

「AI技術は社会技術システムの安全性向上に寄与するか？」

東北大学大学院工学研究科
技術社会システム専攻

高橋 信

社会技術システムにおける安全性の向上に関する基盤技術の研究

航空システムの安全性向上



航空管制システム

- 航空管制業務のタスク分析
- コミュニケーションエラー防止のための訓練手法
- ラインオペレーションにおける良好事例の分析

安全を支える熟練技能の
解明と応用

高度自動化システムと人間の協調

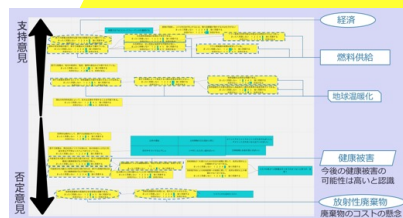


- ドライビングシミュレータ(DS)の構築
- DSによる自動運転解除時の運転挙動の解析
- 生理指標に基づく運転時でのドライバーの感情推定

状況適応的システムの開発

大規模システムの 安全性向上

原子力発電に関する主観的 論理構造の分析

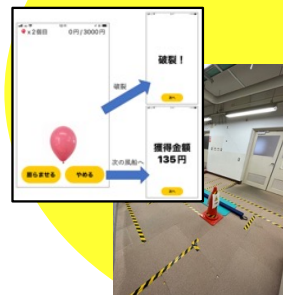


オンラインボードによる可視化

- アンケートと構造化インタビューにより意見を抽出
- 原子力に対する態度の背景にある論理構想を可視化
- 日米の比較も実施予定

対話の場の構築に向けて

安全性向上のための危険感受性・ 敢行性の評価に関する研究



- 単純タスクにより危険感受性/敢行性を評価
- 実タスクにおける「行動指標」との関連付け

安全教育に対する適用

AI技術の現状-AIとAI技術の違い-

Artificial Intelligence: 人工知能

- 人間の脳の仕組みを解明し知能が実現される原理を数学的にあきらかにして工学的に実現する方法
- 脳科学の分野ではfMRI等により測定が進歩
- 特定の認知活動と脳の部位を関連づける機能局在に関する理解は大きく進歩
- 知能そのものに関する理解は道半ば



AI技術

- 知能の基本的な原理の解明は後回しにして、工学的に人間の知能を模倣する仕組みを作るという方法
- 全て統計的手法に基づいて「**知能があるように振る舞う**」仕組みを実現
- ChatGPTもDeepLの基本的にはAI技術

高度なディープラーニングの手法を使おうが、単純な統計的な判断ロジックの手法であろうが、結果としてそのシステムが提供する判断の内容が「人間的・知的」なものであれば、それはAI技術

AI技術の分類

<強いAI>

- ✓ 人間としての全体的な知能を模擬する
- ✓ 近い将来にAIが自意識を持ち人間の能力を遥かに凌駕するという主張(シンギュラリティ仮説)のベースとなる概念



<弱いAI>

- ✓ 特定の問題領域に対して知的な判断を実現
- ✓ 例：チェス、将棋、囲碁等のゲームを対象にしたAI技術



弱いAI技術の現状

- AI技術の成功例の多くは機械学習(Deep learning)とビッグデータの組み合わせ
- 社会における**特定の分野の問題の解決**に大きく貢献
- 人間にとって知能があるように振る舞い、役に立つ技術としてのAI技術は確実に進展

ゲーム分野：

- チェス：1997年 IBMのスーパーコンピュータデーブブルーが世界チャンピオンカスパロフを破る
- 将棋：Ponanzaは既にプロ棋士を上回る能力を有する
- 囲碁：2017年 AlphaGo (Google Deepmind)が李世石 (イ・セドル) を破る

ルールが明確に決まっている不確実性のない環境での意思決定はAI技術が最も得意する分野



弱いAI技術が現実的に役に立っている例

画像認識による自動会計システム

ものづくり日本大賞 優秀賞 GOOD DESIGN AWARD 2015 ものづくりデザイン賞 / BEST100 世界初! カメラで会計 BakeryScan

新着トピック 製品紹介 導入メリット 開発秘話 導入事例 メディア掲載 お問い合わせ 会社概要

製品紹介

使用方法 |

特徴

パンは焼き釜の位置によって焼き色に差がつくことがあります。ベーカリースキャンはこのような色の違いがあるパンも正確に認識します。



間違っているかもしれないパンは黄色の枠で表示されます。パンをタッチすると次の候補のパン画像が表示されます。正しいパンを選択すると、その結果はデータベースに反映され、識別精度が日々向上していきます。



オールインワンタイプには、ベーカリーショップのレジ業務に特化したPOSソフトが付属しています。バーコード商品の精算、新商品の追加も簡単に行えます。



仙台市薬師堂「ジャンヌダルク」

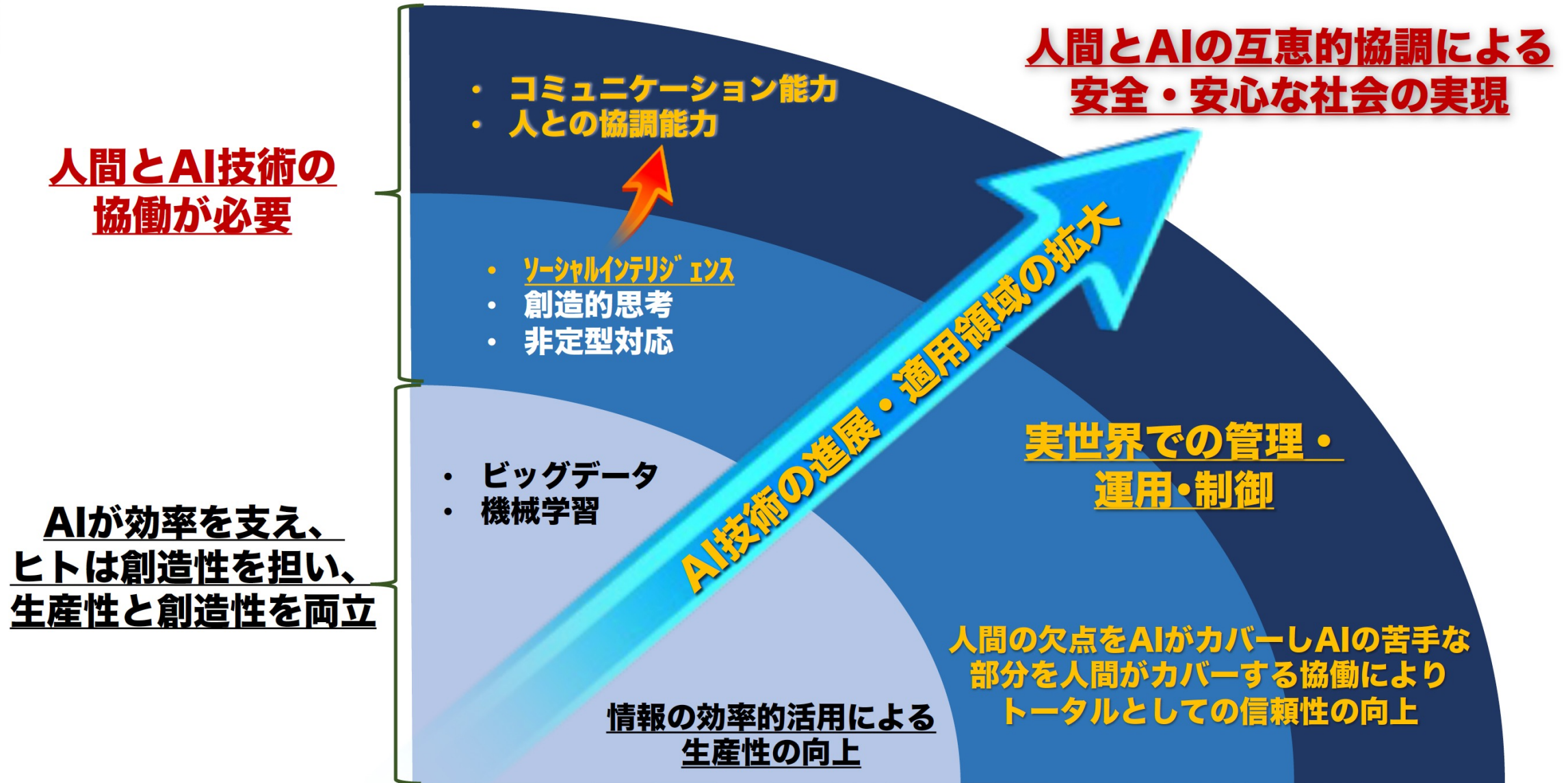
弱いAI技術の医療分野での現状

- 皮膚ガンの診断
 - 10万点以上の良性と悪性の腫瘍、ほくろの画像を診断結果とセットでCNNに与えて学習
 - 皮膚癌の見落としは皮膚科医より少なかった
 - 良性を悪性と「誤診」する割合もCNNのほうが少なかった
 - 皮膚科医が悪性黒色腫を正しく識別できた割合は86.6%
 - CNNの場合はこの割合が95%
- インフルエンザ診断でAIが小児科医を上回る正診率
- 脳出血を鋭敏に捉えるAIアルゴリズムが開発される
- AIを利用し非アルコール性脂肪肝炎を予測
- AIによる心臓病診断システムを臨床現場に導入
- 健康診断における胸部レントゲン読影や心電図解析などは近い将来、AIによって完全に代替される可能性



- ✓ 整理された信頼性の高いビッグデータが必要
- ✓ 画像認識、分類問題はAIというよりも統計解析の問題
- ✓ **データの蓄積により精度向上が実現出来る点が最も重要**
- ✓ 経験に依存しない、バイアスを排除した判断が可能

AIの適用領域の拡大



AIの応用分野

• 信頼性工学分野

- メンテナンスへの応用
- 膨大なセンサーデータを用いてエンジンの故障を予測
- コストの削減を実現

• 投資分野

- 株の売買における意思決定支援
- 人間のトレーダーと同等の能力を発揮
- 単純な統計的手法でも同等の結果
- AIが売り買いの判断をすることが株価の不安定性を増長

• 監視ツール

- バイオメトリクス（生体情報）からの人物同定
 - Gait Recognition(歩き方からの認識)
 - ボディランゲージと表情からの情動推定
- 工場や学校への適用、全体主義的国での利用
- 個人の同定、行動の同定、感情の同定

• 生成的(generative)な応用

- 芸術、音楽、文学
- 本物のように見える、聞こえる、感じる作品を生成
- AI美空ひばり→Deep Fakeへの悪用の可能性

ゴッホのスタイルを
学習して適用

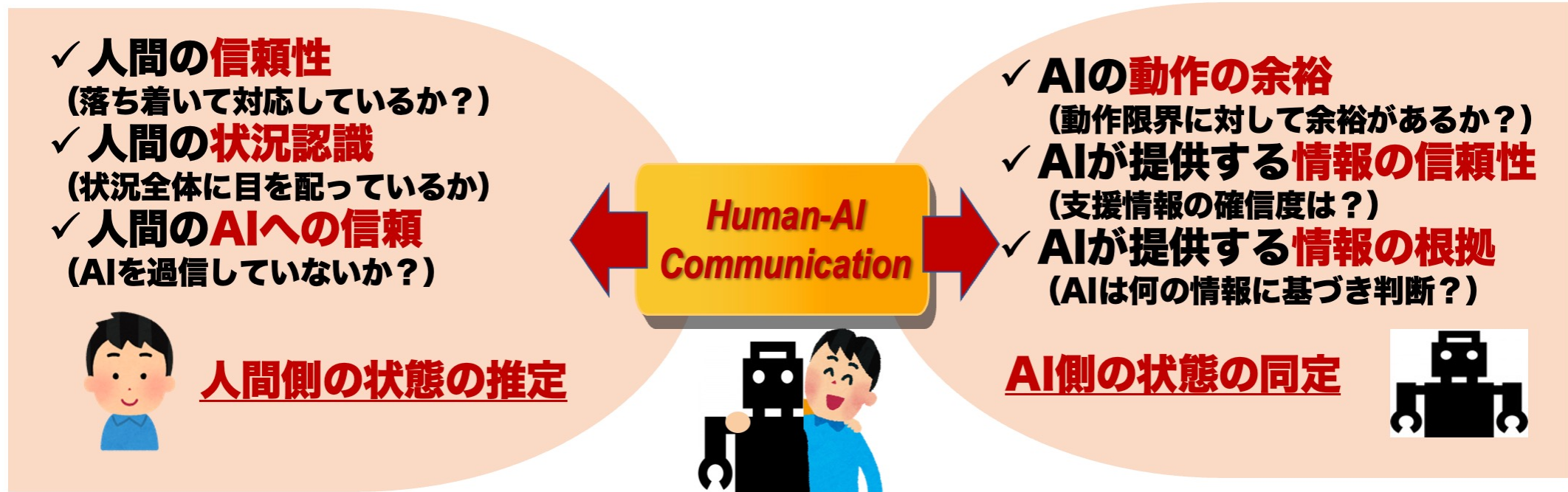


教育分野へのAI技術の適用

- ✓ 中学校における教育へのAI技術の導入に関して調査を実施
- ✓ 最も先進的にAI技術を導入している中学校を対象にインタビューと視察を実施
- ✓ 導入されているのは以下の教材
 - ✓ AI型教材
 - ✓ 自動採点ソフト
 - ✓ 授業動画配信サービス
- ✓ 教材（課題）の準備等における負担軽減には寄与
- ✓ 採点ソフトにより採点のかかわる教員の負担も軽減
- ✓ 生徒の理解度に基づくコーチング等は依然として教師の役割
- ✓ 導入に係わる資金の問題大きい



AI技術と人間の協調の問題



- AI技術側がブラックボックスであるために出力される判断の根拠が示されないために人間側がAIを信用しきれないという問題
 - AI技術側が何を見て（計測して）、どのような根拠で判断を下しているか、どの程度自信を持って判断を下しているかが、人間側にはほとんど伝わらない
- 人間側の情報もAI側にほとんど伝わっていない

AI技術と人間の間の特限の移譲

- 人間機械協調における中心的課題である自動化においては、これまで「人間中心の自動化：Human Centered Automation」がその中心的原則
- 最終的な特限は常に人間側にあり、いつでも人間が機械をオーバーライドできる
- 高度なAI技術をベースに進化している車の自動運転においても、現状のシステムは全て人間中心の自動化であり、最終的な特限、そして責任はドライバーにあることが大原則

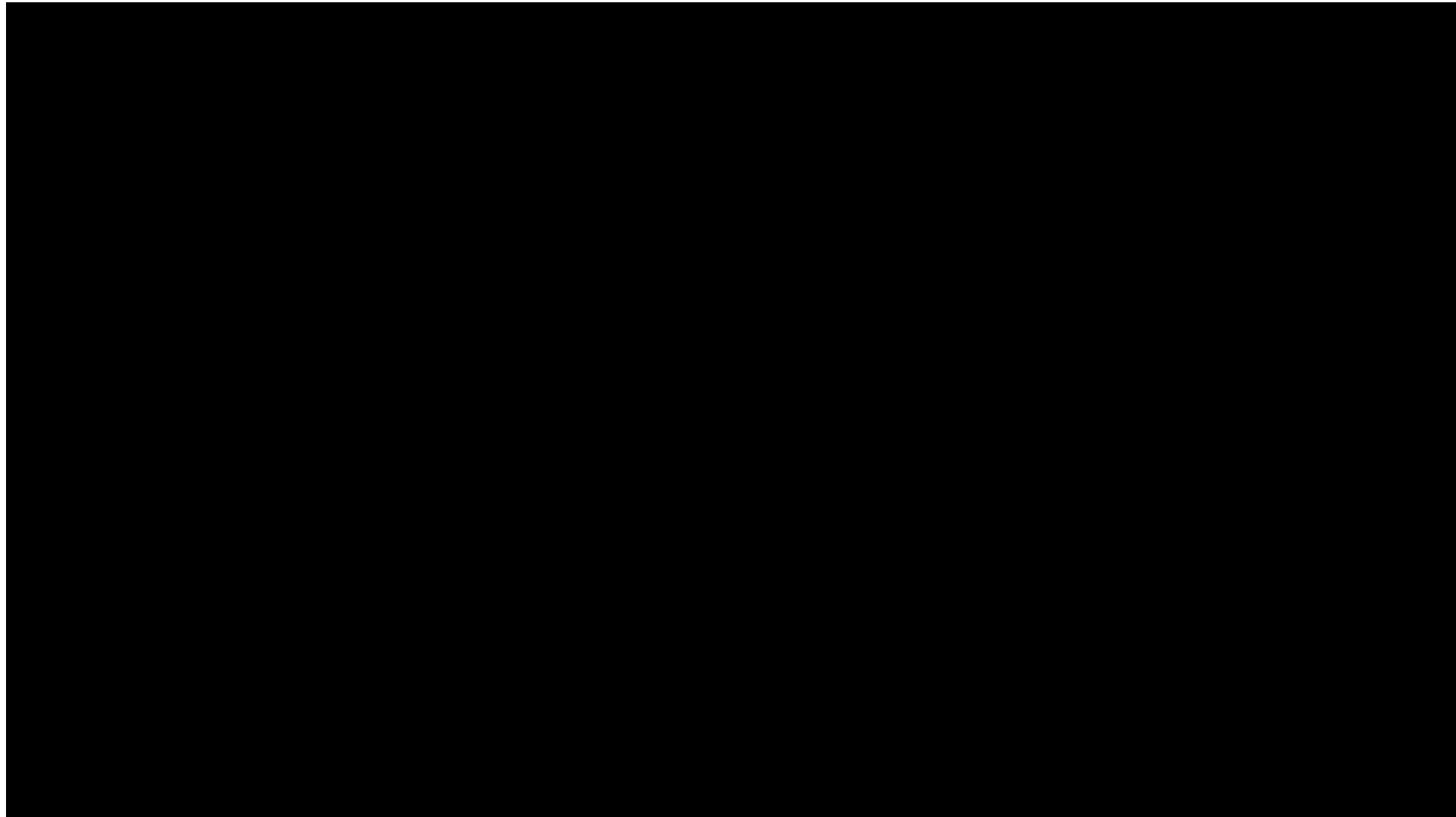
<AI技術主導での問題解決の困難さを示す例>

自動運転車「e-Palette」のパラリンピック選手村での事故（2021年8月）



オペレータの不適切な再スタートの判断を
AIは止めなかった
(オペレータは書類送検(2022年1月))

最先端の自動運転の動作



AI技術への過剰な期待 –AIが苦手なこと–

➤ AI技術の成功例の多くは機械学習(Deep Learning)とビッグデータの組み合わせ

➤ 学習データの量と質に大きく依存

➤ データに偏りがあれば学習結果も偏る (AIによる採用プロセスの問題)

➤ AIによる未来予測は全て過去のデータに依存

➤ 学習データが存在しないこれまで未経験の事象には対応困難

➤ 異常診断等におけるセンサー故障への対応は困難

➤ ブラックスワンの事象への対応は不可能

➤ 精度の高い学習データの収集コストの増大

➤ 整理され検証された症例データが必要

➤ 画像認識、分類問題はAIというよりも統計解析の問題

➤ データの蓄積により精度向上が実現出来る点が最も重要



AIの問題点をChatGPTに質問してみました

質問：“AIは基本的に過去のデータに基づいて判断をします。過去に未経験の事象に対してAIは無力なのではないでしょうか？”

おっしゃるとおり、従来のAIや機械学習モデルは学習に使用されるデータに強く依存しており、そのデータに含まれていない新しい事象や状況に対しては適切に反応できないことが多いです。具体的には、以下のような課題や限界が考えられます：

- 1.外挿問題:** AIが学習したデータの範囲外の事象に対しては適切に反応できないことがあります。
- 2.過去のバイアス:** AIは学習データに含まれるバイアスをそのまま学び取ってしまうことがあります。これは、不公平な判断や偏見を持つモデルを生み出す原因となることがある。
- 3.ブラックボックス問題:** 多くの高度なAIモデルは、どのようにして判断を下しているのかが人間には理解しにくい。これは、新しい事象や未知の状況での動作が予測しにくくなる原因となります。
- 4.一般化の難しさ:** 特定の状況やタスクに特化して学習されたモデルは、他の状況やタスクに対しては性能を発揮できないことがある。

しかし、以下のようなアプローチでこれらの問題を緩和する試みがなされています：

- 1.転移学習:** 一つのタスクや状況で学習されたモデルを、他のタスクや状況にも適用する技術。
- 2.アンサンブル学習:** 複数のモデルの予測を組み合わせることで、一般化性能を向上させるアプローチ。
- 3.アクティブラーニング:** モデルが自ら不確実性が高いデータを指摘し、それをラベル付けして再学習する方法。
- 4.ゼロショット学習、フューショット学習:** ほとんどまたは全くデータがないカテゴリに対しても予測を行う能力を持つモデルの学習方法。
- 5.シミュレーションと強化学習:** 実際の経験を必要とせず、仮想環境でのトライアンドエラーを通じて学習する方法。

結論として、現在のAIは確かに未経験の事象に対しては限定的な能力しか持っていませんが、技術の進化と共にその限界は徐々に広がっていると言えます。しかし、重要な決定を下す場面や高い安全性が求められる状況では、AIの判断だけに依存せず、人間の専門家の判断や監督が不可欠です。

➤ 人間の判断の非合理性の是正

確証バイアス

自分にとって都合のいい情報ばかりを無意識的に集めてしまい、反証する情報を無視したり集めようとしなかったりする傾向

正常性バイアス

多少の異常事態が起こっても、それを正常の範囲内としてとらえ、心を平静に保とうとする傾向

後知恵バイアス

物事が起きる前よりもその結果を知った後の方が、それが予測可能だったと考えてしまう心理的傾向

アンカリング

判断や予測が難しいときに最初に接した目安の数字や価格、特定の情報が判断に大きく影響する心理現象

内集団バイアス

自分が所属する集団（内集団）のメンバーに対して肯定的に評価したり好意的な態度を示したりする心理的傾向

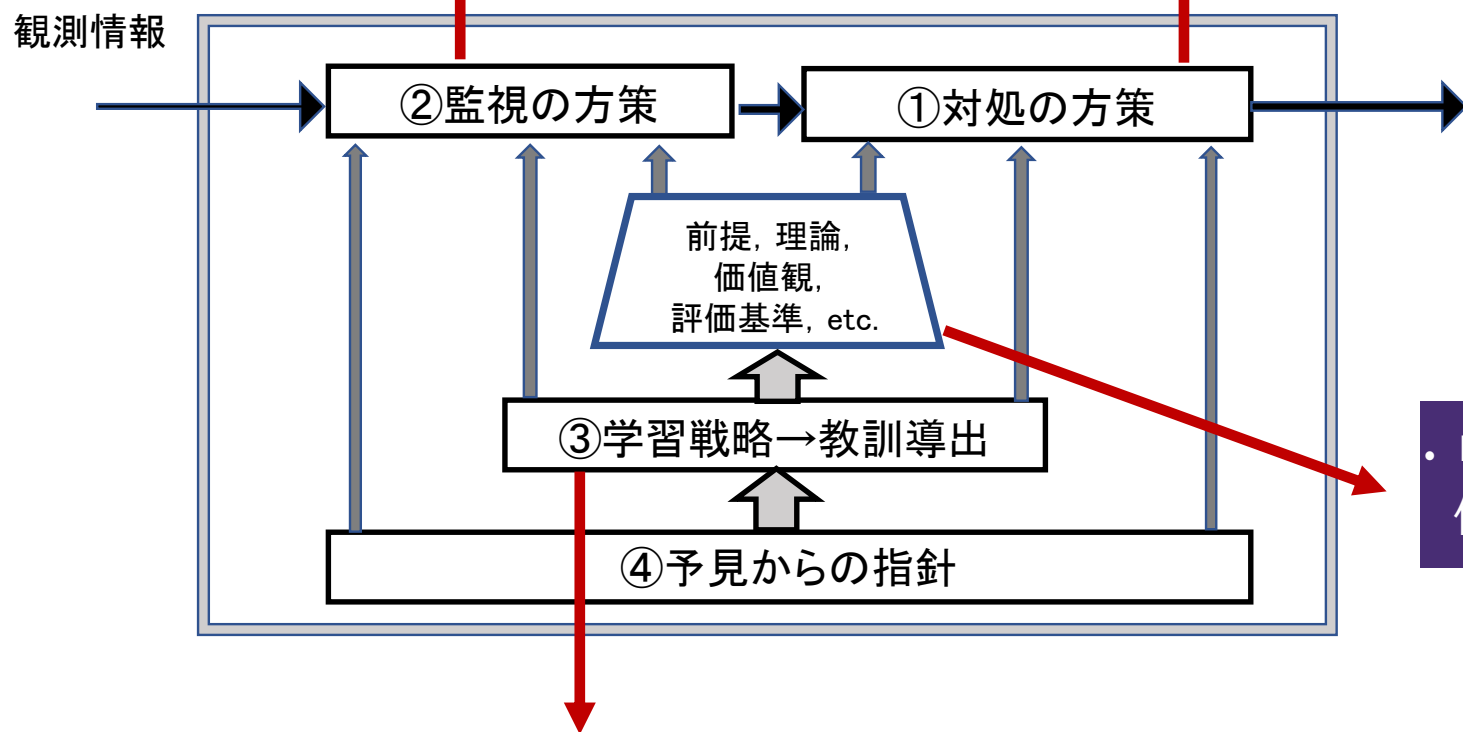
近接性バイアス

近い過去に起こったことが高い頻度で再び起きると感じる傾向

認知バイアスが組織のレジリエンスに与える影響(1)

- 正常性バイアスによる異常認識の遅れ
- 確証バイアスによる参照情報の偏り

- 対処方策は過去の経験から学習
- 優先順位付けに近接性バイアスの影響



- リスク/ベネフィットの認識の偏り

- 過去の事例の解釈におけるバイアス
- 後知恵バイアス/代表性ヒューリスティクス

「対処」に対する近接性バイアスの影響

2011.3.11直前の原子力を廻る情勢

- 「耐震裕度」問題
 - 耐震の問題に関しては2007年7月16日に発生した新潟県**中越沖地震**における東京電力の柏崎刈羽原子力発電所の被災状況の影響を受けて、耐震裕度の向上と再稼働に向けた議論が活発
 - 地震に対する「対処」の優先順位が**近接性バイアス**により上昇 →結果として他の可能性（津波）への対処が遅れることに



現在の原子力を廻る情勢は福島第一原子力発電所の事故の近接性バイアスの影響下にはないだろうか？

- 現在の「対処」方策の大部分は1F事故で起きたことへの対策
- 「津波」「地震」への対処への過度なリソース配分

「予見」への正常性バイアス/確証バイアスの影響

- 安全に対する取り組みを行い、一旦「安全」な状態が達成されると、組織は往々にしてその状態に安住して、その状態がそのまま維持されるように予見してしまう → 「将来も「安全」だろう」
- 外見的には安全な状態が維持されているように見える組織においても、内的な状況は常に変化していて、理想的な状況がそのまま維持されるということはありません

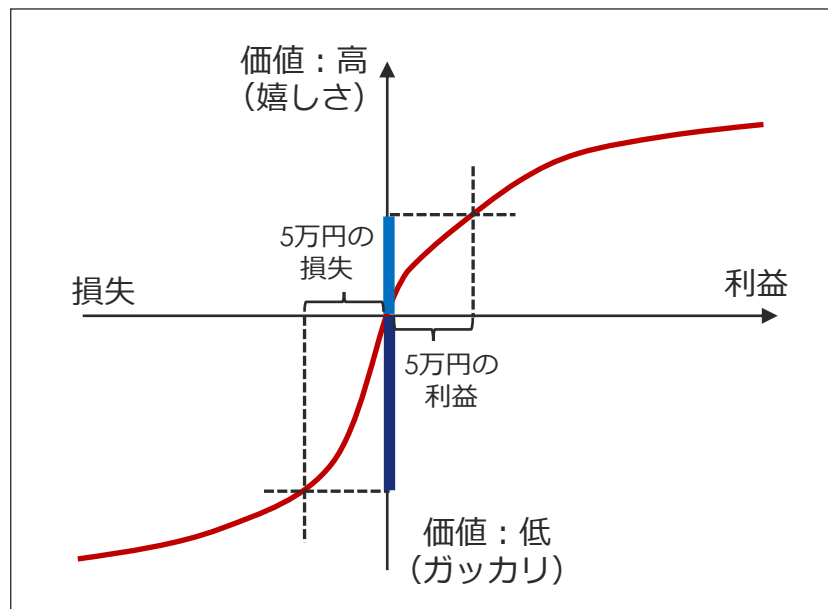


- **正常性バイアス**により組織は「正常」であると認識したい
- **確証バイアス**により「正常」であることを示す情報に目が行く

- 予想される利害額や確率などの条件によって、人間がどのように意思決定を行なうのかをモデル化
- 私たちの意思決定は、必ずしも合理的に行なわれているのではなく、感情や感覚による「ゆがみ」を伴う

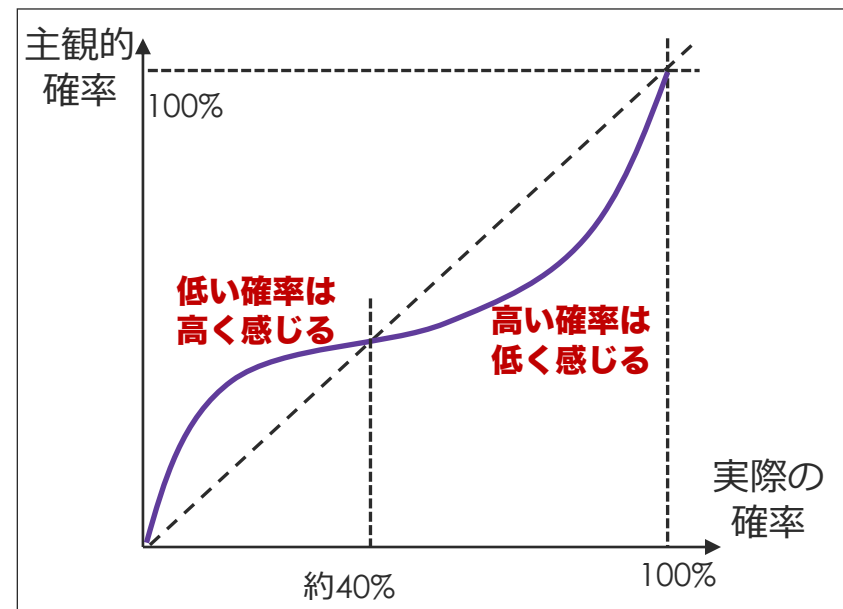
•価値関数：

「価値の感じ方のゆがみ」を表すグラフ



•確率加重関数：

「確率の感じ方のゆがみ」を表すグラフ



損失回避性

損失に対する感応度は同じ額の利得に対する感応度より遥かに強い

→損を感じた瞬間から損失に対して「過剰に反応する」「不必要に保守的になる」「過度の恐怖心を抱く」

感応度逡減性

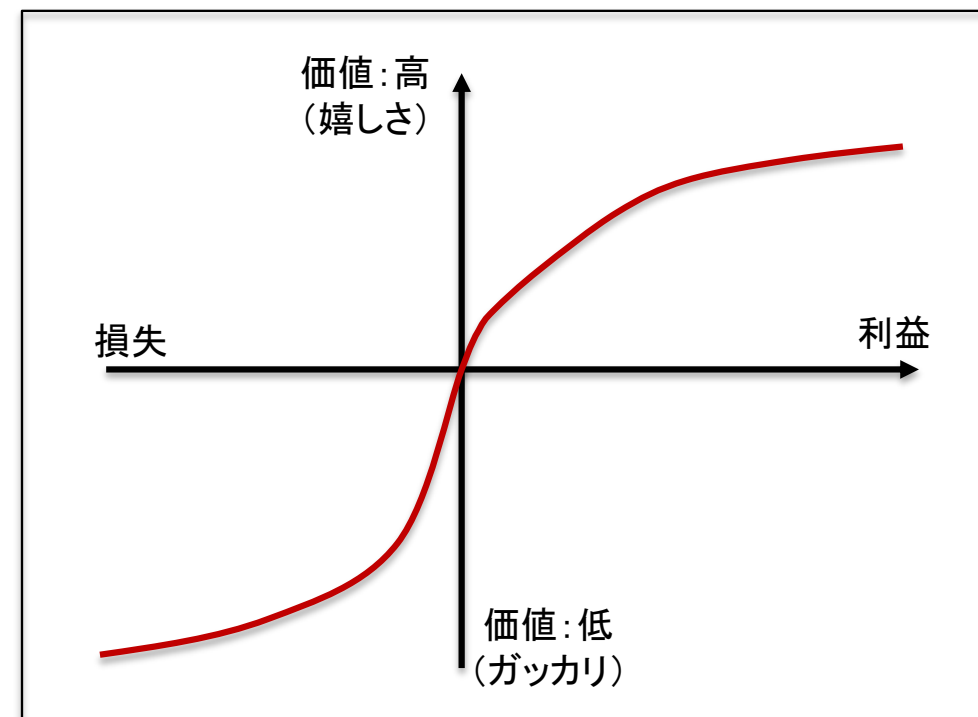
利得、損失いずれについても額が大きくなるほど感応度が逡減

→「安価な買い物は、できるだけ安く手に入れようとする」
「あまりに高額な買い物をする場合、少額の誤差をかんたんに受け入れ、少し高くても購入する」

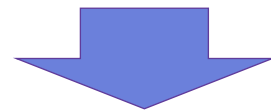
参照点依存性

価値判断は全て参照点に対して行われる

→「金メダルを確実視されていたが、銀メダルを取った」「メダルは困難といわれていたが、銅メダルを取った」場合、銅メダルをとったほうが主観的価値は高い



- ▶ プロスペクト理論に基づく人間の一見非合理的な意志決定可能性
- ▶ 度重なるトラブル事例の隠蔽の背後にある意思決定の様相
 - ▶ **選択肢1**：状況を正直に報告すれば間違いなく大きな損害が発生し会社の社会的責任が厳しく追及される
 - ▶ **選択肢2**：状況を隠蔽すれば確率は低いがバレずに済み、大きな損害を回避できるかもしれない



- コンプライアンス的には当然**選択肢1**を選ぶべきであるが、プロスペクト理論的には人は**選択肢2**を選ぶ傾向がある
- 善悪の問題ではなく、このような**誤った選択をする傾向を一般的な人が持っている**ということを認識することが必要

AI技術が貢献できる可能性

- 人間は、避けることのできない「認知バイアス」「意志決定の歪み」の影響下にある
- 自分の意思決定を以下のような視点でメタ認知することが重要

- **監視**

→ 「自分の仮説を支持する観たい情報だけを見ていないか？」（**確証バイアス**）

- **対処**

→ 「最近に起こったことだけに焦点を当てた対処になっていないか？」（**近接性バイアス**）

- **学習**

→ 「後知恵に基づく教訓から学習していないか？」（**後知恵バイアス**）

- **予見**

→ 「将来の様相を希望的な観測に基づいて予見していないか？」（**正常性バイアス**）

AI技術はこのメタ認知を支援できる可能性