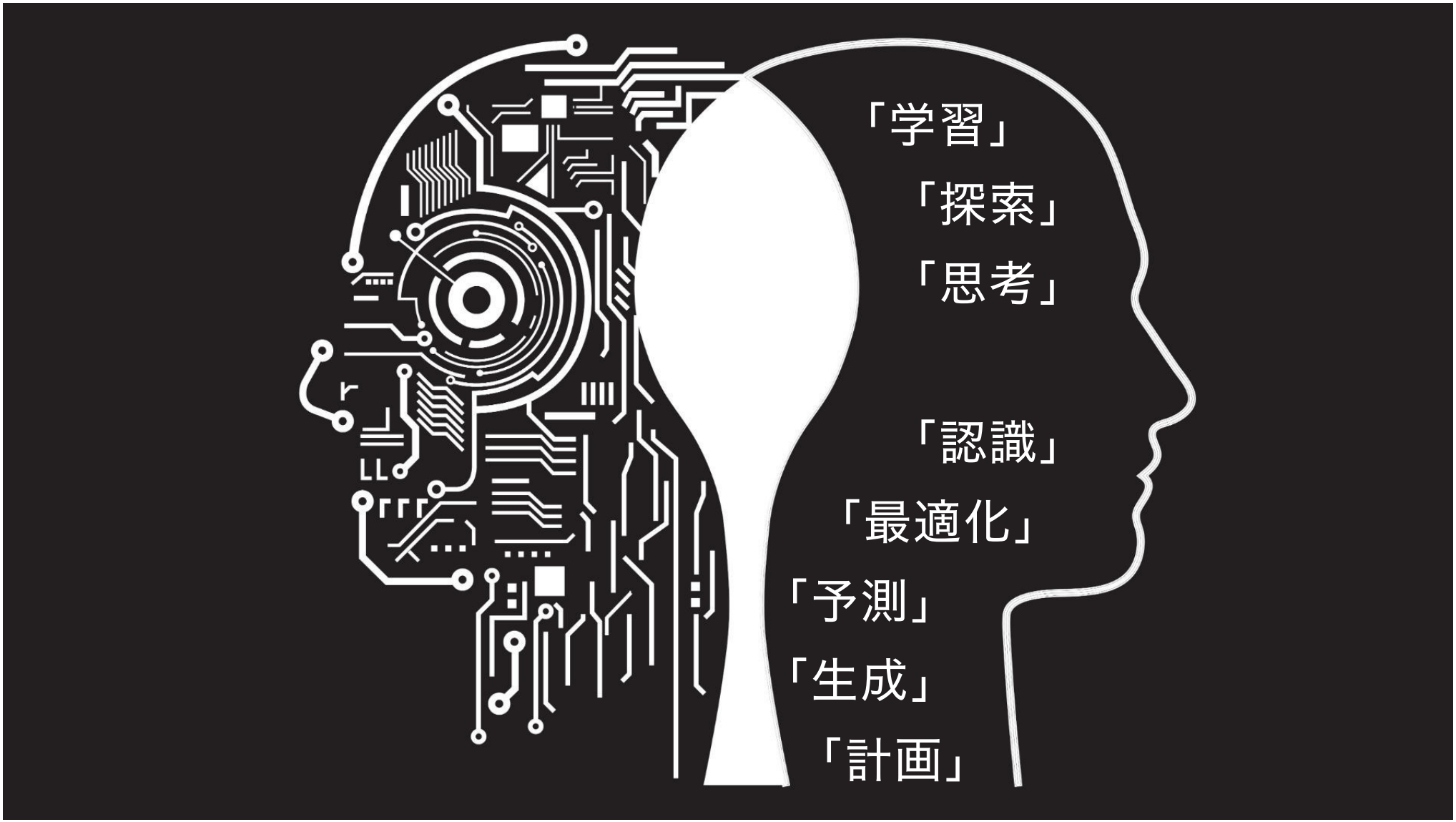
（松尾豊：東京大学）

“ 人間の思考プロセスと同じよう な形で動作するコンピュータプログラムや，コンピュータ上で知的判断を下せるシステム等 ”

（AI事業者ガイドライン案）

AI，機械学習，ディープラーニング（深層学習）




「学習」
「探索」
「思考」
「認識」
「最適化」
「予測」
「生成」
「計画」

AIが出力結果に基づき判断（制御）する

認　識

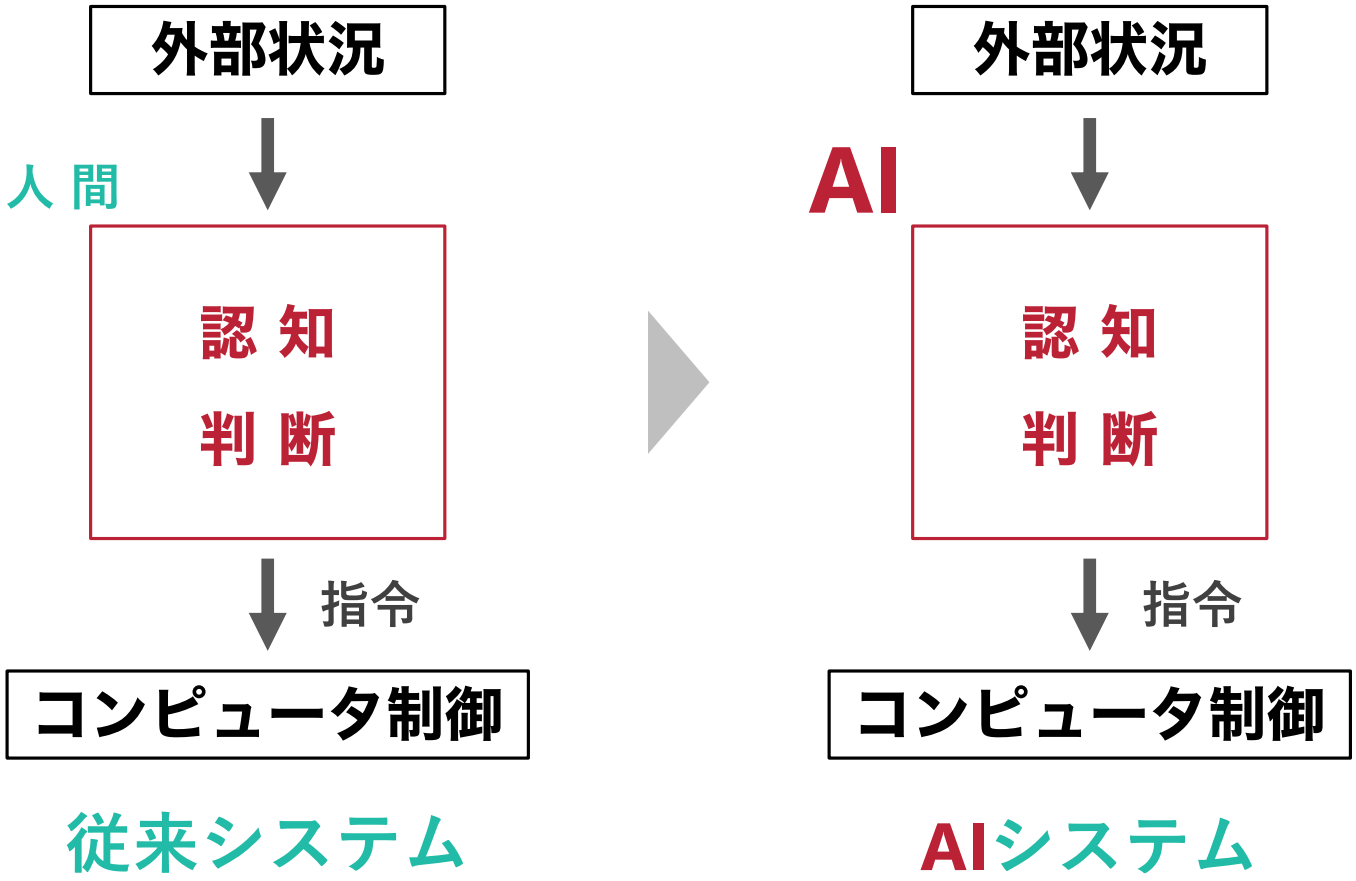
予　測

提　案

決　定

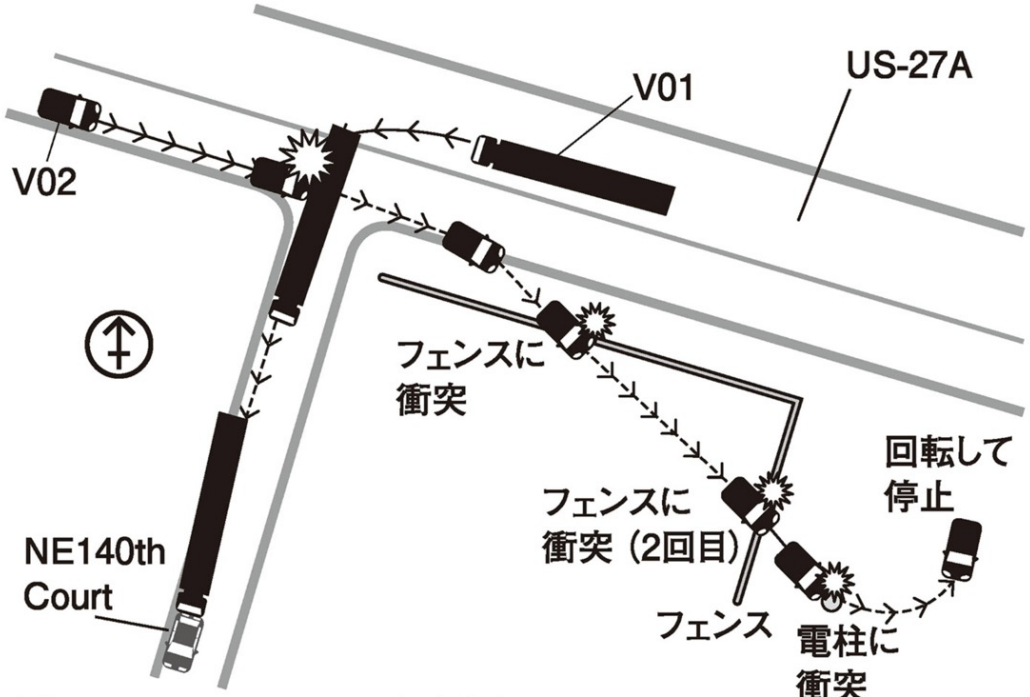
生　成

従来システムとAIシステム



テスラ社の事故①（2016年5月）

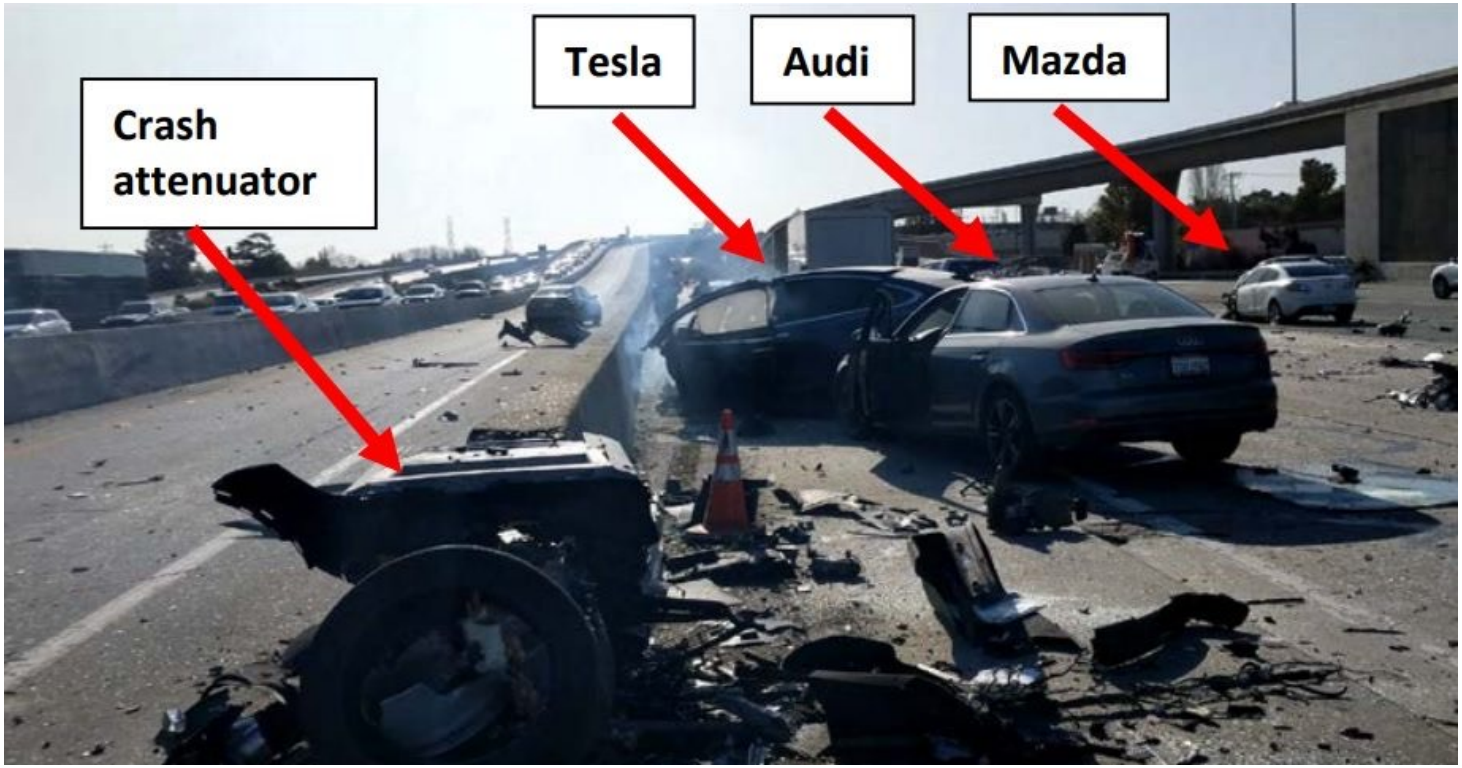
図1　テスラの自動運転車が引き起こした衝突（死亡）事故の様子



出典：Florida Highway Patrol／一部改変

出典｜小林雅一，AIが人間を殺す日，集英社，2017

テスラ社の事故 ②（2018年3月）



出典｜ https://jidounten-lab.com/u_tesla-cause-semi-autonomous-ntsb

講演の内容

01. AI（人工知能）と安全

02. 課題と解決に向けて

03. 視点と論点（開発者・利用者）

AIの課題：安全性が保証されない

判断が予測不可能

（決定論的・非決定論的AI）

判断を間違える（信頼性）

（不完全知覚問題）

判断の根拠が不明

（ブラックボックス問題）

判断を間違える（不完全知覚問題）

リス

出典｜ https://arxiv.org/pdf/1907.07174.pdf

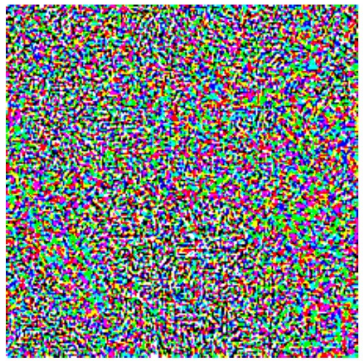
判断を間違える（敵対的攻撃）

元の画像

パンダ

ノイズを乗せる

$+ .007 \times$



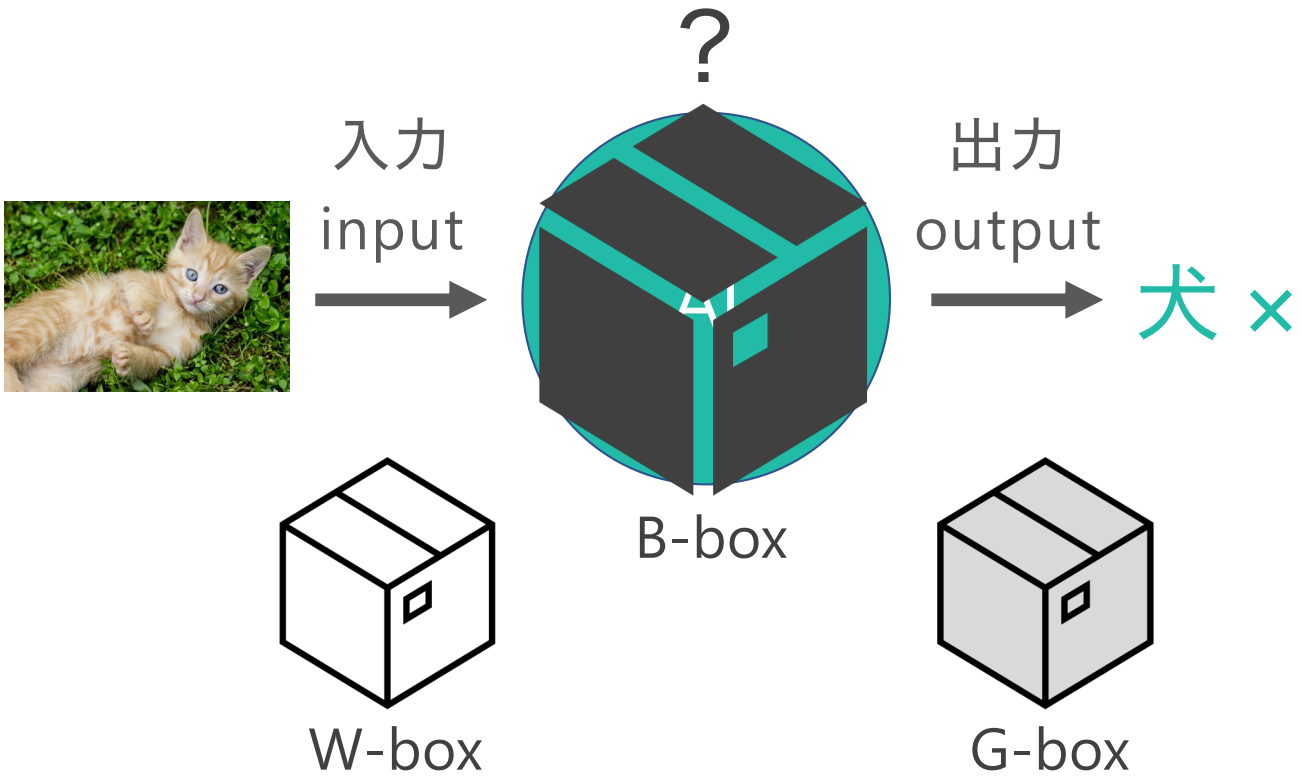
ノイズが乗った画像

$=$

テナガザル？

出典｜ https://arxiv.org/pdf/1412.6572.pdf

判断理由がわからない（ブラックボックス問題）



解決に向けて（技術的）

STAMP/STPA

FRAM（機能共鳴分析手法）

説明可能AI
(XAI：eXplainable AI)

B-box・G-box・W-boxの使い分け

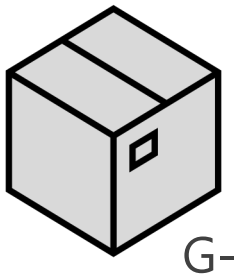
B-box・G-box・W-boxの使い分け

B-box

安全に関与しない用途

→ e.g. 利用者の混雑率のモニタ

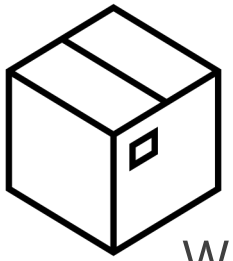
次列車の到着時刻の予測



G-box

直接は安全に関与しない用途

→ e.g. シミュレーションへの活用



W-box

安全に関与する用途

→ e.g. 車両の監視や制御

解決に向けて（ガイドライン）

人工知能搭載システムの安全設計ガイドライン

SEAMS Project（2020年）

エッセンシャル版｜https://www.seams-p.jp/data/SEAMS_Guideline_essential_20200403.pdf

AI事業者ガイドライン（案）

総務省 経済産業省（2024年1月）

本編｜https://www.soumu.go.jp/main_content/000923348.pdf

解決に向けて（ガイドライン）

人工知能搭載システムの安全設計ガイドライン

SEAMS Project（2020年）

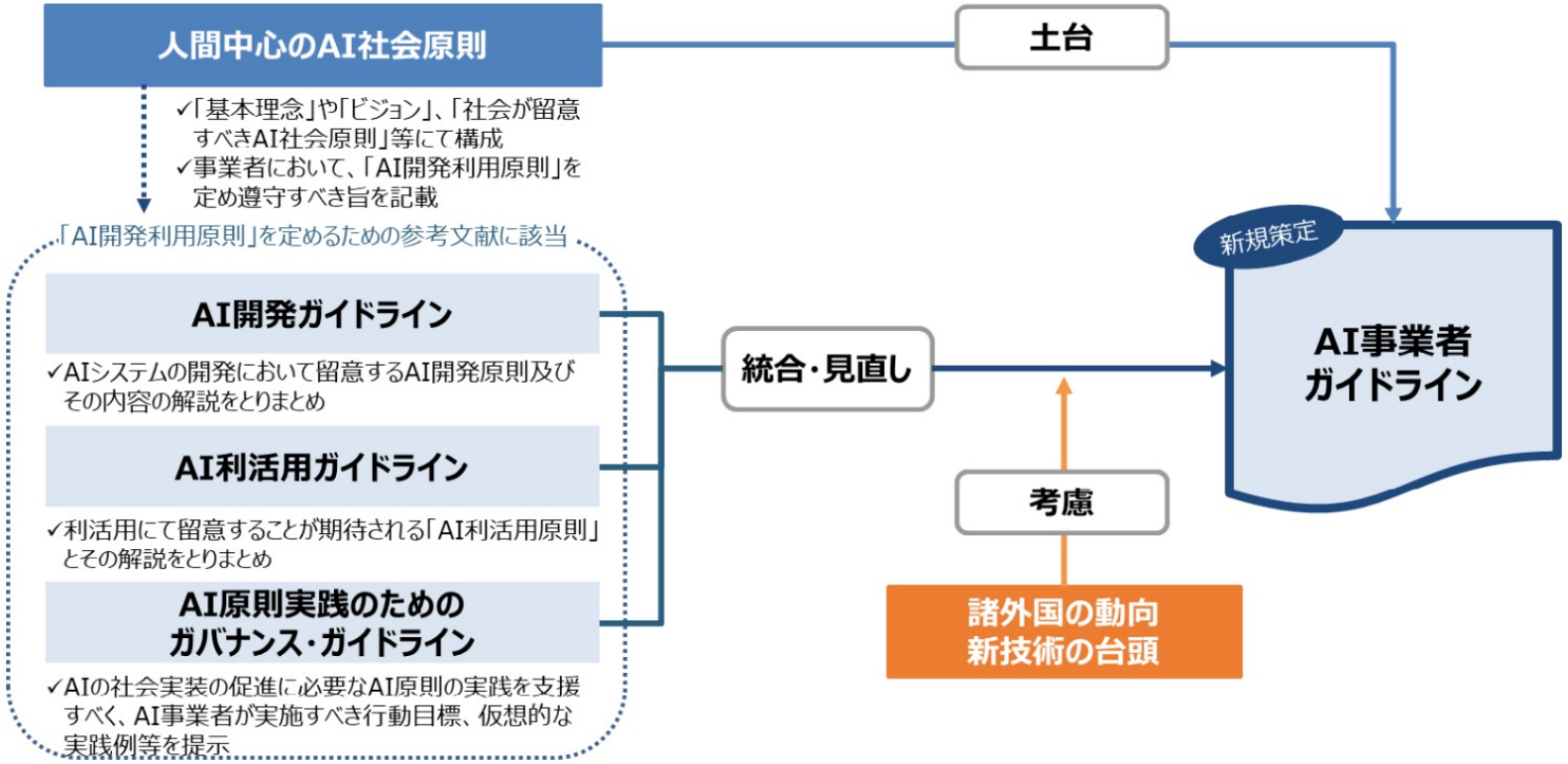
エッセンシャル版｜https://www.seams-p.jp/data/SEAMS_Guideline_essential_20200403.pdf

AI事業者ガイドライン（案）

総務省 経済産業省（2024年1月）

本編｜https://www.soumu.go.jp/main_content/000923348.pdf

AI事業者ガイドラインの位置づけ



出典｜https://www.soumu.go.jp/main_content/000923348.pdf

AI事業者ガイドラインにおける本編・別添の構成

出典｜https://www.soumu.go.jp/main_content/000923348.pdf

AIの事業活動を担う主体

- **AI開発者（AI Developer）**

  AIシステムを開発する事業者（AIを研究開発する事業者を含む）

  AIモデル・アルゴリズムの開発，データ収集（購入を含む），前処理，AIモデル学習，検証を通してAIモデルおよび AIモデルのシステム基盤や入出力等を含む AIシステムを構築する役割を担う.

AIの事業活動を担う主体

- **AI提供者（AI Provider）**

  AIシステムをアプリケーションや製品もしくは既存のシステムやビジネスプロセス等に組み込んだサービスとして AI利用者（AIBusiness User），場合によっては業務外利用者に提供する事業者

  AIシステム検証，AIシステムの他システムとの連携の実装，AIシステム・サービスの提供，正常稼働のための AIシステムにおける AI利用者（AIBusiness User）側の運用サポートや AIサービスの運用自体を担う．AIサービスの提供に伴い，様々なステークホルダーとのコミュニケーションが求められることもある．

AIの事業活動を担う主体

- **AI利用者（AI Business User）**
  事業活動において，AIシステム又は AIサービスを利用する事業者

  AI提供者が意図している適正な利用を行い，環境変化等の情報を AI提供者と共有し正常稼働を継続することや，必要に応じて提供された AIシステムを運用する役割を担う．また，AIの活用において業務外利用者に何らかの影響が考えられる場合は，当該者に対する AIによる意図しない不利益の回避，AIによる便益最大化の実現に努める役割を担う．

各主体が取り組む事項（安全性）

2) 安全性

各主体は、AI システム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすべきである。加えて、精神及び環境に危害を及ぼすことがないようにすることが重要である。

① 人間の生命・身体・財産、精神及び環境への配慮

- ✧ AI の出力の正確性を含め、要求に対して十分に動作している（信頼性）
- ✧ 様々な状況下でパフォーマンスレベルを維持し、無関係な事象に対して著しく誤った判断を発生させないようにする（堅牢性（robustness））
- ✧ AI の活用や意図しない AI の動作によって生じうる権利侵害の重大性、侵害発生の可能性等、当該 AI の性質・用途等に照らし、必要に応じて客観的なモニタリングや対処も含めて人間がコントロールできる制御可能性を確保する
- ✧ 適切なリスク分析を実施し、リスクへの対策（回避、低減、移転、容認）を講じる
- ✧ 人間の生命・身体・財産、精神及び環境へ危害を及ぼす可能性がある場合は、講ずべき措置について事前に整理し、ステークホルダーに関連する情報を提供する
  - ● 関連するステークホルダーが講ずべき措置及び利用規則を明記する
- ✧ AI システム・サービスの安全性を損なう事態が生じた場合の対処方法を検討し、当該事態が生じた場合に速やかに実施できるよう整える

出典｜https://www.soumu.go.jp/main_content/000923348.pdf

各主体が取り組む事項（安全性）

2) 安全性

各主体は、AI システム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすべきである。加えて、精神及び環境に危害を及ぼすことがないようにすることが重要である。

② 適正利用

- ✧ 主体のコントロールが及ぶ範囲で本来の目的を逸脱した提供・利用により危害が発生することを避けるべく、AI システム・サービスの開発・提供・利用を行う

③ 適正学習[13]

- ✧ AI システム・サービスの特性や用途を踏まえ、学習等に用いるデータの正確性・必要な場合には最新性（データが適切であること）等を確保する
- ✧ 学習等に用いるデータの透明性の確保や法的枠組みの遵守、AI モデルの更新等を合理的な範囲で適切に実施する

出典｜https://www.soumu.go.jp/main_content/000923348.pdf

各主体が取り組む事項（安全性）

表 1.「共通の指針」に加えて主体毎に重要となる事項

	第 2 部. C.共通の指針	「共通の指針」に加えて主体毎に重要となる事項		
		第 3 部. AI 開発者 (D)	第 4 部. AI 提供者 (P)	第 5 部. AI 利用者 (U)
1) 人間中心	① 人間の尊厳と個人の自律 ② AI による意思決定・感情の操作等への留意 ③ 偽情報等への対策 ④ 多様性・包摂性の確保 ⑤ 利用者支援 ⑥ 持続可能性の確保	-	-	-
2) 安全性	① 人間の生命・身体・財産、精神及び環境への配慮 ② 適正利用 ③ 適正学習	ⅰ. 適切なデータの学習 ⅱ. 人間の生命・身体・財産、精神及び環境に配慮した開発 ⅲ. 適正利用に資する開発	i. 人間の生命・身体・財産、精神及び環境に配慮したリスク対策 ii. 適正利用に資する提供	i. 安全を考慮した適正利用

出典｜https://www.soumu.go.jp/main_content/000923348.pdf

講演の内容

01. AI（人工知能）と安全

02. 課題と解決に向けて

03. 視点と論点（開発者・利用者）

視点と論点：**AI開発者**の視点

- **論点１**：技術者はどのようにしてAIを利用して安全性を確保するか？
- **論点２**：安全なAIシステムの開発において倫理的な課題にどのように対処するべきか？
- **論点３**：技術者が直面するセキュリティの課題とその解決策は？
- **論点４**：技術者がAIの透明性を確保し，利用者に理解可能な形で説明できる手段は？
- **論点５**：AIのリスクを適切に管理するための技術的な手法やベストプラクティスは？

視点と論点：AI利用者の視点

- **論点１**：利用者はAIを使ったシステムの安全性に対してどのような期待を持っているか？
- **論点２**：利用者が感じるセキュリティやプライバシーに関する懸念は？
- **論点３**：利用者がAIの意思決定を理解しやすくするためにはどのような手段が効果的か？
- **論点４**：利用者が直面する技術的なリスクに対する認識や期待の向上にはどうすれば良いか？
- **論点５**：利用者がセキュリティに対して積極的に参加し，協力するための手段は？

ディスカッション