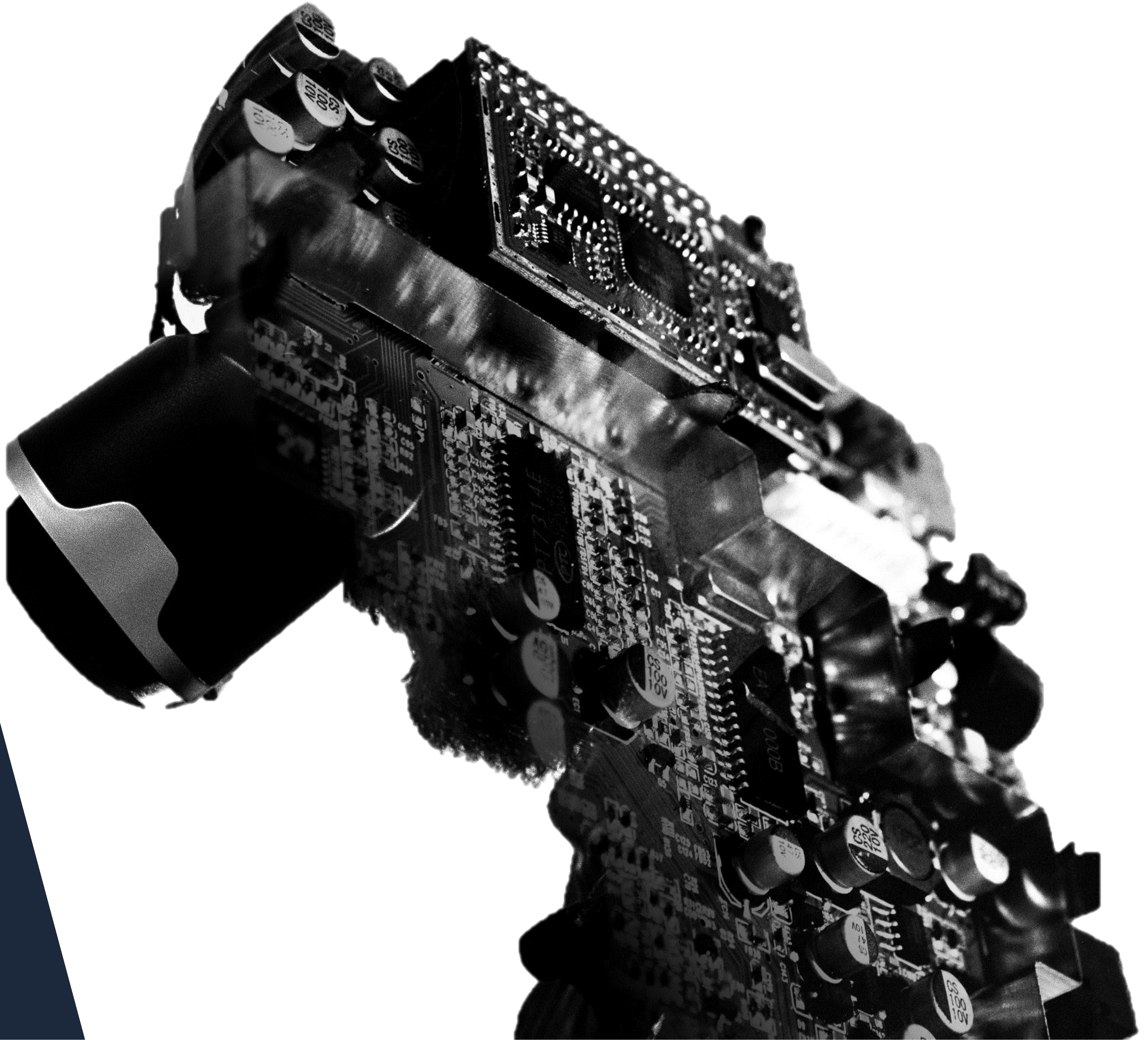
EXPLAINABLE AI IS DEAD! LONG LIVE EXPLAINABLE AI!

WHY YOUR AI TOOL PROBABLY DOESN'T WORK FOR USERS

AND WHY IT IS SO &*%* HARD TO GET IT TO DO SO

Tim Miller

School of Electrical Engineering and Computer Science
The University of Queensland, Australia
timothy.miller@uq.edu.au
@tmiller_uq



XAI IS DEAD

T. MILLER, EXPLAINABLE AI IS DEAD, LONG LIVE EXPLAINABLE AI! HYPOTHESIS-DRIVEN DECISION SUPPORT., IN *PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT)*, 2023. https://arxiv.org/pdf/2302.12389.pdf

EXPLAINABLE AI IS DEAD, LONG LIVE EXPLAINABLE AI!
HYPOTHESIS-DRIVEN DECISION SUPPORT

A PREPRINT

Tim Miller
School of Computing and Information Systems
The University of Melbourne, Melbourne, Australia
tmiller@unimelb.edu.au

March 14, 2023

ABSTRACT

In this paper, we argue for a paradigm shift from the current model of explainable artificial intelligence (XAI), which may be counter-productive to better human decision making. In early decision support systems, we assumed that we could give people recommendations and that they would consider them, and then follow them when required. However, research found that people often ignore recommendations because they do not trust them; or perhaps even worse, people follow them blindly, even when the recommendations are wrong. Explainable artificial intelligence mitigates this by helping people to understand how and why models give certain recommendations. However, recent research shows that people do not always engage with explainability tools enough to help improve decision making. The assumption that people will engage with recommendations and explanations has proven to be unfounded. We argue this is because we have failed to account for two things. First, recommendations (and their explanations) take control from human decision makers, limiting their agency. Second, giving recommendations and explanations does not align with the cognitive processes employed by people making decisions. This position paper proposes a new conceptual framework called **Evaluative AI** for explainable decision support. This is a machine-in-the-loop paradigm in which decision support tools provide evidence for and against decisions made by people, rather than provide recommendations to accept or reject. We argue that this mitigates issues of over- and under-reliance on decision support tools, and better leverages human expertise in decision making.

Keywords Explainable AI · Cognitive Processes · Abductive Reasoning · Decision Support · Cognitive Forcing · Evidence · Hypotheses

Long Live the Queen

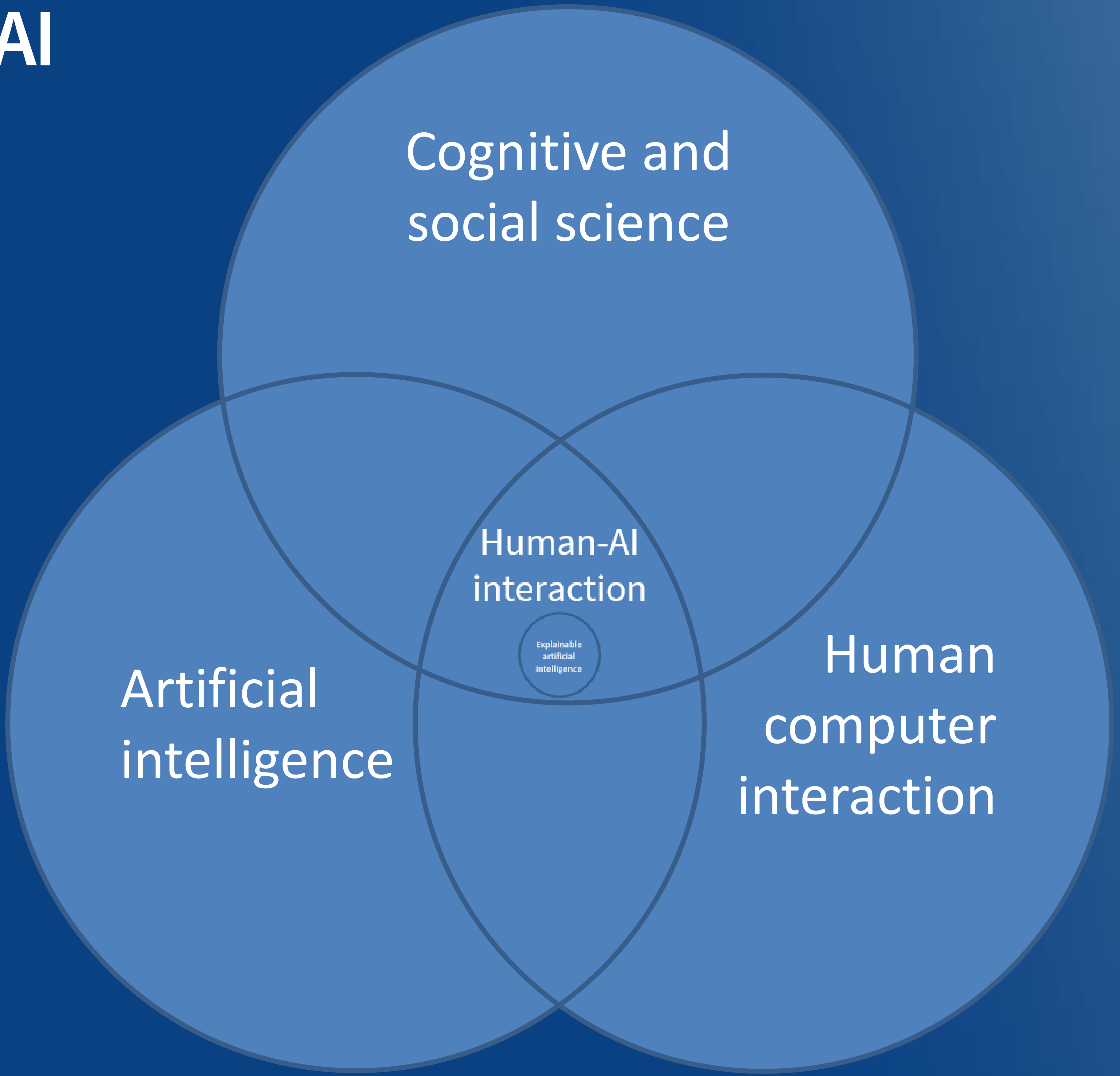
THE SCOPE OF XAI

Artificial intelligence

Cognitive and social science

Human computer interaction

THE SCOPE OF XAI



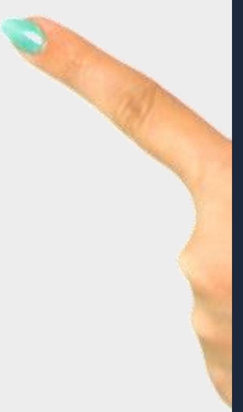
Human-AI interaction

Explainable artificial intelligence

INFUSING THE SOCIAL SCIENCES

Symptom	Cause	Prob
Weight gain	Stopped exercising	80%
Fatigue	Mononucleosis	50%
Nausea	Stomach virus	50%
Weight gain, fatigue, nausea	Pregnancy	15%

S. J. READ, A. MARCUS-NEWHALL, EXPLANATORY COHERENCE IN SOCIAL EXPLANATIONS: A PARALLEL DISTRIBUTED PROCESSING ACCOUNT, JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 65 (3) (1993)



THE BEST EXPLANATION?

1) Stopped exercising
   Mononucleosis
   Stomach virus; or
2) Pregnancy

INFUSING THE SOCIAL SCIENCES

Artificial Intelligence 267 (2019) 1–38

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

A R T I C L E I N F O

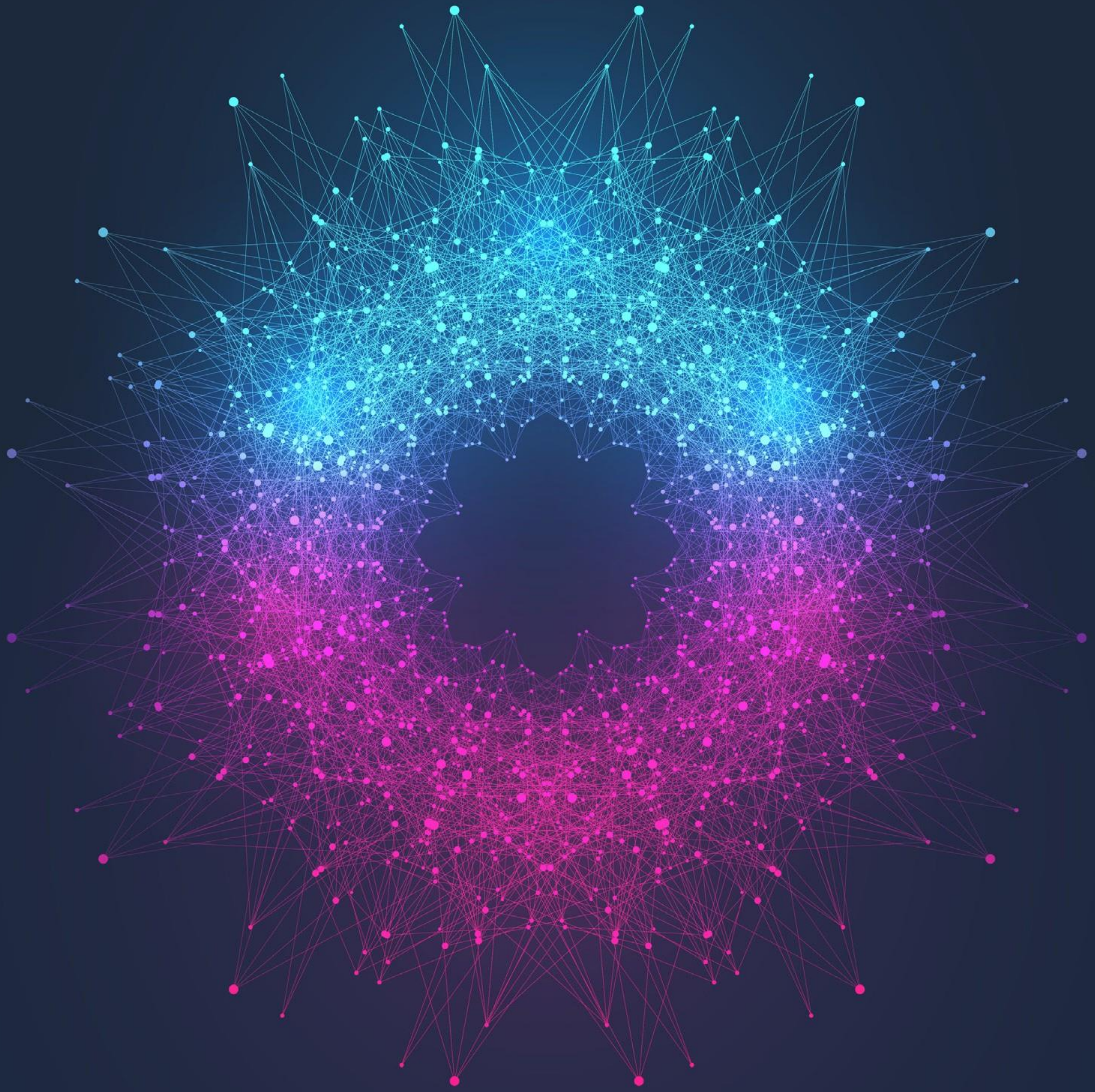
Article history:
Received 22 June 2017
Received in revised form 17 May 2018
Accepted 16 July 2018
Available online 27 October 2018

Keywords:
Explanation
Explainability
Interpretability
Explainable AI
Transparency

A B S T R A C T

There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain cognitive biases and social expectations to the explanation process. This paper argues that the field of explainable artificial intelligence can build on this existing research, and reviews relevant papers from philosophy, cognitive psychology/science, and social psychology, which study these topics. It draws out some important findings, and discusses ways that these can be infused with work on explainable artificial intelligence.

WHAT ARE THE KEY LESSONS?



EXPLANATIONS ARE CONTRASTIVE

“The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case”

DENIS HILTON: CONVERSATIONAL PROCESSES AND CAUSAL EXPLANATION, PSYCHOLOGICAL BULLETIN. 107(11):65-81, (1990)

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY?

CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY?

CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY?

CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

CONTRASTIVE EXPLANATION
THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

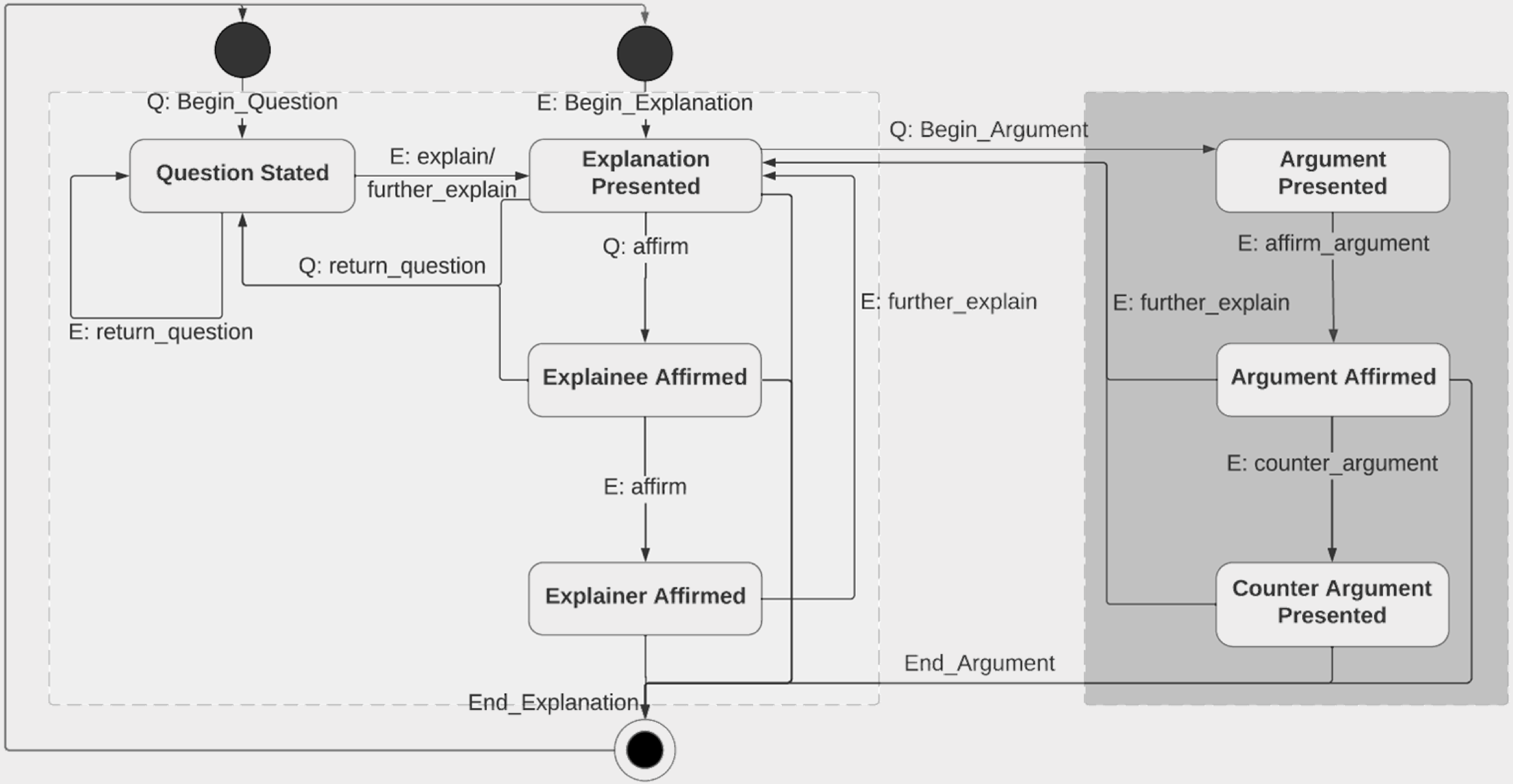
CONTRASTIVE EXPLANATION: A STRUCTURAL-MODEL APPROACH.
T. MILLER, KNOWLEDGE ENGINEERING REVIEW, 36 E14: 2021

EXPLANATIONS ARE SOCIAL

"Causal explanation is first and foremost a form of social interaction. The verb to explain is a three-place predicate: ***Someone*** *explains* ***something*** *to* ***someone****. Causal explanation takes the form of conversation and is thus subject to the rules of conversation."*
[Emphasis original]

DENIS HILTON: CONVERSATIONAL PROCESSES AND CAUSAL EXPLANATION, PSYCHOLOGICAL BULLETIN. 107(11):65-81, (1990)

SOCIAL EXPLANATION



A GROUNDED INTERACTION PROTOCOL FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE. MADUMAL, P.; MILLER, T.; SONENBERG, L.; AND VETERE, F. IN PROCEEDINGS OF AAMAS 2019, 2019.

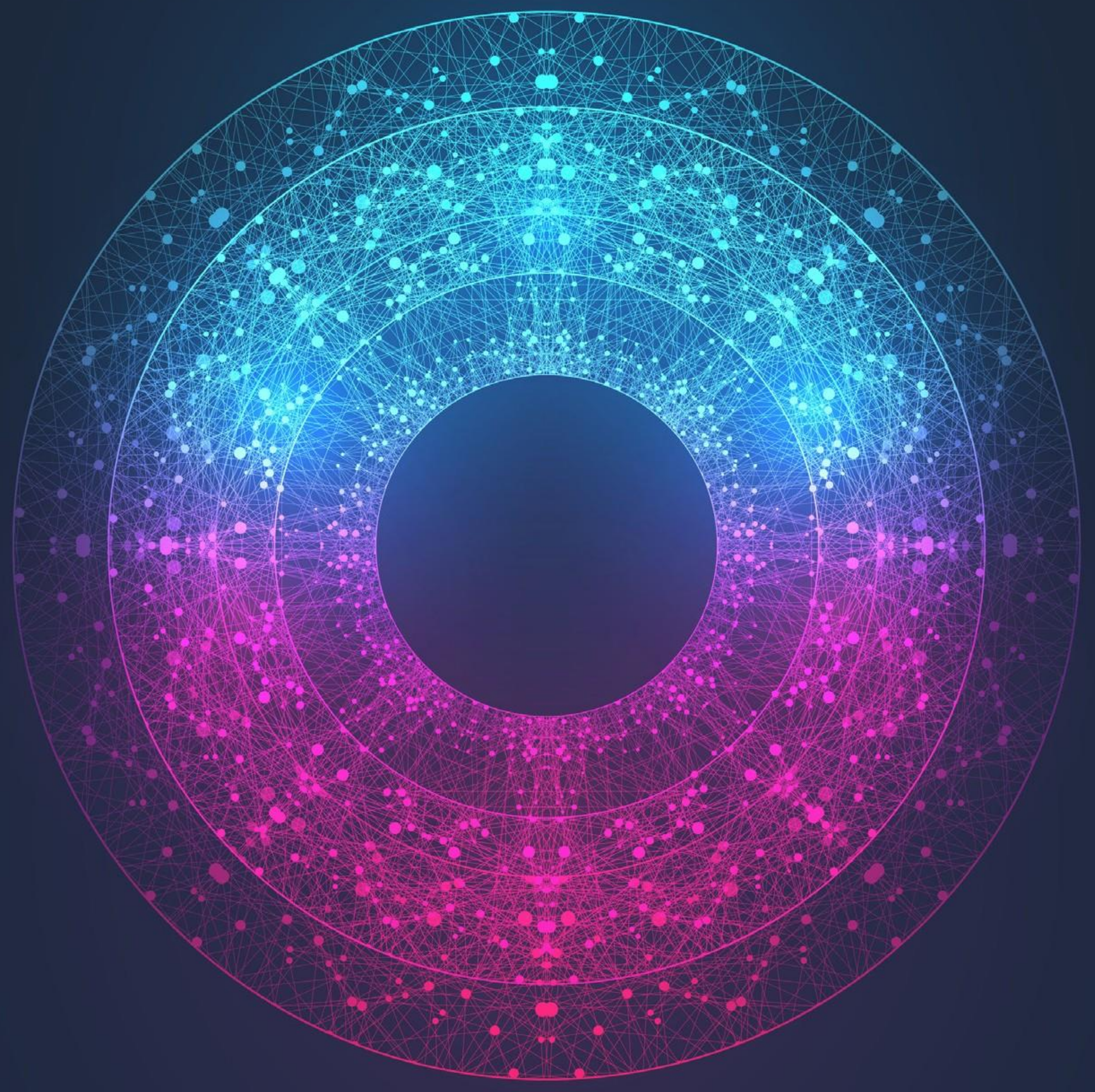
EXPLANATIONS ARE SELECTED

“The accident occurred at a major intersection. The light turned amber as Mr. Jones approached. Witnesses noted that he braked hard to stop at the crossing, although he could easily have gone through. His family recognized this as a common occurrence in Mr. Jones driving. As he began to cross after the light changed, a light truck charged into the intersection at top speed, and rammed Mr. Jones' car from the left. On the day of the accident, Mr. Jones left his office at the regular time. He sometimes left early to take care of home chores at his wife's request, but this was not necessary on that day. Mr. Jones did not drive home by his regular route. The day was exceptionally clear and Mr. Jones told his friends at the office that he would drive along the shore to enjoy the view.”

D. KAHNEMAN AND A. TVERSKY, THE SIMULATION HEURISTIC, IN *JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES*, NEW YORK: CAMBRIDGE UNIVERSITY PRESS, 1982.

EXPERIENCE

APPLYING THESE INSIGHTS



OUR EXPERIENCE

INSIGHTS

- CONTRASTIVE EXPLANATION
- CAUSALITY
- INTERACTION
- TEMPORAL SELECTION
- HUMAN STUDIES

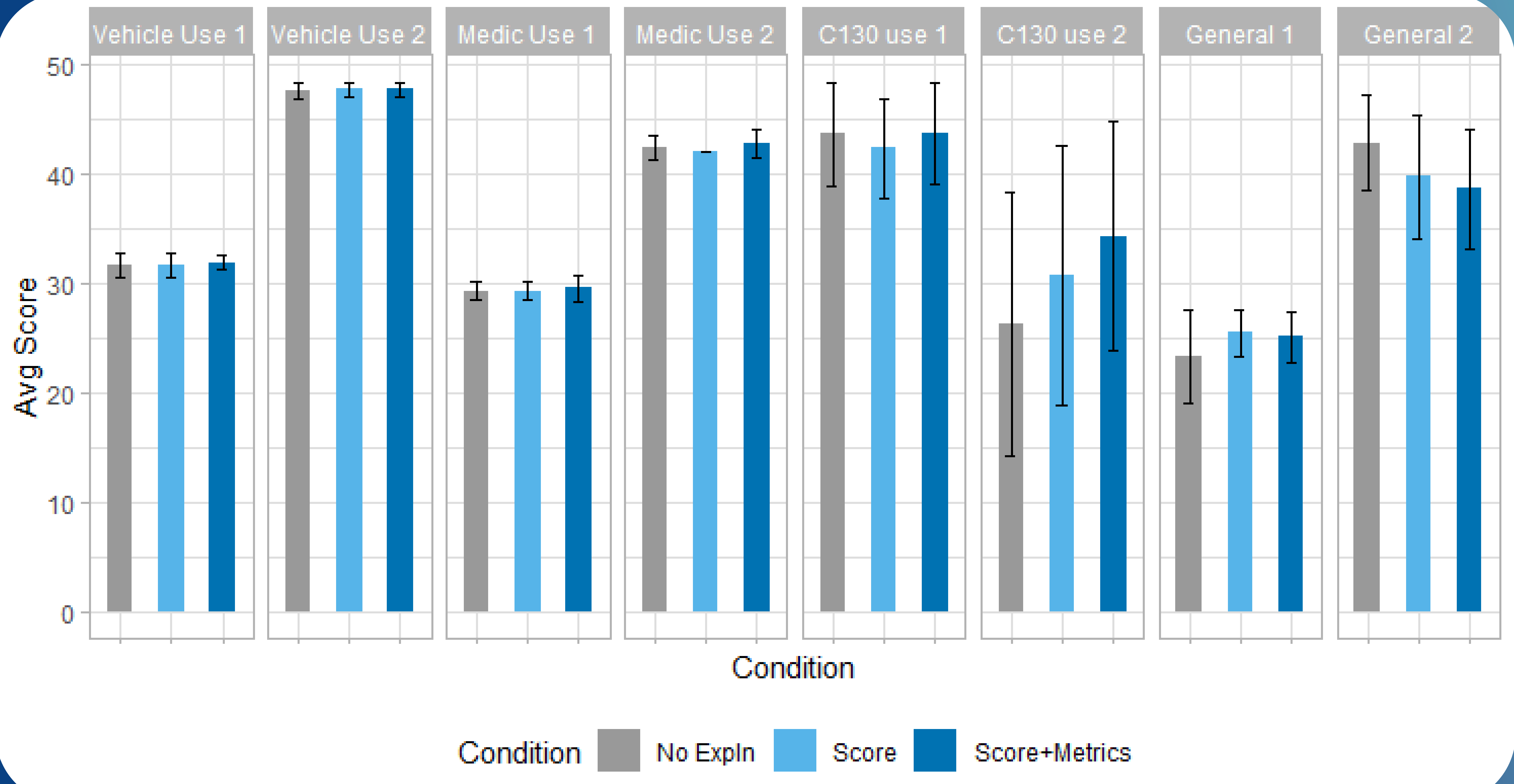
TECHNIQUES

- REINFORCEMENT LEARNING
- AI PLANNING
- MACHINE LEARNING
- COMPUTER VISION
- MULTI-AGENT SYSTEMS

DOMAINS

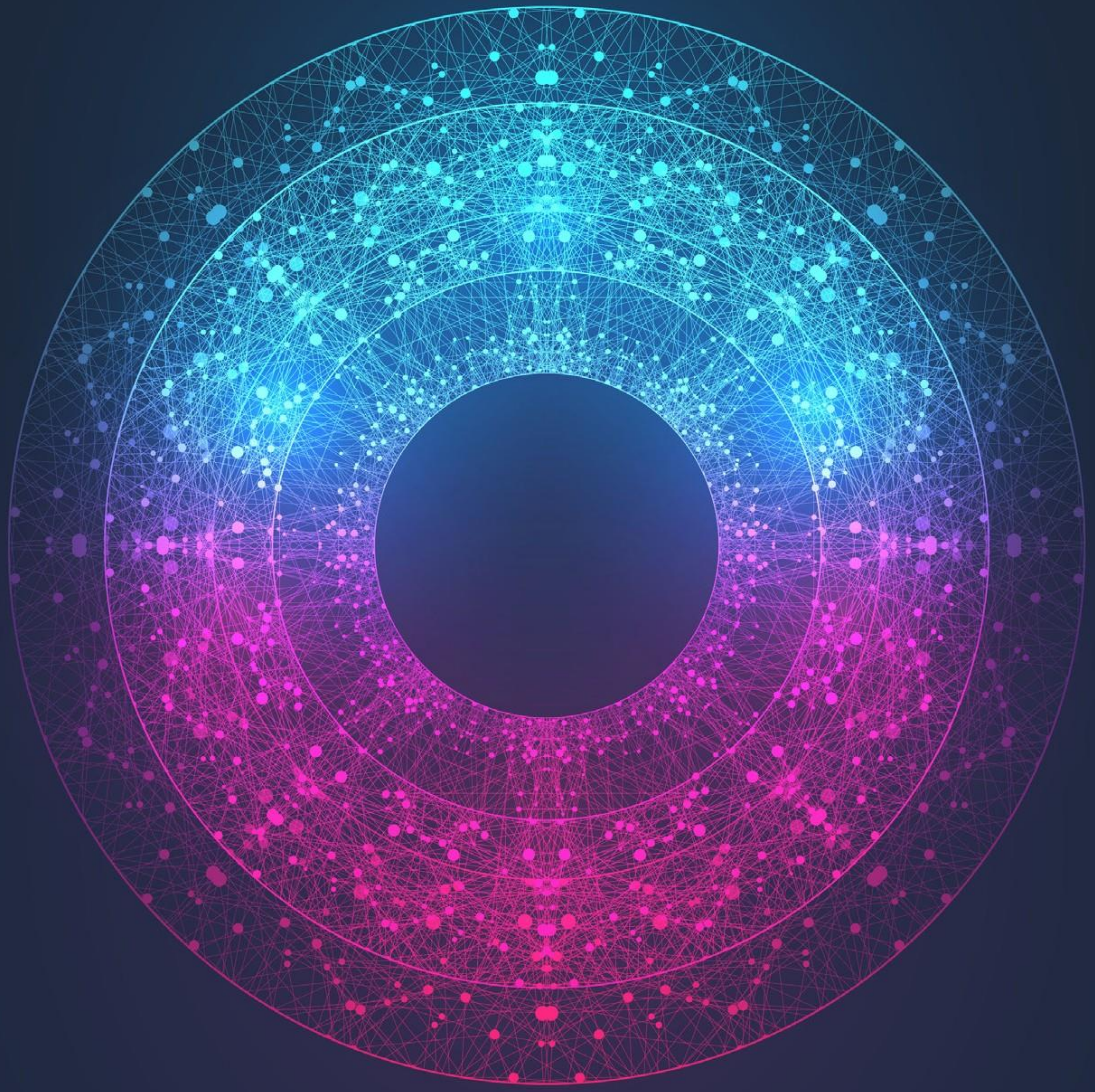
- SEARCH AND RESCUE PLANNING
- CREDIT SCORING
- MEDICAL IMAGING
- ILLEGAL FISHING
- GAME PLAYING

EXPERT DECISION MAKING

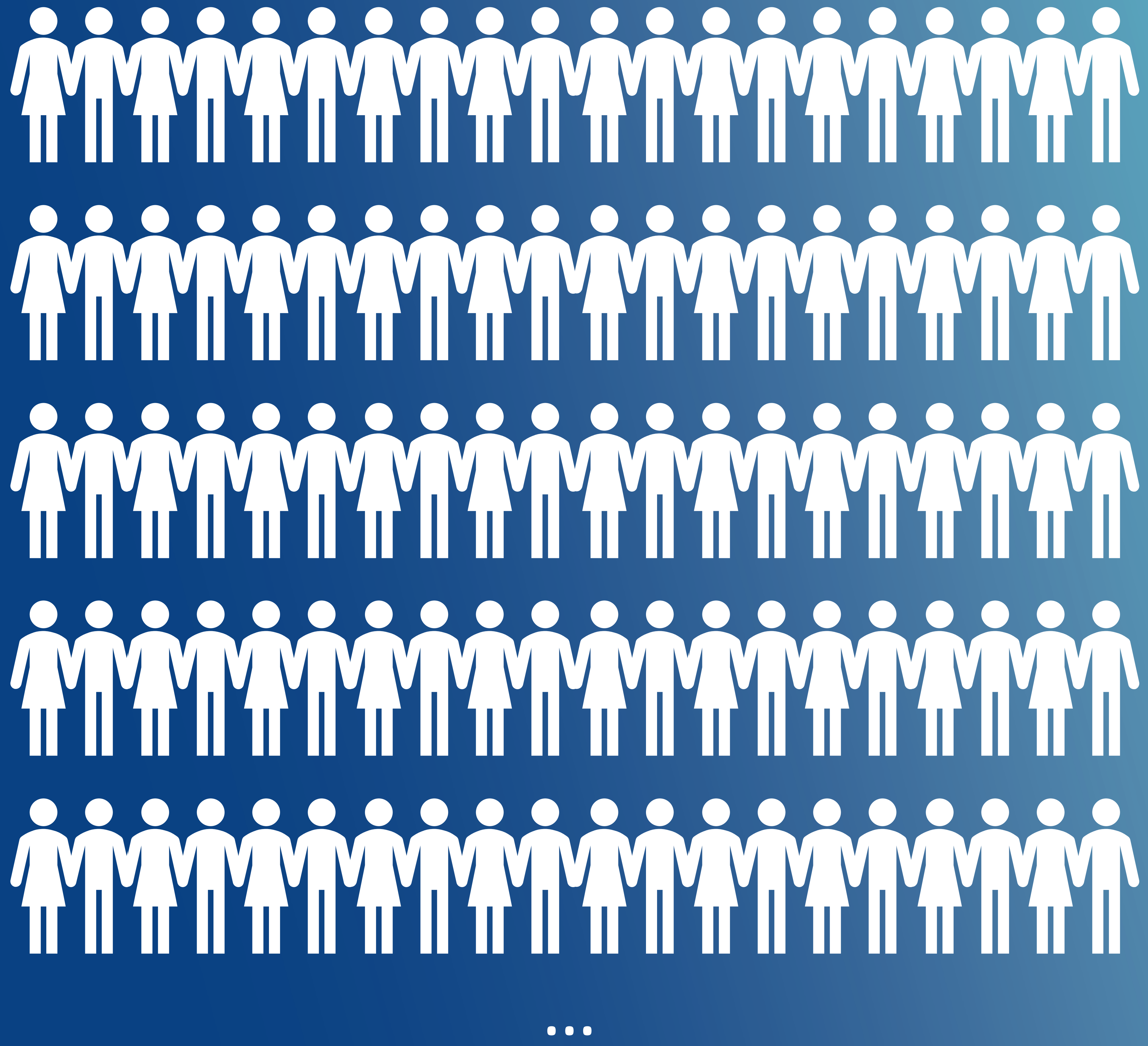


IS EXPLAINABLE AI DEAD?

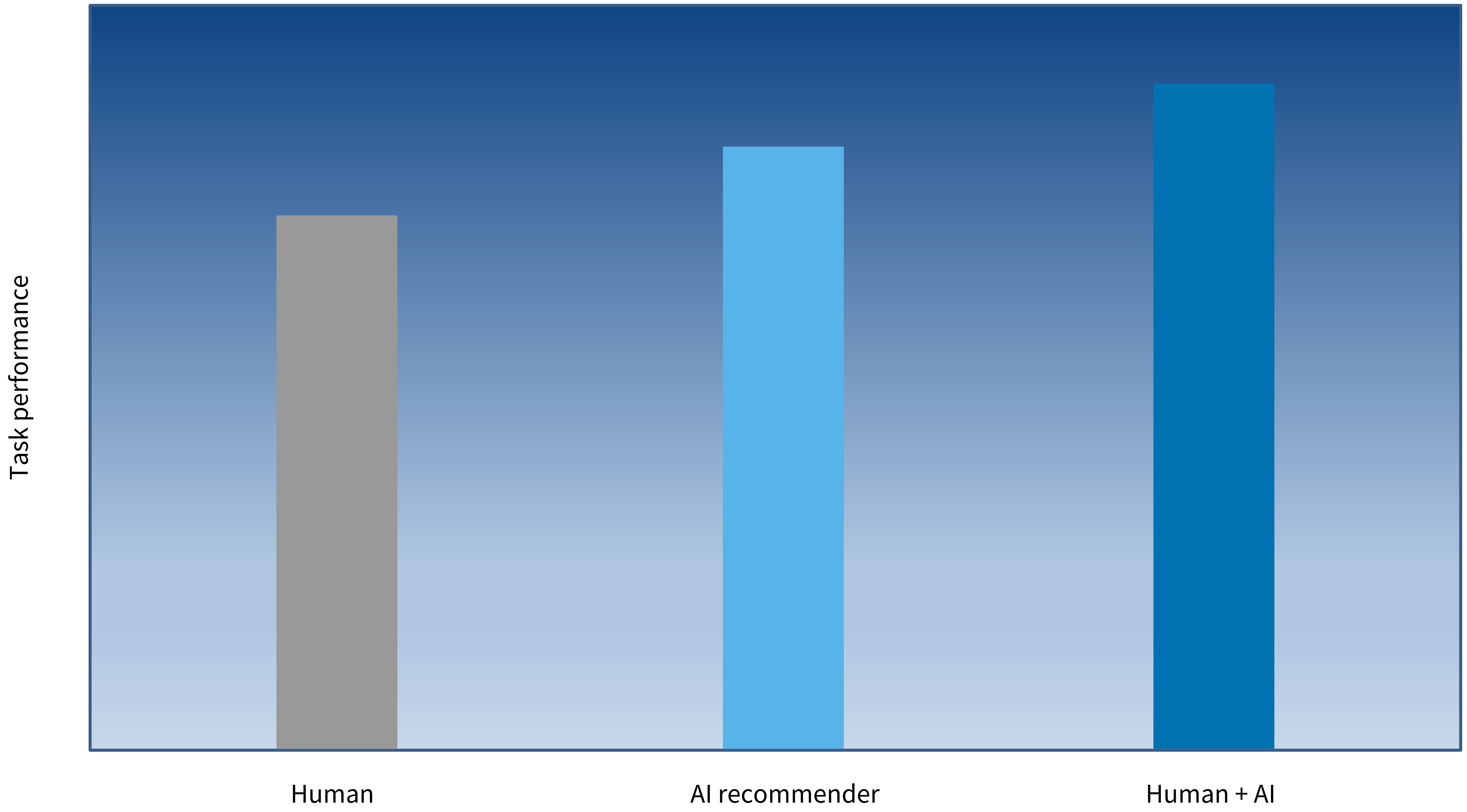
A QUICK SURVEY



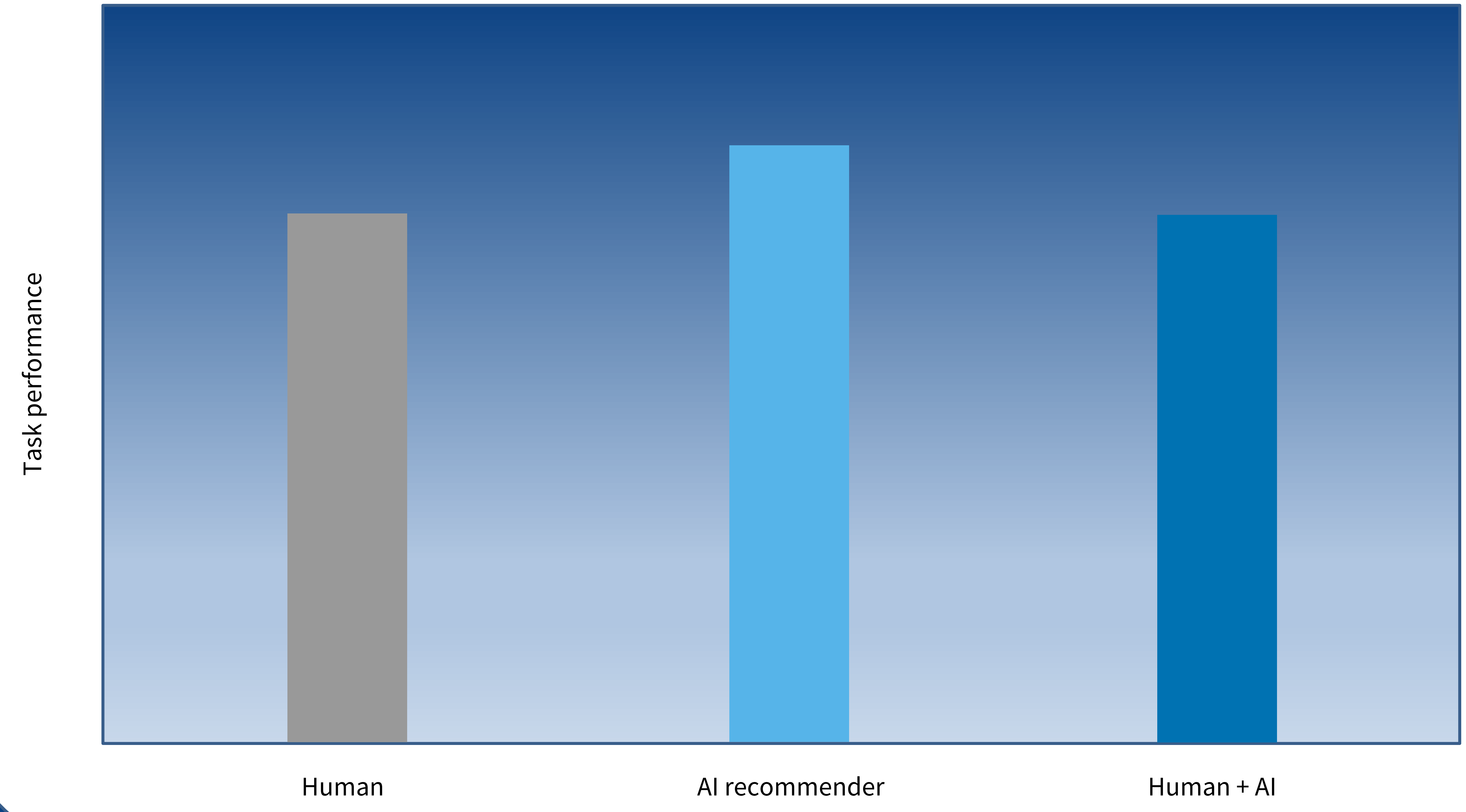
BLUSTER VS. PRUDENCE



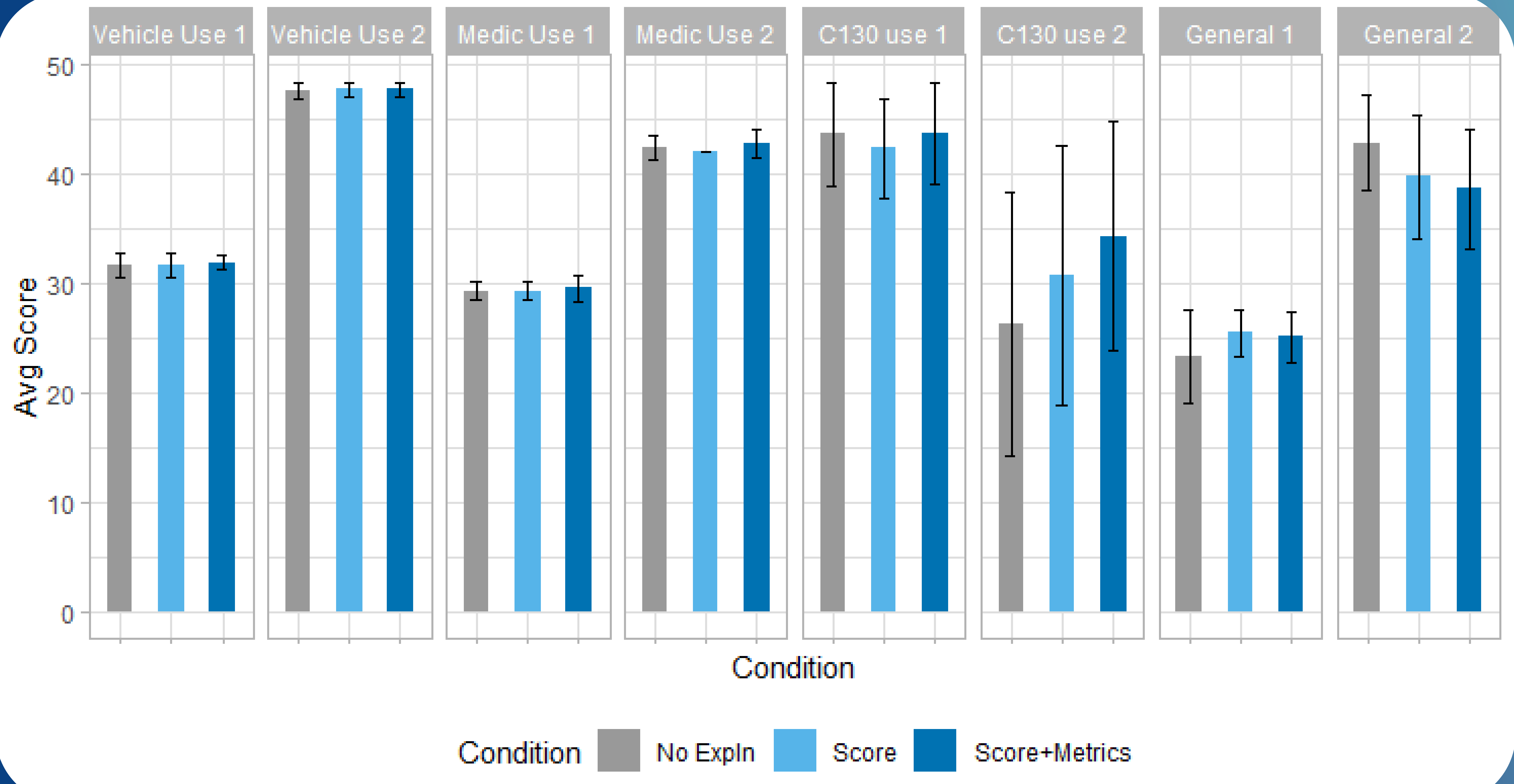
DECISION AIDS



DECISION AIDS



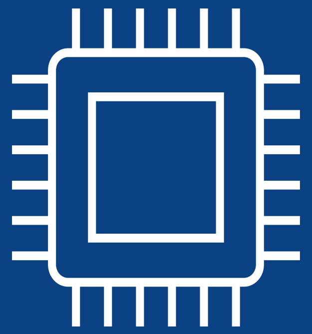
EXPERT DECISION MAKING



RECOMMENDATION-DRIVEN
EXPLAINABLE AI

HYPOTHESIS-DRIVEN
EVALUATIVE AI

(DIS)TRUST AND (UNDER-)RELIANCE



TRUSTWORTHY

TRUSTED

NOT TRUSTWORTHY

DISTRUSTED

(DIS)TRUST AND (UNDER-)RELIANCE

	TRUSTED	DISTRUSTED
TRUSTWORTHY	WARRANTED TRUST/RELIANCE	UNWARRANTED DISTRUST/ UNDER-RELIANCE
NOT TRUSTWORTHY	UNWARRANTED TRUST/ OVER-RELIANCE	WARRANTED DISTRUST/ RELIANCE

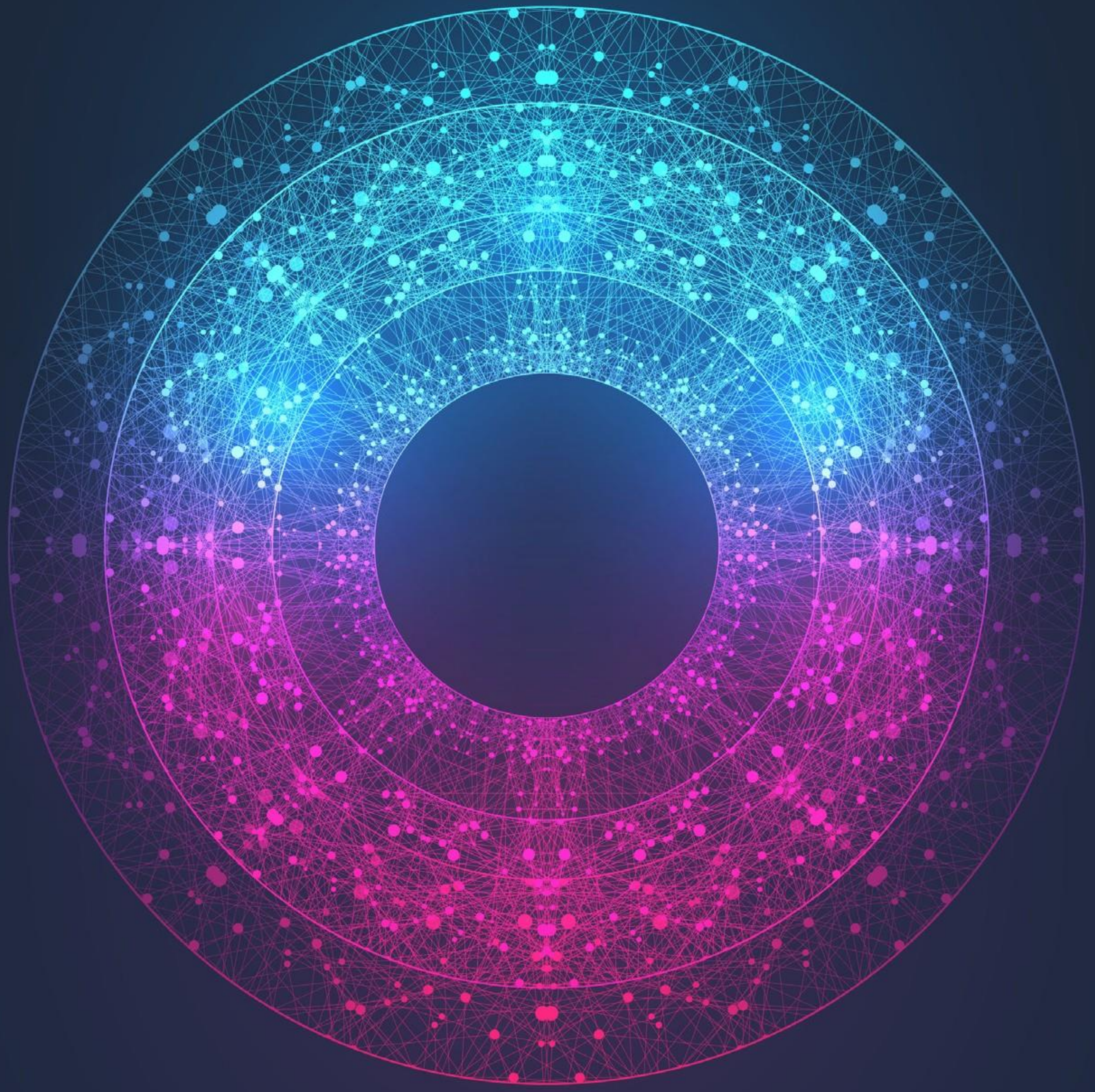
JACOVI, A., MARASOVIĆ, A., MILLER, T., & GOLDBERG, Y. FORMALIZING TRUST IN ARTIFICIAL INTELLIGENCE: PREREQUISITES, CAUSES AND GOALS OF HUMAN TRUST IN AI. IN *PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY* (FAccT), pp 624-635, 2021.

(DIS)TRUST AND (UNDER-)RELIANCE

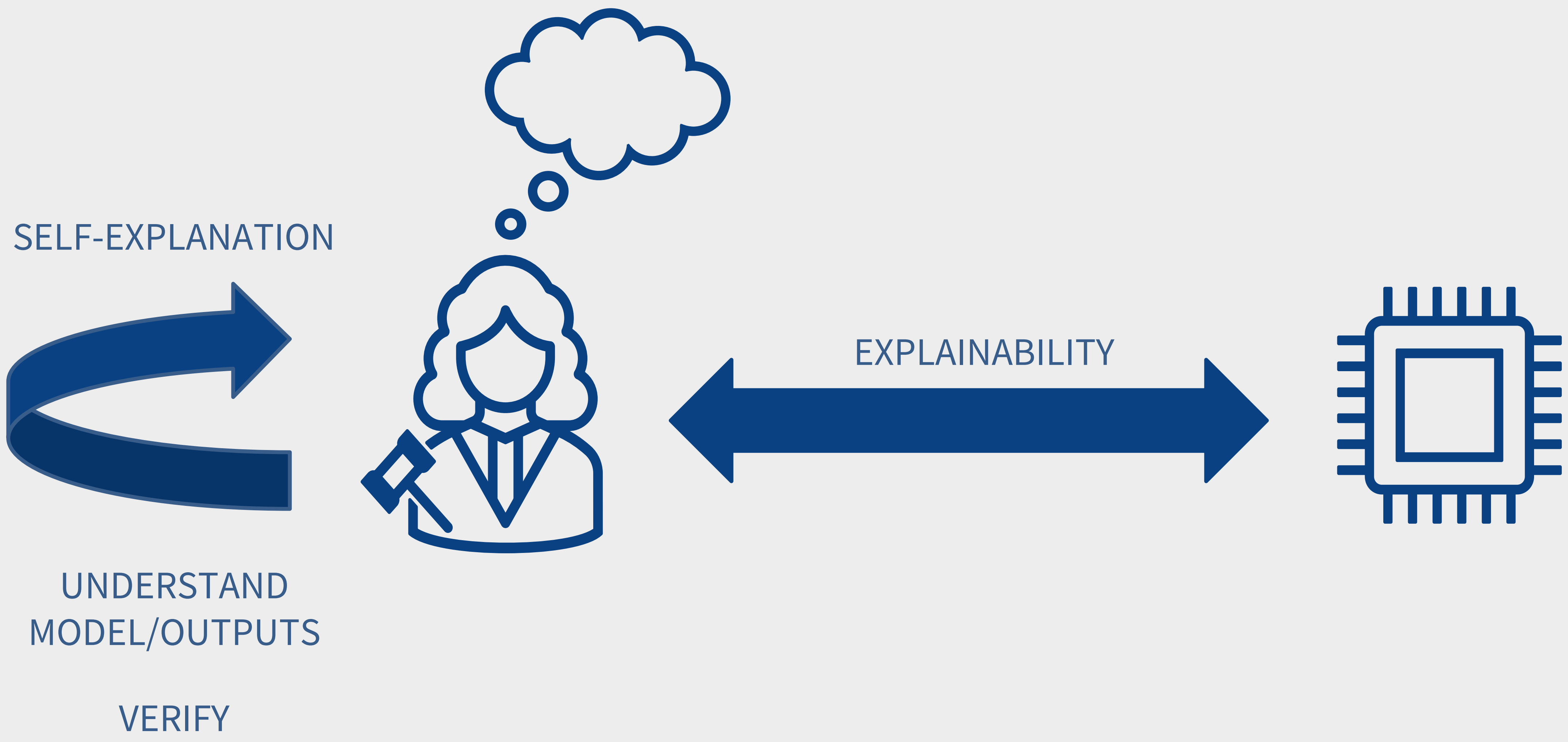
	TRUSTED	DISTRUSTED
TRUSTWORTHY	**WARRANTED TRUST/RELIANCE**	~~UNWARRANTED DISTRUST/ UNDER-RELIANCE~~
NOT TRUSTWORTHY	~~UNWARRANTED TRUST/ OVER-RELIANCE~~	**WARRANTED DISTRUST/ RELIANCE**

JACOVI, A., MARASOVIĆ, A., MILLER, T., & GOLDBERG, Y. FORMALIZING TRUST IN ARTIFICIAL INTELLIGENCE: PREREQUISITES, CAUSES AND GOALS OF HUMAN TRUST IN AI. IN *PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY* (FAccT), pp 624-635, 2021.

SELF-EXPLANATION IN DECISION MAKING



SELF EXPLANATION



ABDUCTIVE REASONING AND VERIFICATION

PROCESS	REQUIREMENTS
1. Observe event	
2. Generate hypotheses	
3. Judge plausibility	
4. Resolve explanation	
5. Extend explanation	

PSYCHOLOGY AND AI AT THE CROSSROADS: HOW MIGHT COMPLEX SYSTEMS EXPLAIN THEMSELVES? R. HOFFMAN, T. MILLER, W. CLANCEY
AMERICAN JOURNAL OF PSYCHOLOGY, 135(4), pp. 365-378, 2022.

ABDUCTIVE REASONING AND VERIFICATION

PROCESS	REQUIREMENTS
1. Observe event	Design interfaces to determine what has happened Design interfaces to highlight unusual events
2. Generate hypotheses	Help to construct (likely) hypotheses
3. Judge plausibility	
4. Resolve explanation	
5. Extend explanation	

PSYCHOLOGY AND AI AT THE CROSSROADS: HOW MIGHT COMPLEX SYSTEMS EXPLAIN THEMSELVES? R. HOFFMAN, T. MILLER, W. CLANCEY
AMERICAN JOURNAL OF PSYCHOLOGY, 135(4), pp. 365-378, 2022.

ABDUCTIVE REASONING AND VERIFICATION

PROCESS	REQUIREMENTS
1. Observe event	Design interfaces to determine what has happened
	Design interfaces to highlight unusual events
2. Generate hypotheses	Help to construct (likely) hypotheses
3. Judge plausibility	Help to explore how causes affect outputs
	Find evidence to support and refute hypotheses
4. Resolve explanation	Identify and record important information
5. Extend explanation	

PSYCHOLOGY AND AI AT THE CROSSROADS: HOW MIGHT COMPLEX SYSTEMS EXPLAIN THEMSELVES? R. HOFFMAN, T. MILLER, W. CLANCEY
AMERICAN JOURNAL OF PSYCHOLOGY, 135(4), pp. 365-378, 2022.

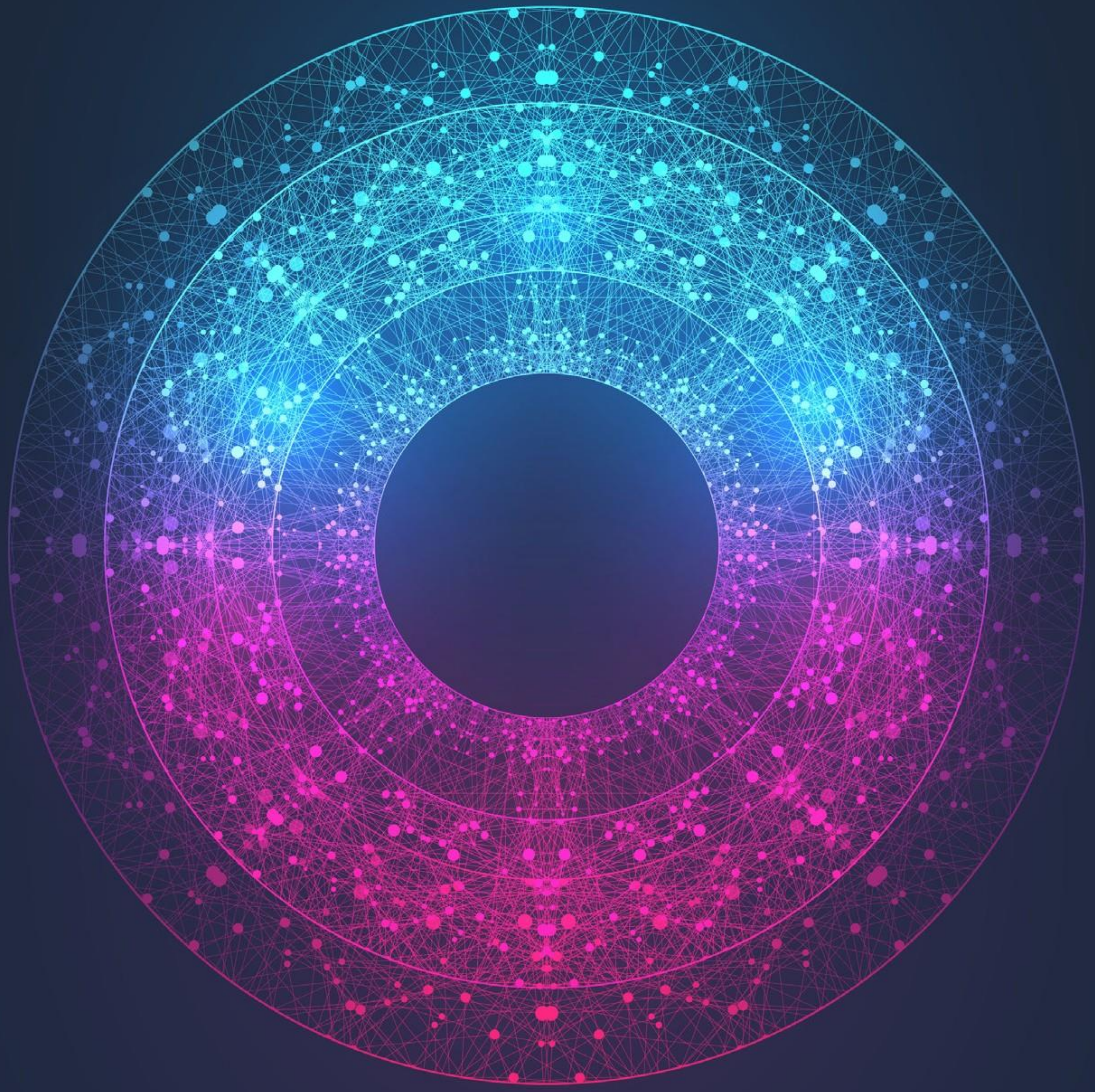
ABDUCTIVE REASONING AND VERIFICATION

PROCESS	REQUIREMENTS
1. Observe event	Design interfaces to determine what has happened Design interfaces to highlight unusual events
2. Generate hypotheses	Help to construct (likely) hypotheses
3. Judge plausibility	Help to explore how causes affect outputs Find evidence to support and refute hypotheses
4. Resolve explanation	Identify and record important information
5. Extend explanation	Support hypothesis revision Support interactive exploration

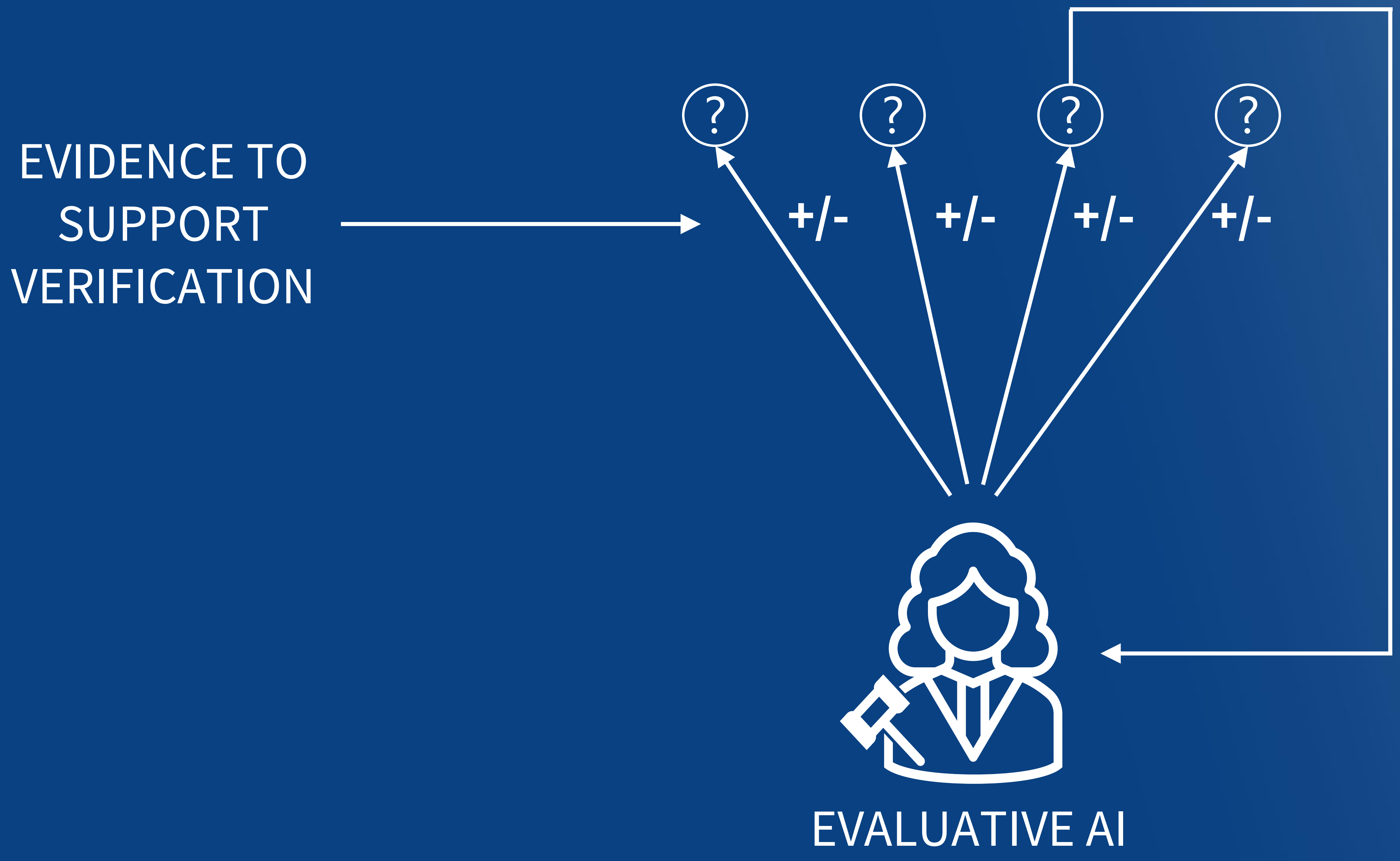
ABDUCTIVE REASONING AND VERIFICATION

PROCESS	REQUIREMENTS
1. Observe event	Design interfaces to determine what has happened Design interfaces to highlight unusual events
2. Generate hypotheses	Help to construct (likely) hypotheses
3. Judge plausibility	Help to explore how causes affect outputs Find evidence to support and refute hypotheses
4. Resolve explanation	Identify and record important information
5. Extend explanation	Support hypothesis revision Support interactive exploration

DECISION SUPPORT = VERIFICATION



EVALUATIVE AI



Lesion

Notes

Patient reports itchiness and bleeding.
Lesion has changed colour.

Lesion location

- ○ Head
- ○ Face
- ◉ Back
- ○ Front Torso
- ○ Upper arm
- ○ Hand/Lower Arm
- ○ Upper Leg
- ○ Foot/Lower Leg

Your hypothesis

Melanoma

Melanocytic Nevus

Basal Cell Caricinoma

Actinic Keratosis

Benign Keratosis

Dermatofibroma

Vascular Lesion

Evidence for

Lesion location

Colour

Scarred

Bleeding

Evidence against

Asymmetric shape

Changed colour

Itchiness

Lesion

Notes

Patient reports itchiness and bleeding.
Lesion has changed colour.

Lesion location

- ○ Head
- ○ Face
- ◉ Back
- ○ Front Torso
- ○ Upper arm
- ○ Hand/Lower Arm
- ○ Upper Leg
- ○ Foot/Lower Leg

Your hypothesis

- **Melanoma**
- Melanocytic Nevus
- **Basal Cell Caricinoma**
- **Actinic Keratosis**
- Benign Keratosis
- Dermatofibroma
- Vascular Lesion

Evidence for

- Asymmetric shape
- Changed colour
- Itchiness
- Bleeding
- Colour

Evidence against

- Scarred
- Legion location

IS EXPLAINABLE AI DEAD?

LONG LIVE EXPLAINABLE AI!

KEY TAKEAWAYS

EXPLAINABLE AI

Explainable decision aids don't really improved decision making (much)

Some false assumptions

- People look to machine recommendations
- People look to machine explanations
- Intuition needs to be overridden

HOWEVER ....

Evaluative AI provides the framework

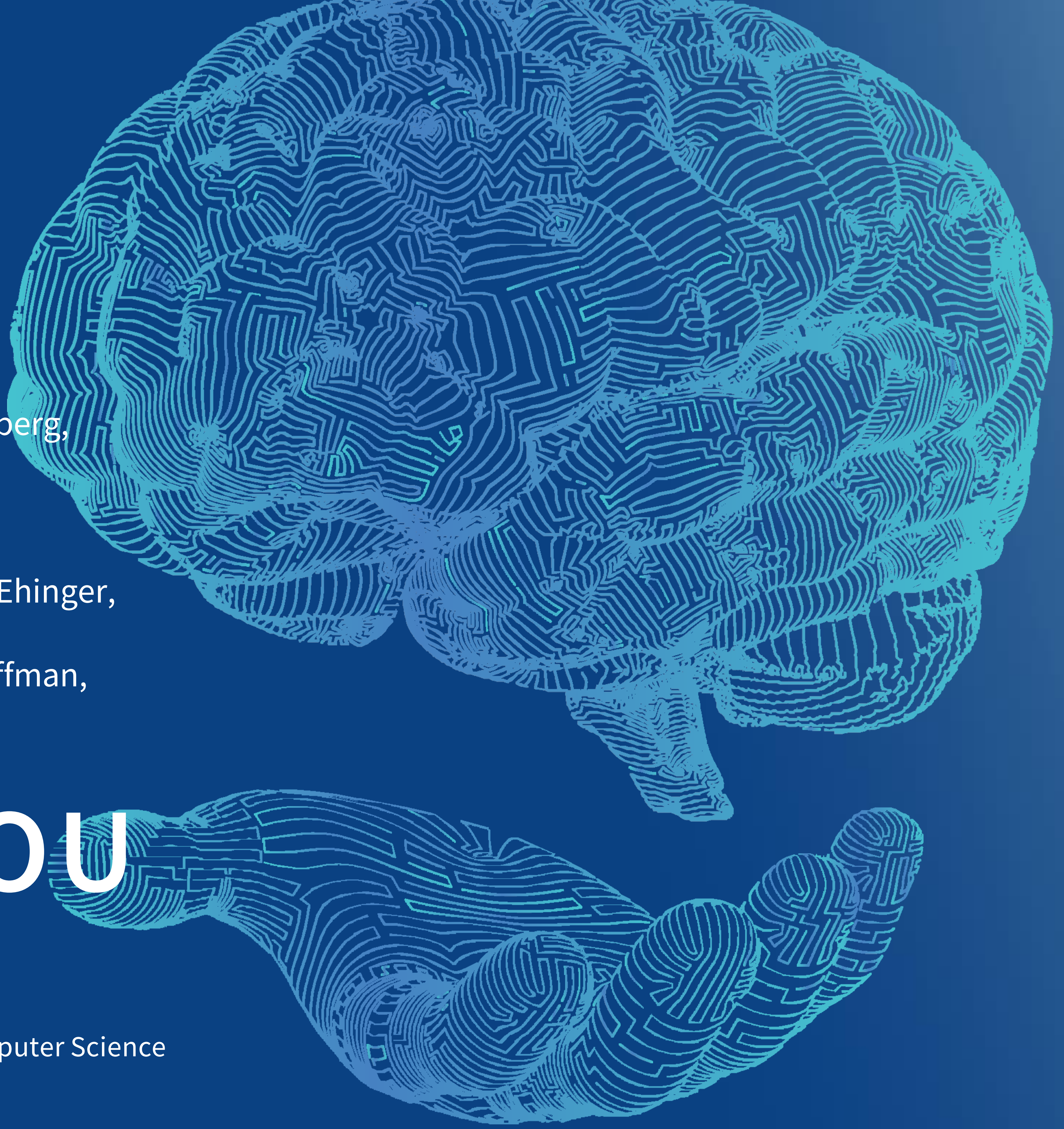
- **Support the human decision-making loop**
- Build on expertise and expert intuition

Focus on the user and their tasks/roles

Explainable AI is dead! ...

... Long live explainable AI!

Thanks to : Prashan Madumal,
Piers Howe, Ronal Singh, Liz Sonenberg,
Eduardo Velloso, Mor Vered,
Frank Vetere, Abeer Alshehri,
Ruihan Zhang, Emma Baillie,
Henrietta Lyons, Paul Dourish, Kris Ehinger,
Ben Rubinstein, Michelle Blom,
Thao Le, Rick Tompkins, Robert Hoffman,
Gary Klein, William Clancey



THANK YOU

Tim Miller

School of Electrical Engineering and Computer Science
The University of Queensland, Australia
@tmiller_uq

KEY TAKEAWAYS

EXPLAINABLE AI

Explainable decision aids don't really improved decision making (much)

Some false assumptions

- People look to machine recommendations
- People look to machine explanations
- Intuition needs to be overridden

HOWEVER ....

Evaluative AI provides the framework

- **Support the human decision-making loop**
- Build on expertise and expert intuition

Focus on the user and their tasks/roles

Explainable AI is dead! ...

... Long live explainable AI!