ACM CHI 2023 参加レポート
Doctoral Consortium / Case Study 編

筑波大学大学院 博士課程3年
矢倉 大夢 (Hiromu Yakura)

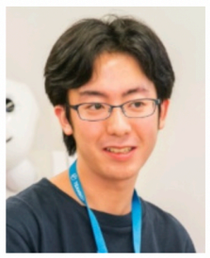
2023.05.22

自己紹介

- **所属**: 筑波大学大学院 博士後期課程3年
  - 連携大学院制度で、研究は産業技術総合研究所
    メディアインタラクション研究グループで実施
- **興味分野**: Human-Computer Interaction、機械学習
- Google / Microsoft Research の Ph.D Fellow に選出される
- ACM CHI '19 @ Glasgow より参加

Microsoft | Research Our research Programs & events Blogs & podcasts About Sign up: Research Newsletter All Microsoft

2021 Fellows



Hiromu Yakura

University of Tsukuba

Supervisors: Masataka Goto

Research interests: Human-Computer Interaction, Machine Learning

Long-term research goal: As a researcher in HCI, my research focus is on expanding the application area of computers, especially, machine learning techniques. Given that we cannot deny the possibility of the mistakes of machine learning models, we need to

Google AI Blog

The latest news from Google AI

Announcing the 2020 Google PhD Fellows

Thursday, October 8, 2020

Posted by Susie Kim, Program Manager, University Relations

Google created the PhD Fellowship Program in 2009 to recognize and support outstanding graduate students who seek to influence the future of technology by pursuing exceptional research in computer science and related fields. Now in its twelfth year, these Fellowships have helped support approximately 500 graduate students globally in North America and Europe, Africa, Australia, East Asia, and India.

It is our ongoing goal to continue to support the academic community as a whole, and these Fellows as they make their mark on the world. We congratulate all of this year's awardees!

Human Computer Interaction
Abdelkareem Bedri, *Carnegie Mellon University*
Brendan David-John, *University of Florida*
Hiromu Yakura, *University of Tsukuba*
Manaswi Saha, *University of Washington*
Muratcan Cicek, *University of California, Santa Cruz*
Prashan Madumal, *University of Melbourne*

目次

Doctoral Consortium とは

Doctoral Consortium

Quick Facts

The CHI'23 Doctoral Consortium is a CHI'23 event for mentoring and sharing research among a small group of Ph.D. students. The full CHI'23 conference is a blend of physical on-site attendance in Hamburg, Germany, and virtual attendance on-line.

- 博士課程の学生が研究発表を行い、メンタリングを受ける場
  - ACM CHI に限らず幅広い国際会議で実施されており
    SIGCHI 系でも C&C, CSCW, DIS, IUI, UIST などがある
- 参加する学生に学びの機会を与えながら
  コミュニティを育て、拡げていくことが目的にある

Case Study とは

Case Studies of HCI in Practice

What is a Case Study?

Case Studies are compelling stories about applied HCI practice based on real-world experiences that will be instructive and of interest to other community members. Based on the concrete research and design cases, HCI practitioners and researchers will learn how they can apply HCI principles and methods in practical HCI work.

- フルペーパーと別の、HCI の応用に主眼をおいたトラック
  - 新規性というよりも、実世界での応用や
    そこからの HCI コミュニティに共有すべき学びについて

Importantly, Case Studies need to make a contribution beyond the study itself. A writeup of a single usability study, for example, would not make for a Case Study. Submissions need to reflect on methods or situations and be largely interesting to the broader HCI Community.

目次

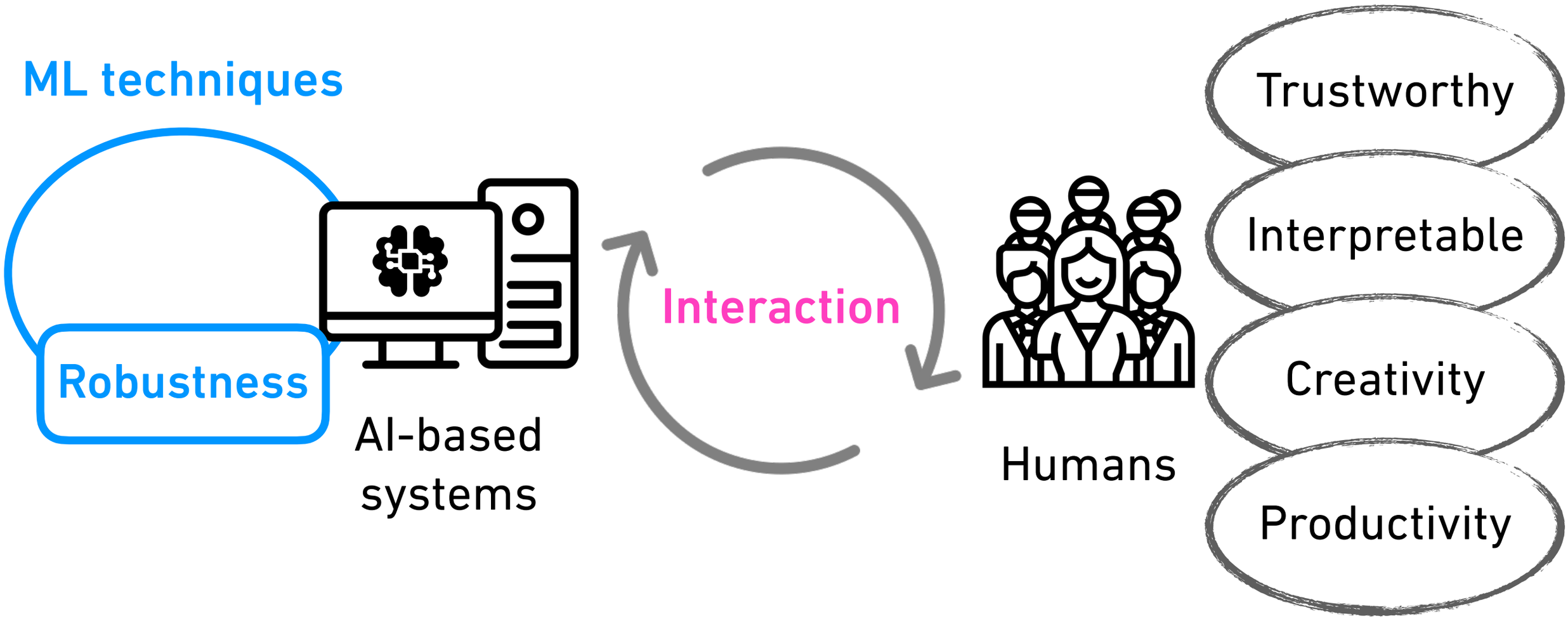
A Generative Framework for Designing Interactions to Overcome the Gaps between Humans and Imperfect AIs Instead of Improving the Accuracy of the AIs

Hiromu Yakura

University of Tsukuba, Japan
Google / Microsoft Research Ph.D. Fellow

Research context

- Research interest: **ML** + **HCI** (human-computer interaction)
  - How to apply **machine learning** in a **human-centric** manner?



Generate (non-software) Bugs to Fool Classifiers

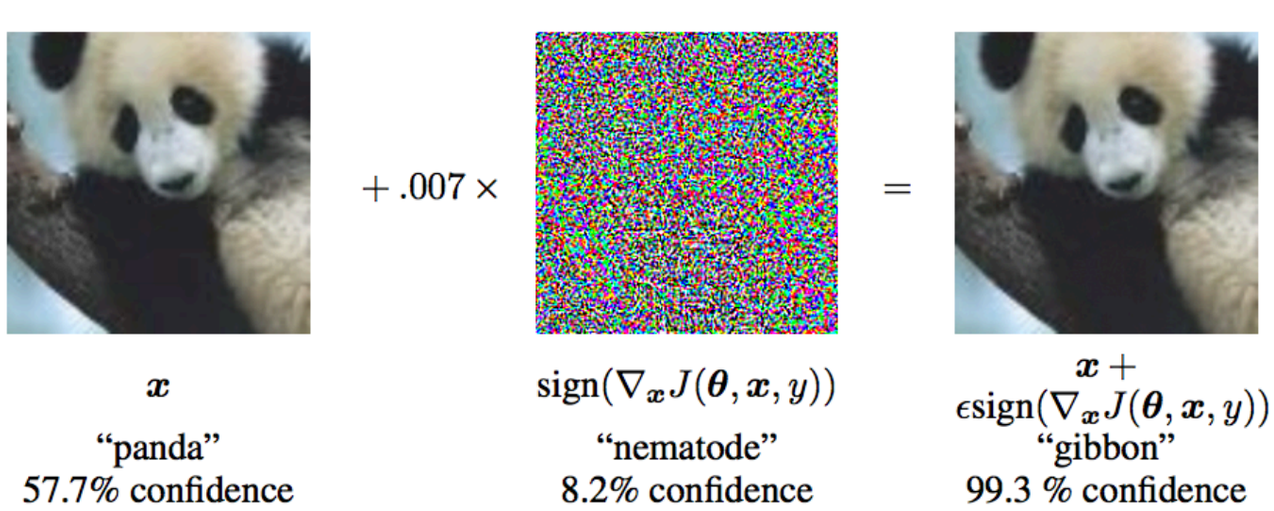
Hiromu Yakura[*†], Youhei Akimoto[*†], Jun Sakuma[*†]

[*] University of Tsukuba, Japan
[†] RIKEN Center for Artificial Intelligence Project, Japan

AAAI 2020

Background: Risk of abusing AI using adversarial example

- It is known that **we can mislead ML models** by intentionally adding a small noise to the inputs.
  - In this case, the right image is classified as gibbon.



Using this adversarial example, we can abuse AI-based systems without being noticed by humans.

Goodfellow, I. J., Shlens, J., & Szegedy, C.: Explaining and harnessing adversarial examples. In *Proc. of ICLR.* (2015)

Proposal: Attacking self-driving cars with moth-like stickers

- This mechanism is also applicable to deceive self-driving cars.
  - **What if the cars recognize a STOP sign as Speed 80?**
    - This example can cause such a mistake
      but looks too suspicious not to be noticed by humans.

[Chen+, '18]

64x64 pixel

128x128 pixel

Proposed

We showed that these moth-like stickers can mislead ML models without making humans feel suspicious.

S. Chen, et al. Shapeshifter: Robust physical adversarial attack on faster RCNN object detector. ECML PKDD 2018.

Failure-Resistant Intelligent Interaction

Basic idea: How to overcome the imperfection of AI models

No matter how much technical improvements we make,
AI-based systems will make mistakes.

Design an interaction in which humans and AI-based systems can collaborate effectively even when AIs make a mistake

Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation

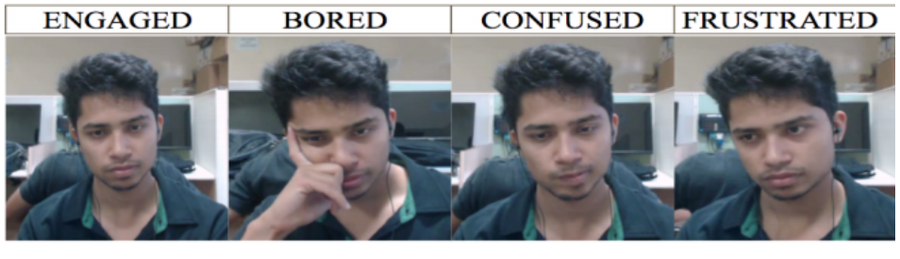
Riku Arakawa† and Hiromu Yakura‡
(equal contribution)

† The University of Tokyo, Japan
‡ University of Tsukuba, Japan

ACM CHI 2021

Background: Limitation of alerting intervention

- ML models can detect the moment when people are not engaging.



[Gupta+, '16]

  - It is possible to alert distracted students in video lectures.
- But, **misinformed alerts caused by false positives disrupt the students**, which leads them to distrust the system.

- What is the best way to intervene in distracted students when **we have the risk of false positive alerts**?

A. Adams, et al. Mindless computing: designing technologies to subtly influence behavior. *UbiComp* (2015).

Demo: Mindless Attractor

Proposal: Mindless Attractor

Humans often intentionally or unconsciously change the volume or pitch of speech to draw listeners' attention.

- **Our system changes the volume or pitch for a moment** to draw attention without consuming conscious awareness.
  - This is **machine learning-friendly because it won't be frustrating even when activated by false-positive detection**.
- In our study, we confirmed its effectiveness to help refocus even when the students are unconscious of the changes.

CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models

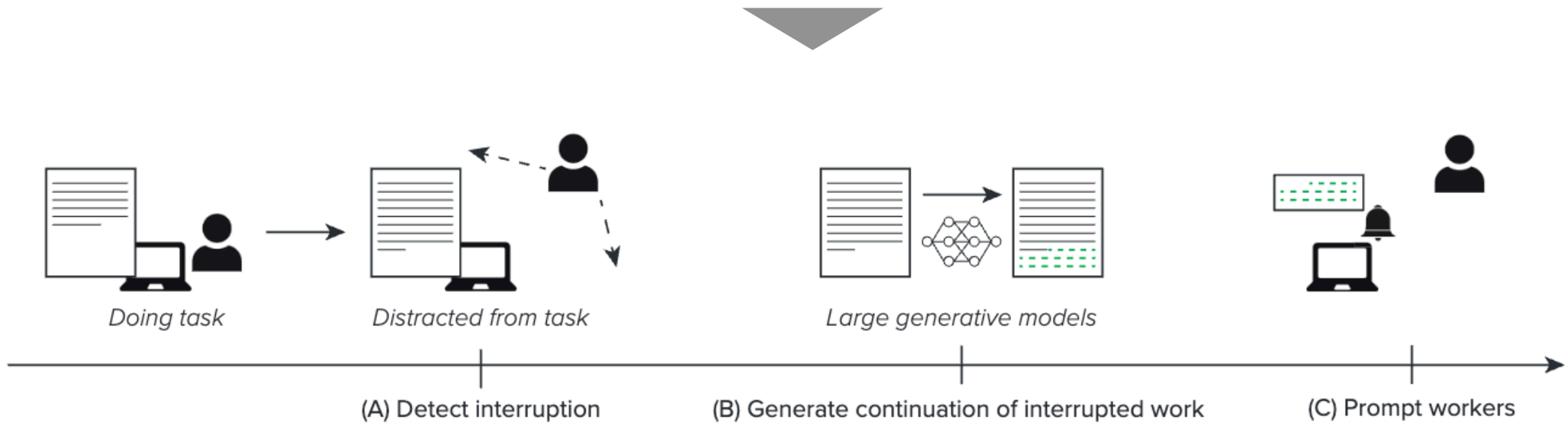
Riku Arakawa[†*], Hiromu Yakura[‡*], and Masataka Goto[¶]
([*]equal contribution)

[†] The University of Tokyo, Japan
[‡] University of Tsukuba, Japan
[¶] National Institute of Advanced Industrial Science and Technology, Japan

ACM CHI 2023

Background: Use generative models to avoid procrastination

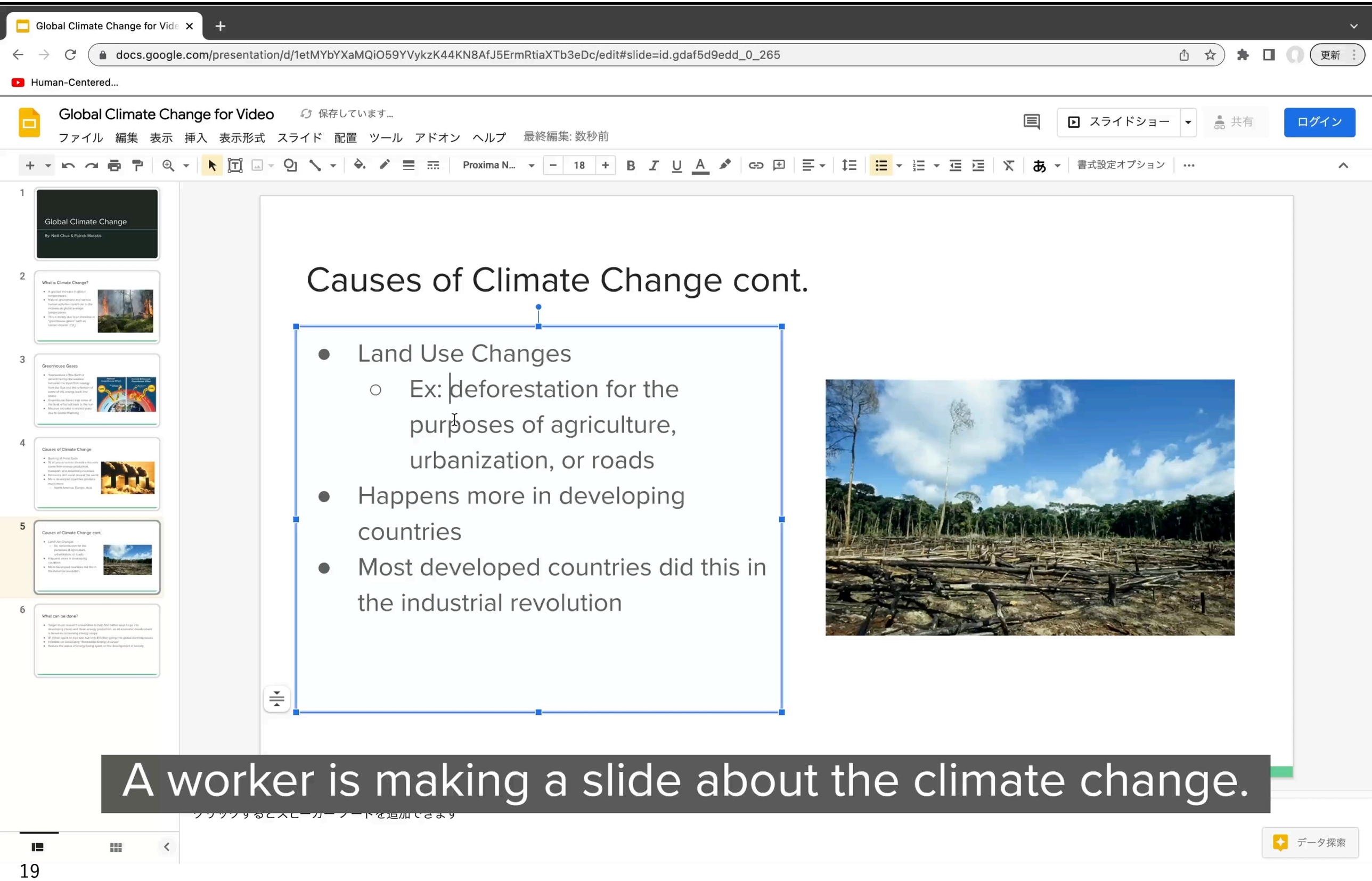
- Large language models not always perfect to alternate our intellectual tasks (e.g., writing domain-specific documents).
  - Is it possible to benefit from them in such cases?



- We assumed that **they can help people avoid procrastination**.
  - **Even imperfect content generated can be used to guide users' interests to their tasks.**

A. Adams, et al. Mindless computing: designing technologies to subtly influence behavior. *UbiComp* (2015).

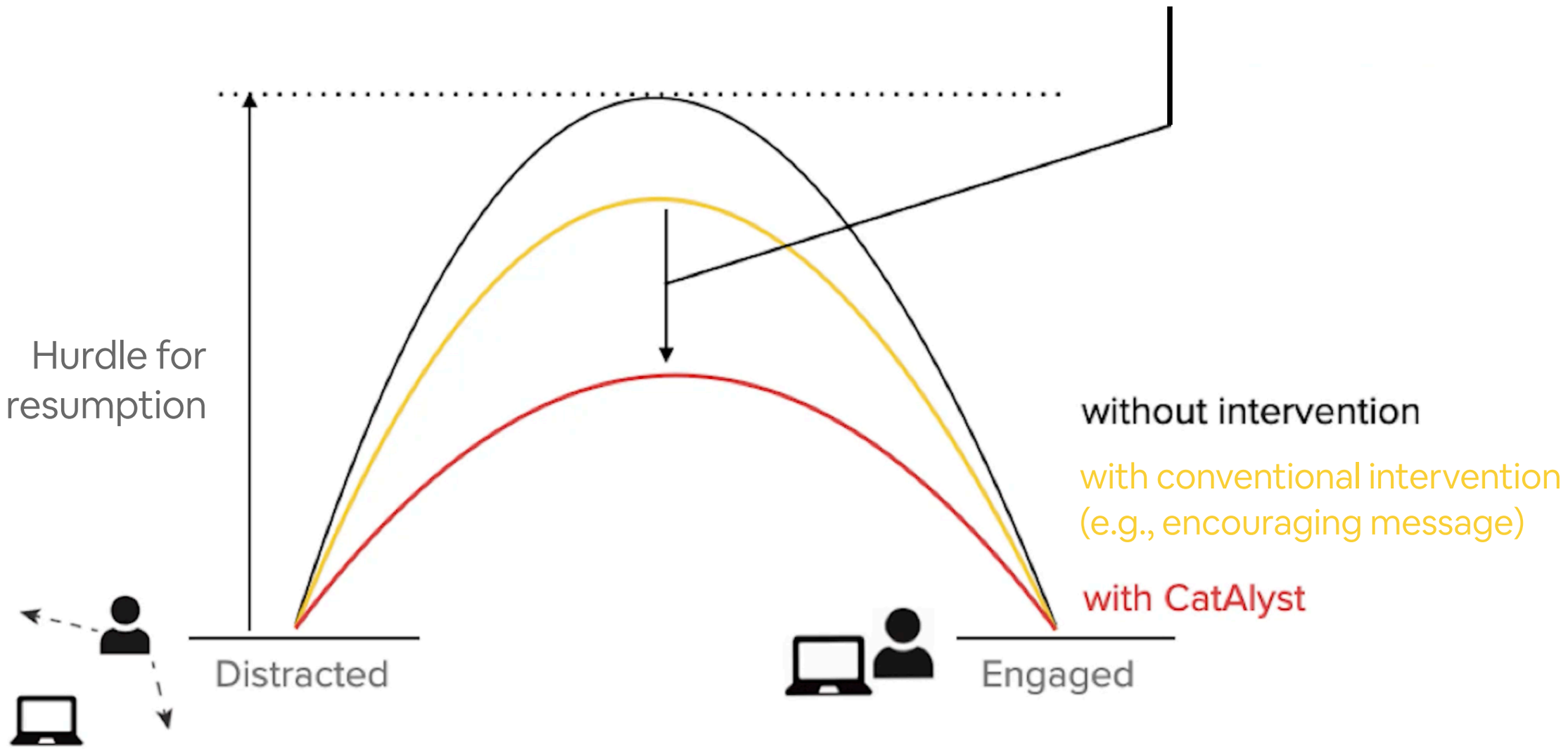
Demo: CatAlyst



Proposal: CatAlyst

Generated contents that are not sufficient to alternate our tasks can serve as a context-aware intervention to draw our interests.

- In addition, they sometimes provide inspirational ideas.



Evaluation: CatAlyst

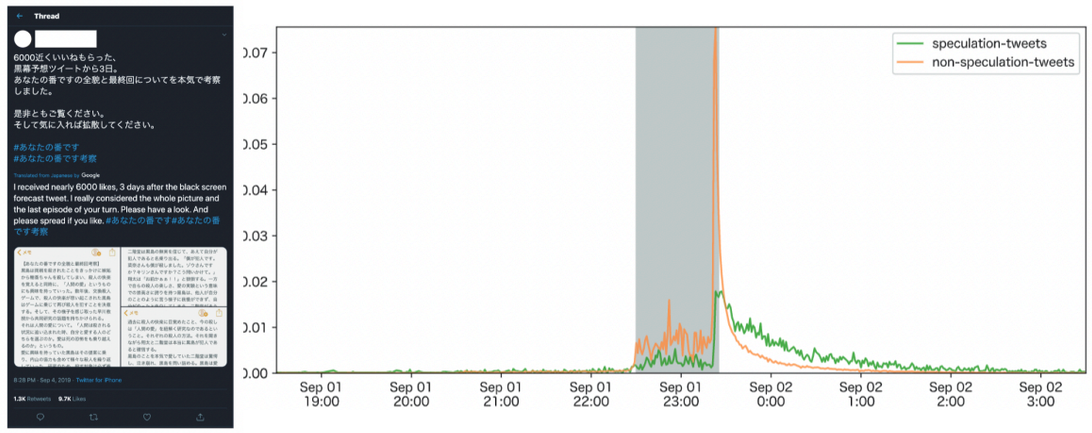
- We evaluated its effectiveness from various aspects:
  - **Ignorance rate**: a rate of notifications ignored
  - **Interest retrieval time**: duration passed before resumption
  - **Progress after resumption**: progress made within 45 s after resumption
  - **Total time**: time spent on completing the assigned task
  - **Subjective quality**: product quality rated by crowdworkers
  - **Cognitive load**: NASA-TLX score responded by participants

CatAlyst exhibited better performance than conventional intervention, **except for total time**

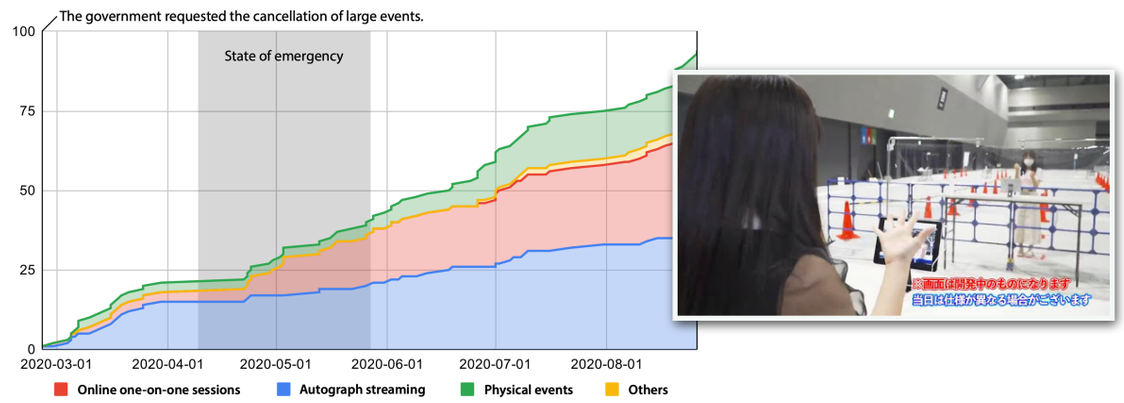
Summary: Failure-Resistant Intelligent Interaction

- My research approach is:
  1. **Clarify the limitation of the current ML models**
  2. **Find a situation in which even such models are beneficial**

- For this purpose, I also conducted some qualitative research focusing on new grassroots situations of using computers.



Speculation-driven new tweeting behavior for TV drama series



Use of computer-mediated communication by Japanese idol groups under COVID-19

Toward my thesis

- My thesis will extract design principles from these projects:

  Q. What is the meta strategy to find situations
  in which even imperfect AIs are beneficial?

  Q. How do the principles relate to other HCI concepts?

  ⋮

- Happy to hear your comments and questions!

hiromu.yakura@aist.go.jp

@hiromu1996

【Case Study】

Honorable Case Study Recognition

AI for human assessment: What do professional assessors need?

Riku Arakawa[†]

Carnegie Mellon University

Hiromu Yakura[†]

University of Tsukuba / AIST

[†] Equal contribution

In collaboration with ACES Inc.

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

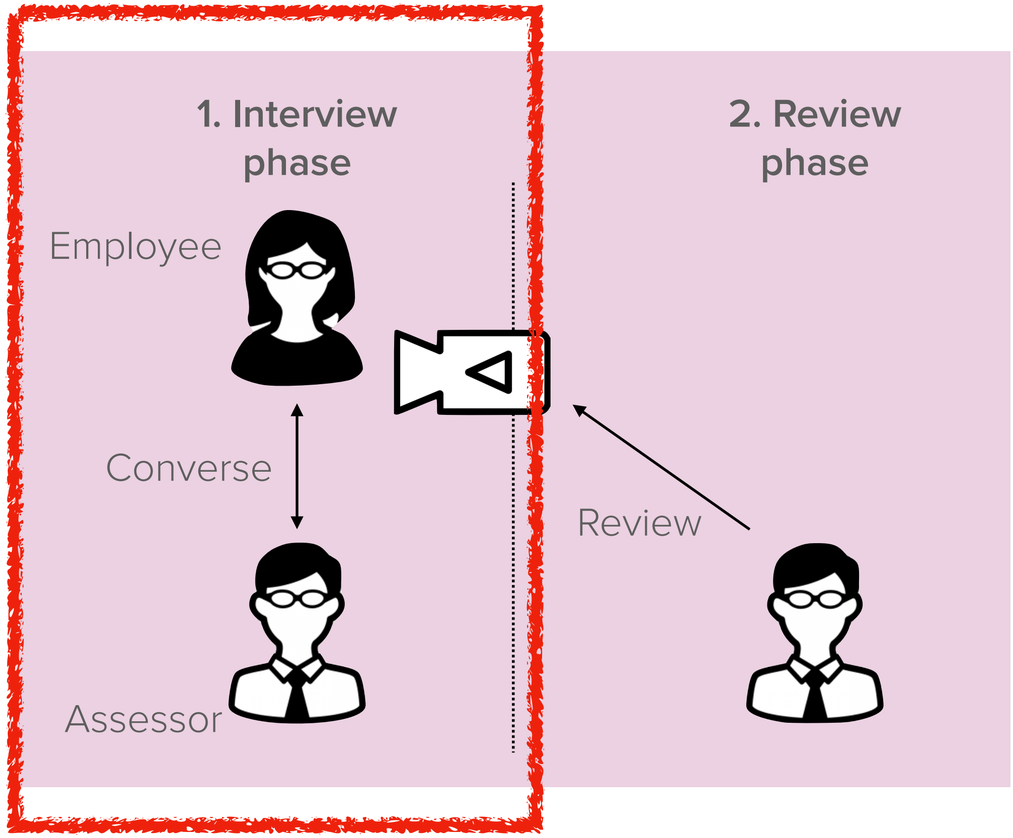
Human assessment

Evaluate candidates regarding their suitability for certain types of employment, mostly through interviews by professional assessors

Human assessment

Human assessment typically consists of two phases:

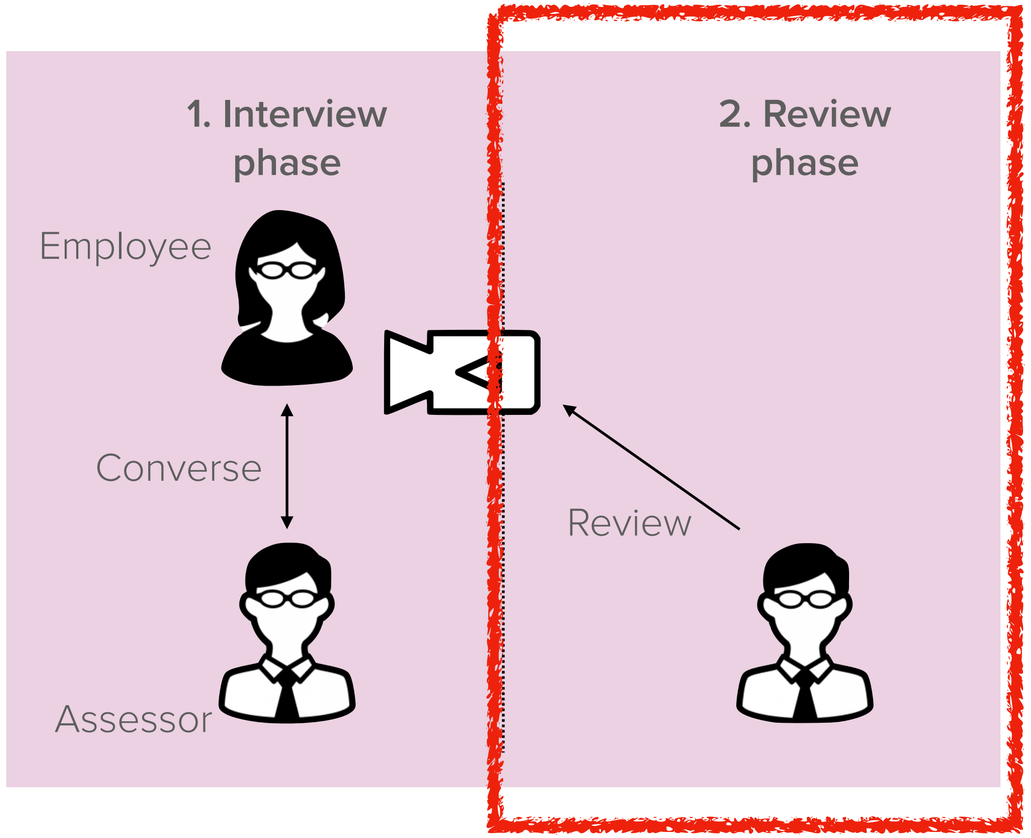
1. **Interview phase**
   - An assessor plays a certain role in an one-on-one interview.
   - The conversation is video-recorded.
2. **Review phase**
   - The assessor playbacks the recorded video.
   - The video is used to find verbal and nonverbal cues for evaluating the employee as a manager.



Human assessment

Human assessment typically consists of two phases:

1. Interview phase
   - An assessor plays a certain role in an one-on-one interview.
   - The conversation is video-recorded.
2. **Review phase**
   - The assessor playbacks the recorded video.
   - The video is used to find verbal and nonverbal cues for evaluating the employee as a manager.



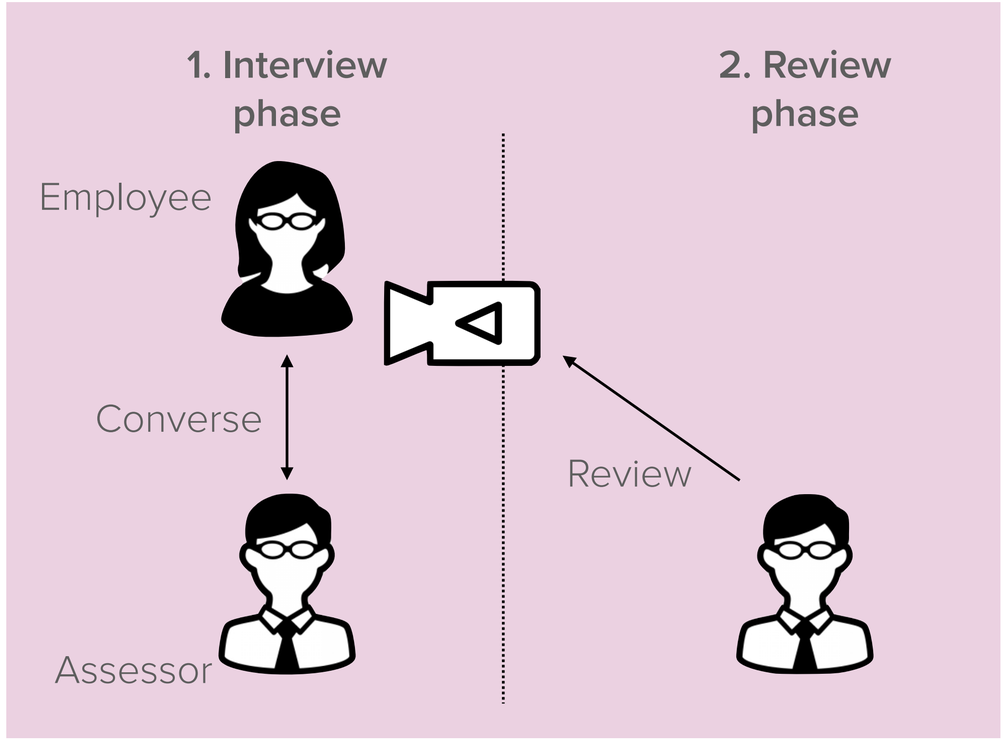
How can we support human assessment with AIs?

[1] Arakawa and Yakura, REsCUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection, CHI'19

[2] Arakawa and Yakura, INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching, CHI'20

Initial workshop

- We conducted a workshop with 2 professional assessors:
- Difficulties
  - The review phase is time-consuming ⏳
  - Assessors’ subjectivity can lead to a wrong decision 🧐

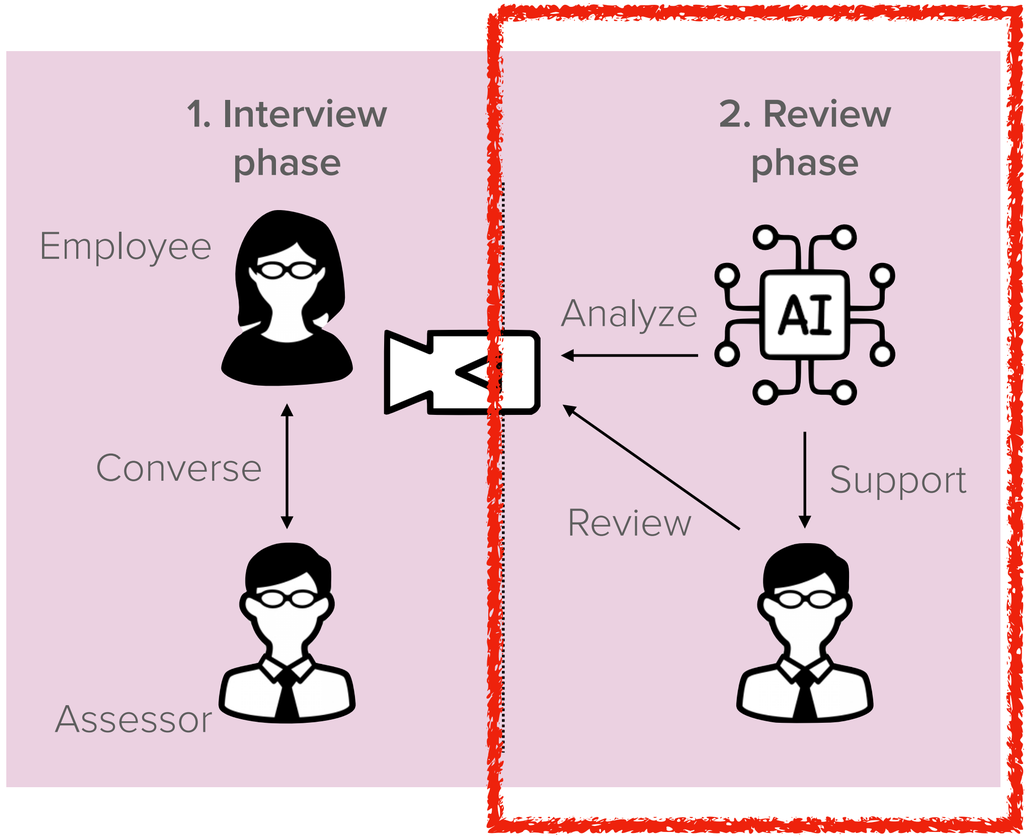


Initial workshop

- We conducted a workshop with 2 professional assessors:
- Difficulties
  - The review phase is time-consuming ⏳
  - Assessors' subjectivity can lead to a wrong decision 🧐

⬇

Q: How can AI systems support professional assessors' decision-making in review-phase?



System requirements

- The assessors were **skeptical about AI-based end-to-end decision making** because human assessment should consider various factors specific to each employee.
  - They are highly human-contextual and difficult to be captured by computers.

System requirements

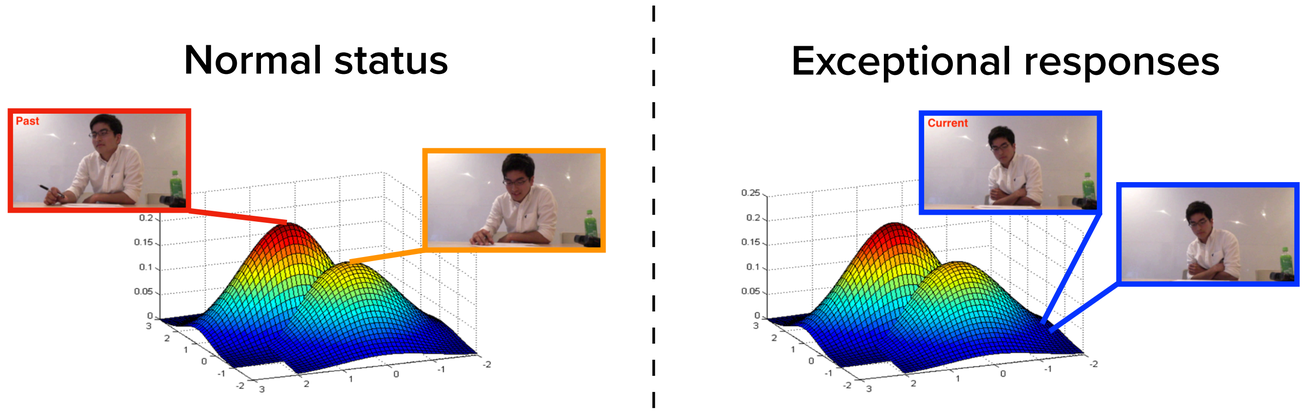
- The assessors were **skeptical about AI-based end-to-end decision making** because human assessment should consider various factors specific to each employee.
  - They are highly human-contextual and difficult to be captured by computers.
- The assessors **expected AI systems to help them not miss important behavior cues due to their subjectivity or mental demands.**
  - Then, the assessors can revise their judgment by taking the contextual meaning of such AI-detected cues into consideration.

▼

Hypothesis: Separating observation (by AI) and judgment (by professionals)

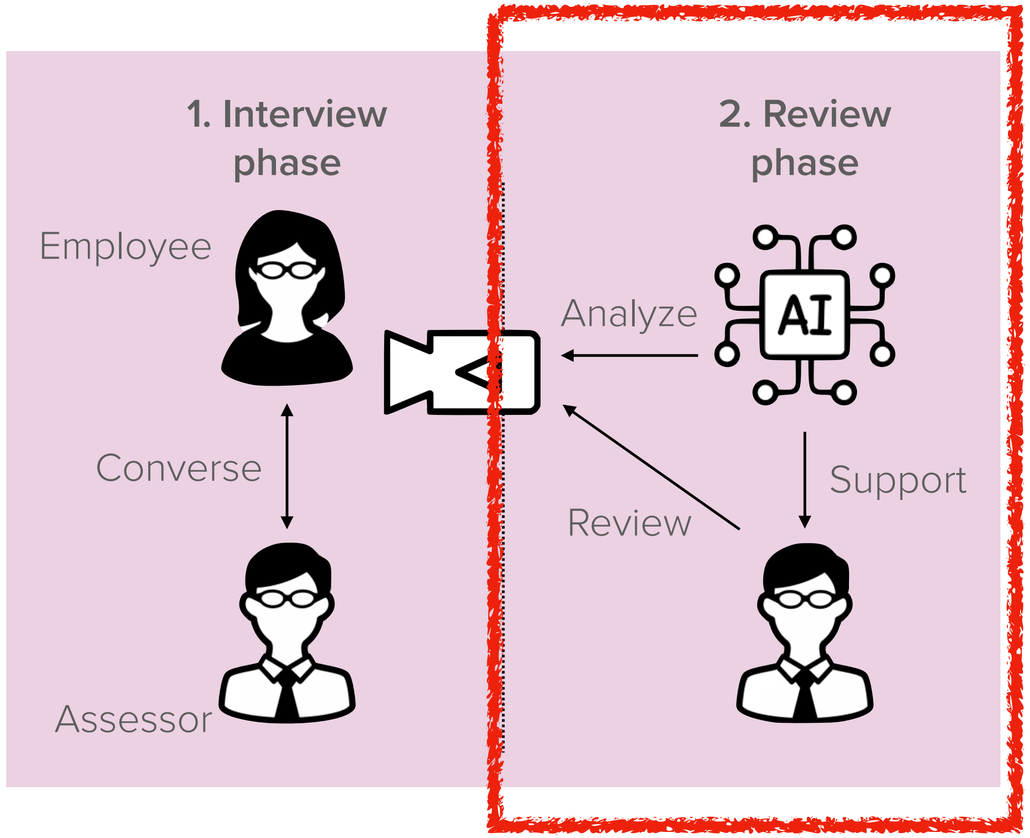
Feasibility study

- We adopted nonverbal behavior analysis algorithm, REsCUE [1] used in executive coaching.
  - It can extract anomalous cues of people in conversation.
  - It provides clear visualization of the cues based on GMM.



Feasibility study

- 20 interview videos
- Two assessors annotated important scenes manually
- Our algorithm also extracts anomaly scenes



Findings

- We examined the agreement between the algorithm and the assessors

  and found that the algorithm does not completely replicate their annotation.

  - The discrepancy was attributed to both **false-positive detection** and assessors' subjectivity.

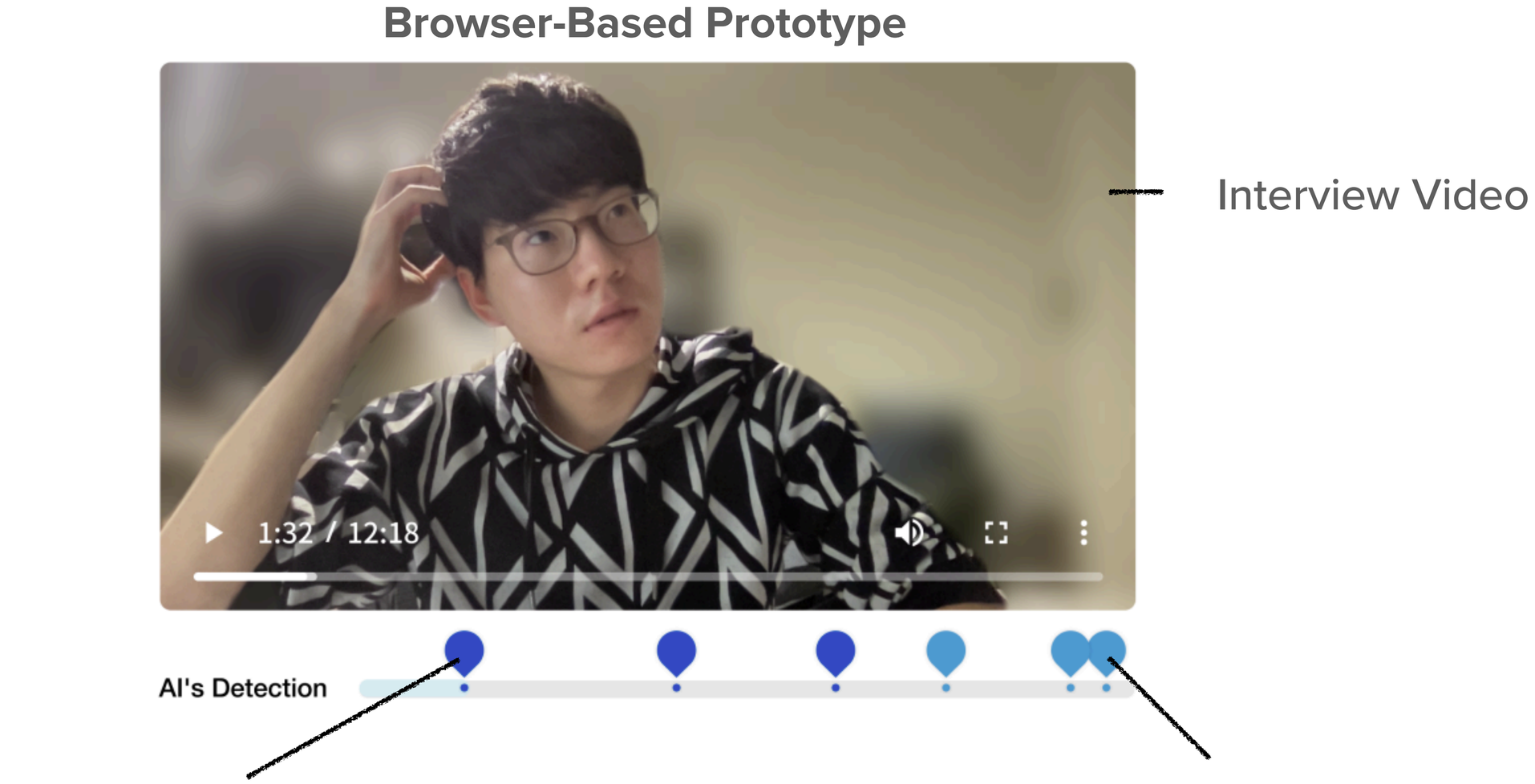
Findings

- We examined the agreement between the algorithm and the assessors and found that the algorithm does not completely replicate their annotation.
  - The discrepancy was attributed to both **false-positive detection** and assessors' subjectivity.
- However, the assessors found that the algorithm would facilitate their assessment.
  - **The interpretable output of the anomaly-detection-based algorithm guided them to infer the reason behind the detection, questioning their decisions.**
  - It helped maintain the assessors' trust in the case of false-positives 🤝

▼

A: The separation contributed to the trust in this highly contextual domain.

Usability study — Prototype



Usability study — Procedure

- 6 professional assessors who had not participated in our first study:
  - 2 junior assessors, 4 senior assessors
- Each assessor reviewed randomly chosen four videos with the prototype.
- We conducted semi-structured interviews after they reviewed all videos to ask about usability of the prototype.

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “rethought such cases but could easily resolve the conflict by referring to other signals”

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “rethought such cases but could easily resolve the conflict by referring to other signals”

Usability study — Result

- **Deepened quality of assessment**
  - enhanced objectivity (← false-positive)
  - gain confidence (← true-positive)
  - not lose confidence (← false-negative)

  “*rethought such cases but could easily resolve the conflict by referring to other signals*”

Usability study — Result

- **Room for improvement of the prototype**
- **Potential use scenarios**

Please refer to the paper!

Lessons learned

- It is neither recommended nor feasible to train an AI model that replicates assessors' decision-making process.
  - Inevitable inconsistency among their processes

    (= different assessors look at different cues while having the same assessment result)
  - Lack of interpretability and validity in its output.

- Our design of **separating observation and judgment** is a promising approach in such highly contextual domains.
  - Importantly, our goal is not replacing human decision, but helping them.

Lessons learned

- It is neither recommended nor feasible to train an AI model that replicates assessors' decision-making process.
  - Inevitable inconsistency among their processes

    (= different assessors look at different cues while having the same assessment result)
  - Lack of interpretability and validity in its output.

- Our design of **separating observation and judgment** is a promising approach in such highly contextual domains.
  - Importantly, our goal is not replacing human decision, but helping them.

目次

Doctoral Consortium に採択されるまで

- 以下の4つの書類を用意して、PCS で投稿
  - 6ページの Research Statement
  - DC で何を得たいか、どう DC に貢献できるかのエッセイ
  - 指導教員からの推薦状
  - Curriculum Vitae (履歴書)
- Research Statement については過去の参加者分が
  ACM Digital Library で公開されているので参考にするとよい
  - 複数の研究をまとめている人から、1つの研究について
    あるいはその計画のみの人までフェーズに応じて様々

Doctoral Consortium に採択されるまで

- 2人の査読があり、それを元にして
  参加可否が決まる
  - 内容のクオリティに加えて
    博士課程として有効活用できそうか
    参加タイミングの適切性も見られる
- CHI '23 では115件の提出があり
  20件がオンライン開催枠に
  10件がオフライン開催枠に採択された

2AC review (reviewer 1)

Timing

A little too late but still fine to be in DC

Submission Quality

5.0 - Strongly agree

2AC review (reviewer 2)

Timing

Too late

Submission Quality

3.0 - Neutral

Case Study に採択されるまで

- フルペーパーと似たような形で査読がある
  - シングルカラムで 4-10p と短めなのと
    査読者が2名であることが大きな違い

Reviewer	Role	Score
Reviewer 2	1AC	5.0
Reviewer 1	reviewer	4.0

- 共著者の荒川と一緒に、1年ほど前から戦略的に投稿を準備
  - ACM CHI '19 で発表した手法の実応用を進める中で
    せっかくであればそこでの発見を共有できればと考えた
  - ただ、技術的新規性が大きい訳では無いということもあり
    どうしようかと考える中で Case Study に思い至る

Case Study に採択されるまで

- ACM CHI '22 のワークショップで In progress 版を発表
- Case Study のスコープに合わせて改訂の上で投稿

- Design to support a specific type of experience, discussing its rationale and lessons learned
- Research of a specific domain, user group, or experience, discussing its insights and lessons learned
- Domain-specific topics, significantly lesser known but essential fields of interest
- Management and strategy of research (either academic research or user research) and design in organizations
- Pilot studies preceding and informing larger-scale investigations
- Application, critique, or evolution of a method, process, or tool
- Innovation through Research or Design (disruptive or otherwise)
- Practical issues associated with HCI Teaching and Learning in education, training, or knowledge sharing

- Lessons Learned というセクションを追加して
  HCI コミュニティへの知見という点の接続性を追加
- HCI の応用領域としての可能性をより強調

目次

Doctoral Consortium に参加するまで

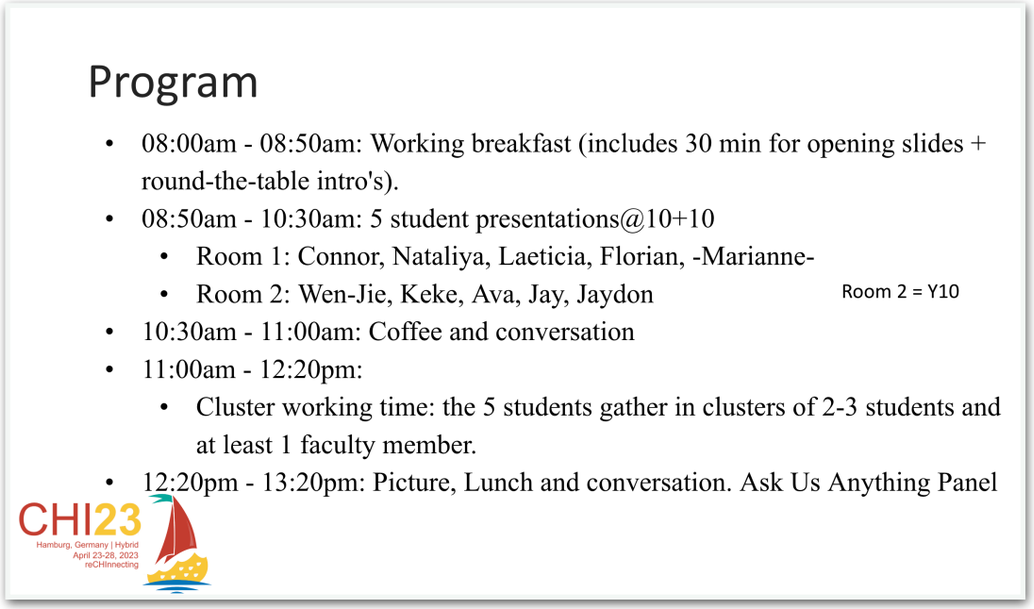
- 事前にメンバーリストが届き、ピアメンタリングがスタート
  - アメリカ 7名, イギリス 4名, フランス・ドイツ 2名に加え
    カナダ・イタリア・日本・南アフリカ・ルクセンブルク
- 提出の内容だけでなく、地理的多様性を考慮して
  選んだ（らしい）

Doctoral Consortium に参加するまで

- 事前にメンバーリストが届き、ピアメンタリングがスタート
  - 1人にメンターになってもらい、別の1人のメンターになる
    - プレゼンをシェアして Zoom で練習＆フィードバック
    - **当日時点で知っている人が2人いるというのがありがたい**
- また、Committee の先生がメンターとしてアサインされて事前コミュニケーションできる
  - University College London の Prof. Brumby が担当してくれることに

Doctoral Consortium の参加の様子

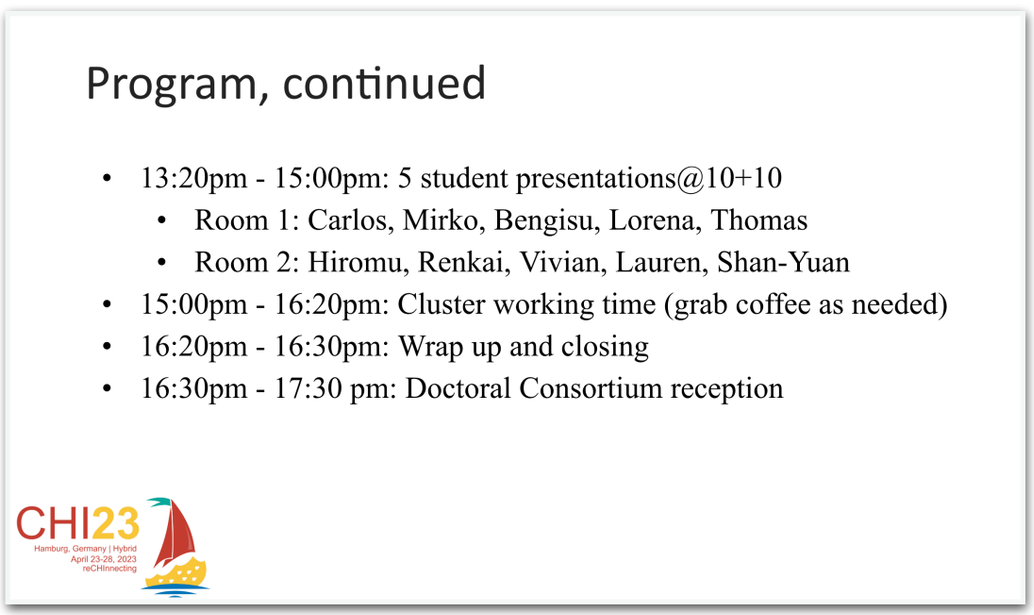
- 前日に参加者ディナーがあり、そこからスタート
- 当日は朝から晩まで
  インテンシブなスケジュール
  - 各自が10分プレゼン＋
    10分質疑
  - そのあとはメンターの先生を
    囲みながら3人くらいで
    ディスカッションタイム
  - ランチ中も Ask Us Anything


Program

- 08:00am - 08:50am: Working breakfast (includes 30 min for opening slides + round-the-table intro's).
- 08:50am - 10:30am: 5 student presentations@10+10
  - Room 1: Connor, Nataliya, Laeticia, Florian, -Marianne-
  - Room 2: Wen-Jie, Keke, Ava, Jay, Jaydon
- 10:30am - 11:00am: Coffee and conversation
- 11:00am - 12:20pm:
  - Cluster working time: the 5 students gather in clusters of 2-3 students and at least 1 faculty member.
- 12:20pm - 13:20pm: Picture, Lunch and conversation. Ask Us Anything Panel

Room 2 = Y10

CHI23 Hamburg, Germany | Hybrid April 23-28, 2023 reCHInnecting


Program, continued

- 13:20pm - 15:00pm: 5 student presentations@10+10
  - Room 1: Carlos, Mirko, Bengisu, Lorena, Thomas
  - Room 2: Hiromu, Renkai, Vivian, Lauren, Shan-Yuan
- 15:00pm - 16:20pm: Cluster working time (grab coffee as needed)
- 16:20pm - 16:30pm: Wrap up and closing
- 16:30pm - 17:30 pm: Doctoral Consortium reception

CHI23 Hamburg, Germany | Hybrid April 23-28, 2023 reCHInnecting

Doctoral Consortium の参加の様子

- 質疑＆フィードバックもかなりしっかりしたやりとり
  - Prof. Burnett: 研究者の採用面接だと思ってやります
- プレゼンのフォントサイズからストーリー構成まで
  研究室内の発表練習かのようにコメントをもらえる
  - **Q. あなたの HCI に対する貢献を1文でいうとなんですか？**

（参加者の間では賛否が分かれていたが
個人的には芯を食ったいい質問だと思う）

＋

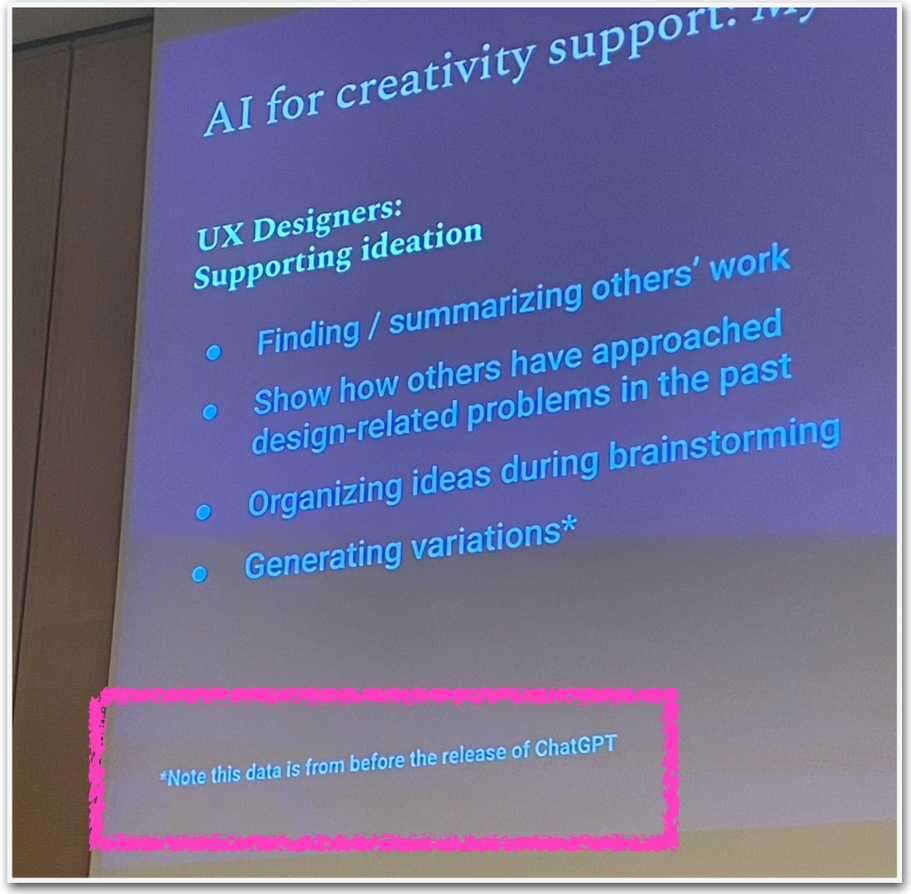
- 期間中もポスターセッションでの発表の機会を貰える

Doctoral Consortium に参加してよかったこと

- なによりもいろんな人からフィードバックをもらえる
  - Neurodiversity とつなげるといいと思う、など
- 改めて HCI の分野の幅広さを体感できる
  - VR や AI から色覚異常の支援や自画撮り防止なども
- CHI コミュニティの中に近い世代の知り合いが増える
  - DC 後もご飯に行ったり、パーティーで何時間も話したり

Case Study の参加の様子

- フルペーパーと同じようにパラレルセッションの1つとして3-4本の発表が行われる
  - 発表したセッションでは Google と Microsoftからの Case Study があった
  - Google による UX Designer が AI ツールをどう活用できるか調べた発表の注釈が面白かった



まとめ

- Doctoral Consortium はいいぞ！
  - 実績よりも教育やコミュニティ形成に焦点をおいており
    日本からの投稿は通りやすそう（？）
  - いろんな研究アプローチの人と交流し、仲良くなることで
    学べることや発見も大きい
- Case Study もおすすめ
  - フルペーパーと毛色は違うが、Lessons Learned があれば
    しっかり評価してもらえる
- **聞きたいことや見たい資料などあれば**
  気軽に連絡ください！

hiromu.yakura@aist.go.jp
@hiromu1996