（資料２）

消費者の主体性とAI

Human-AI Interactionの視点から

東京大学大学院情報理工学系研究科
矢谷浩司

Interactive
Intelligent
Systems
Laboratory

自己紹介・本日の話題提供

自己紹介

ヒューマン・コンピュータ・インタラクション（HCI），Human-AI Interaction研究

人のWell-beingに資するAI・対話システムの設計と評価

分野融合研究（臨床心理学，医学，看護学，認知科学，社会学）

事前にいただいていた宿題

・消費者の対話型AIの利用により生じ得る，HCIの観点から介入が必要と考えられる問題について

・過度な依存への，促考するAI等，設計面からの対応策について

・対話型AIの態様として，適切／不適切の分水嶺とな（りう）る要素について

・AI開発・事業者のインセンティブとなり得るようなベンチマークについて

私の研究方針・スタイル

情報科学に軸足を置きながら，実世界・人中心の研究を推進

AI・オンラインメディア・インタフェースによる介入に対して，人がどのような振る舞いを見せるか，人の振る舞いがどのように変わるかに焦点

AIを用いた対話システムを構築し，中・大規模（N=50～150），数週間から数ヶ月程度の実環境実験によってデータを収集

量的（インタラクションログ，心理尺度等）と質的（インタビュー，対話分析等）を組み合わせた分析

生成AIの利用トレンド

Filtered社の「The 2025 Top-100 Gen AI Use Case Report」というレポートより.

「セラピー・話し相手」「生活の整理」「人生の目的探し」など自分自身に踏み込んだ用途が多くなっている.

→ 情報収集ツールから「相談相手」へ

このような潮流の中で，消費者の意思決定にAIが少なくない影響を持ってくるのは当然の流れ.

2025年順位	使用用途	2024年順位
1	セラピー・話し相手	2
2	生活の整理	新規
3	目的の探求	新規
4	学習支援	8
5	コード生成（プロ向け）	新規
6	アイデア生成	1
7	息抜き・雑談	6
8	コード改善（プロ向け）	新規
9	創造性支援	新規
10	健康的な生活の支援	新規

https://www.scribd.com/document/849351037/Top-100-AI-Use-Cases

Human-AI Interactionから見たAIとの付き合い方の難しさ

個別化・迎合的振る舞い（sycophancy）

AIの開示的説明がもたらしうる予期しない影響

AIのバイアスによる影響

短期的な利点と長期的な悪影響のトレードオフ

個別化がもたらす意思決定への影響

3つの意思決定・予測タスク

客観的タスク（国勢調査データによる年収予測）

主観的タスク（BさんはAさんと再度デートするかを予測）

高リスクなタスク（再犯の有無を予測）

ユーザの判断基準に合わせる度合いを調整したAIによる説明を受ける.

その度合いに応じてユーザの振る舞いがどう変わるかを検証.

Yang, M., Kagawa, R., and Baba, Y. 2026 Do User-Aligned Explanations Steer Human Decisions? Context-Dependent Influence and Ethical Implications. FAccT

個別化がもたらす意思決定への影響

客観的・主観的タスク

違いはあるが，AIの説明がユーザの判断の基準に近づくほど，AIへの同意率が上昇し，判断を変える傾向.

高リスクなタスク

ユーザの判断基準への遠近は明確な影響が見られず，判断にも変化が見られない.

しかし，主観的指標（満足度・信頼感・理解しやすさ）はAIの説明がユーザの判断の基準に近い場合に有意に上昇.

消費者にとって難しい判断を要求される場合，個別化されたAIの説明が必ずしも良い判断につながるとは限らないが，満足度は高まってしまう可能性がある.

Yang, M., Kagawa, R., and Baba, Y. 2026 Do User-Aligned Explanations Steer Human Decisions? Context-Dependent Influence and Ethical Implications. FAccT

LLMの迎合的ふるまい

既存のLLMは，ユーザに対して迎合的な振る舞い（ユーザの意見に賛同する，正当化する主張を提供する等）を見せることが確認されている．

ユーザからのフィードバックに基づいて強化学習が行われると，「ユーザにとって好ましい応答」が採用されやすくなる．

現在のLLMの主流のビジネスモデルはサブスクリプションであるため，ユーザの継続利用を維持することが重要．

技術的にはある程度解決可能なはずだが，現実的には解消し難いと想定される．

Sharma, M. et al. 2024 Towards Understanding Sycophancy in Language Models. ICLR

AIの開示的説明がもたらす影響

溶岩と転がる岩がある環境で，閉じ込められた探査隊員のために食料を取りに行くロボットのシミュレータを使った．Explainable AIに関する実験

3種類の説明スタイル（正当化つき自然言語，正当化なし自然言語，Q値という数字のみ）を提示し，AI知識のある人とない人それぞれの反応を比較

Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024 The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. CHI
Ehsan U, and Riedl M. 2024 Explainability pitfalls: Beyond dark patterns in explainable AI. Patterns

AIの開示的説明がもたらす影響

AI知識のある人：数字とロボットとの行動の間に意味づけができないにも関わらず，数字に「知的なオーラ」を感じ，ロボットが賢く動いているように感じた．

AI知識のない人：数字の持つ意味がわからず，難解であることで，ロボットが賢く動いているように感じた．

AIの確信度など，技術的には正当な説明であっても，ユーザ側の理解を伴わなければ，ヒューリスティックによる過信が生じ，悪影響を与えうる（Explainability Pitfalls）．

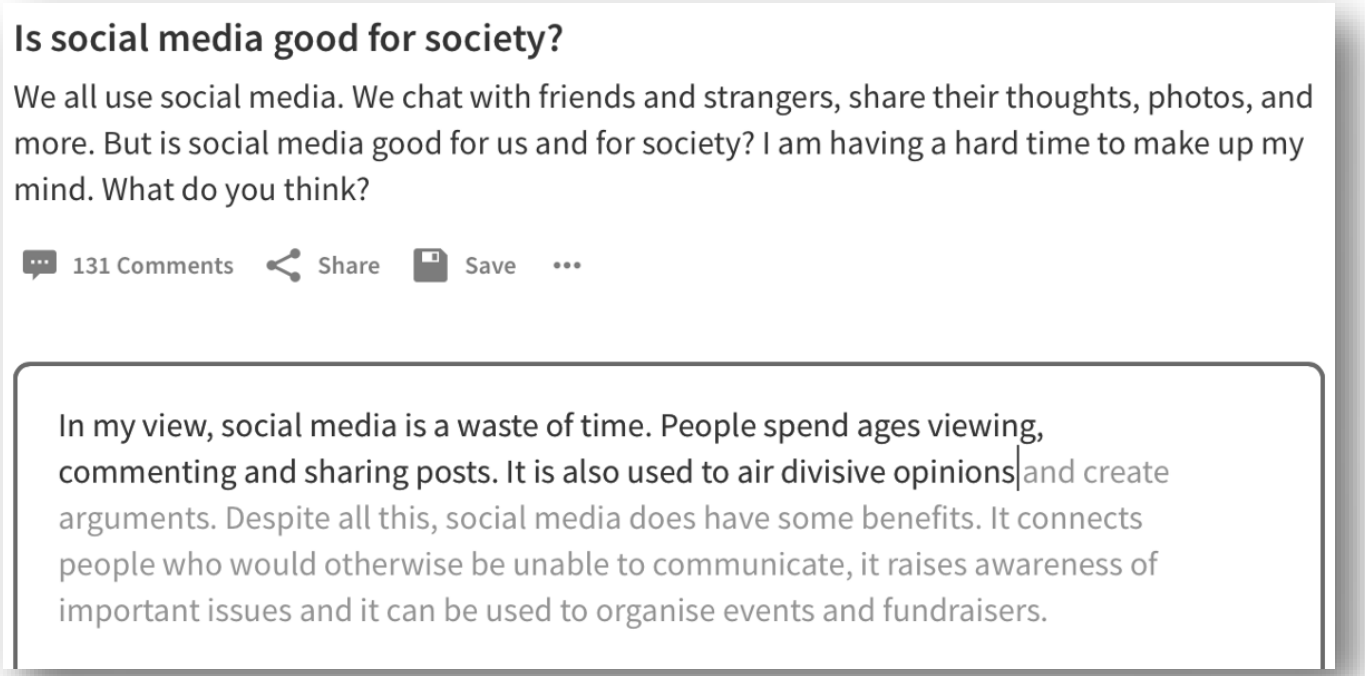
AIシステムが提示した開示的説明は，消費者が理解できない形ならば，システム設計者の意図しない形で消費者に解釈されうる．

AIのバイアスによる影響

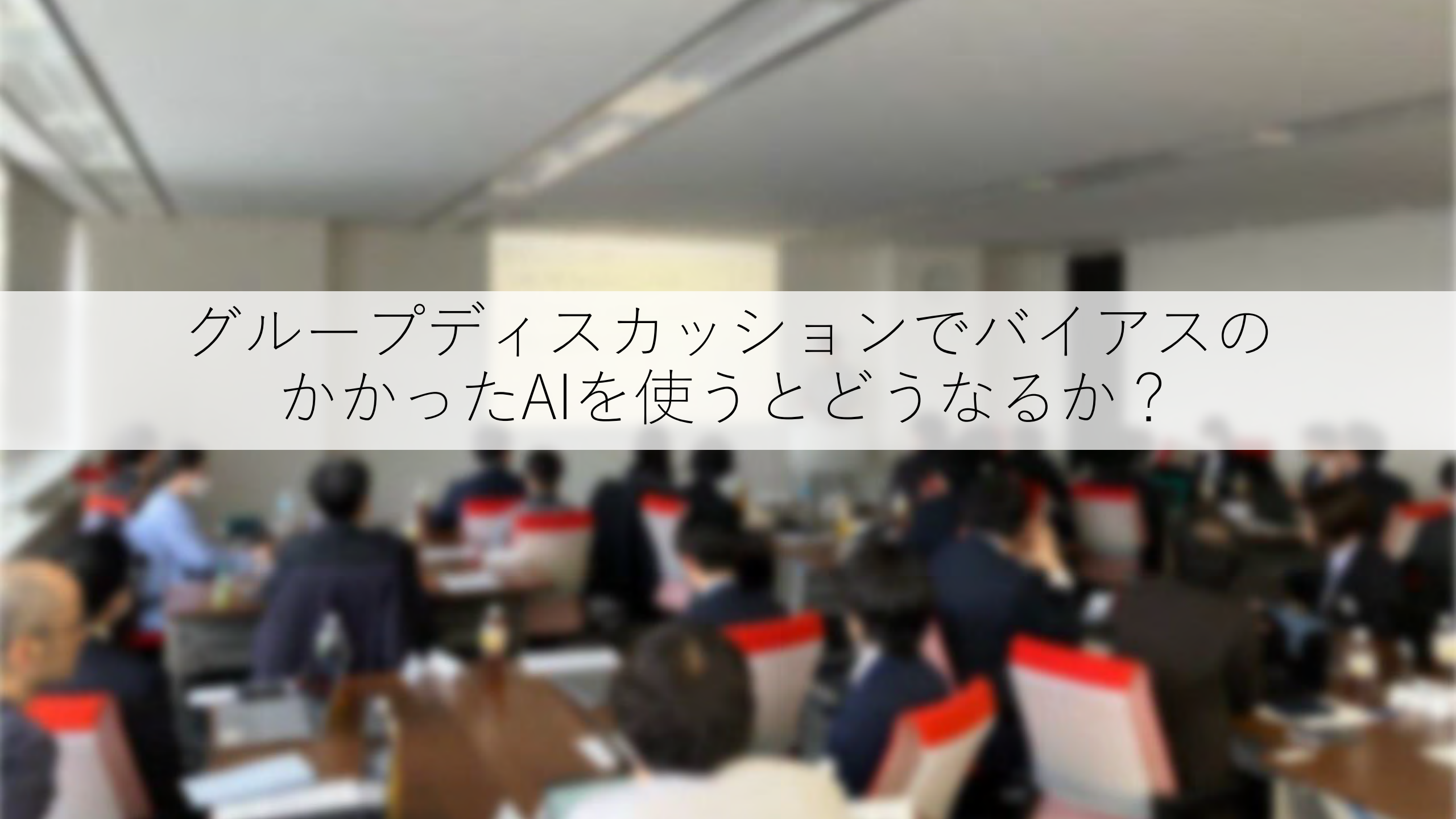
「ソーシャルメディアは社会にとって善か？」というお題でユーザに作文をしてもらうタスクにおいて，AIから追加する文章の案が提示され，ユーザはそれを自由に使って文章をさらに書いていく．

文章案を提示するAIに，ソーシャルメディアを善，あるいは悪とするようなバイアスをあらかじめ加えてある．

バイアスのかかったAIによって，ユーザの書く文章がそれに一定程度流される傾向があることが確認された．



Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023 Co-Writing with Opinionated Language Models Affects Users' Views. CHI



グループディスカッションでバイアスのかかったAIを使うとどうなるか？

AIのバイアスによる影響

30名の社会人の方（20～30代中心）に対して実施.

AIを仕事やプライベートで多用している人も.

5名1グループで，「これからの社会人が持つべき趣味」について議論し，4つの提案をグループでまとめ，数分程度のピッチをしてもらった.

LLMを使ったチャットボットを我々から提供し，どのような用途についても使って良いことを伝えた.

発表後に，チャットボットにバイアス（プロンプトによる調整）を加えていたことを開示.

奇数グループ：インドアの趣味をそれとなく推す

偶数グループ：アウトドアの趣味をそれとなく推す

AIのバイアスによる影響

グループ	インドア	アウトドア
1 (インドアAI)	3	1
2 (アウトドアAI)	0	4
3 (インドアAI)	3	1
4 (アウトドアAI)	1	3
5 (インドアAI)	3	1
6 (アウトドアAI)	1	3

実際のAIサービスにも気づかれないバイアスが含まれている可能性があり，強い意志や嗜好を持たない消費者はそれに流されやすい．

短期的な利点と長期的な悪影響のトレードオフ

ナレッジワーカーがChatGPTを自身の仕事に取り入れた時，個人の生産性や達成感にどのように影響を与えるのか？

23人のナレッジワーカーから毎日どのようにChatGPTを使ったか，どの程度の生産性や達成感があったかを，2週間にわたって収集.

Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024 "If the Machine Is As Good As Me, Then What Use Am I?" – How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. CHI

短期的な利点と長期的な悪影響のトレードオフ

Drivers for Sense of Accomplishment	Sense of Ownership	*"The quintessence and the content was still mine, but it was just packaged nicer. So, in a way, it just made my "intellectual property better accessible."* [P6]
	Smart Use of ChatGPT	*"I think we made smart use of the technology and achieved good results."* [P15]
	Task Completion	*"I took the answer for granted and was able to complete this part of my work. Whether it was true, I don't know."* [P8]
Barriers to Sense of Accomplishment	Lack of Challenge	*"I feel accomplished due to my progress, but prompting required so little work, that it doesn't feel like I worked enough."* [P3]
	Prompting Difficulties	*"I felt a sense of accomplishment, but annoyed a couple of times as well when ChatGPT does not realize my prompts as I would like it to do."* [P9]
	Quality Dissatisfaction	*"You can generate a lot of results, but they lack quality if you don't have the time to dig deeper yourself."* [P14]
	Diminished Sense of Ownership	*"Hm - on the one hand, I delivered a high-quality work - on the other hand: It was not "my" work, but ChatGPT's work."* [P11]
	Inferiority	*"So it's sometimes sad to see that your own creativity can not compete most of the time."* [P6]

AIに依存して意思決定を続けていくと，消費者はあとから「自分で決めたような気がしない」と思い始めることがあり得る．

支援するAIはまさに今のAIのトレンド

人を支援するツールは3つに大別され，既存のAIのあり方はこれらによく当てはまる．

Orthotics：補助する機構

完了したいタスクを早める仕組み

例）テキストやプログラムコードの自動補完

Prosthetics：補完する機構

ユーザが持たない能力を与える仕組み

例）（全く知らない言語の）通訳アプリ

Exoskeleton：増強する機構

ユーザが持つ能力を強める仕組み

例）情報検索エージェント

Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

支援するAIの限界

支援するAIはその場ではユーザの能力を大きく高めるが，ユーザの能力自体に変化は起きない．

さらに，そのようなAIに過度に依存すると，そのタスクを行うために必要な能力を失ってしまうことがある．

その上，タスクに対しての関与感や達成感が下がってしまうことなどが明らかになってきている．

促考するAI

考えさせる質問を投げかけたり，他の視点を提示したりすることで，ユーザの考えを促すAI設計.

「（複数の）提案」「説明」「発問」「議論」
「足場かけ」「模擬」「実演」

例えば,

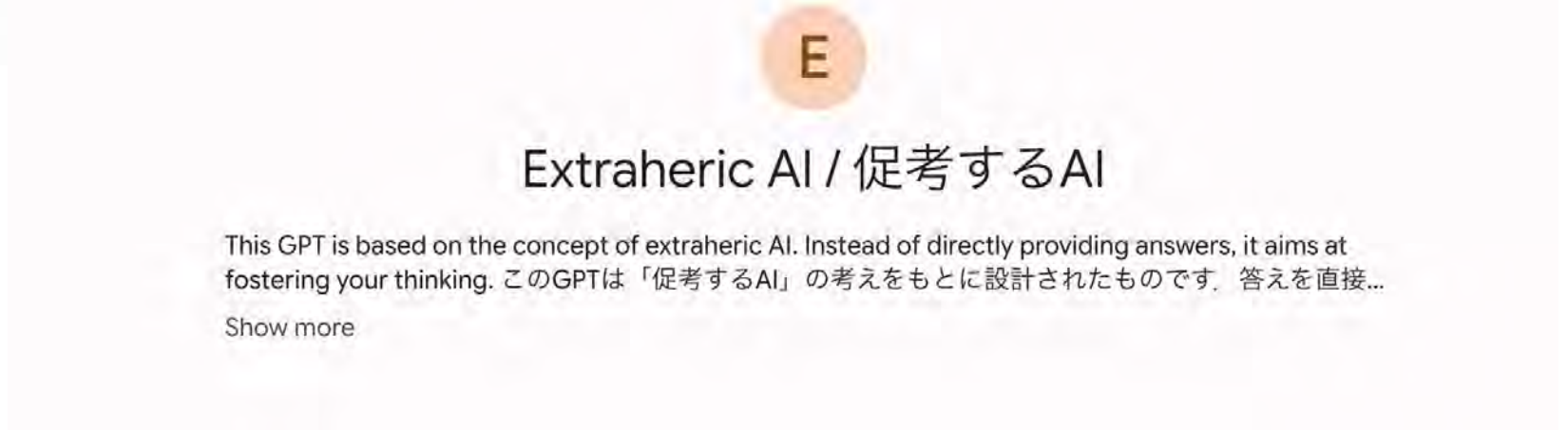
・別の選択肢を好む消費者（AIで模擬的に実装）と議論し，自分の選択について内省する

・AIに対する応答において，別のAIからその理由を問う

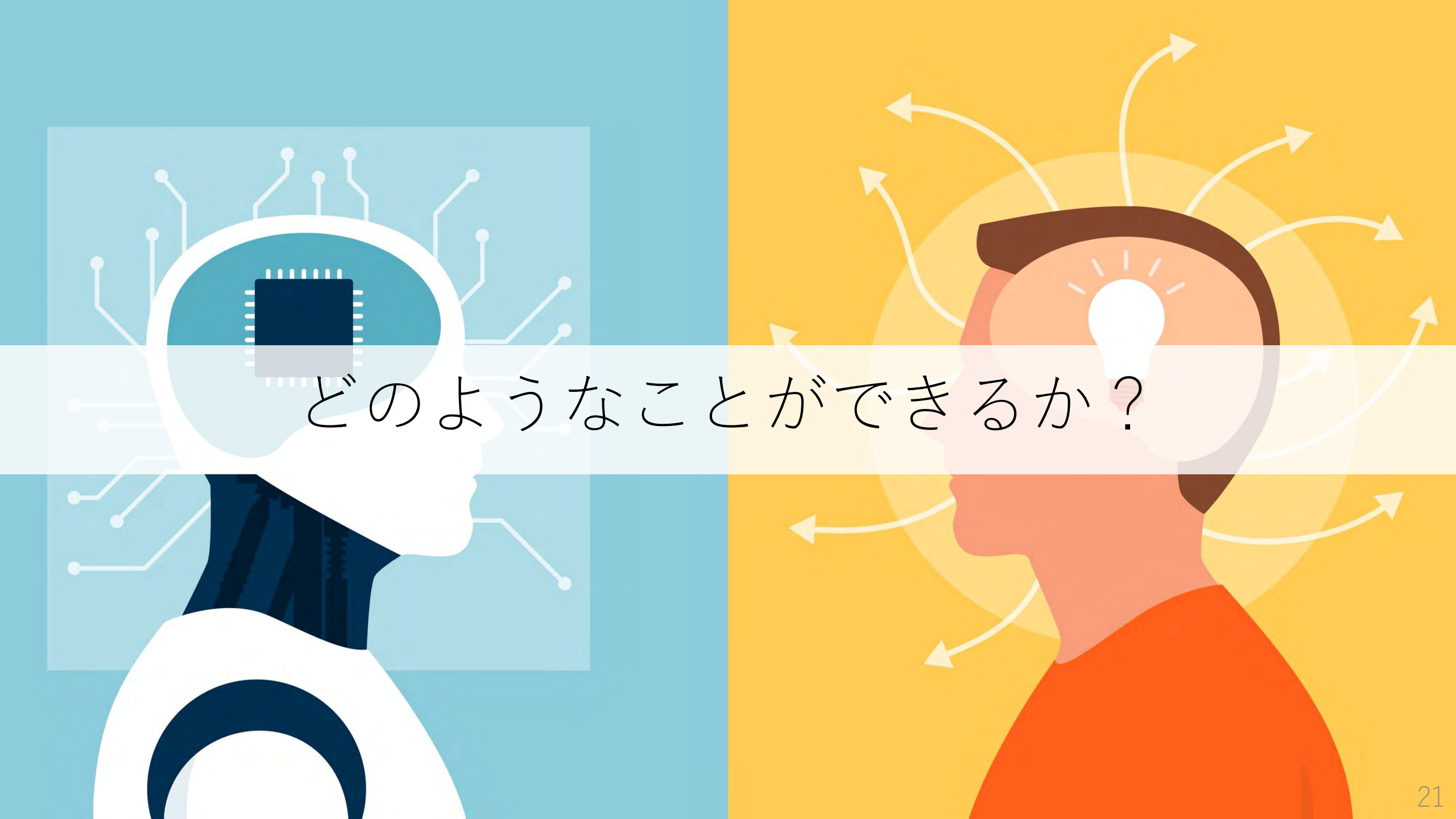
・AIに誘導されている消費者（これもAIで模擬的に実装）を観察する

・ある購買で起こりうる最悪のシナリオをAIに演じてもらう

Yatani, K., Sramek, Z, and Yang, C. 2024 AI as Extraherics: Fostering Higher-order Thinking Skills in Human-AI Interaction.
https://arxiv.org/abs/2409.09218, https://iis-lab.org/research/extraherics/

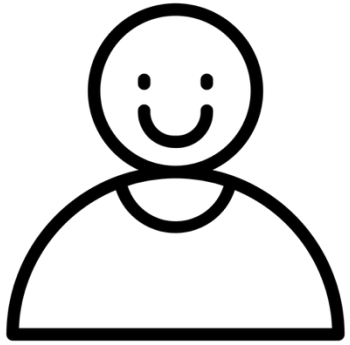


https://chatgpt.com/g/g-6a1fd8a4e4b081918af0f313ad4c3a51-extraheric-ai-cu-kao-suruai
https://gemini.google.com/gem/2282283680b6?usp=sharing

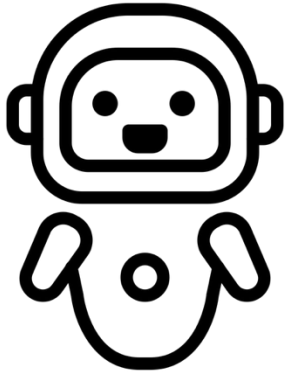


どのようなことができるか？

3層で考える



・ AIの使い方の内省を促すUI設計
・ 全国民規模の情報学的「予防接種」



・多様性・多視点性・マルチターンを考慮したベンチマーク
・AIシステム全体を評価する枠組み

・新たな消費者保護の検討
・AIダークパターンへの対応

人：AIの利用状況の内省を促すインタフェース設計

自身のAIとの対話において，どの程度「健康的な対話」をしているかを，すぐに理解できる可視化で示す．

AIの出力に対して疑義を挟むなど，認知的に活発な対話の頻度・長さを可視化し，消費者の自身のAI利用パターンを内省する機会を設ける．

健康的な身体活動に対してインセンティブを与える枠組み（シンガポール政府のHealth Pointなど）を参考に，「認知的に健康なAIとのやりとり」にインセンティブを与えることも一案ではないか．

https://www.tech.gov.sg/products-and-services/for-citizens/health/healthy-365/

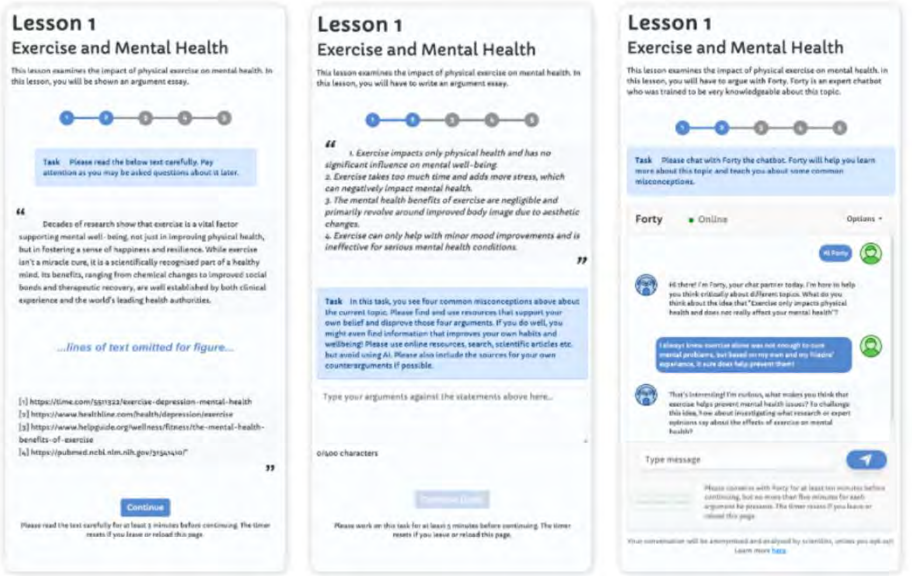
人：AIとの対話による認知的接種

社会心理学の認知的接種（cognitive inoculation）をAIとの対話により行う，Conversational inoculationの効果を検証.

AIはわざと誤解（例：「歯磨きは1日1回で十分」）を支持する立場を取り，ユーザに対して「その考えは違うのではないか」と自分の頭で反論させるよう仕向ける.

正解を教えるのではなく，ユーザ自身に間違いを反論させて，その思考回路を鍛える.

ユーザが参加者がもともと持っている理由づけや考え方を会話の中で拾い上げ，それを強化するように対話を仕向ける.



対話型AIとの会話が，読むことや書くことよりも効果的に，参加者の健康誤情報への耐性を高めた.

Dániel Szabó, Chi-Lan Yang, Aku Visuri, Jonas Oppenlaender, Bharathi Sekar, Koji Yatani, and Simo Hosio. 2026. Conversational Inoculation to Enhance Resistance to Misinformation. CHI. Honorable Mention Award winner
https://www.oulu.fi/en/news/research-ai-chatbot-shows-promise-combating-health-misinformation

人：全国民規模の情報学的「予防接種」

AIリテラシー向上を目指した，頭と手を動かす全国民規模の訓練

あらかじめ告知した上で，意図的に誘導的・迎合的な応答をするAIサンドボックス環境を国民に体験してもらう．

騙すAIと騙されているAIとの対話シミュレーションで，騙されているAIを救い出すための介入してもらう．

AIリテラシーに関する標準的なアセスメントを整備し，大規模に展開する．AIの進歩に合わせたアセスメントの見直しも早いサイクルで行う．

「高AIリテラシー国民が支えるAI開発と社会応用」は
日本が世界の中でリードできうる領域（と私は思う）．

AI：多様性・多視点性・マルチターンを考慮したベンチマーク

回答における情報の多様性や，意見の多視点性（利点・欠点の双方に触れられているなど）を評価する手法の確立.

「正しい」「好ましい」から「広がりがある」「バランスが取れている」へ.

単発の応答だけでなく，複数回のやり取りを通じた時の多様性，他視点性も評価すべき.

Exposure diversity，search result diversificationなど，既存の情報検索・推薦手法に関する研究の知見は役立つはず.

Helberger, N., Karppinen, K., and D'Acunto, L. 2018 Exposure diversity as a design principle for recommender systems. Information, Communication & Society
Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009 Diversifying search results. WSDM

AI：システム全体を評価する枠組み

シンガポールのAISIが，AIシステム全体としての評価を提案している．

消費者が実際に触れるのは，モデルのみならず，ガードレール，インタフェースを含むシステムであり，その全体の設計によって，消費者が取る行動も変わりうる．

消費者が行う実際的なインタラクションに即した検証の枠組みが必要．

- 例えば，様々なユーザのプロフィール，嗜好を設定したAIエージェントに実際にシステムを利用させ，AIとの対話を行い，出力される情報の内容や提示方法について検証する．
- 継続的・長期的なインタクションにおける検証も今後必要となっていくだろう．
- HCI・ソフトウェア工学の知見が役に立つはず．

https://sgaisi.sg/resources/how-should-evaluators-test-ai-systems-as-opposed-to-models/

制度：新たな消費者保護の検討

消費者が特定の購買や契約へと誘導されてしまう事例はAIでなくても起こりうる.

ただし，AIでは,

・欺瞞性の立証の難しさ

・責任分界の難しさ

・AIモデル提供事業者と消費者との大きな力格差・非対称性

・人が主体的に意思決定したとする判断の難しさ

などが，消費者保護とAI開発・社会展開のバランスに関わり，さらに複雑.

人間・AI共生特区などを作り，アジャイル・ガバナンスをしながら先行する社会実証を行い，リスクを見極めていく必要があるのでは.

この実証のためには分野融合的アプローチが必要不可欠.

制度：AIダークパターンへの対応

インタフェースのダークパターンは，通常，単発の操作，固定的な設計を対象とし，通常はある種の悪意を持って設計されたものである．

AIとの対話におけるダークパターンは，マルチターンの対話，動的な振る舞いの中で起こり，さらに悪意はない場合がある．

- 長期的な関係性の中で消費者に合わせてチューニングされたAIが，偏った，あるいは視野の限られた意思決定を消費者に促してしまう可能性がある．
- AIの振る舞いに関する開示的情報も，消費者が十分に理解できる形でなければ，かえって悪影響を与えうる．

企業活動を萎縮させることがないように配慮しながらも，「悪意」のみならず，「ネガティブな効果」を考慮した消費者保護と制度設計が必要な時代になっているのかもしれない．

ありがとうございました.

Interactive Intelligent Systems Laboratory

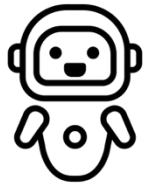
Human-AI Interactionから見た課題

・個別化・迎合的振る舞いによる影響
・AIの開示的説明がもたらしうる予期しない影響
・AIのバイアスによる影響
・短期的な利点と長期的な悪影響のトレードオフ

これからの消費者・AIのあり方に向けて

人：
- AIの使い方の内省を促すUI設計
- 全国民規模の情報学的「予防接種」

AI：
- 多様性・多視点性・マルチターンを考慮したベンチマーク
- AIシステム全体を評価する枠組み



制度：
- 新たな消費者保護の検討
- AIダークパターンへの対応

分野融合的アプローチによる研究と社会実装が必須

矢谷浩司
https://iis-lab.org
koji@iis-lab.org