Partnering with Explainable AI

Michael Winikoff

Te Herenga Waka - Victoria University of Wellington

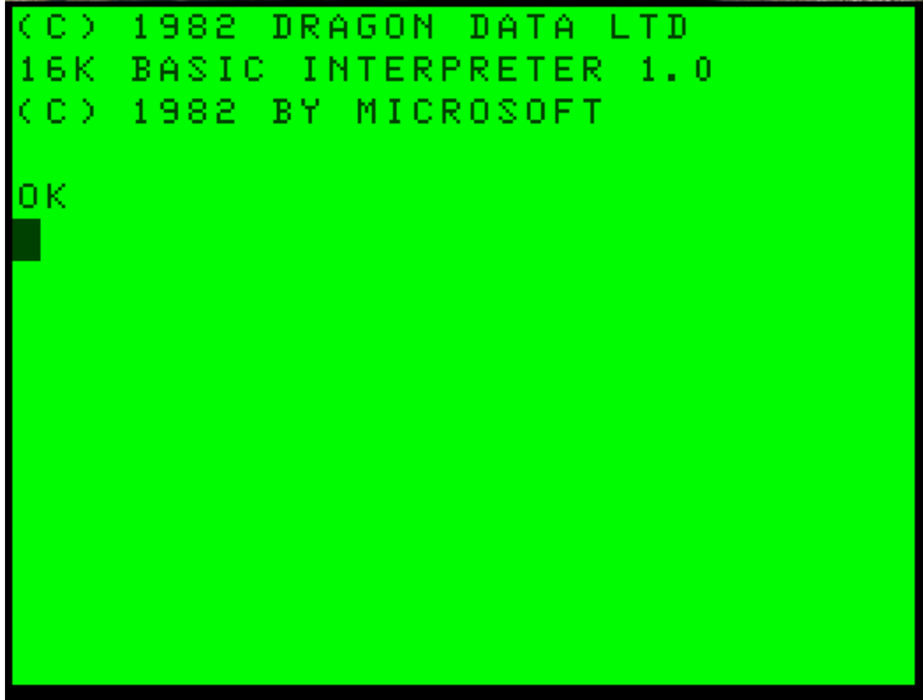
New Zealand

1897

VICTORIA UNIVERSITY OF
WELLINGTON
TE HERENGA WAKA

Introductions …

- Background:
  - Programming, especially *declarative*: functional (1993), logic programming (1994-97)
  - Formal methods (1998-1999)
  - Engineering autonomous systems (1999-)
- Interests: software engineering (process, notations, languages, …), explanation, assurance, trust, logic & formal methods
- Non-work interests: my family, reading, music (choir, piano, composing)



```
(C) 1982 DRAGON DATA LTD
16K BASIC INTERPRETER 1.0
(C) 1982 BY MICROSOFT

OK
```

Agenda: briefly cover four papers …

1. Partnering with AI: the case of digital productivity assistants (JRSNZ, 2023)
2. Artificial Intelligence and the Right to Explanation as a Human Right. (IEEE Internet Computing, 2021)
3. A Scoresheet for Explainable AI (AAMAS, 2025)
4. Evaluating Contrastive Explanations of Autonomous Systems. (under review)

Over-arching narrative: role of AI as partner, need for explanation, and how to specify and assess explanation.

Partnering with AI: the case of digital productivity assistants

JRSNZ, 2023, joint work with Jocelyn Cranefield, Yi-Te Chiu,
Yevgeniya Li, Cathal Doyle & Alex Richter

doi: 10.1080/03036758.2022.2114507

Journal of the Royal Society of New Zealand, 53:1, 95-118

Also: Michael Winikoff, Jocelyn Cranefield, Jane Li, Alexander Richter, and Cathal Doyle. The Advent of Digital Productivity Assistants: The Case of Microsoft MyAnalytics, HICSS 2021.

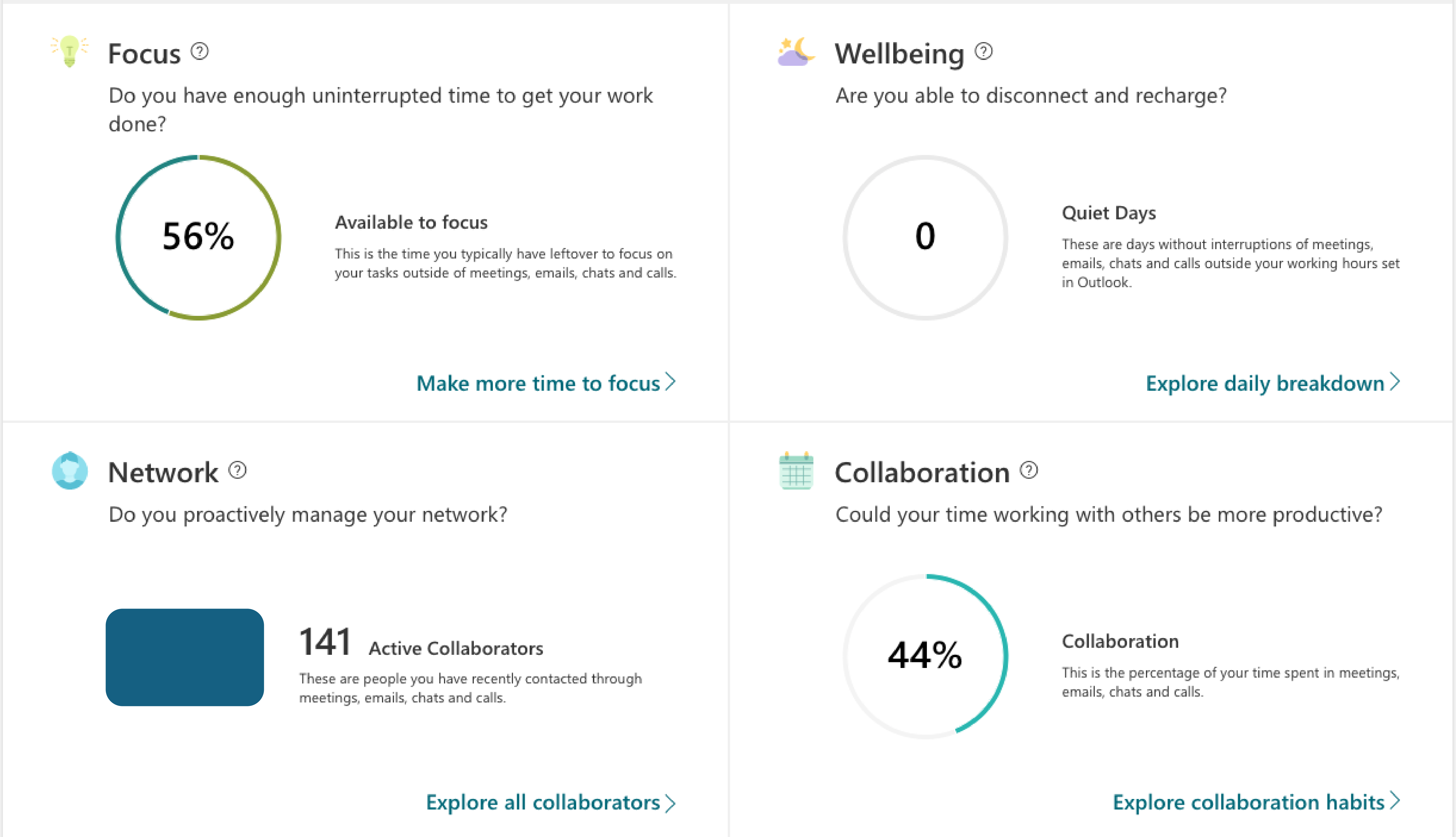
- **Aim**: Understand human-AI relationship in practice

- **Context**: productivity and well-being at work: Digital Productivity Assistant (DPA)

- **Research Questions**:
  1. What *opportunities* do DPA’s offer to improve productivity & wellbeing?
  2. What *barriers* are experienced?

Digital Productivity Assistant (DPA)

- Use personal **workplace data** to provide **insight** and **persuasion** to help workers improve their productivity and wellbeing.
- Example: Microsoft MyAnalytics (MMA) (since renamed Viva Insights)
  - **Data**: email, chats, calendar details, activity, etc.
  - Provide **overview of behavior**, and **actionable advice** ("AI-powered suggestions")
  - Persuasion: e.g. rhetorical questions ("do you have enough uninterrupted time …"), normative suggestions (report on how any meeting invites sent less than 24 hours before meeting)



Focus ⓘ

Do you have enough uninterrupted time to get your work done?

56%

Available to focus

This is the time you typically have leftover to focus on your tasks outside of meetings, emails, chats and calls.

Make more time to focus >

Wellbeing ⓘ

Are you able to disconnect and recharge?

0

Quiet Days

These are days without interruptions of meetings, emails, chats and calls outside your working hours set in Outlook.

Explore daily breakdown >

Network ⓘ

Do you proactively manage your network?

[redacted]

141 **Active Collaborators**

These are people you have recently contacted through meetings, emails, chats and calls.

Explore all collaborators >

Collaboration ⓘ

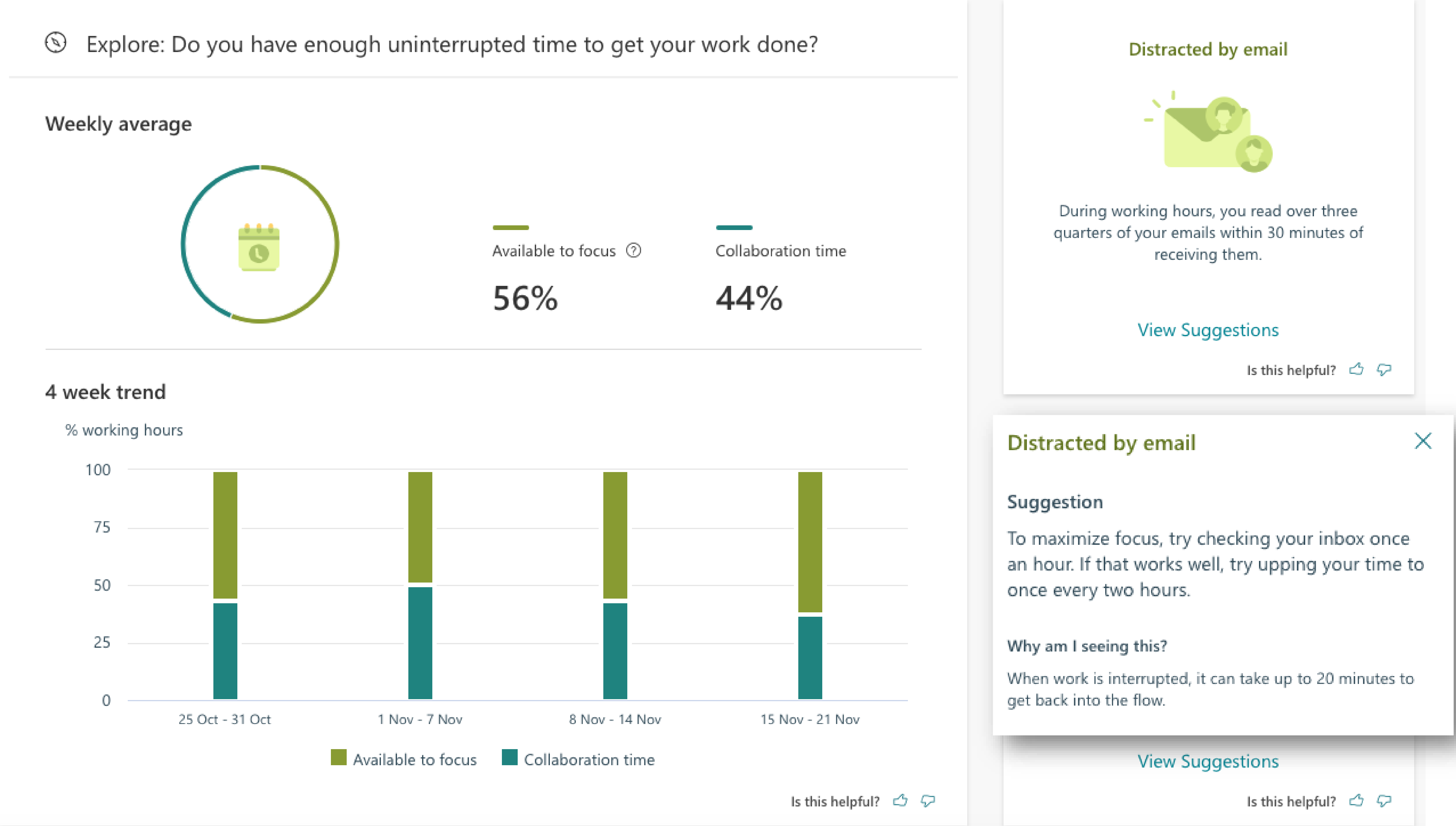
Could your time working with others be more productive?

44%

Collaboration

This is the percentage of your time spent in meetings, emails, chats and calls.

Explore collaboration habits >



Explore: Do you have enough uninterrupted time to get your work done?

Weekly average

Available to focus ⓘ

56%

Collaboration time

44%

4 week trend

Is this helpful?

Distracted by email

During working hours, you read over three quarters of your emails within 30 minutes of receiving them.

View Suggestions

Is this helpful?

Distracted by email

Suggestion

To maximize focus, try checking your inbox once an hour. If that works well, try upping your time to once every two hours.

Why am I seeing this?

When work is interrupted, it can take up to 20 minutes to get back into the flow.

View Suggestions

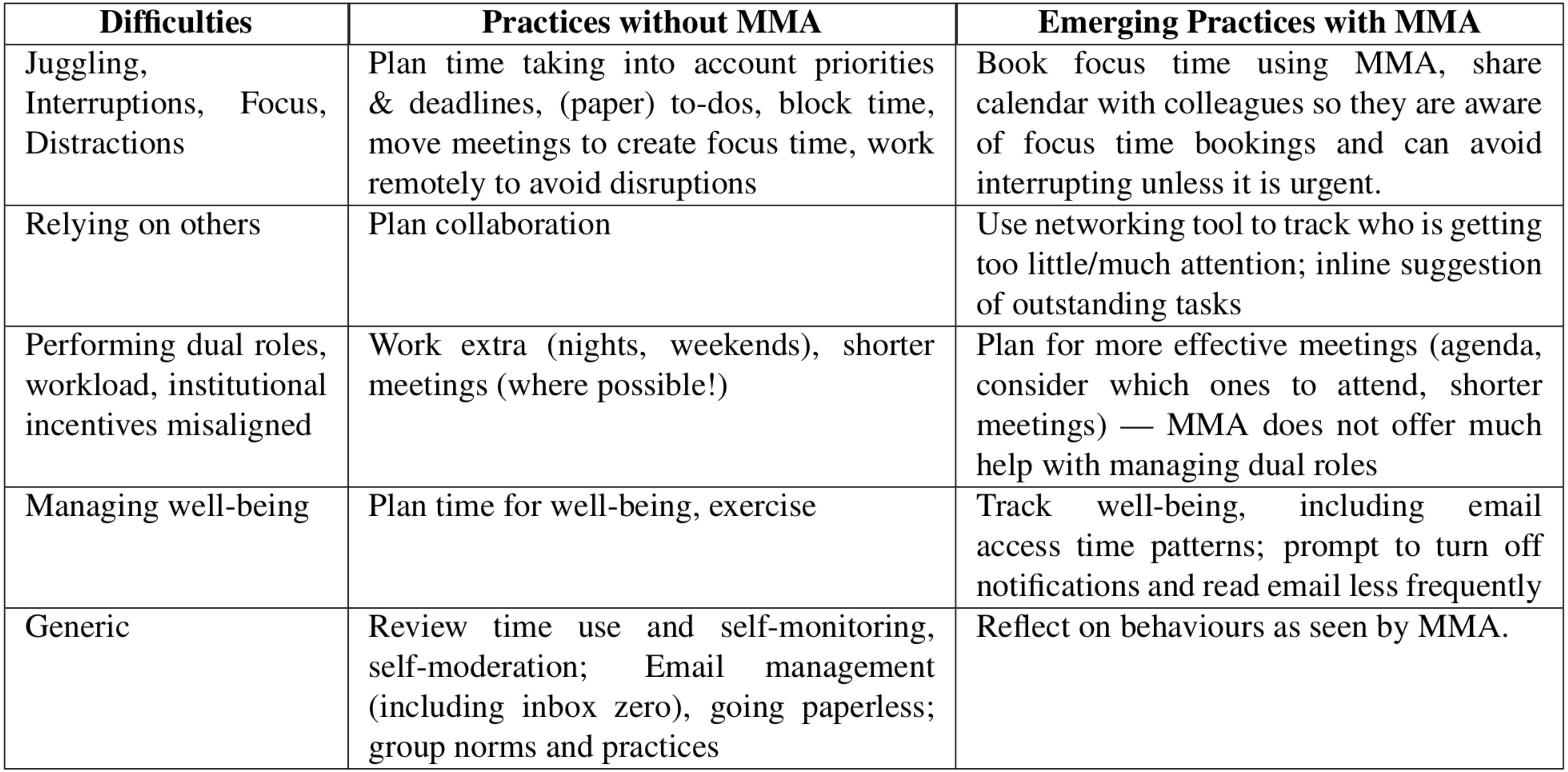
Is this helpful?

Methodology

- Data set 2: in-depth interviews with 28 workers in 3 organisations (12 academics, 4 academic support, 12 IT professionals)
    1. Explore general approach to self-regulation of time at work
    2. Explore their understanding of Microsoft MyAnalytics and its value
    3. Asking for comment on the most recent report from MyAnalytics
- Data analysed iteratively, inductively to develop themes
- Also (first) reviewed own use of MyAnalytics to build understanding of its functionality and user interaction

RQ1: What *opportunities* do DPA's offer to improve productivity & wellbeing?

Most participants did not make use of DPA affordances, but a few did



Difficulties	**Practices without MMA**	**Emerging Practices with MMA**
Juggling, Interruptions, Focus, Distractions	Plan time taking into account priorities & deadlines, (paper) to-dos, block time, move meetings to create focus time, work remotely to avoid disruptions	Book focus time using MMA, share calendar with colleagues so they are aware of focus time bookings and can avoid interrupting unless it is urgent.
Relying on others	Plan collaboration	Use networking tool to track who is getting too little/much attention; inline suggestion of outstanding tasks
Performing dual roles, workload, institutional incentives misaligned	Work extra (nights, weekends), shorter meetings (where possible!)	Plan for more effective meetings (agenda, consider which ones to attend, shorter meetings) — MMA does not offer much help with managing dual roles
Managing well-being	Plan time for well-being, exercise	Track well-being, including email access time patterns; prompt to turn off notifications and read email less frequently
Generic	Review time use and self-monitoring, self-moderation; Email management (including inbox zero), going paperless; group norms and practices	Reflect on behaviours as seen by MMA.

RQ2: What **barriers** are experienced?

1. **Perceived inaccuracy** of the tool
   - e.g. tool cannot see *ad hoc* meetings
   - e.g. tool assumes meeting with no participants = focus time (but lecture?), and meeting with participants is collaboration (but writing group?) … also ignores non-meeting collaboration e.g. async

2. **Lack of relevance of categories**
   - e.g. focus/meeting – vs. teaching/research/service
   - e.g. focus time/collaboration duality not exhaustive
3. **Tool creates work**
   - e.g. new role demand to interact with tool, learning curve
   - e.g. change use of calendar to give tool more accurate data

But privacy & ethics not seen as issues

Lessons

- *Perceived* accuracy is important
- Transparency …
  - Concepts, e.g. “quiet days”, “focus time”
  - Processing, e.g. “why is this 68%?”
  - Assumptions, e.g. appointment with no invitees = focus time
  - (implicit) norms & values, e.g. meetings should be reduced
    - Norm conflict: reduce meetings, reduce emails
- Co-regulation lens: missing feedback loop – unidirectional!
  - Including configurability (specific classification, categories)

Discussion & Conclusion

- **Digital Productivity Assistant** – emerging class of intelligent tools
- DPAs can help knowledge workers change habits and manage their time and well-being …
- … but there are **barriers** for the use of DPAs
  - Key feedback loop missing (user to tool)
  - Transparency important!
- We provide strategies to overcome the barriers (see paper)

Artificial Intelligence and the Right to Explanation as a Human Right

IEEE Internet Computing, 2021, joint work with Julija Sardelić (*)

doi:10.1109/MIC.2020.3045821

IEEE Internet Computing, 25(2):108-112.

* Michael Sardelic Winikoff & Julija Sardelić Winikoff

Key Questions:

Suppose we develop good XAI techniques …

How can we encourage organisations to use them?

… without having to create new laws

Key Questions:

Suppose we develop good XAI techniques …

How can we encourage organisations to use them?

… without having to create new laws

in what situations can the right to explanation be positioned as a human right?

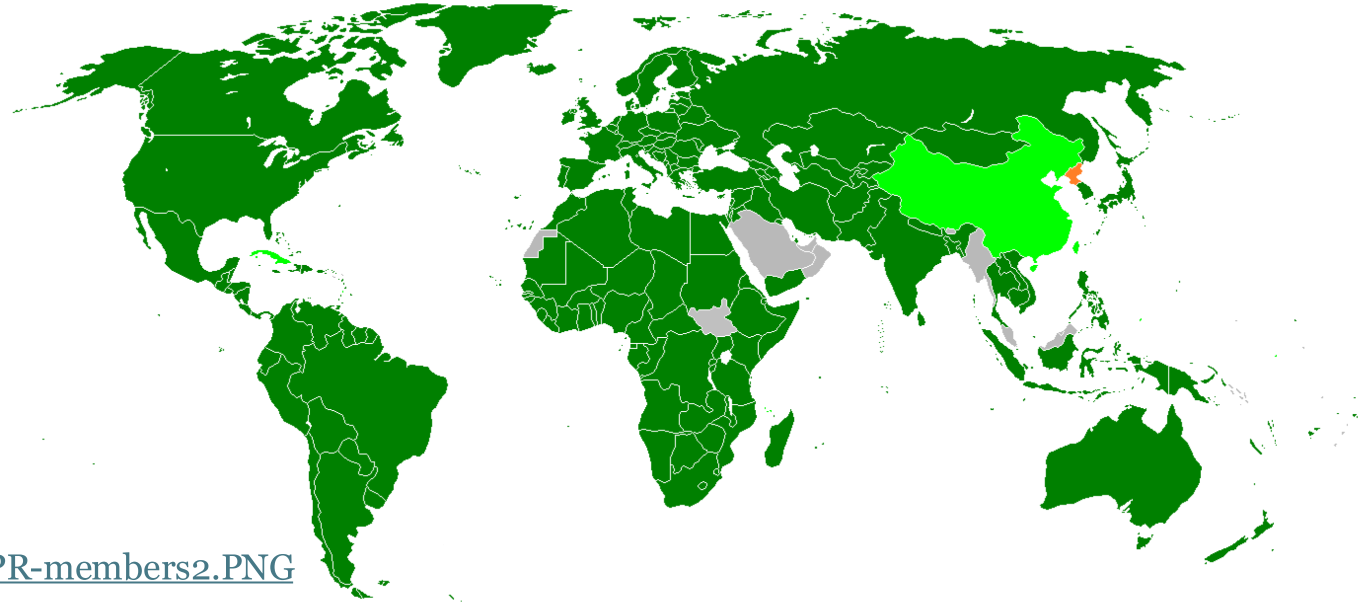
Human Rights Legislation

- Universal Declaration on Human Rights (UDHR) – aspirational, not legally binding, but …

https://www.un.org/en/about-us/universal-declaration-of-human-rights

- International Covenant on Civil and Political Rights (ICCPR) and International Covenant on Economic, Social and Cultural Rights (ICESCR) *are* legally binding on states that have ratified (dark green)

GDPR? Debatable - “an explanation of the decision” appears in the non-binding recital. Binding clauses (13-15) have “meaningful information about the logic involved”, not an explanation of a specific decision made.



ICCPR status
https://commons.wikimedia.org/wiki/File:ICCPR-members2.PNG

XAI and Human Rights: Discrimination

- “the right to social security” (UDHR article 22)
- “the right to a standard of living … and well-being … including food, clothing, housing and medical care and necessary social services” (Article 25)
- “higher education shall be equally accessible to all on the basis of merit” (Article 26)
- Cannot use prohibited criteria “such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status” (Article 2) ***… or proxies for these!***
- ***Explanation can help detect (and avoid) such use***

XAI and Human Rights: Judicial applications

- Legal rights prominent in human rights documents (Articles 6-11 UDHR)
- Judicial system expected to be highly accountable
- Consequences of legal processes can affect human rights, including: physical freedom (UHDR 3&9), privacy (12), property ownership (17), citizenship removal (15).
- “arbitrary” – need explanation to assess
- Also “without any discrimination” (7), and a fair hearing “by an independent and impartial tribunal” (10).
- Also right “To be informed promptly and in detail in a language which he [or she] understands of the nature and cause of the charge against him [or her]” (ICCPR 13)

XAI and Human Rights: Other Cases

- “right to freedom of opinion and expression” (19) – what about AI-curated social media?
- Health

Summary

- Q: How can we motivate organisations to deploy explanation mechanisms?
- A: Range of cases in which the right to explanation can be argued to be a natural consequence of existing human rights obligations

A Scoresheet for Explainable AI

AAMAS, 2025, joint work with John Thangarajah and Sebastian Rodriguez

arXiv:2502.09861v1

In Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, IFAAMAS, 10 pages. Also available (with supplementary material) from arXiv:2502.09861v1.

Why?

- Explainability important for e.g. transparency, accountability, understandability, trust calibration
- Given a choice of systems, how to compare explainability?
- Existing work (e.g. ALTAI[*], IEEE P7001) does not provide adequate guidance for assessing explainability

[*] Assessment List for Trustworthy Artificial Intelligence

What?

- A *scoresheet* for Explainable AI
- Can be used to assess systems
- … or to *specify* explainability requirements

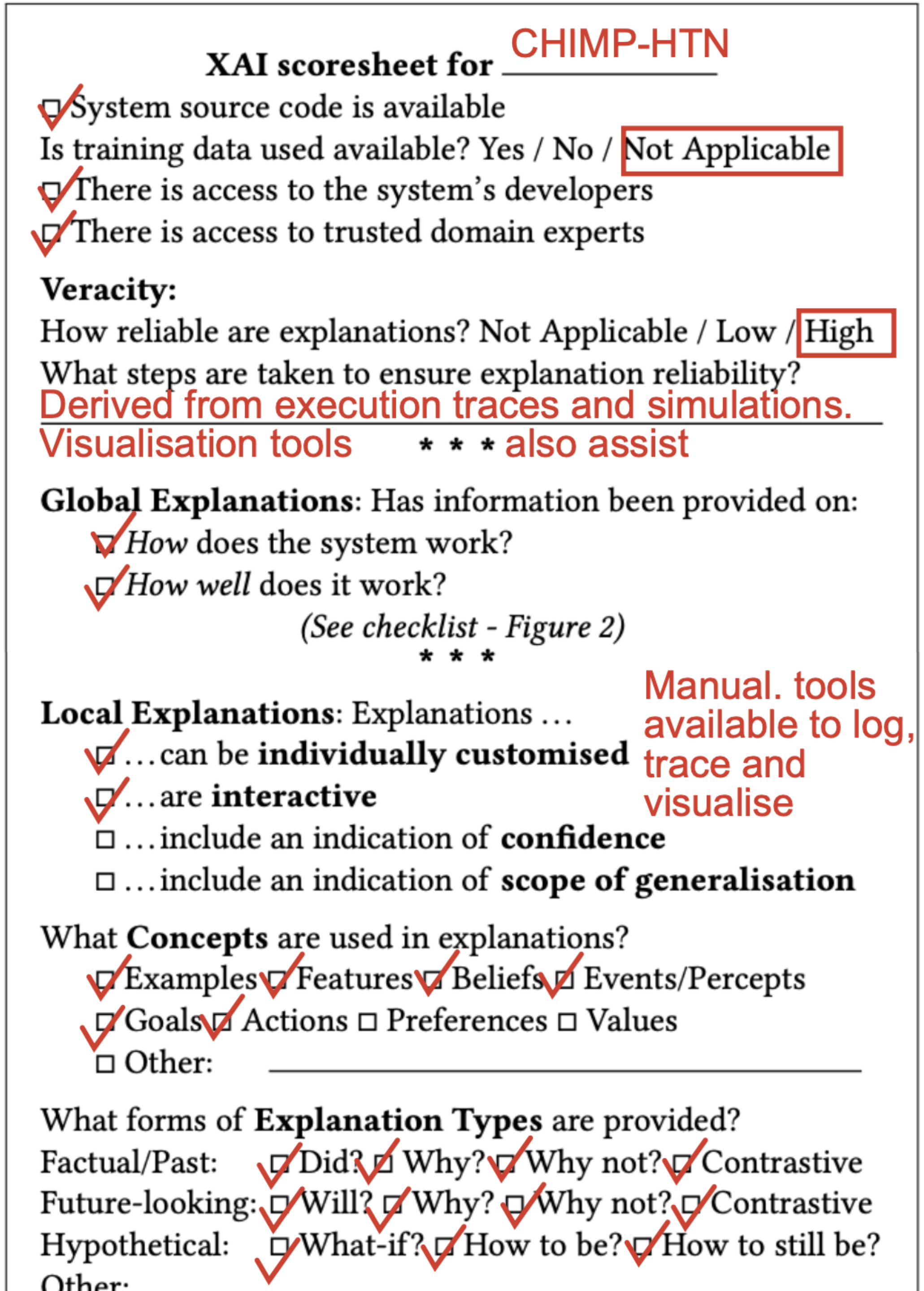
Contributions

- A ***scoresheet*** for Explainable AI (next slide)
- … with ***justification***,
- … ***guidance*** for how to complete it (including a ***checklist*** for global explanations), and
- … ***demonstration*** on a range of systems

  → *usable*, *generic*, *useful*

Scoresheet covers …

- Basic information
- Veracity
- Global Explanations: how, how well?
- Local Explanations
  - Features (e.g. customization, interaction)
  - Concepts used
  - Explanation Types
  - Automation



XAI scoresheet for CHIMP-HTN

☑ System source code is available

Is training data used available? Yes / No / [✓ Not Applicable]

☑ There is access to the system's developers

☑ There is access to trusted domain experts

Veracity:

How reliable are explanations? Not Applicable / Low / [✓ High]

What steps are taken to ensure explanation reliability?

Derived from execution traces and simulations.

Visualisation tools * * * also assist

Global Explanations: Has information been provided on:

☑ *How* does the system work?

☑ *How well* does it work?

(See checklist - Figure 2)

* * *

Local Explanations: Explanations …

Manual. tools available to log, trace and visualise

☑ …can be **individually customised**

☑ …are **interactive**

□ …include an indication of **confidence**

□ …include an indication of **scope of generalisation**

What **Concepts** are used in explanations?

☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts

☑ Goals ☑ Actions □ Preferences □ Values

□ Other: ____________

What forms of **Explanation Types** are provided?

Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive

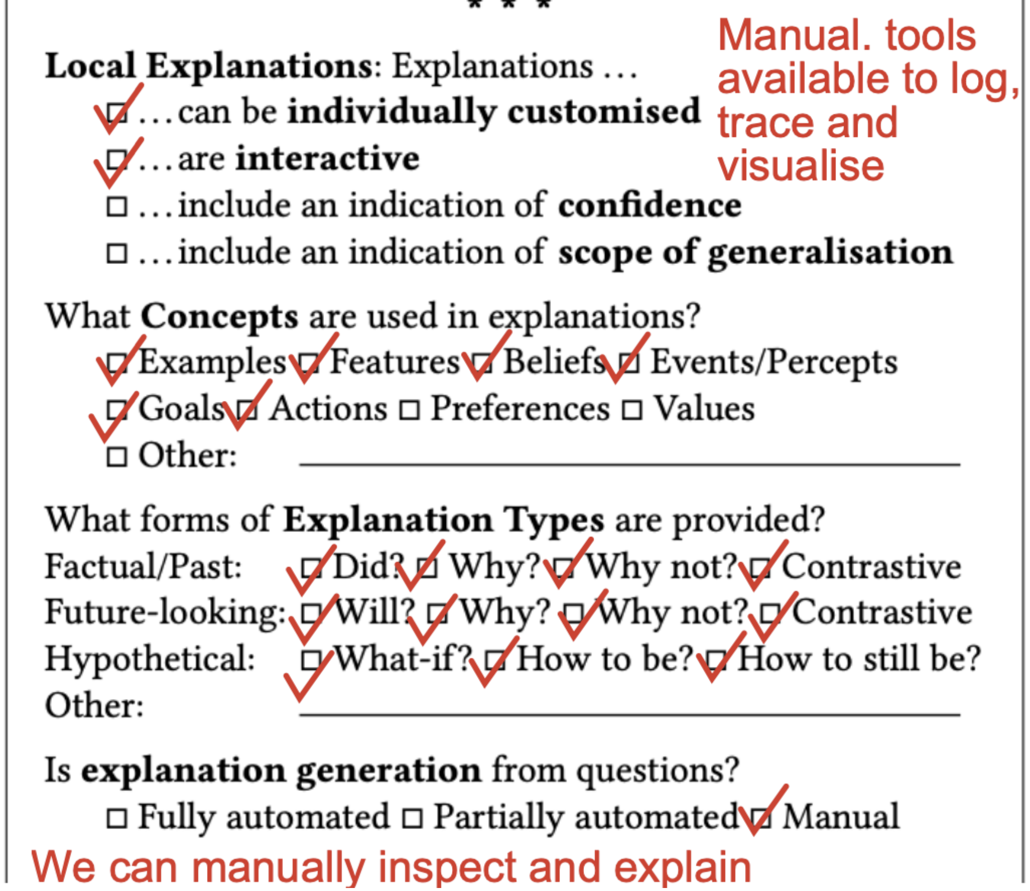
Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive

Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?

Other:

Scoresheet covers …

- Basic information
- Veracity
- Global Explanations: how, how well?
- Local Explanations
  - Features (e.g. customization, interaction)
  - Concepts used
  - Explanation Types
  - Automation



Local Explanations: Explanations …

☑ … can be **individually customised**

☑ … are **interactive**

□ … include an indication of **confidence**

□ … include an indication of **scope of generalisation**

Manual. tools available to log, trace and visualise

What **Concepts** are used in explanations?

☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts

☑ Goals ☑ Actions □ Preferences □ Values

□ Other: ______________________

What forms of **Explanation Types** are provided?

Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive

Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive

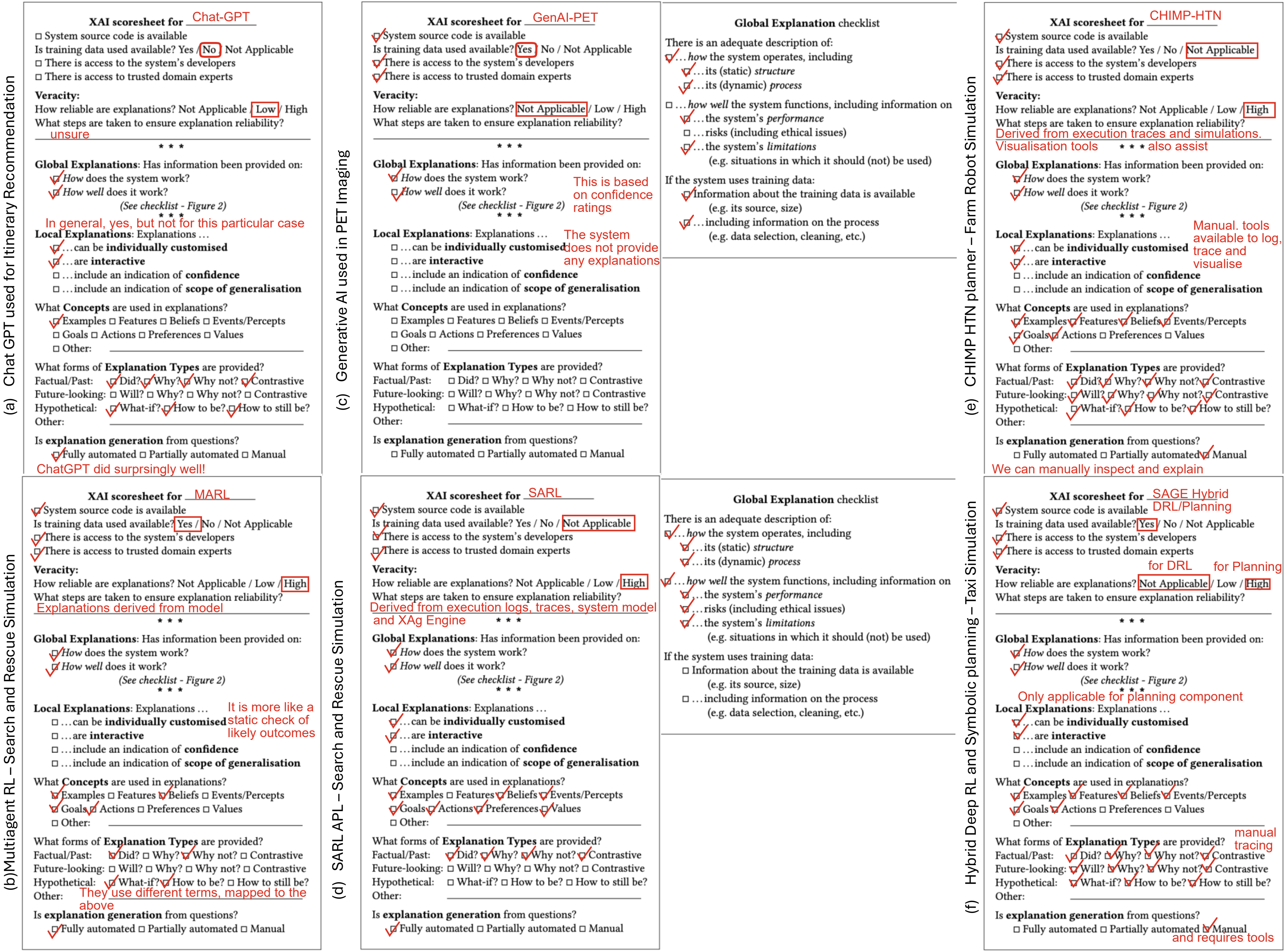
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?

Other: ______________________

Is **explanation generation** from questions?

□ Fully automated □ Partially automated ☑ Manual

We can manually inspect and explain



(a) Chat GPT used for Itinerary Recommendation

XAI scoresheet for Chat-GPT

□ System source code is available
Is training data used available? Yes / [✓ No] / Not Applicable
□ There is access to the system's developers
□ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / [✓ Low] / High
What steps are taken to ensure explanation reliability?
unsure

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

* * *

In general, yes, but not for this particular case

Local Explanations: Explanations …
☑ …can be **individually customised**
☑ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

What **Concepts** are used in explanations?
☑ Examples □ Features □ Beliefs □ Events/Percepts
□ Goals □ Actions □ Preferences □ Values
□ Other:

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: □ Will? □ Why? □ Why not? □ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?
Other:

Is **explanation generation** from questions?
☑ Fully automated □ Partially automated □ Manual
ChatGPT did surprsingly well!

(b)Multiagent RL – Search and Rescue Simulation

XAI scoresheet for MARL

☑ System source code is available
Is training data used available? [✓ Yes] / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / [✓ High]
What steps are taken to ensure explanation reliability?
Explanations derived from model

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

* * *

Local Explanations: Explanations …
□ …can be **individually customised**
□ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

It is more like a static check of likely outcomes

What **Concepts** are used in explanations?
☑ Examples □ Features ☑ Beliefs □ Events/Percepts
☑ Goals ☑ Actions □ Preferences □ Values
□ Other:

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? □ Why? ☑ Why not? □ Contrastive
Future-looking: □ Will? □ Why? □ Why not? □ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? □ How to still be?
Other: They use different terms, mapped to the above

Is **explanation generation** from questions?
☑ Fully automated □ Partially automated □ Manual

(c) Generative AI used in PET Imaging

XAI scoresheet for GenAI-PET

☑ System source code is available
Is training data used available? [✓ Yes] / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? [✓ Not Applicable] / Low / High
What steps are taken to ensure explanation reliability?

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

This is based on confidence ratings

* * *

Local Explanations: Explanations …
□ …can be **individually customised**
□ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

The system does not provide any explanations

What **Concepts** are used in explanations?
□ Examples □ Features □ Beliefs □ Events/Percepts
□ Goals □ Actions □ Preferences □ Values
□ Other:

What forms of **Explanation Types** are provided?
Factual/Past: □ Did? □ Why? □ Why not? □ Contrastive
Future-looking: □ Will? □ Why? □ Why not? □ Contrastive
Hypothetical: □ What-if? □ How to be? □ How to still be?
Other:

Is **explanation generation** from questions?
□ Fully automated □ Partially automated □ Manual

Global Explanation checklist

There is an adequate description of:
☑ …*how* the system operates, including
☑ …its (static) *structure*
☑ …its (dynamic) *process*

□ …*how well* the system functions, including information on
☑ …the system's *performance*
□ …risks (including ethical issues)
☑ …the system's *limitations*
(e.g. situations in which it should (not) be used)

If the system uses training data:
☑ Information about the training data is available
(e.g. its source, size)
☑ …including information on the process
(e.g. data selection, cleaning, etc.)

(d) SARL APL – Search and Rescue Simulation

XAI scoresheet for SARL

☑ System source code is available
Is training data used available? Yes / No / [✓ Not Applicable]
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / [✓ High]
What steps are taken to ensure explanation reliability?
Derived from execution logs, traces, system model and XAg Engine

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

* * *

Local Explanations: Explanations …
☑ …can be **individually customised**
☑ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

What **Concepts** are used in explanations?
☑ Examples □ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions ☑ Preferences ☑ Values
□ Other:

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: □ Will? □ Why? □ Why not? □ Contrastive
Hypothetical: □ What-if? □ How to be? □ How to still be?
Other:

Is **explanation generation** from questions?
☑ Fully automated □ Partially automated □ Manual

Global Explanation checklist

There is an adequate description of:
☑ …*how* the system operates, including
☑ …its (static) *structure*
☑ …its (dynamic) *process*

☑ …*how well* the system functions, including information on
☑ …the system's *performance*
☑ …risks (including ethical issues)
☑ …the system's *limitations*
(e.g. situations in which it should (not) be used)

If the system uses training data:
□ Information about the training data is available
(e.g. its source, size)
□ …including information on the process
(e.g. data selection, cleaning, etc.)

(e) CHIMP HTN planner – Farm Robot Simulation

XAI scoresheet for CHIMP-HTN

☑ System source code is available
Is training data used available? Yes / No / [✓ Not Applicable]
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / [✓ High]
What steps are taken to ensure explanation reliability?
Derived from execution traces and simulations. Visualisation tools also assist

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

* * *

Local Explanations: Explanations …
☑ …can be **individually customised**
☑ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

Manual. tools available to log, trace and visualise

What **Concepts** are used in explanations?
☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions □ Preferences □ Values
□ Other:

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?
Other:

Is **explanation generation** from questions?
□ Fully automated □ Partially automated ☑ Manual
We can manually inspect and explain

(f) Hybrid Deep RL and Symbolic planning – Taxi Simulation

XAI scoresheet for SAGE Hybrid DRL/Planning

☑ System source code is available
Is training data used available? [✓ Yes] / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? [✓ Not Applicable] for DRL / Low / [✓ High] for Planning
What steps are taken to ensure explanation reliability?

* * *

Global Explanations: Has information been provided on:
☑ *How* does the system work?
☑ *How well* does it work?
(*See checklist - Figure 2*)

* * *

Only applicable for planning component

Local Explanations: Explanations …
☑ …can be **individually customised**
☑ …are **interactive**
□ …include an indication of **confidence**
□ …include an indication of **scope of generalisation**

What **Concepts** are used in explanations?
☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions □ Preferences □ Values
□ Other:

What forms of **Explanation Types** are provided? manual tracing
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?
Other:

Is **explanation generation** from questions?
□ Fully automated □ Partially automated ☑ Manual and requires tools

Evaluating Contrastive Explanations of Autonomous Systems

(under review)

Contrastive Explanations

- Explanation: “Why did you do *X*?”
- Contrastive: “Why did you do *X* ***instead of Y***?”
- Terminology: X is the *fact*, Y is the *foil*?
- Evidence that humans ask contrastive questions
- … but sometimes implicitly (i.e. foil is implicit)

Human-subject evaluation

- Have previously shown that contrastive explanations are shorter (see paper)
- But are they ***preferred*** (by humans) and are they ***effective*** (at supporting appropriate trust and transparency)?

We hypothesise that …

preference

- H1 contrastive explanations are preferred to full explanations.
- H2: the *perceived quality* of explanation is higher for contrastive than for full explanations.
- H3: contrastive explanations are more likely to be considered to have the right level of detail.

effectiveness

Expl. type

- H4: contrastive explanations yield higher trust than full explanations.
- H5: contrastive explanations yield higher belief in understanding of the system than full explanations.
- H6: contrastive explanations yield more confidence in the system's correct behaviour than full explanations.

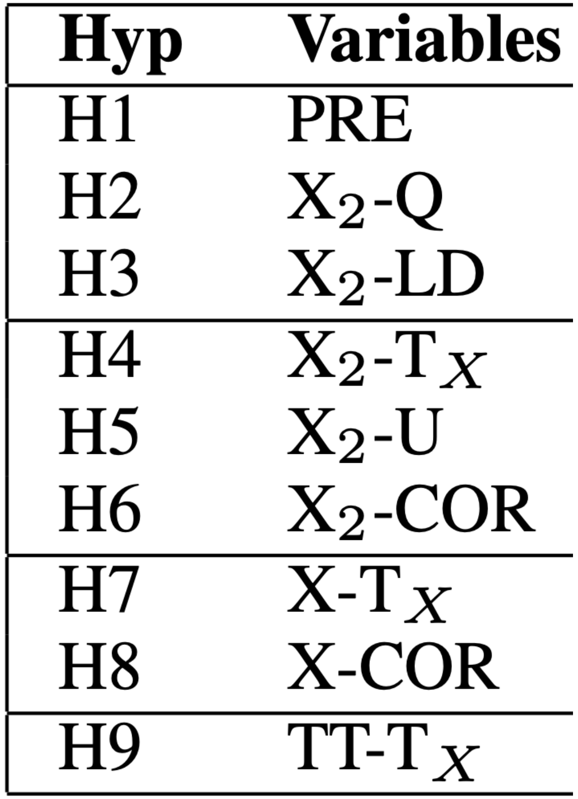
No expl.

- H7: both types of explanation yield higher trust than no explanation.
- H8: having an explanation (either full or contrastive) yields more confidence in the system’s correct behaviour than not having an explanation.

- H9: there is a correlation between trust in technology in general and trust in each of the two systems, but that the strength of the correlation is not high.

Methodology

- Used Prolific to recruit gender-balanced sample of adults fluent in English
- Split into three groups (**X**): full, contrastive, none
- Survey: (sections repeated: 2 systems, 3 scenarios each)
  - Technology Trust (**TT**),
  - Present System (pancake robot and search-and-rescue)
    - Present Scenario (including action and explanation [except for None group])
      - Explanation quality (**Q**), understanding (**U**) and level of detail (**LD**) [skip for None]
      - Trust (short) and belief in correctness (**COR**) [for each scenario]
    - Trust (long) [for each system] $\mathbf{T_{Pan}}$ and $\mathbf{T_{SAR}}$
  - Preferred explanation (**PRE**) for six scenarios
  - Demographics



Hyp	**Variables**
H1	PRE
H2	X_2-Q
H3	X_2-LD
H4	X_2-T_X
H5	X_2-U
H6	X_2-COR
H7	X-T_X
H8	X-COR
H9	TT-T_X

$\mathbf{X_2}$ – full & contrastive only

Responses

- 161 responses
- Filtered using attention check questions (12 responses failed both, 18 failed one)
- Filtered using inconsistent short-long trust responses (27 responses)
- 104 responses analysed
- Also checked quick completions
- Cronbach’s alpha checked, high enough (>0.8)

Explanation Type Preferences

- H1 contrastive explanations are preferred to full explanations
  - **Partially confirmed**: scenarios 1-4 show difference, 1&2 prefer contrastive, 3&4 prefer full explanations
- H2: the *perceived quality* of explanation is higher for contrastive than for full explanations.
  - **Not confirmed**: no significant difference
- H3: contrastive explanations are more likely to be considered to have the right level of detail.
  - **Not confirmed**: no significant difference

Explanation Type Effectiveness

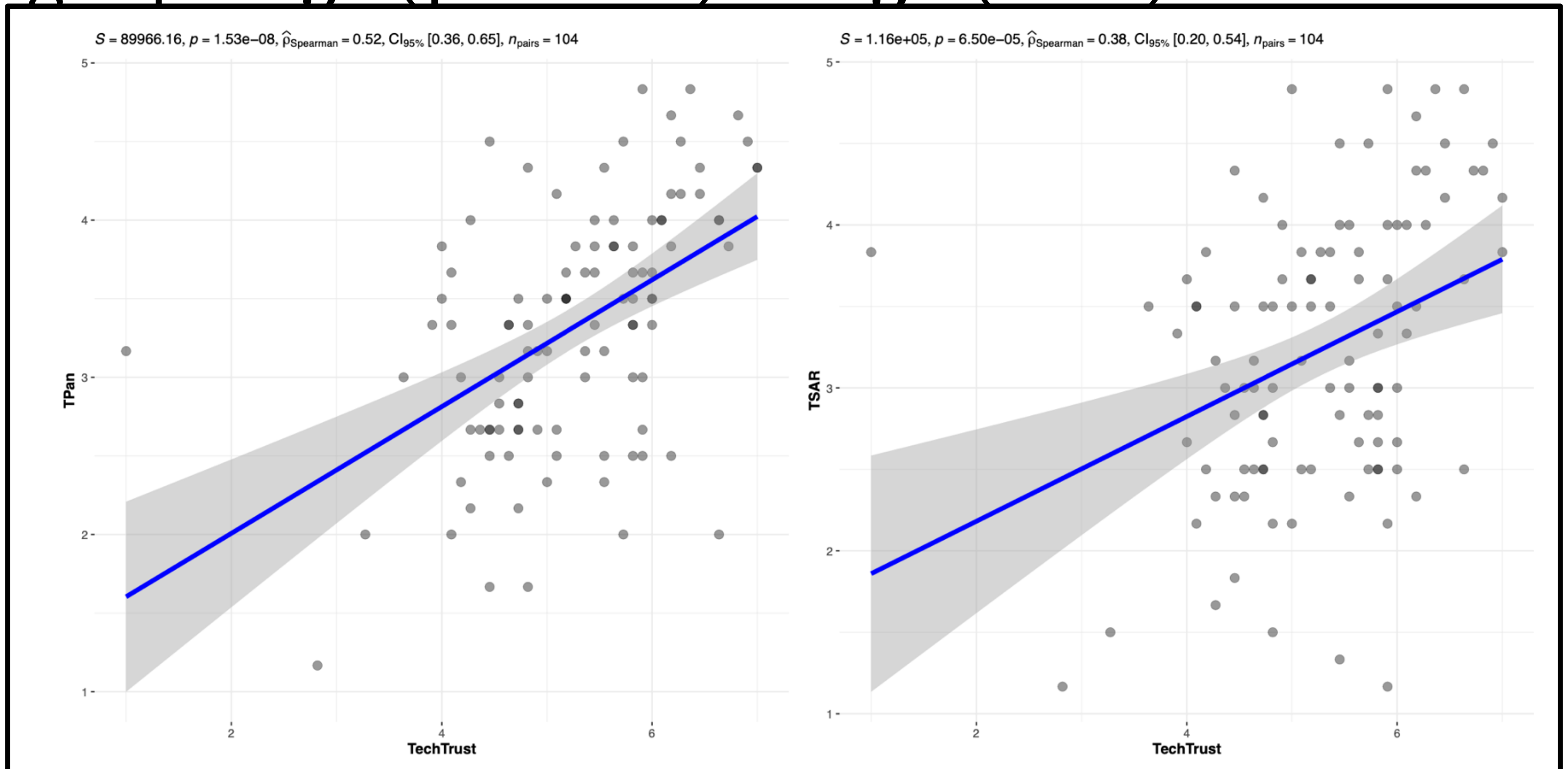
- H4: contrastive explanations yield higher trust than full explanations.
  - **Confirmed**: median trust for contrastive 3.5 (pancake) & 3.67 (SAR) vs. full explanation 3.0 (pancake) & 3.17 (SAR)
- H5: contrastive explanations yield higher belief in understanding of the system than full explanations.
  - **Confirmed**, but only for scenario 2 (mean 3.676 vs. 4.225 for contrastive)
- H6: contrastive explanations yield more confidence in the system's correct behaviour than full explanations.
  - **Confirmed**, but only for scenarios 1&2

Effects of not having explanations

- H7: both types of explanation yield higher trust than no explanation.
  - **Not confirmed**: no significant difference (however, median was higher for no explanation!)
- H8: having an explanation (either full or contrastive) yields more confidence in the system's correct behaviour than not having an explanation.
  - **Not confirmed**: significant difference for scenarios 3&4 with lower score for full than no explanation

Trust vs. Trust in Technology

- H9: there is a correlation between trust in technology in general and trust in each of the two systems, but that the strength of the correlation is not high.
  - **Confirmed**: highly significant (p<<0.01) but moderate strength ρ=0.52 (pancake) & 0.38 (SAR)



Discussion

- Difference between preference and effectiveness (cf. Amitai *et al*): contrastive not consistently preferred, but did give higher trust
- For some scenarios providing full explanations reduced trust (compared with no explanation). Kaptein *et al* also found counter-intuitive results (recommendation explanations resulted in participants being *less likely* to follow them)
- Human decision-making is complex! Speculation: too-long explanations might reduce system trust?
- Scenario dependent results consistent with prior work (Harbers *et al*)
- Possible issue with foil mismatch …

Y. Amitai, Y. Septon, and O. Amir. *Explaining reinforcement learning agents through counterfactual action outcomes*. AAAI, 2024. doi:10.1609/AAAI.V38I9.28863

M. Harbers. *Explaining Agent Behavior in Virtual Training*. SIKS dissertation series no. 2011-35, SIKS (Dutch Research School for Information and Knowledge Systems), 2011.

Harbers, K. van den Bosch, and J. C. Meyer. *Design and evaluation of explainable BDI agents*. IAT, 2010. doi:10.1109/WI-IAT.2010.115

F. Kaptein, J. Broekens, K. V. Hindriks, and M. A. Neerincx. *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults*. Ro-MAN, 2017. doi:10.1109/ROMAN.2017.

F. Kaptein, J. Broekens, K. V. Hindriks, and M. A. Neerincx. *Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes*. ACII, 2019. doi:10.1109/ACII.2019.8925526.

Implications

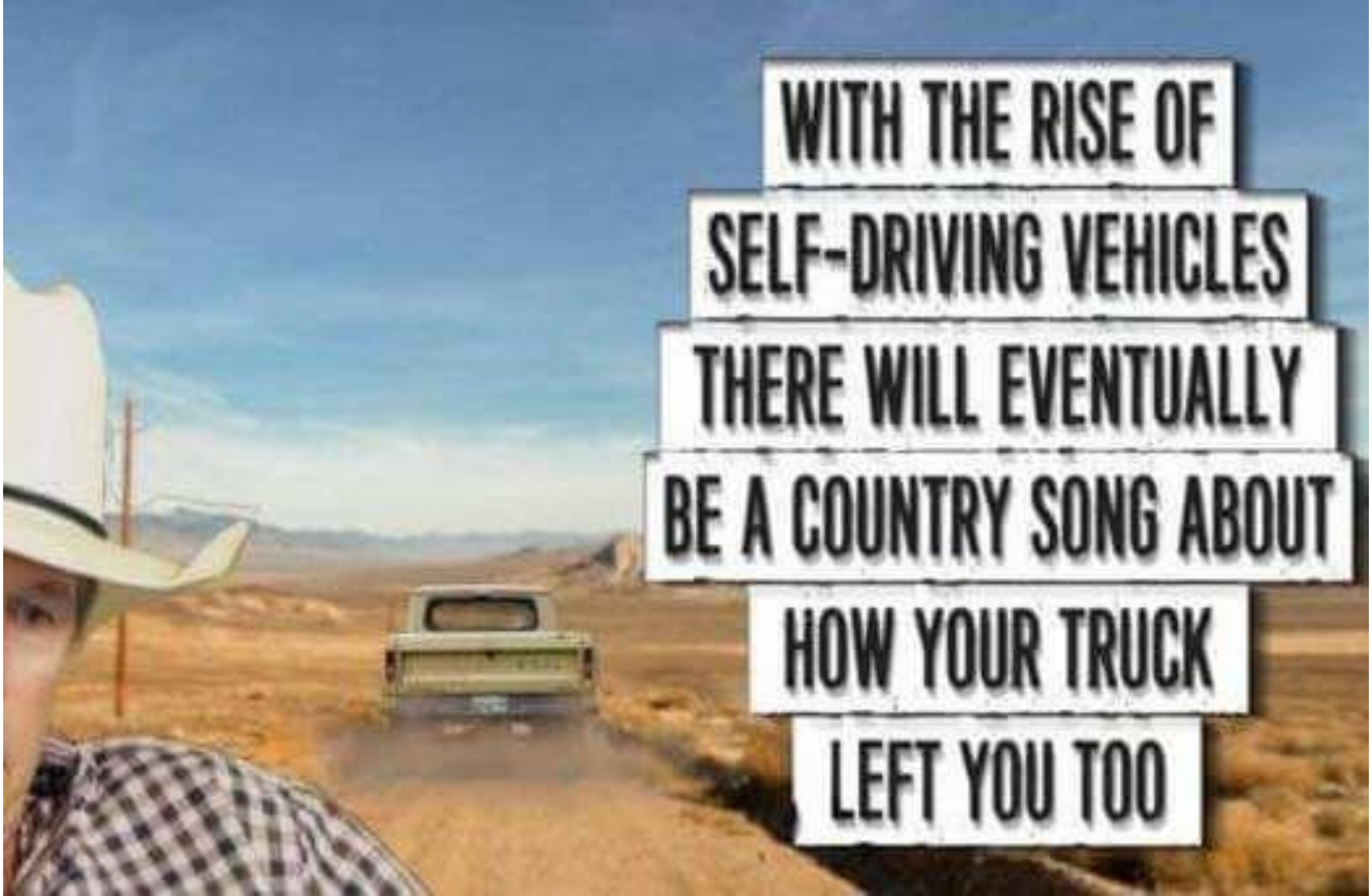
- Important to ensure foil matches user's expectations – ask user
- Providing explanations is not risk-free – poor quality or too-long explanations can *reduce* trust
- Human behaviour is complex. Need to iteratively guide development & deployment with (careful) user evaluation

Future Work

- Making explanations interactive
- Adding *hypothetical* explanations (What if? How to be? How to still be?)
- Extending to non-BDI agents (cf. Gimenez-Abalos *et al* policy graph work for “Why?”)
- Further evaluation (more scenarios, systems)
    - One approach for handling foil mismatch: ask what system should do and then use that as the foil … but low trust …

Summary

- AI tools sometimes need to be viewed as *partners*
- Explanation is important for transparency and trust
- Explanation can be arguably motivated under existing laws
- A *scoresheet* can be used to succinctly describe the explanatory capabilities of a system
- Providing effective explanations can be tricky


WITH THE RISE OF
SELF-DRIVING VEHICLES
THERE WILL EVENTUALLY
BE A COUNTRY SONG ABOUT
HOW YOUR TRUCK
LEFT YOU TOO

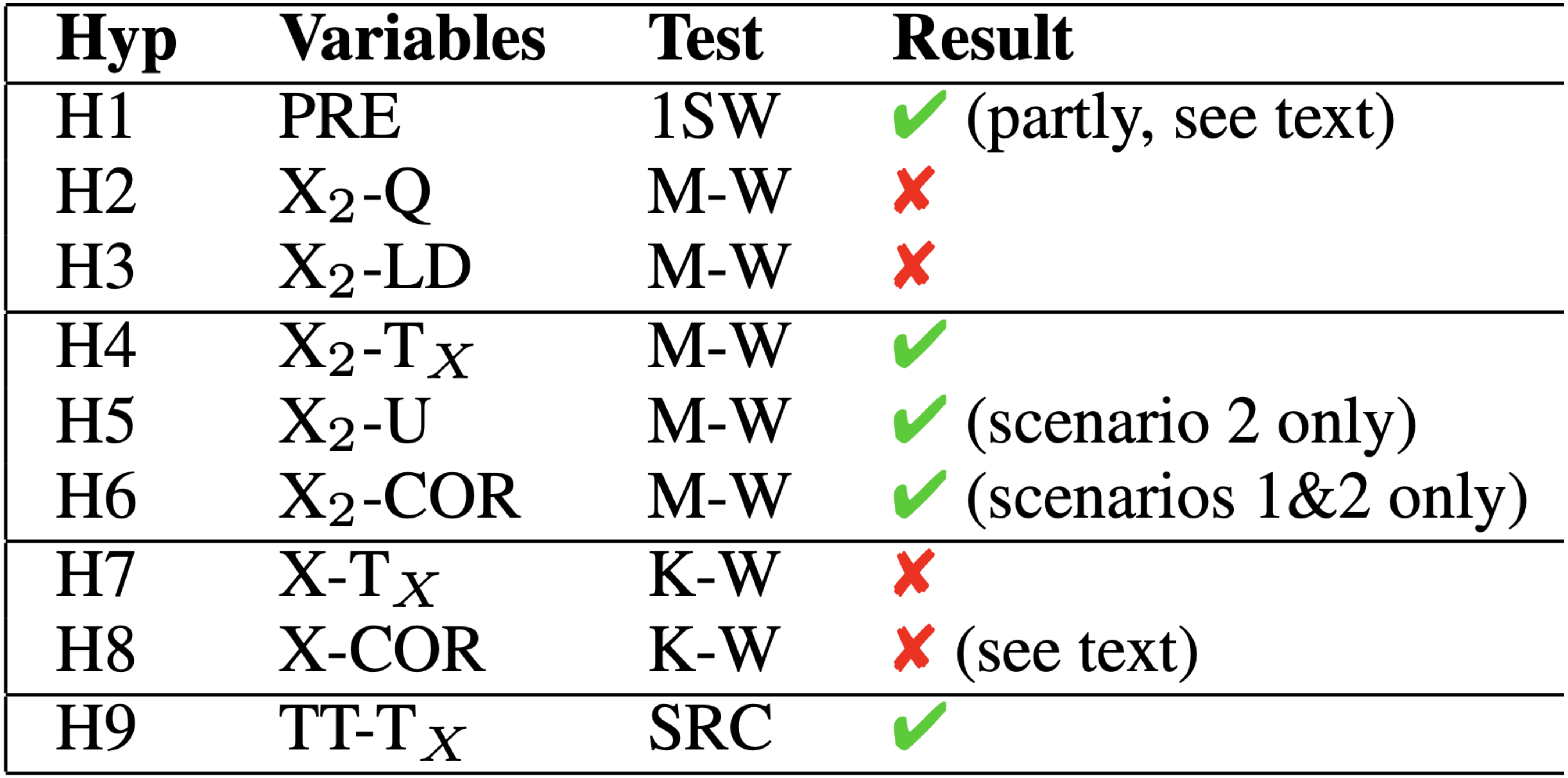
https://youtu.be/ESNDtH6suUo

https://youtu.be/mxhcQgvBD2Y

Demographics

- Gender: 50 Male, 54 Female.
- Age: 23 participants were aged 18-24, 49 were aged 25-34, 15 (35-44), 10 (45-54), 5 (55-64), 2 (65-74), 0 (75+).
- Education: 22 (completed high school), 56 (completed undergrad degree), 23 (Masters), 2 (PhD), 1 (declined to answer).
- Ethnicity: 40 (African), 36 (European), 8 (North American), 7 (South American), 5 (Asian), 3 (Other), 2 (Australian), 2 (declined), 1 (New Zealander).
- Programming experience: 38 (None), 28 (hobby), 12 (studied at high school), 12 (currently studying degree), 10 (completed degree), 4 (other).

Summary of Results



Hyp	Variables	Test	Result
H1	PRE	1SW	✔ (partly, see text)
H2	X_2-Q	M-W	✘
H3	X_2-LD	M-W	✘
H4	X_2-T_X	M-W	✔
H5	X_2-U	M-W	✔ (scenario 2 only)
H6	X_2-COR	M-W	✔ (scenarios 1&2 only)
H7	X-T_X	K-W	✘
H8	X-COR	K-W	✘ (see text)
H9	TT-T_X	SRC	✔