

XAI FUNDAMENTALS CHALLENGE

Dr Lina Kyrimi

e.kyrimi@qmul.ac.uk

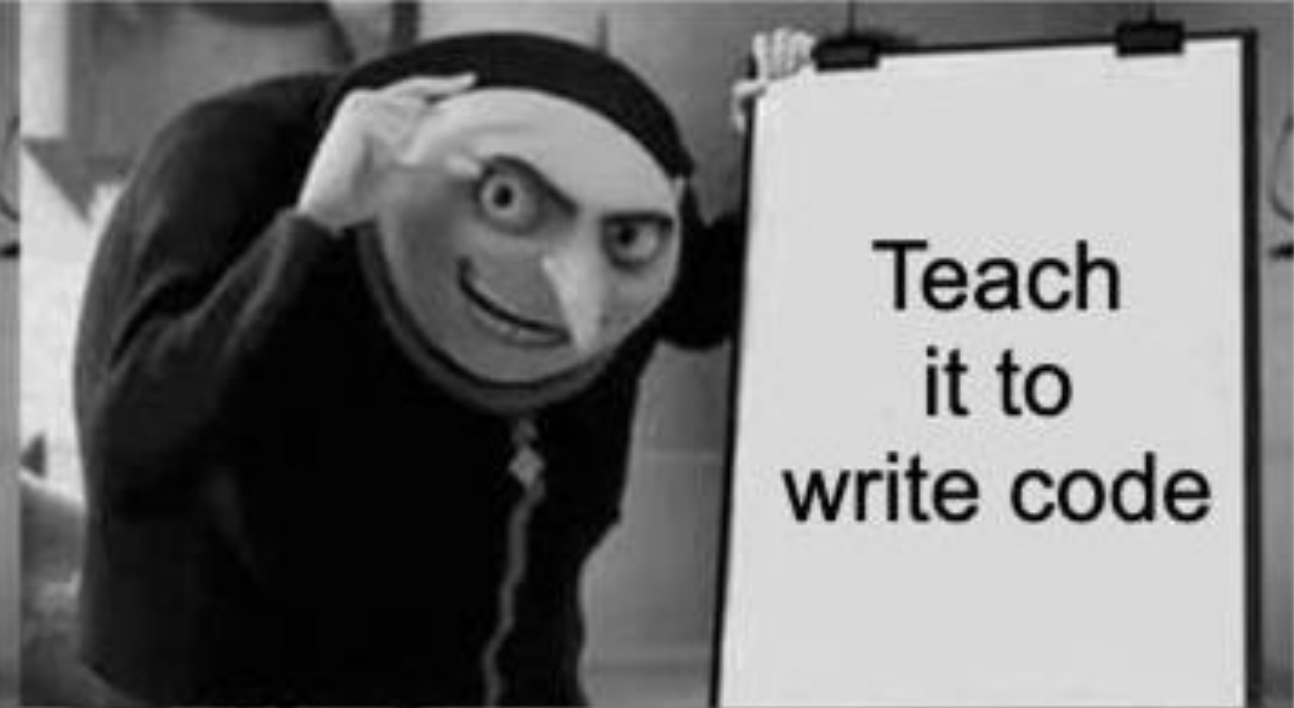
OUTLINE

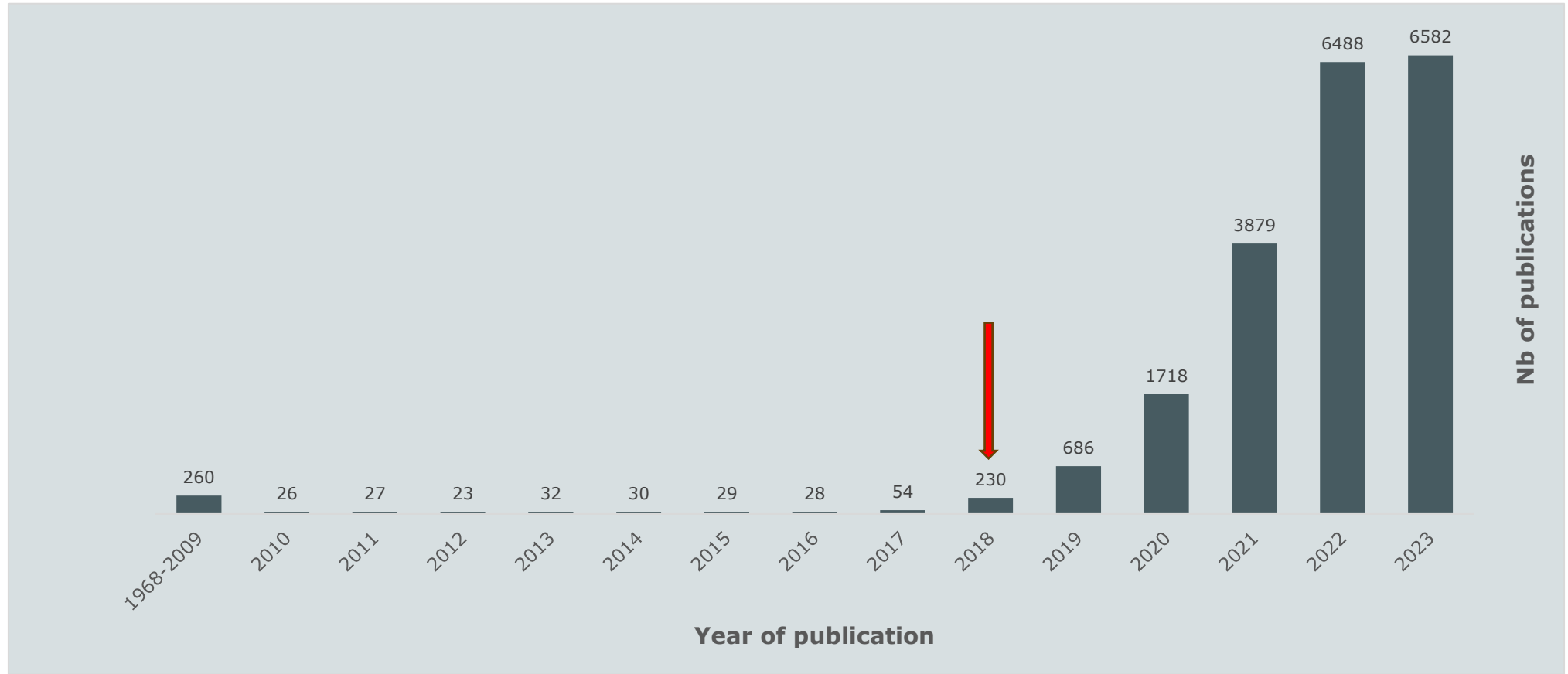
Research Interest

Fundamentals Challenge

Existing work

ExAIDSS Objective









FUNDAMENTALS CHALLENGE

Lack of formal definition of what an explanation in AI is

Lack of a set of global attributes for a “good” explanation in AI

WHY THIS IS AN ISSUE?

Communication

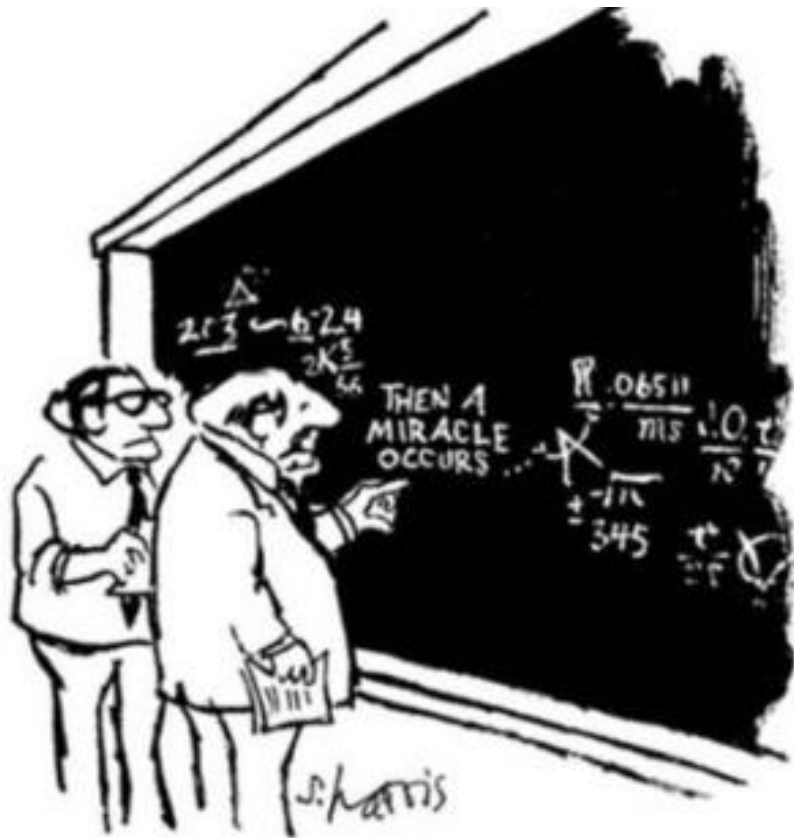
Unproductive
Disjoint goals

Development

No user focus

Validation

Unable to compare the
quality of different
explanations



"I think you should be more explicit here in step two."

EXISTING WORK

"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"

"the details or reasons that someone gives to make something clear or easy to understand"

"the details or reasons a model gives to make its functioning clear or easy to understand"

"making intelligible and providing insight into the outcome of AI systems"

"the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts"

LIMITATIONS

Do not provide
guidance

“the inmates
running the
asylum”

Circular
definitions

Domain-specific

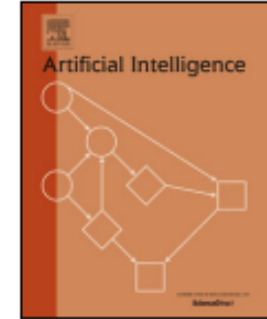


ELSEVIER

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

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



EXISTING WORK - CHARACTERISTICS

Contrastive

Selected

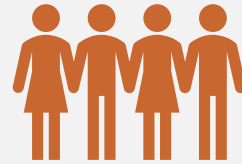
Causal

Social

LIMITATIONS



Based on literature



Human stakeholders not involved



Restricted to a specific domain

ExAIDSS

Explainable AI to ensure trust in clinical Decision Support Systems

[Home](#)

[About](#)

[Team](#)

[Publications](#)

[Events](#)



**Royal Academy
of Engineering**



Queen Mary
University of London

EXAIDSS OBJECTIVES



INVESTIGATE THE
FUNDAMENTALS
CHALLENGE



DEVELOP CAUSAL
EXPLANATION
ALGORITHMS



CREATE USER-
SPECIFIC
EXPLANATION
OUTPUTS



PROPOSE AN
EVALUATION
PROTOCOL



INTEGRATE THE
EXPLANATION
ALGORITHMS INTO
EXISTING
PLATFORMS

EXAIDSS: FUNDAMENTALS CHALLENGE

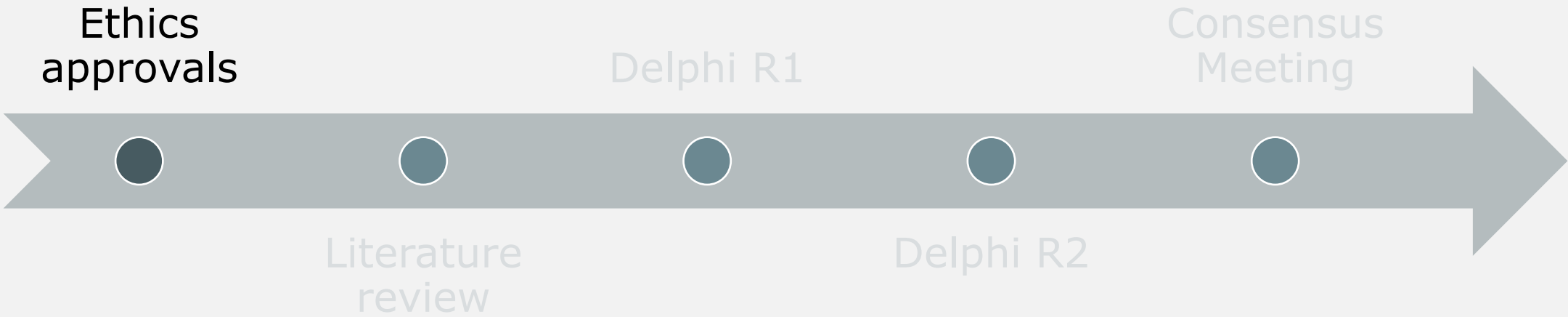


Provide a definition and a global list of characteristics of a good explanation for health-AI.

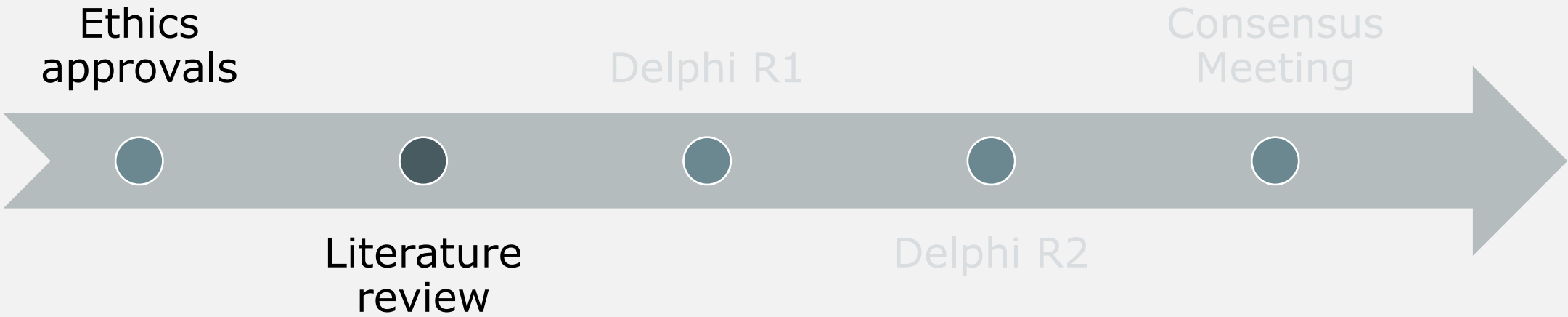


Synthesis published data and expert opinions from a diverse research background

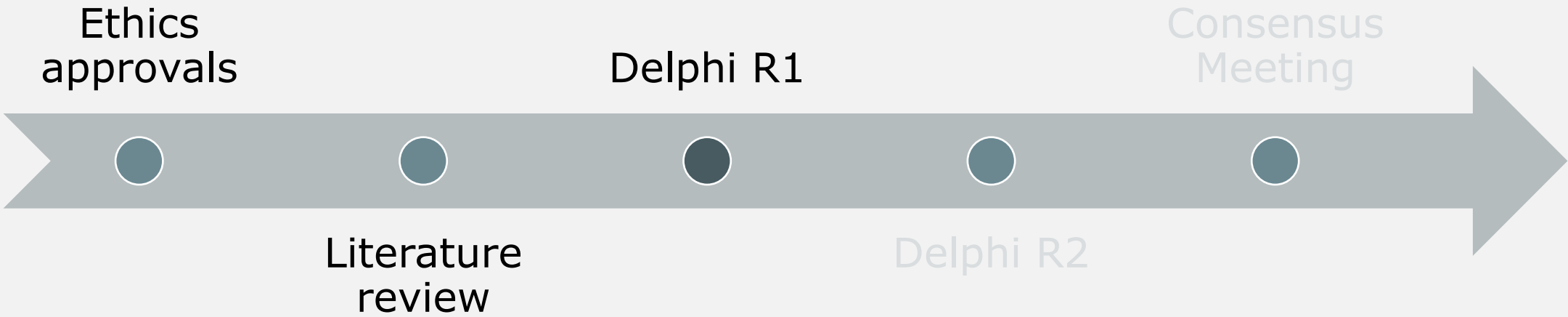
ACTION PLAN



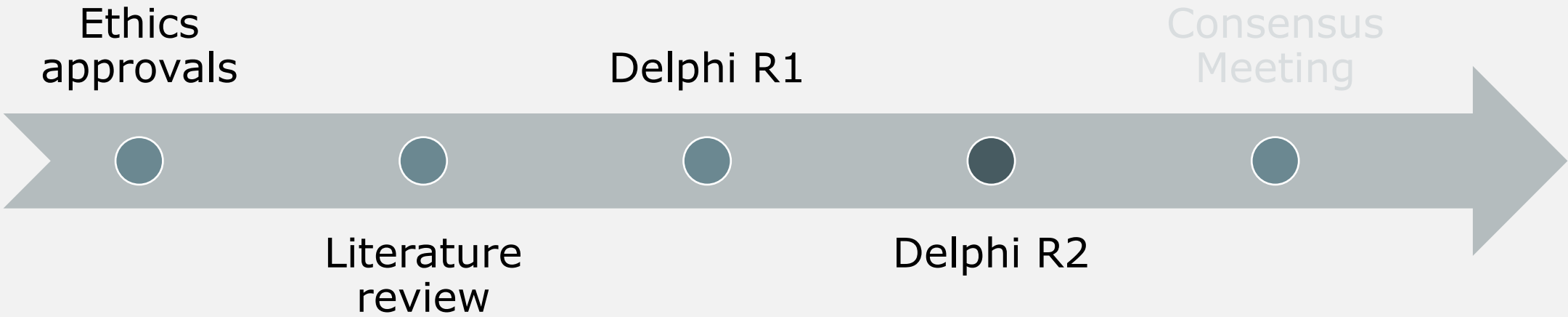
ACTION PLAN



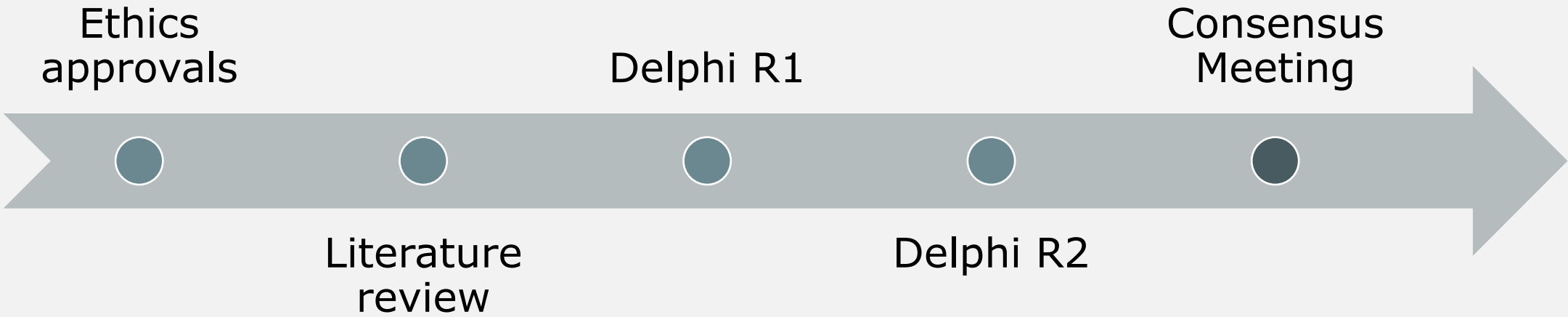
ACTION PLAN

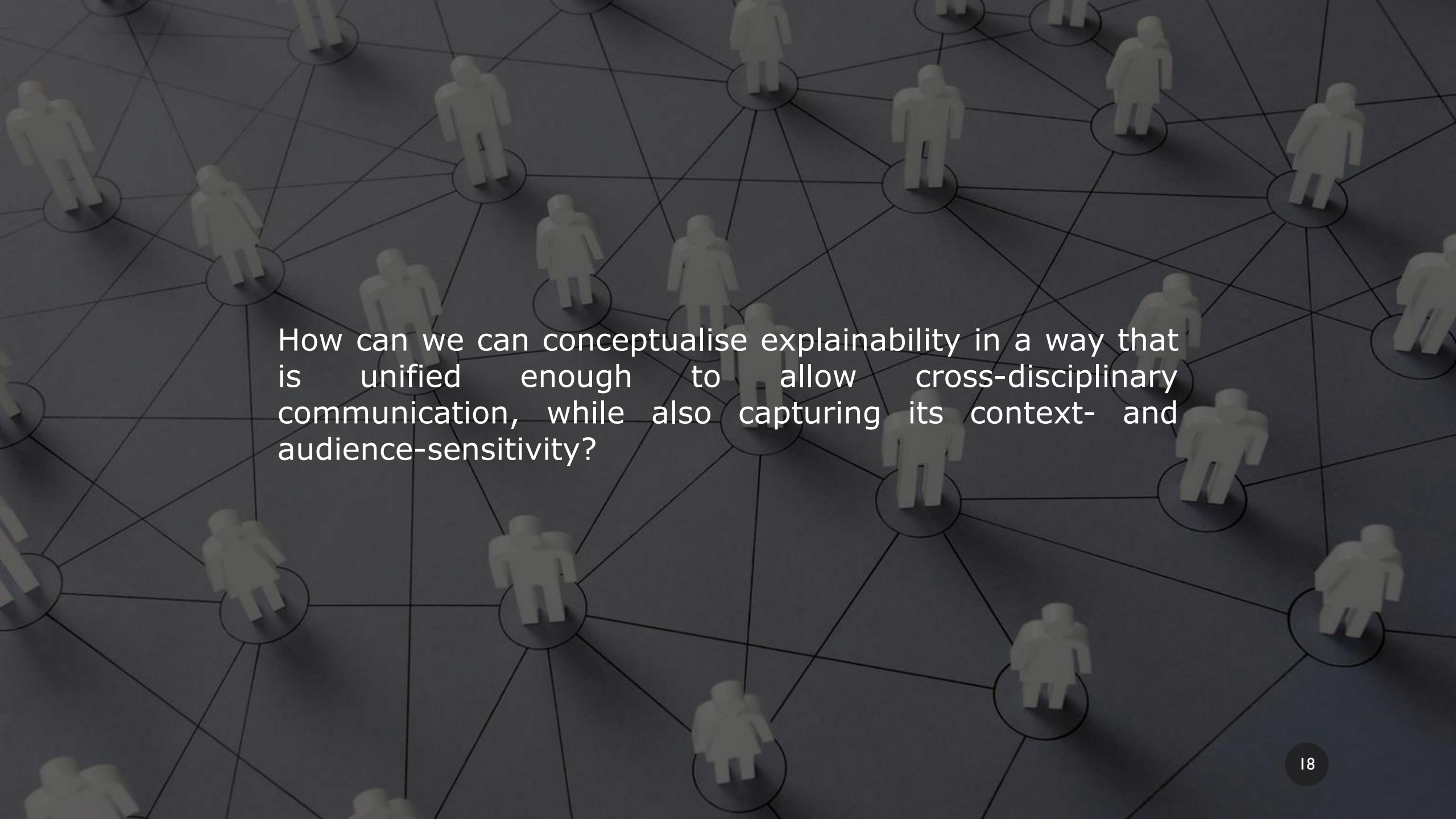


ACTION PLAN



ACTION PLAN





How can we can conceptualise explainability in a way that is unified enough to allow cross-disciplinary communication, while also capturing its context- and audience-sensitivity?

