産総研
ともに挑む。つぎを創る。

# AIと意見が割れた。そのとき人は正しく意思決定できるのか？

## ～AI精度の受け止め方・適切な信頼・人とAIの協働から考えるAIの安全性～




2026年10月16日（金）
第２回 AIセーフティシンポジウム 本会議

国立研究開発法人 産業技術総合研究所
人工知能研究センター 人間AI協働機構研究チーム
研究チーム長　野里博和

# はじめに

●本日の発表は、次の2つの報告を基にしています。

## 1. AIRCテクニカルレポート「人間-AIチーミングシステムの検証と課題抽出」第1版

Nosato Hirokazu, Kagawa Rina, 2026, AIRCテクニカルレポート「人間-AIチーミングシステムの検証と課題抽出」第1版: 産業技術総合研究所.
https://aist.repo.nii.ac.jp/records/2003489

※AIRCテクニカルレポート

## 2. 人工知能学会全国大会論文集「ユーザーが主観的に見積もるAIの正確性が人-AI協働意思決定に与える影響」

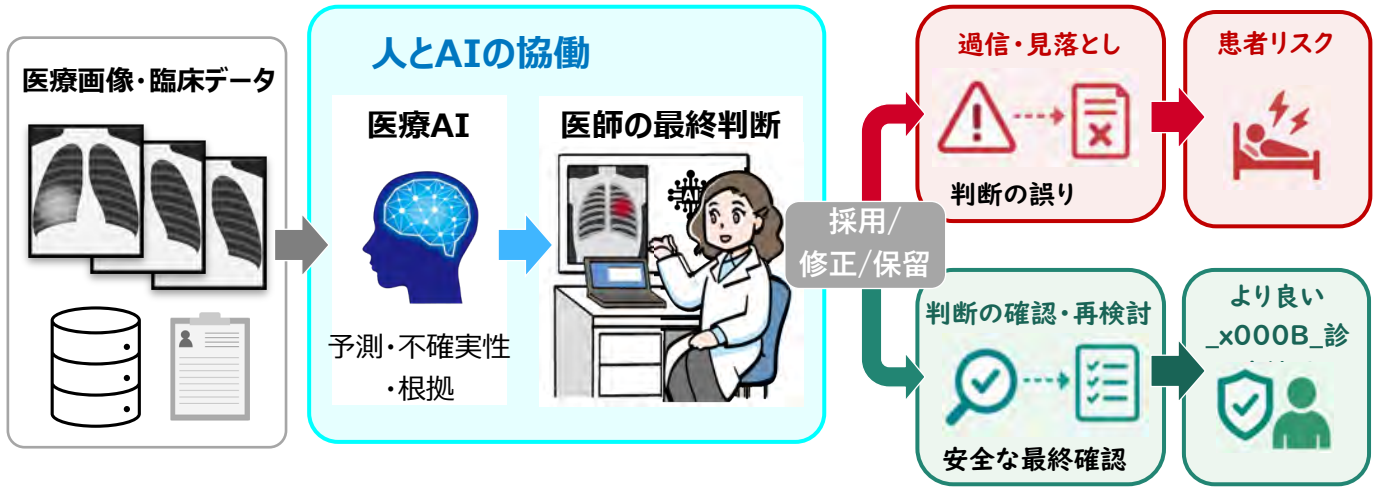
竹内 恵祐, 野里 博和, 香川 璃奈, ユーザーが主観的に見積もるAIの正確性が人-AI協働意思決定に与える影響, 人工知能学会全国大会論文集, 2026, JSAI2026 巻, 第40回 (2026), セッションID 1H3-OS-40-01, p. 1H3OS4001.
https://www.jstage.jst.go.jp/article/pjsai/JSAI2026/0/JSAI2026_1H3OS4001/_article/-char/ja

※JSAI2026

# AIが間違ったとき、本当に危険なのか？

※AIRCテクニカルレポート

## 医療AIの安全性は『医師の最終判断』まで含めて考える




AIの精度だけでなく、判断の過程全体を評価

**AIの誤り**

誤分類・未知の入力
偏り・不確実性

**人の介入**

気づく／無視する
採用する／修正する

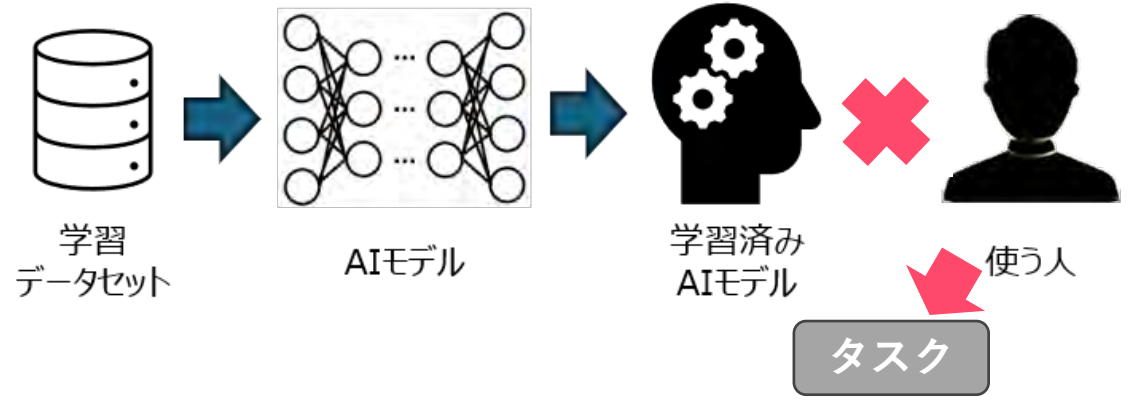
**最終決定**

安全な判断
または判断の誤り

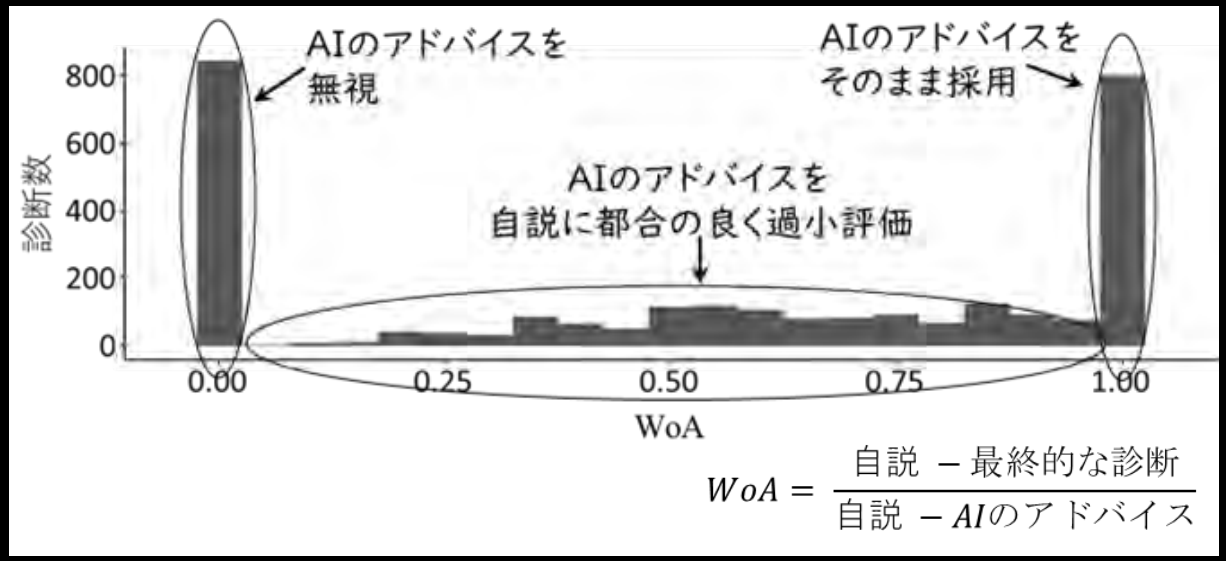
**AIの誤りが、そのまま最終判断の誤りにつながるとは限らない**

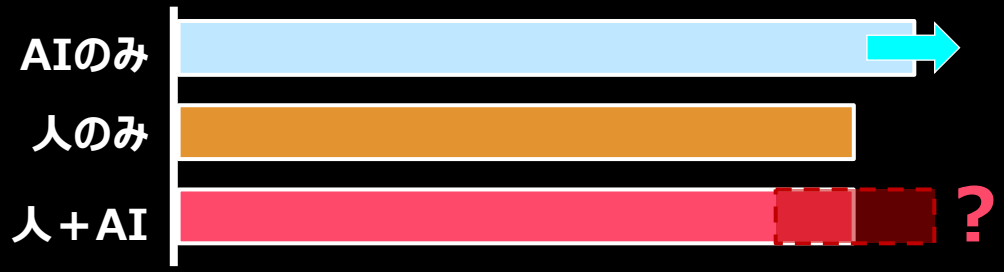
# AI単体の精度 ≠ 人+AIの性能

※JSAI2026




本発表で考える安全性の評価対象




AIの性能だけでは協働の成果を予測できない

**医師が誤り、AIが正しかった場合の調査結果**

Pálfi, B., Arora, K., & Kostopoulou, O. (2022).
Algorithm-based advice taking and clinical judgement:
impact of advice distance and algorithm information.
*Cognitive research: principles and implications*, *7*(1), 70.

AIが正確でも、使う人が納得しなければ
助言は活用されない

**重要なのは「AIを使った結果、人の最終判断が良くなったかどうか」**

# 人とAIの協働（Human-AI Teaming）

※AIRCテクニカルレポート

本発表の焦点




検索エンジン
ナビゲーション
自動運転（レベル4）

医療診断支援
AIによる自動検査

自動運転（レベル3）

自動運転（レベル2）
ゲームAI

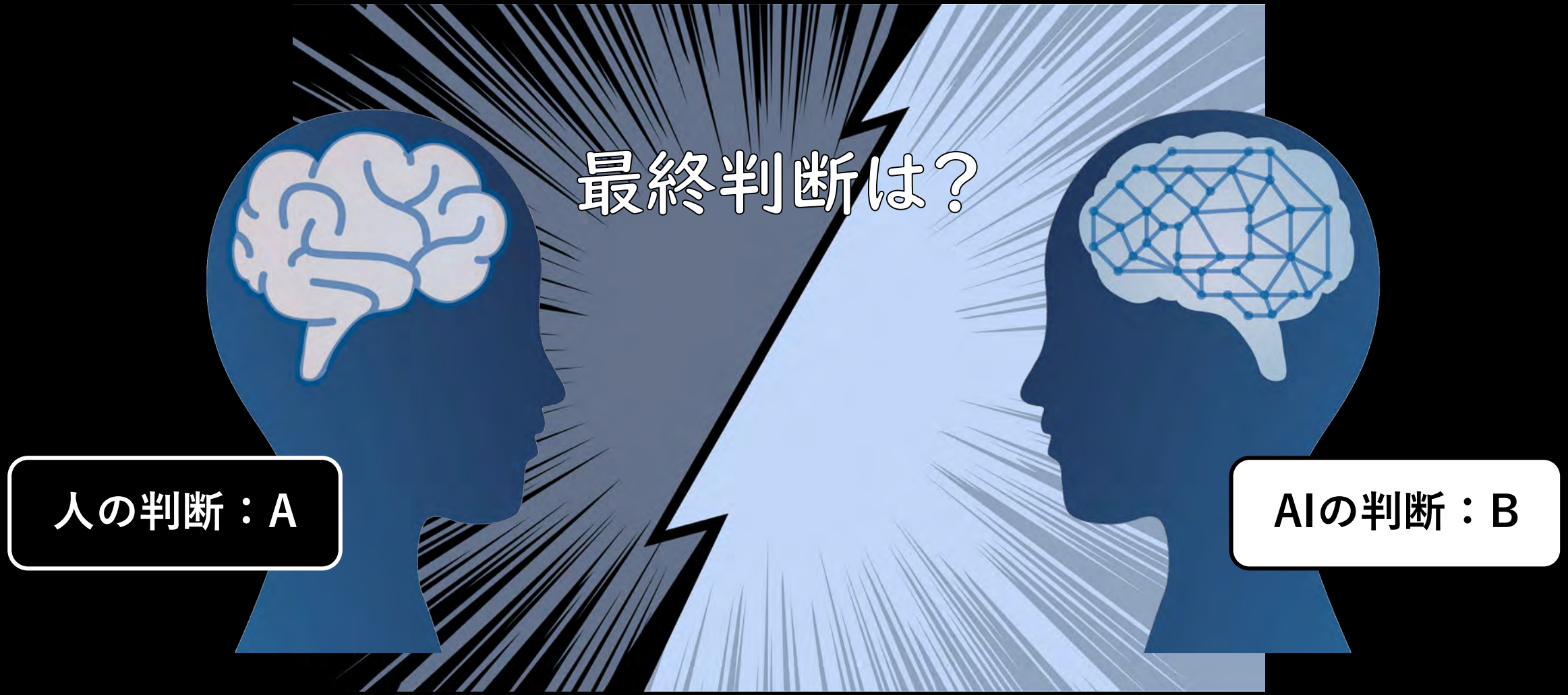
クラウドソーシング

※丸山文宏,「人とAIの関係性を考える」,『AIデジタル研究』第7号（2023年2月）を参考に改編

**本発表では、AIが助言し、人が責任をもって最終決定する関係に注目**

# 最も重要なのは「意見が割れた瞬間」

※AIRCテクニカルレポート




**人を優先？ AIを優先？ 根拠を比較？ 第三者に相談？**

# 医療AIで考えてみる

※AIRCテクニカルレポート




**医師**

「異常なし／別の疾患」

**AI**

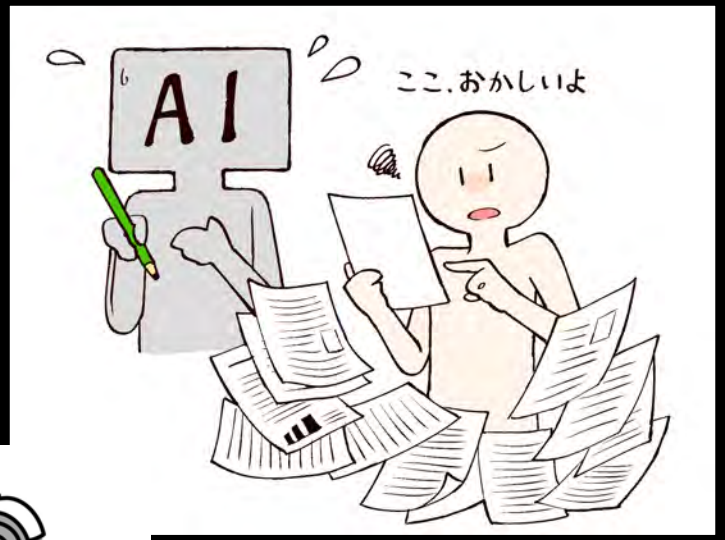
「疾患の疑いあり」

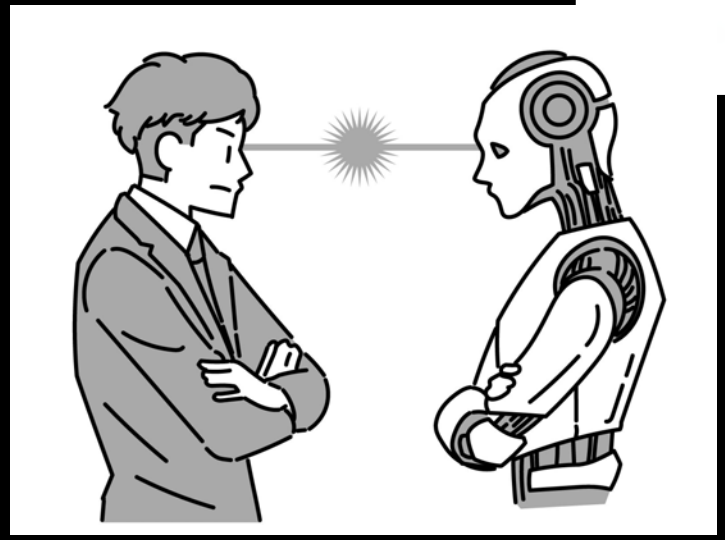
**単純な多数決では**
**決められない**

**「どちらが正しいか」だけではなく、「どう決めるか」を設計する**

# なぜ人とAIの意見は食い違うのか？

※AIRCテクニカルレポート



**AIの誤判定**
モデルの限界・不確実性

**人の誤判定**
認知・経験・見落とし

**利用環境の変化**
施設・装置・処理方法の違い

**判断基準の違い**
学習時の正解と利用者の基準

**正解の曖昧さ**
専門家の判断も一致しない

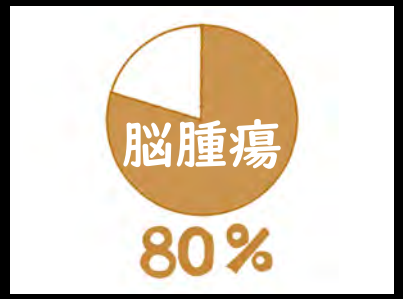
**人とAIの意見が違っても、AIが間違っているとは限らない**

# AIは「答え」だけを出せばよいのか？

※AIRCテクニカルレポート




判定

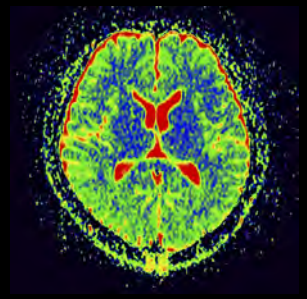



確信度・不確実性






類似事例・学習データ



背景情報・判断根拠

**AIの答えに加え、人が判断の適切さを確かめられる情報を示す**

# しかし、説明すれば安全になるのか？

※AIRCテクニカルレポート
※JSAI2026





## 説明の利点

判断の適切さを確かめやすい
判断基準を比較できる

## 説明の副作用

もっともらしい説明による過信
情報が多すぎると理解の負担に

Explainability ≠ Safety

**説明の有無だけでなく、人の判断がどう変わるかを評価する**

# AIへの信頼「Trust」が最終判断に影響する

※JSAI2026




AIへの理解・過去の成功や失敗・提示方法・確信度など

**AIの性能への信頼が、人の行動と最終判断に影響する**

# 人はAIの精度を正しく認識できるのか？

産総研
ともに挑む。つぎを創る。

※JSAI2026

Actual / Observed Accuracy

**観測時AI精度**

AIが実際にどれだけ
正しく答えたか

≠

Perceived Accuracy

**知覚AI精度**

人が「どれくらい正しい」と
感じたか

**実際のAIの精度と、人が感じる精度は別のもの**

# 検証実験：AI精度を0～100%に変えて調査

※JSAI2026

1 **初期判断**
人が画像を見て回答

2 **AI提案**
AIの分類結果を提示

3 **最終判断**
AIの助言も踏まえて判断

4 **事後評価**
知覚精度・Trust









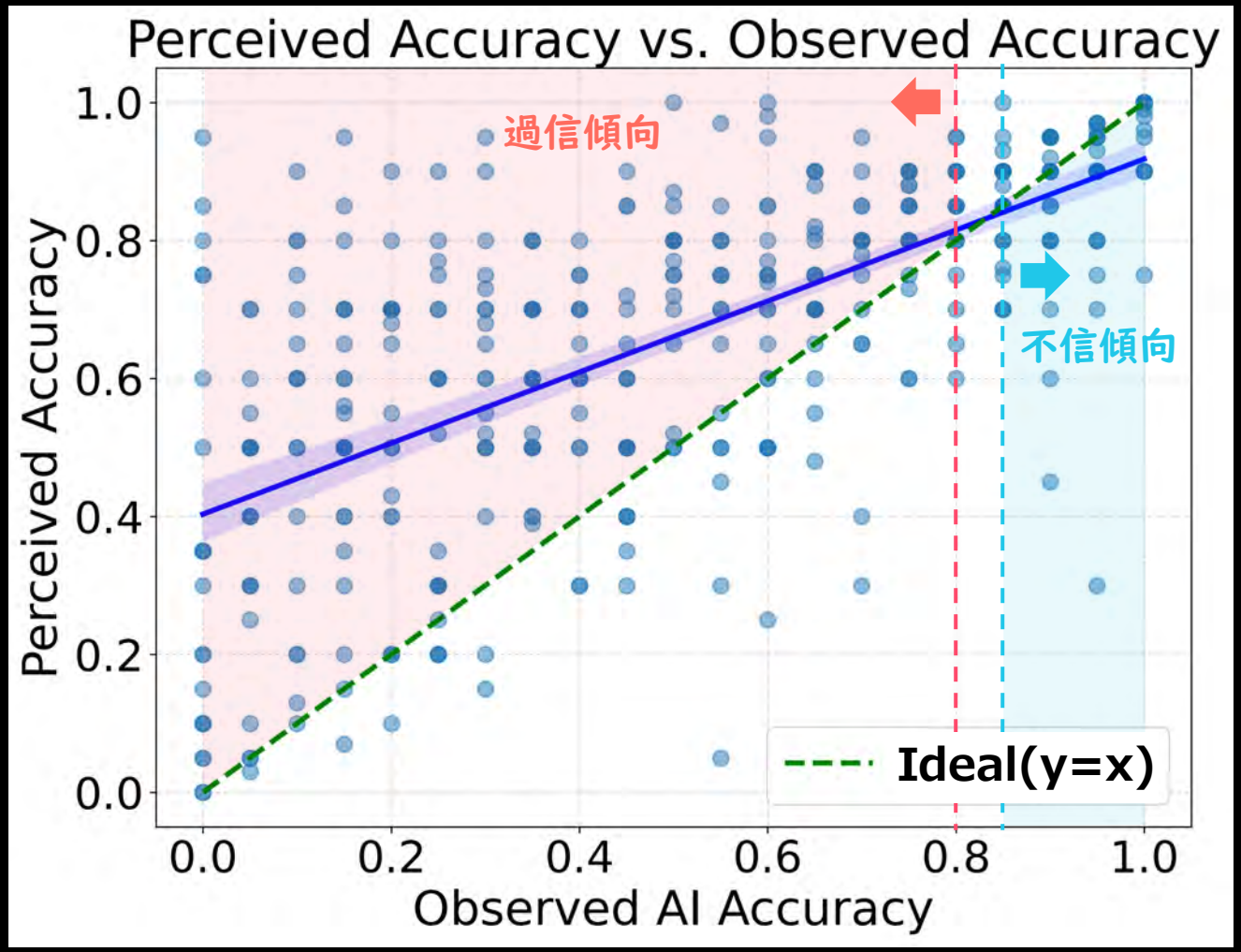

**アンケート調査**
- 使ったAIの精度の評価
- 使ったAIへの信頼
- 親近感
- AI全般への信頼傾向

420名、AI精度0～100%（21段階）、100問と事後アンケート

**人の判断とAIの助言を経て、最終判断・知覚精度・Trustを調査**

# 結果：人はAIの精度を「そのまま」見ていない

※JSAI2026




**0～80%**

実際より高く評価

**過信傾向**

**85～100%**

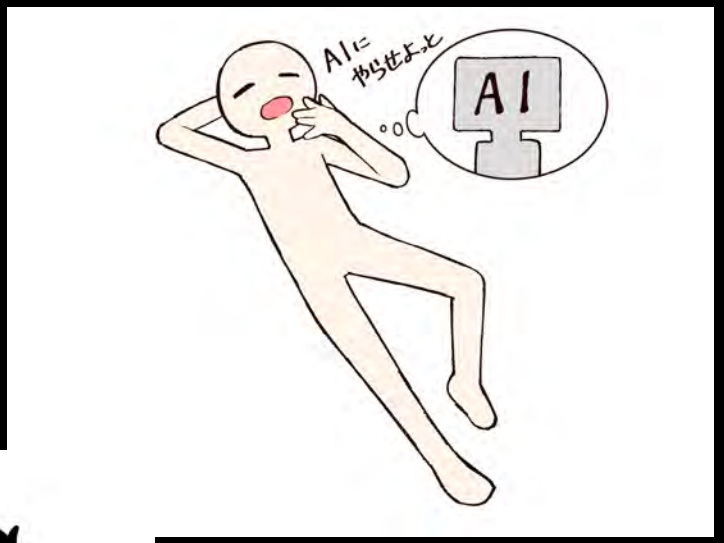
実際より低く評価

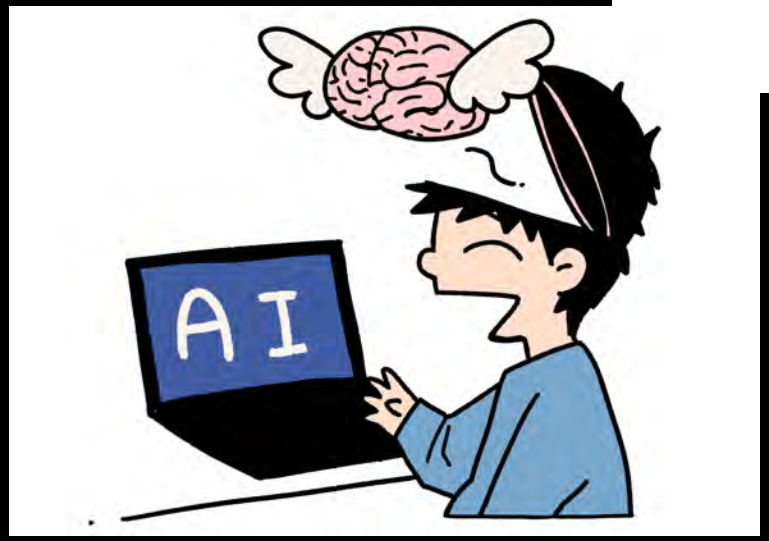
**不信傾向**

**人が感じるAIの精度は、実際の精度とは異なる**

# 危険①：AIを信じすぎる

※AIRCテクニカルレポート
※JSAI2026




**初期判断**

人：A（正しい）

↓

**AI助言**

AI：B（誤り）

↓

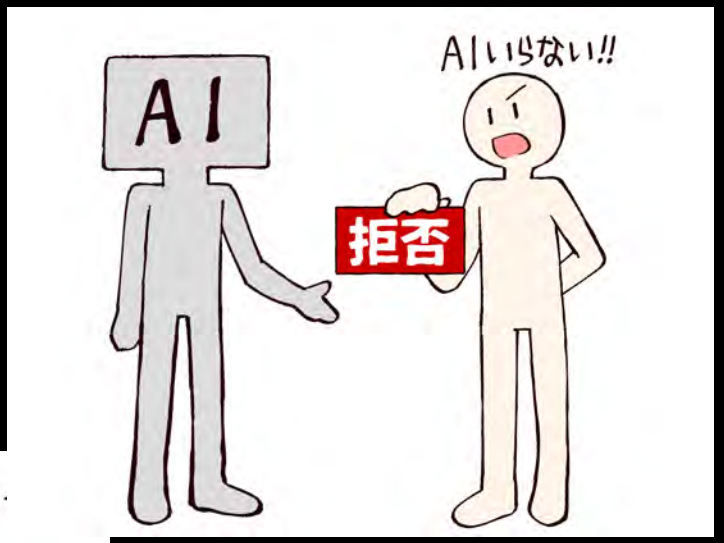
**最終判断**

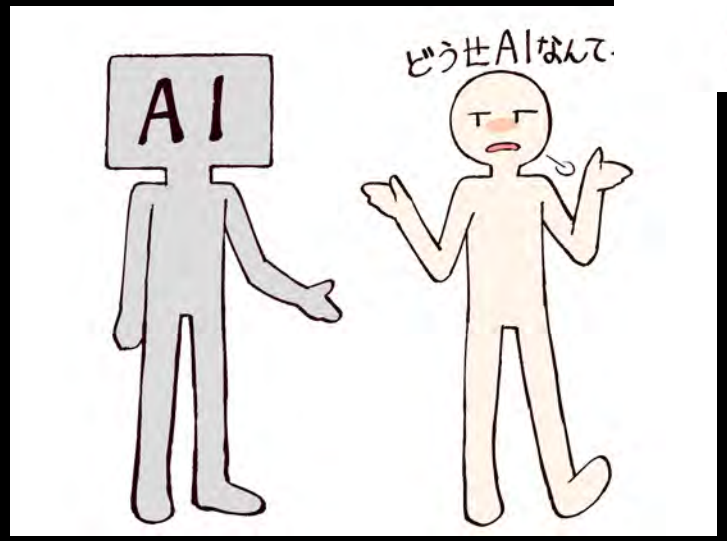
人：Bに変更

**正しかった人の判断も、AIの誤った助言で変わることがある**

# 危険②：AIを信じなくなる

※AIRCテクニカルレポート




**一度の目立つ誤り**

「このAIは使えない」

↓

**信頼の急落**

その後の助言を採用しない

↓

**役立つ助言を逃す**

正しいAIの助言も活用しない

**不信もまたSafety Risk**

# では「AIを信頼する」のが正解なのか？

※JSAI2026

AIを頼らなさすぎる　　能力に見合った信頼　　AIを頼りすぎる



**正しいAIの助言は採用し、誤った助言は採用しない。**
**この適切な対応の割合（Appropriate Reliance）を**
**全試行の中で高めることを目指す**

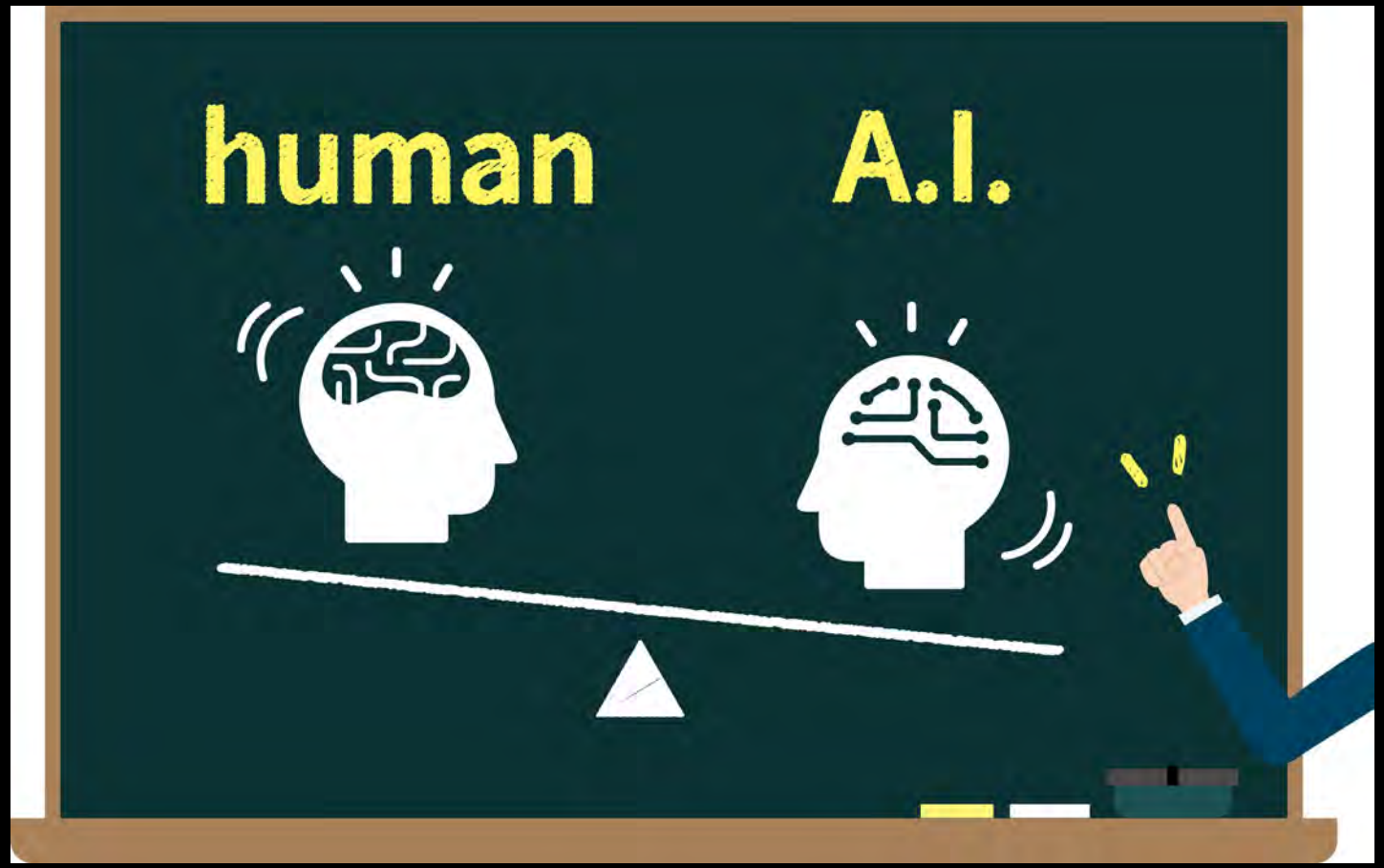
## 信頼（Trust）を高めることだけが目標ではない

AIの能力に合わせて信頼の程度を調整する

## 目標は、AIの能力に見合った信頼（Trust Calibration）

# 人とAIの「不一致」は本当に危険なのか？

※AIRCテクニカルレポート




## 一致したまま誤る

Human = AI = Wrong
誰も異常に気づかない

## 不一致で立ち止まる

Human ≠ AI
再検討・追加確認を起動

Safety Signal

**意見の不一致にはリスクがある一方、判断を見直すSafety Signalになり得る**

# 人とAIが協働するときの安全性の設計

※AIRCテクニカルレポート



### 1 AIモデル
精度・不確実性・適用限界

### 2 提示する情報
根拠・類似事例・確信度

### 3 利用者
信頼・感じる精度・AIへの理解

### 4 やり取り
意見が違うときの再検討手順

### 5 画面・使いやすさ
過信や不信を招かない見せ方

### 6 運用
監視・記録・第三者への相談

**安全性は、AIモデルだけでは確保できない**

# AIの安全性を評価する範囲の拡張

※AIRCテクニカルレポート
※JSAI2026



AIから、人とAIの協働、社会へ

AI

AIは安全か？

人
AI
操作画面
業務手順
組織

人とAIの協働は
安全か？

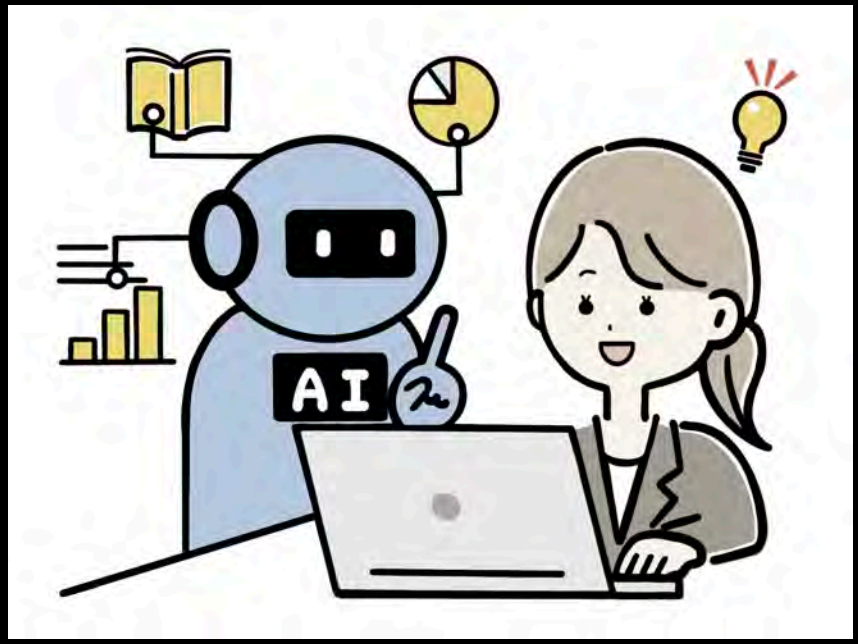
**人・技術・組織を含む仕組み全体で、AIの安全性を評価する**

# Take-Home Message

※AIRCテクニカルレポート
※JSAI2026

# AI Safetyの最後の1メートルは「人の意思決定」である

- 高精度AI ≠ 高性能なHuman-AI Team
- Actual Accuracy ≠ Perceived Accuracy
- 過信も不信も危険
- 不一致はSafety Signalになり得る
- 目標はTrust Calibration




AIとあなたの意見が割れたとき、
あなたは何を根拠に最終決定しますか？

# ご清聴ありがとうございました。

