Explainable AI: Beware of Inmates Running the Asylum

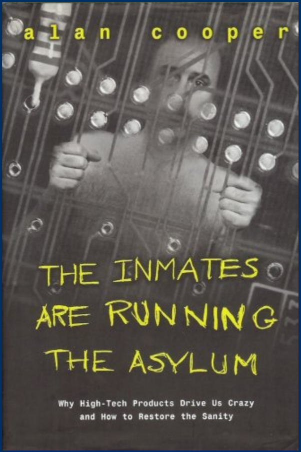
Or: How I Learnt to Stop Worrying and Love the Social Sciences

Tim Miller

School of Computing and Information Systems
Co-Director, Centre for AI & Digital Ethics
The University of Melbourne, Australia
tmiller@unimelb.edu.au

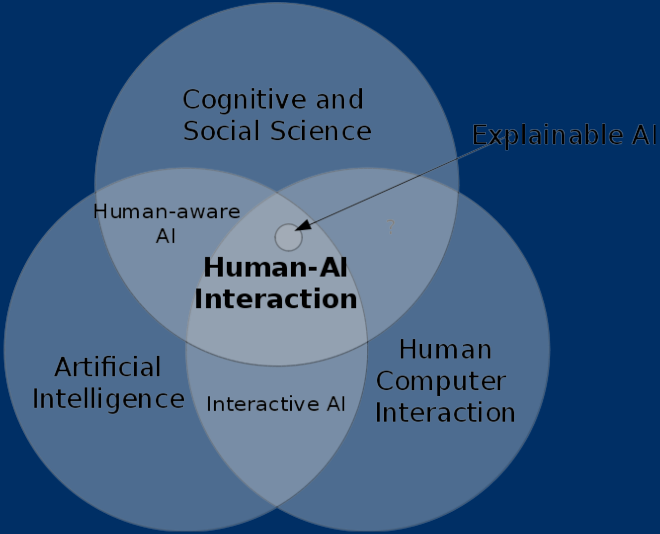
7 May, 2021

Inmates...



Alan Cooper (2004): *The Inmates Are Running the Asylum*
Why High-Tech Products Drive Us Crazy and
How We Can Restore the Sanity

Explainable Artificial Intelligence



Explanation in Artificial Intelligence

Explanation is answering a *why-question*.

Explanation in Artificial Intelligence

Explanation is answering a *why-question*.

This is: philosophy, cognitive psychology/science, and social psychology.

Infusing the Social Sciences

A patient has: (1) weight gain; (2) fatigue; and (3) nausea.

GP infers the following most likely causes

Cause	Symptom	Prob.
Stopped Exercising	Weight gain	80%
Mononucleosis	Fatigue	50%
Stomach Virus	Nausea	50%
Pregnancy	Weight gain, fatigue, nausea	15%

Infusing the Social Sciences

A patient has: (1) weight gain; (2) fatigue; and (3) nausea.

GP infers the following most likely causes

Cause	**Symptom**	**Prob.**
Stopped Exercising	Weight gain	80%
Mononucleosis	Fatigue	50%
Stomach Virus	Nausea	50%
Pregnancy	Weight gain, fatigue, nausea	15%

The 'Best' Explanation?

A) Stopped exercising and mononucleosis and stomach virus

OR

B) Pregnant

Infusing the Social Sciences

Artificial Intelligence 267 (2019) 1–38

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

ELSEVIER

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.

https://arxiv.org/abs/1706.07269

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." — D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin. 107 (1) (1990) 65–81.

Contrastive Explanation — The Difference Condition

Why is it a fly?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *arXiv preprint arXiv:1811.03163*, 2019. `https://arxiv.org/abs/1811.03163`

Contrastive Explanation — The Difference Condition

Why is it a fly?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *arXiv preprint arXiv:1811.03163*, 2019. `https://arxiv.org/abs/1811.03163`

Contrastive Explanation — The Difference Condition

Why is it a fly rather than a beetle?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *arXiv preprint arXiv:1811.03163*, 2019. `https://arxiv.org/abs/1811.03163`

Contrastive Explanation — The Difference Condition

Why is it a fly rather than a beetle?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	**2**	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	**5**	✔	2

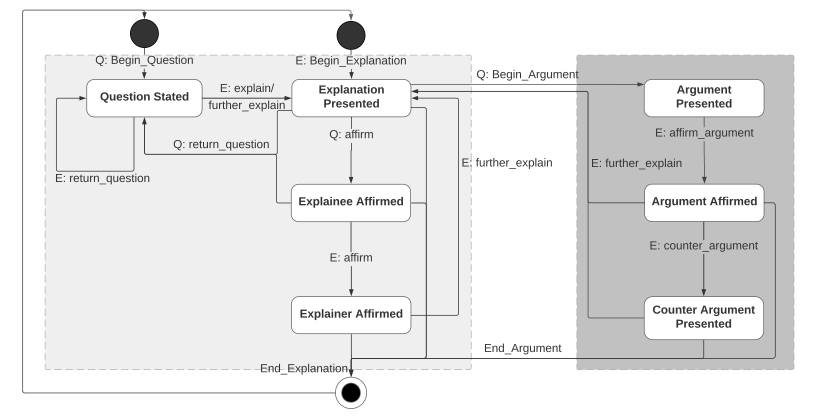
T. Miller. Contrastive Explanation: A Structural-Model Approach, *arXiv preprint arXiv:1811.03163*, 2019. `https://arxiv.org/abs/1811.03163`

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 (1) (1990) 65–81.

Social Explanation



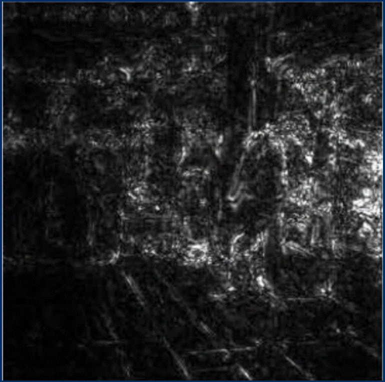
P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of AAMAS 2019*. https://arxiv.org/abs/1903.02409

Explanations are *Selected*

"There are as many causes of x as there are explanations of x. Consider how the cause of death might have been set out by the physician as 'multiple haemorrhage', by the barrister as 'negligence on the part of the driver', by the carriage-builder as 'a defect in the brakelock construction', by a civic planner as 'the presence of tall shrubbery at that turning'. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others."

N. R. Hanson, Patterns of discovery: An inquiry into the conceptual foundations of science, *CUP Archive*, 1965.

(Not) Infusing Human-Centered Studies



Source: Been Kim: Interpretability – What now? Talk at Google AI.
Saliency map generated using SmoothGrad

Evaluating XAI models

Metrics for Explainable AI: Challenges and Prospects

Robert R. Hoffman
Institute for Human and Machine Cognition [rhoffman@ihmc.us]
Shane T. Mueller
Michigan Technological University [shanem@mtu.edu]
Gary Klein
Macrocognition, LLC [gary@macrocognition.com]
Jordan Litman
Institute for Human and Machine Cognition [jlitman@ihmc.us]

Abstract

The question addressed in this paper is: If we present to a user an AI system that explains how it works, how do we know whether the explanation works and the user has achieved a pragmatic understanding of the AI? In other words, how do we know that an explainable AI system (XAI) is any good? Our focus is on the key concepts of measurement. We discuss specific methods for evaluating: (1) the goodness of explanations, (2) whether users are satisfied by explanations, (3) how well users understand the AI systems, (4) how curiosity motivates the search for explanations, (5) whether the user's trust and reliance on the AI are appropriate, and finally, (6) how the human-XAI work system performs. The recommendations we present derive from our integration of extensive research literatures and our own psychometric evaluations.

https://arxiv.org/abs/1812.04608

Our experience

We have used these insights over a range of techniques:

- Reinforcement learning
- Automated planning
- Computer Vision
- Multi-agent systems

and a range of domains:

- Credit scoring
- Search and rescue
- Illegal fishing
- Starcraft II

Fellow inmates, please consider . . .

Data Driven Models

Generation, selection, and evaluation of explanations is well understood
Social interaction of explanation is reasonably well understood

Fellow inmates, please consider . . .

Data Driven Models

Generation, selection, and evaluation of explanations is well understood
Social interaction of explanation is reasonably well understood

Validation

Validation on human behaviour data is necessary – at some point!

Remember: Hoffman et al., 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* `https://arxiv.org/abs/1812.04608`.

Wardens, please consider . . .

Models

Helping to improve the link between the social sciences and explainable AI.

Wardens, please consider . . .

Models

Helping to improve the link between the social sciences and explainable AI.

Interactions

Helping to study the design of interactions between ‘explainable’ intelligent agents and people.

Funding Acknowledgements

- *Explanation in Artificial Intelligence: A Human-Centred Approach* — Australian Research Council (2019–2021).
- *Catering for individuals' emotions in technology development* — Australian Research Council (2016-2018).
- *Human-Agent Collaborative Planning* — Microsoft Research Cambridge.
- *"Why?": Causal Explanation in Trusted Autonomous Systems* — CERA Next Generation Technologies Fund grant.

Overview

Explainability is a human-agent interaction problem

The social sciences community perhaps already knows more than the AI community about XAI

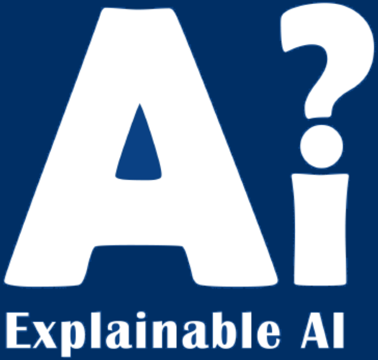
Integrating social science research has been useful for my lab:

1. Contrastive explanation
2. Social explanation
3. Selecting explanations

Cross-disciplinary research teams are important!

Thanks! And Questions....

Thanks: Prashan Madumal, Piers Howe, Ronal Singh, Liz Sonenberg, Eduardo Velloso, Mor Vered, Frank Vetere, Abeer Alshehri, Ruihan Zhang, Henrietta Lyons, Paul Dourish.



Overview

Explainability is a human-agent interaction problem

The social sciences community perhaps already knows more than the AI community about XAI

Integrating social science research has been useful for my lab:

1. Contrastive explanation
2. Social explanation
3. Selecting explanations

Cross-disciplinary research teams are important!