



INFUSING THE SOCIAL SCIENCES



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

ARTICLE INFO

Article history:

Received 22 June 2017

Received in revised form 17 May 2018

Accepted 16 July 2018

Available online 27 October 2018

Keywords:

Explanation

Explanatory

Interpretability

Explainable AI

Transparency

ABSTRACT

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.

© 2018 Elsevier B.V. All rights reserved.

