

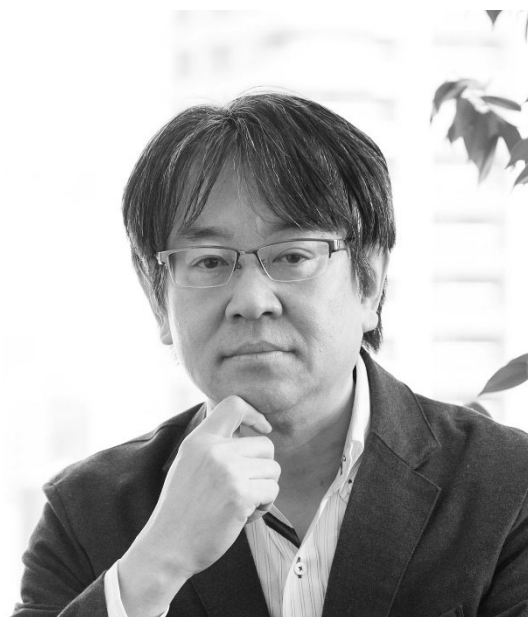
## DAY 1

### 生成系 AI の驚きと葛藤，建築界の『それから』 “Generative AI: Revelations, Frictions, and Beyond...”

2023 年 12 月 7 日（木）15:00 ～ 17:30 | Thu, 7th Dec 2023, 15:00～17:30

#### 人間からの適応を引き出す AI，常識のない AI としての ChatGPT

ChatGPT as AI that elicits adaptations from humans, AI without common sense



山田 誠二

**Seiji YAMADA**

国立情報学研究所／総合研究大学院大学 教授

**Professor, National Institute of Informatics /**

**the Graduate University for Advanced Studies, SOKENDAI**

#### 要旨

ChatGPT は、自然言語による自然な対話、QA がマルチリンガルで可能な世界初の無料 AI サービスである。本講演では、HAI ヒューマンエージェントインタラクション、AI の弱点に関する議論における重要概念である『人間から AI への適応』、『常識のなさ』の観点から ChatGPT を特徴付け、建築における人間-AI 協調設計について触れる。

#### 略歴

1989 年大阪大学大学院博士課程修了、同大学助手、講師、1996 年東京工業大学助教授を経て、2002 年より現職。専門は人工知能、HAI ヒューマンエージェントインタラクション。人工知能学会元会長。

# 人間からの適応を引き出すAI, 常識のないAI としての ChatGPT

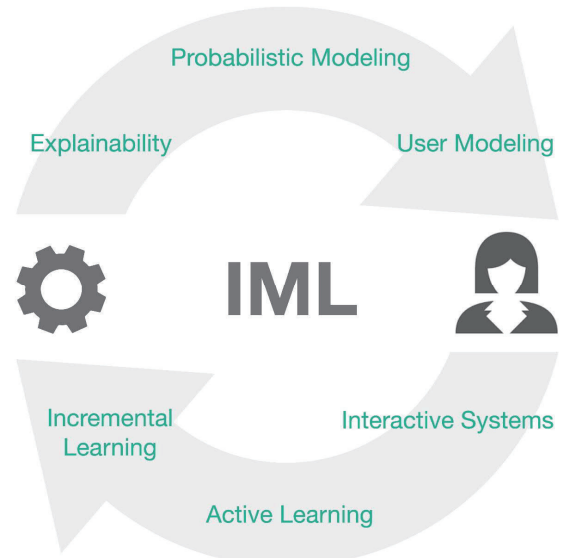
山田 誠二

国立情報学研究所／総研大  
人工知能学会 元会長・顧問

## AI (ML) 研究の重要なトレンド

## AI研究 (ML)のトレンド ～ NeurIPS, ICML, ICRLからの主観的解釈 ～

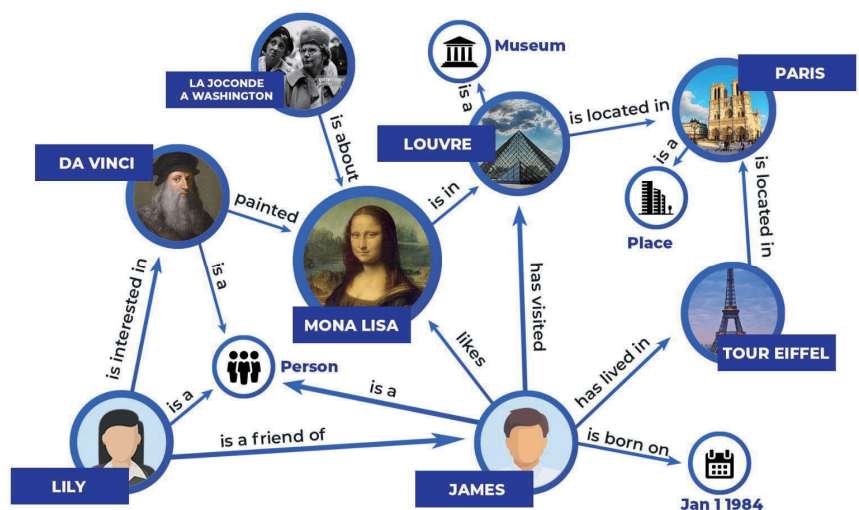
- 人間-AI協調 (ML・意思決定)
  - 人間の知識を導入して学習を高速化 ← ある意味, 80年代回帰
- 人間の知識の利用方法
  - インタラクティブ機械学習・**Human-in-the-loop**



3

## 知識グラフ

- Googleの提唱する知識表現  
(semantic webの **RDF**, 意味ネットワークと類似, 古くは意味ネットワークと類似)
- AIの世界で標準化されつつある



4

## 少数訓練データからの学習

- Zero/few shot learning : 一つ／少数の訓練データからの学習  $\Leftarrow$  事前学習のあとのファインチューニング
  - ChatGTPのプロンプトエンジニアリング
- 自己学習 (self-learning)
  - データ拡張 (data argumentation)により訓練データを水増し
- 半教師あり学習 (semi-supervised learning)
  - 同じクラスター内のデータは同じラベル
- 共通の仮定：類似データは、類似（正解）ラベルをもつ

5

人間からの適応を引き出すAI = ChatGPT

6

# ChatGPTとは？

- OpenAIが開発，ローンチしたLLMの  
エンドユーザ向けフロントエンド.
- 基本，無料で使える.
- 使われているLLM
  - 現在GPT3からGPT4（有料）
  - GPT3
    - webデータは，**2021年9月**まで.
    - 例：Q：「日本の総理は誰？」，A：  
「菅義偉」



7

S 底辺 5 cm，高さ 3 cmの平行四辺形の面積は？



底辺が5cm、高さが3cmの平行四辺形の面積を求めるためには、底辺と高さの値を利用して計算します。

平行四辺形の面積は、底辺の長さbと高さhを掛けたものとなります。したがって、面積Aは以下のように計算されます。

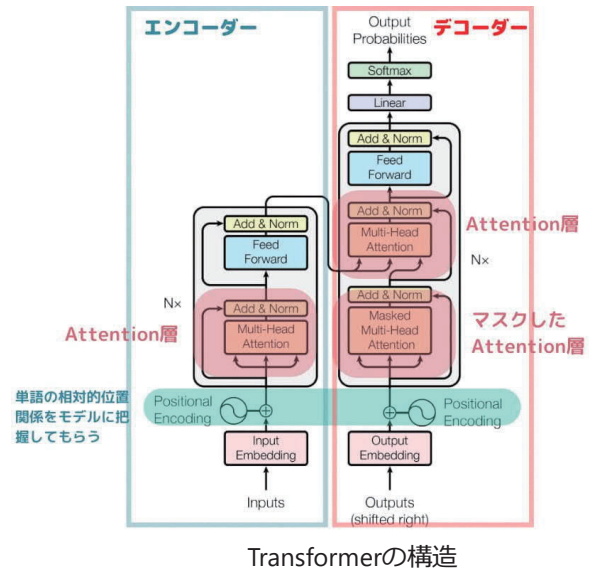
$$\begin{aligned} A &= \text{底辺} \times \text{高さ} \\ &= 5\text{cm} \times 3\text{cm} \\ &= 15\text{cm}^2 \end{aligned}$$

よって、底辺が5cm、高さが3cmの平行四辺形の面積は $15\text{cm}^2$ となります。

8

# ChatGPTの本質とは何か？

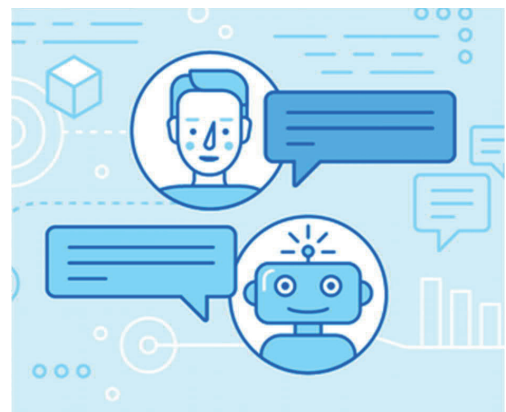
- 膨大なN-gramを使った次単語予測システム LLM (大規模言語モデル)
  - N-gram : N個の単語系列から次の単語を予測する確率モデル
- GPT4
  - 32,767-gram !
  - パラメータ数 > 数兆個
  - 20兆～30兆個の単語による訓練データ
    - この数は、人類がこれまで書いたすべての本にある単語の数に相当か越える。
- 入力：単語の埋め込み表現
- アテンションにより非連続・長期コンテキストを利用可能



9

## ChatGPTの特徴 ～インタラクションの観点から～

1. 入出力が自然言語 (= QAシステム, チャットボット)
  - 誰でも使える (誰でも入力でき, 出力を理解できる)
  - 世界中の一般人がエンドユーザ (億単位! : 米国 > インド > 日本)
2. 多言語 (日本語でも高性能, localization不要)
3. エンドユーザがシステム適応!
  - ⇒ In-context learning (プロンプトエンジニアリング)



10

## In-context learning

- 大規模な言語モデルLLMを用いてタスクを実行する際に、モデルのパラメータを調整することなく、わずかな例文のみを使用して所望の解を得る方法。
  - ラベル付きのデータが限られている領域や取得コストが高い領域で特に有益。
- 機械学習の観点
  - 少数訓練データからの学習：Zero-shot, One-shot, Few-shot
  - GoT (Chain of Thought: 思考の連鎖)

11

### Zero-shot (例なしで) いきなり質問する

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

|   |                              |                    |
|---|------------------------------|--------------------|
| 1 | Translate English to French: | ← task description |
| 2 | cheese =>                    | ← prompt           |

12

## One-shot 1つだけ解答例を与える

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ← prompt
   .....
```

13

## Few-shot いくつか解答例を与える

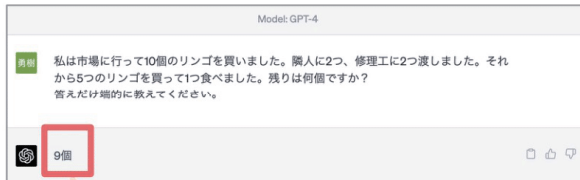
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ← prompt
   .....
```

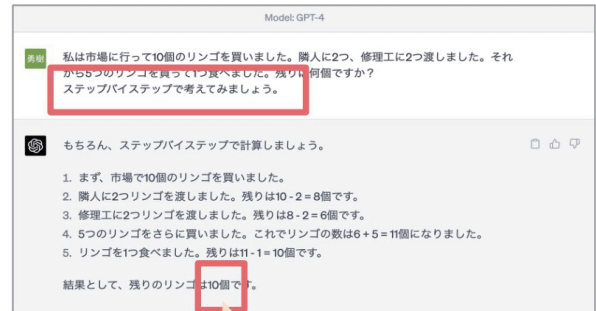
14

## Zero-shot CoT (Chain of Thought) プロンプティング

- プロンプトに「ステップバイステップで考えてみましょう」と追加すると、計算などが正確になりやすいです

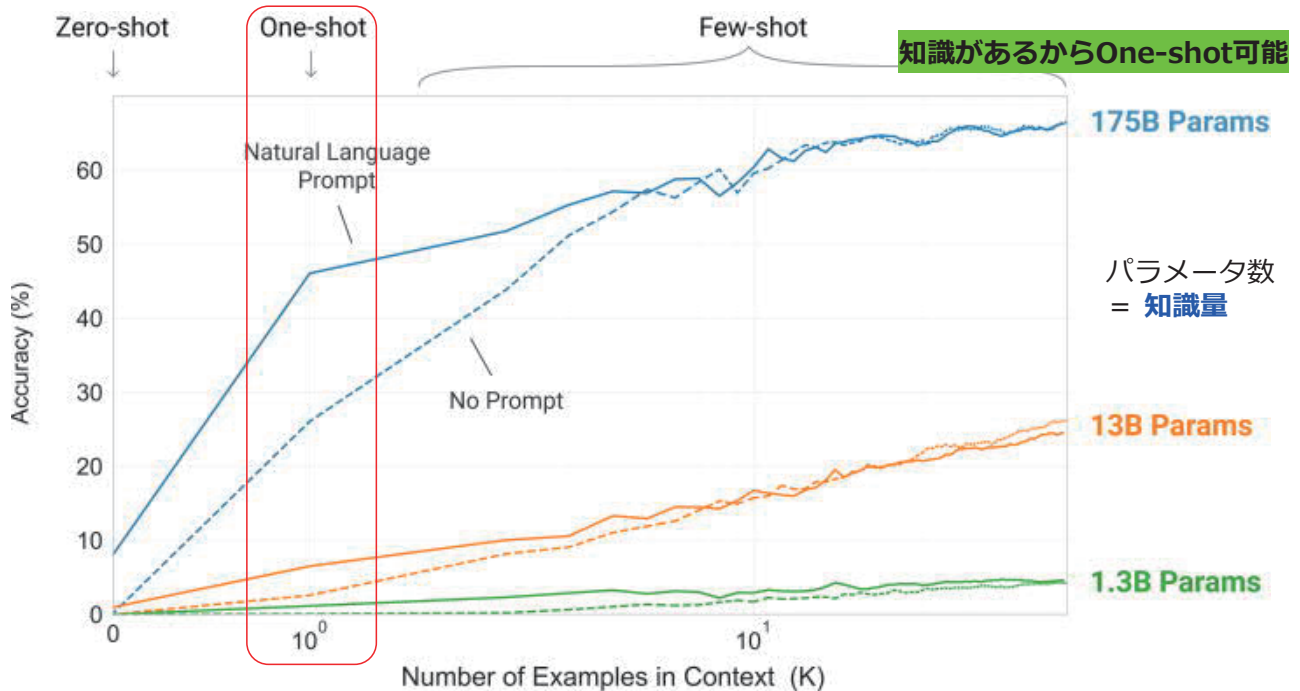


間違い

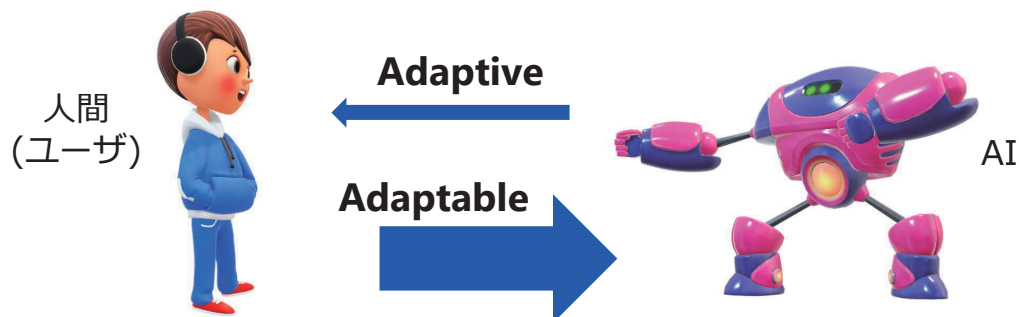


正解

AIモデルへの入力を最適化し、意図した出力を得るために例から学ぶ、プロンプトエンジニアリングの概要 - ログミーTech  
<https://logmi.jp/tech/articles/329255>



## Adaptive & Adaptable



- Adaptive **ユーザ適応** : AIから人間への適応
- Adaptable **システム適応** : 人間からAIへの適応

HAIの仮定 : Adaptive ≪ Adaptable

17

## Adaptive & Adaptable



- Adaptive **In-context learning** 人間への適応
- Adaptable **システム適応** : 人間からAIへの適応

HAIの仮定 : Adaptive ≪ Adaptable

18

## 常識のないAI = ChatGPT

19

## AIは常識がない！？

- **物理的常識** (右図)
  - 物理現象
  - 自然現象
- **社会的常識**
  - 社会通念 (NGワード集)
  - モラル
- **膨大な量の知識**
  - 書き尽くせない
  - 機械学習も非現実的

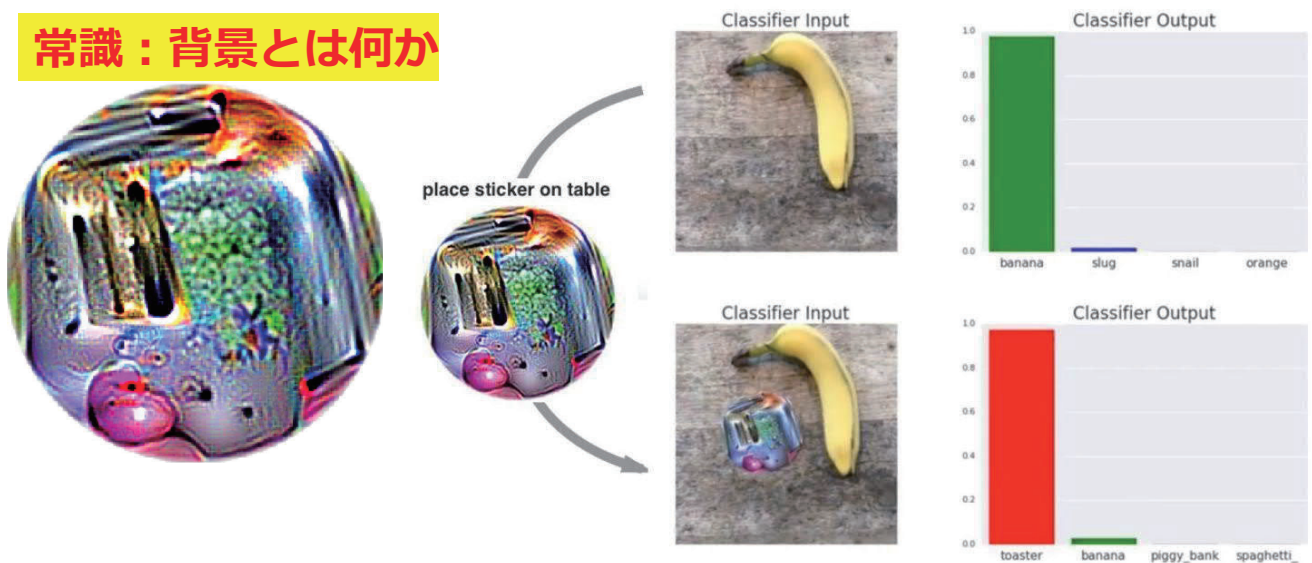


*Communications of the ACM, 58(9) 2015*

20

(実は ...) だまされやすいAI Gullible AI in fact

常識：背景とは何か



T. B. Brown, D. Mané, A. Roy, M. Abadi and J. Gilmer/Neural Information Processing Systems 2017

21

ChatGTPの『常識』は大丈夫？

- 本質的な問い：ChatGTPは常識を獲得したのか？

**No!**

— 口からでまかせ hallucination

22

## 身近な『口からでまかせ』

- Q: こんな論文, 教えて!
- A: 論文リストに『口からでまかせ』が!
  - 複数論文の
    - 筆者を平気でマージ
    - 論文タイトルを平気でマージ
    - DOIを平気でマージ
    - URLを平気でマージ
    - . . . .

『論文』に関する常識  
が全くない!!

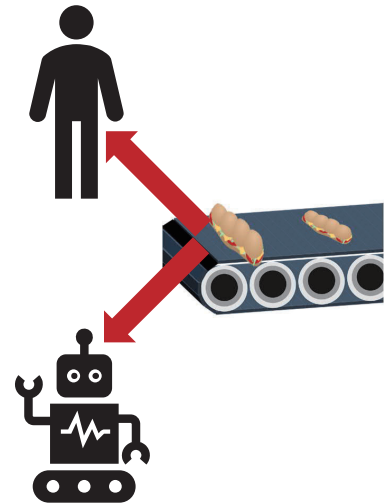
23

人間-AI協調意思決定 ～ 信頼校正 ～

24

## 人間-AI協調意思決定のタスク構造 (リライアンス)

- コンベアで流れてくるタスクを人間が自分で処理するか, AIに任せるかを人間が決定(リライアンス) する
  - AIに任せただけの場合, 人間は何もしない
    - AIの処理結果も見ない
  - **重要な決定は人間が!**
- 様々な協調意思決定がこのタスク構造
  - 様々な自動運転 (レベル3)
  - X線画像の協調読影
  - **人間とChatGPTとの協調意思決定**



25

## 人間とChatGPTとの協調意思決定 = リライアンス

1. 今日のタスク系列  $[T_1, \dots, T_m]$  を始めるかー.
2. このタスク  $T_i$  は, 自分でやるべきか, ChatGPTにやらせるべきか?
  - 要約タスクならChatGPTが上手いが, エゴサーチは下手.
3. 自分でやるならやる, ChatGPTに任せるならやらせて終わり.
  - 実際には, ChatGPTの結果を人間が修正したりする.
4. 次のタスク  $T_{i+1}$  に行こう, Step 1. へ.

26

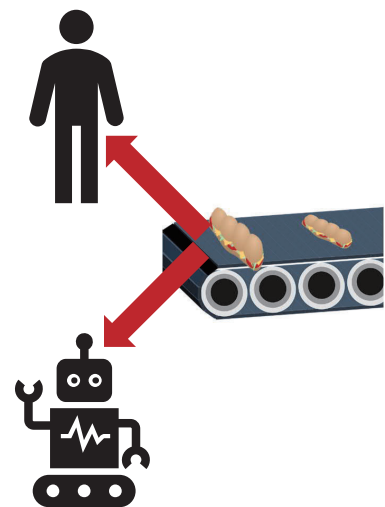
## 人間-AI協調意思決定

- 人間とAIがひとつのタスクを協調して問題解決する系（システム）
- 例：
  1. 自動運転（レベル3：自動車，ドローン，船舶，…）
  2. 人間-AI協調パターン認識（協調読影，手荷物検査，…）
  3. ChatGPTを活用した文書作成
  4. AIロボットとのビル内 協調監視…

27

## 人間-AI協調意思決定のタスク構造（リライアンス）

- コンベアで流れてくるタスクを人間が自分で処理するか，AIに任せるかを人間が決定（リライアンス）する
  - AIに任せただけの場合，人間は何もしない
  - AIの処理結果も見ない
  - 重要な決定は人間！
- 様々な協調意思決定がこのタスク構造
  - 様々な自動運転（レベル3）
  - X線画像の協調読影
  - その他，前述のタスクのほぼすべて



28

## AIに対する 信頼 (trust)

- AIの性能（タスクの成功確率）に対する人間の推定値.
  - Emotional trust (期待を裏切らなさそう)は捨象.
  - 人間の内部状態  $\Rightarrow$  計測不可能

## 適正信頼

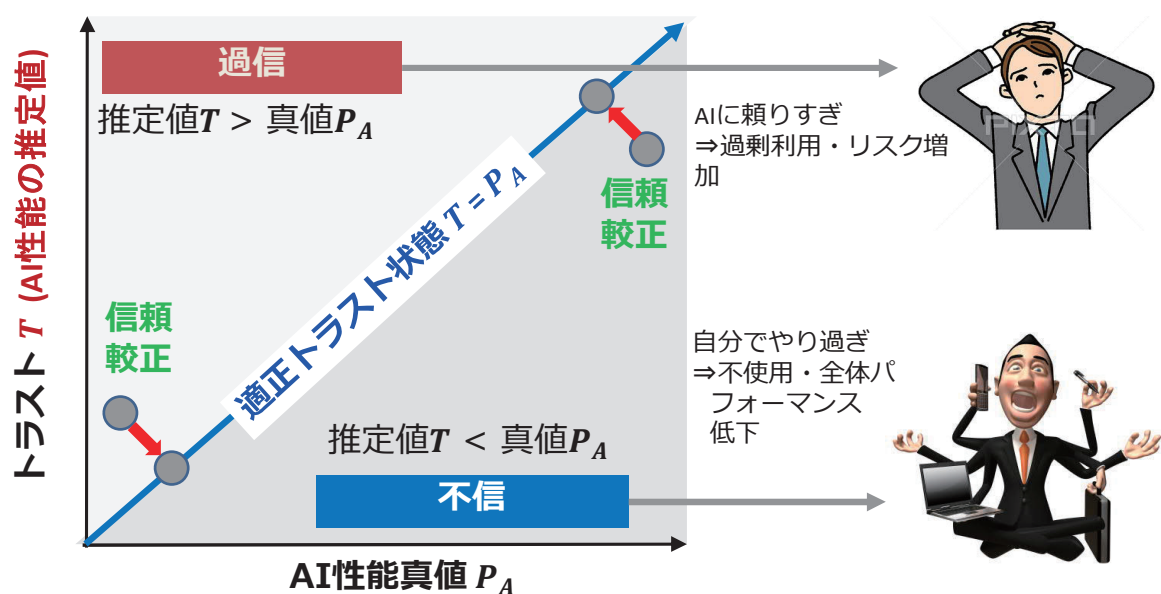
- 信頼 = AIの真の性能

## 信頼較正 (trust calibration)

- 人間が自分で『過信・不信』を正しい信頼へと修正.

29

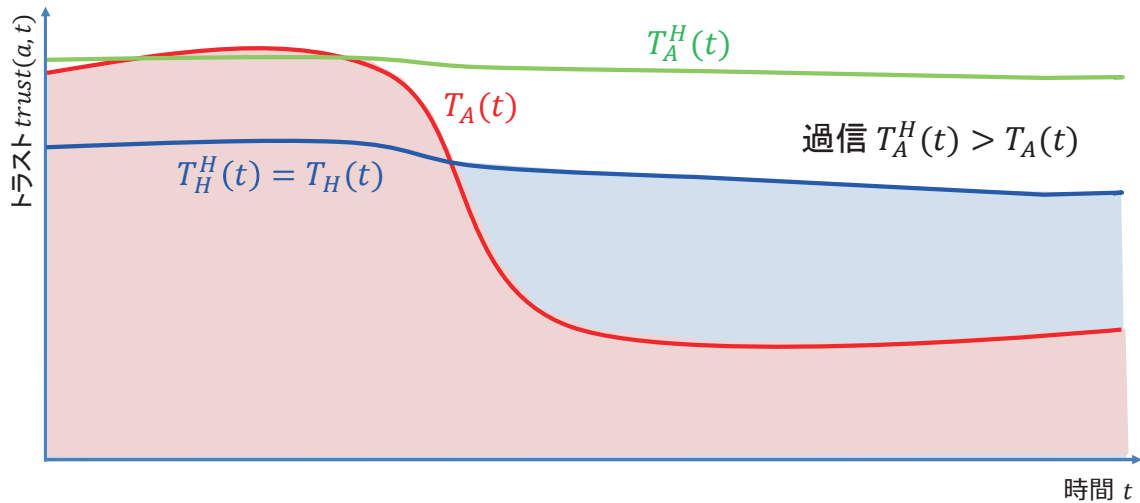
## 人間-AI協調意思決定における 過信・不信



30

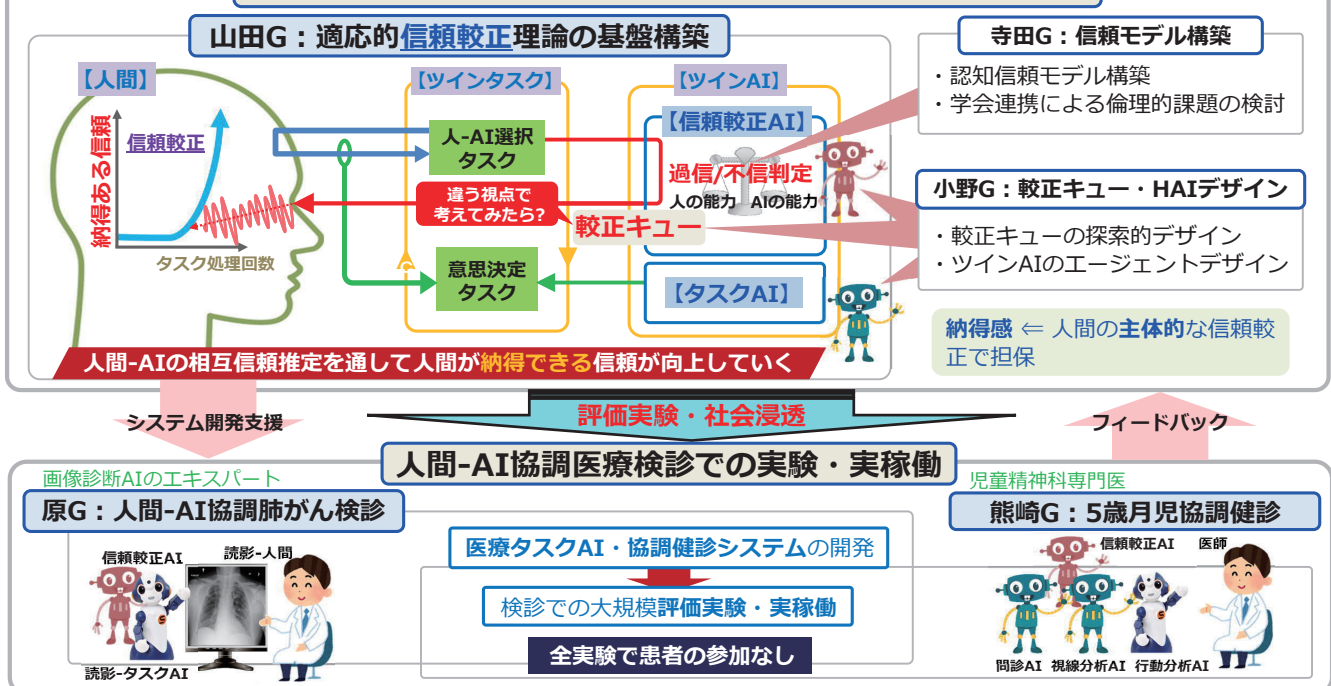
## 最適信頼 = 最大パフォーマンス

$$D^H(t) = \arg \max_{u \in \{A, H\}} T_a^H(t) \quad p(t_i, t_j) = \frac{1}{t_j - t_i} \int_{t_i}^{t_j} D(t) dt$$

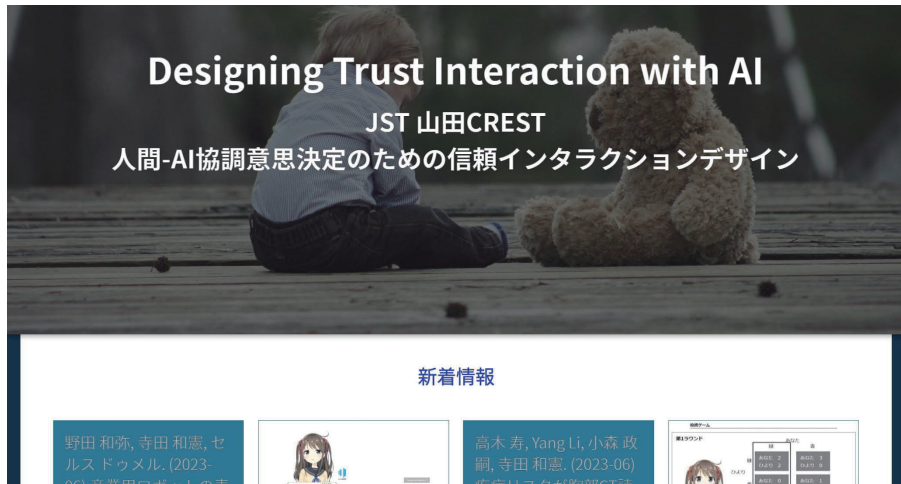


31

## 信頼インタラクションデザインによる人間-AI協調意思決定



## 人間-AI協調意思決定のための信頼インタラクションデザイン



以上です