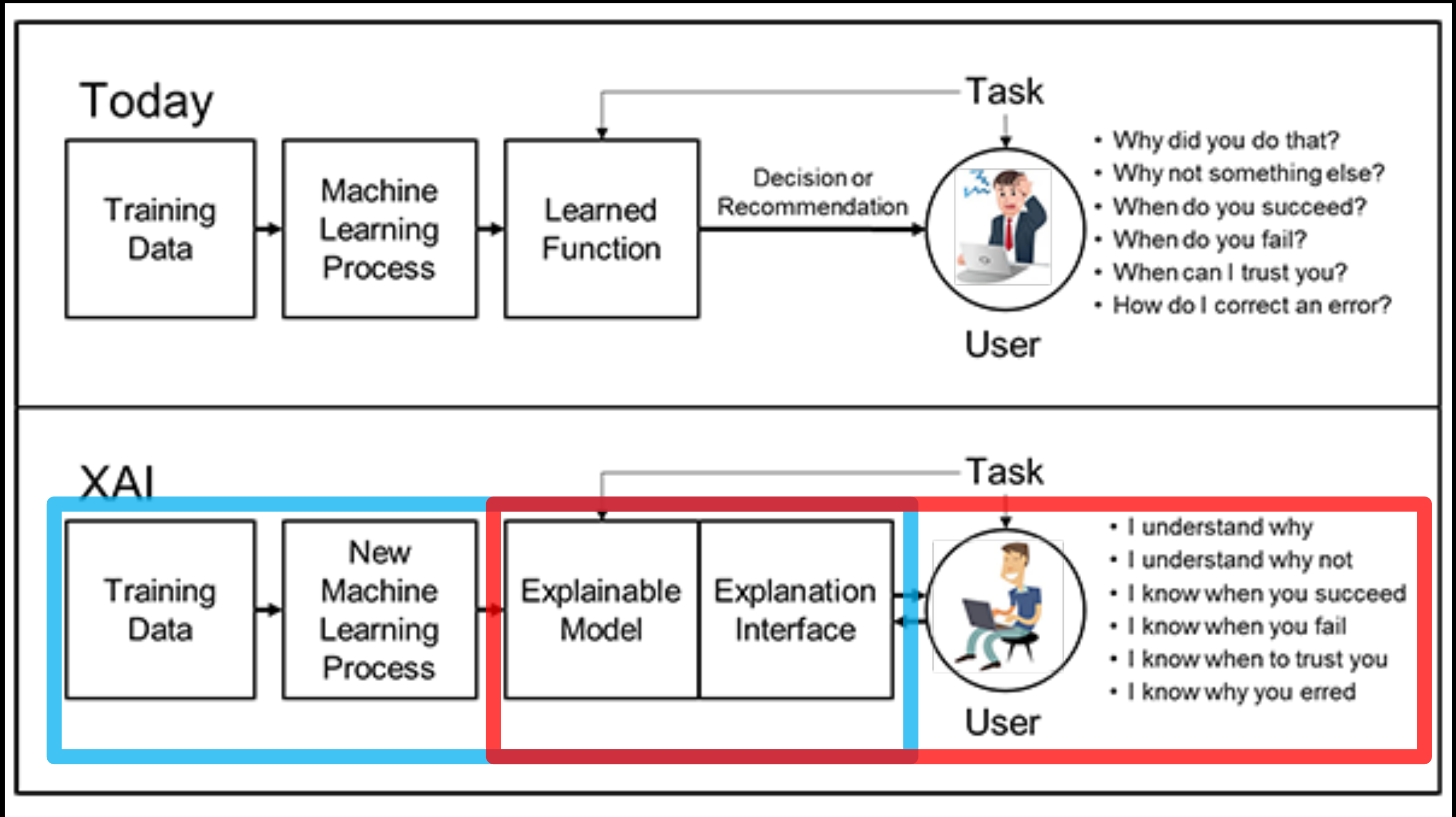
When explanations might do more harm than good

Dr Simone Stumpf
Centre for HCI Design
City, University of London

Simone.Stumpf.1@city.ac.uk
@DrSimoneStumpf

XAI vision



"Appropriate trust"

Technical

Human

Unfortunate news for XAI

- No explanations desired for certain tasks and contexts

[Bunt et al. IUI 2012]

- Different people need different explanations

[Gunning et al. Science Robotics 2019]

- Perceived control increases user satisfaction

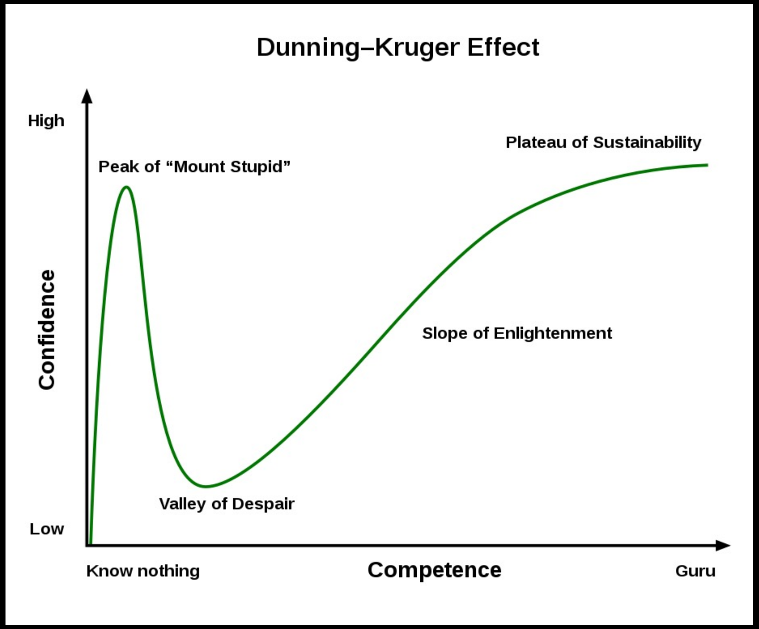
[Smith-Renner et al. CHI 2020]

- “Placebic” explanations and persuasive force

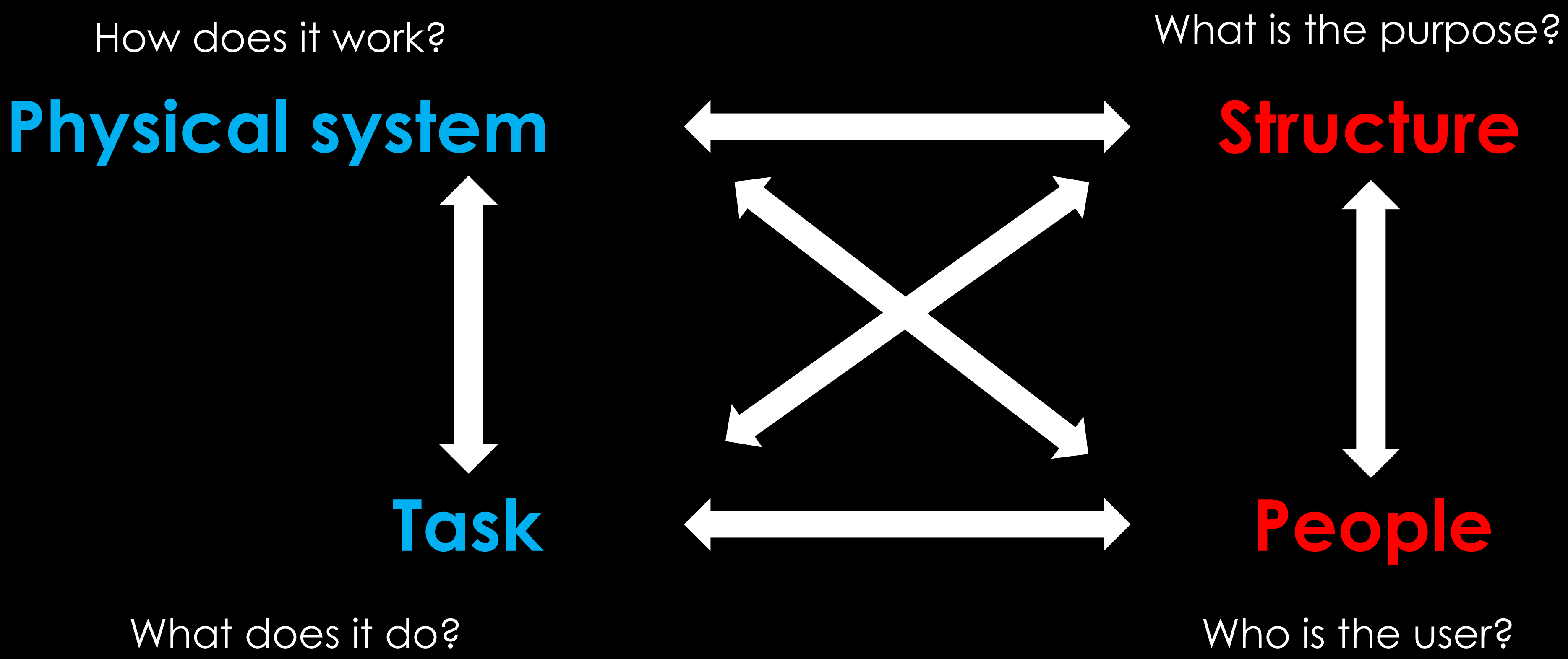
[Eiband et al. CHI 2019, Bussone et al. ICMI 2015]

- Explanations might be outside of the system itself

[Ehsan et al. CHI 2021]



Complex socio-technical system



How does it work?

Physical system

What is the purpose?

Structure

Task

People

What does it do?

Who is the user?

Structure

- Why explain?
  - Increased adoption / trust / satisfaction
  - Better use / appropriate trust
  - Spot the mistakes / biases
  - Better training data

People

- Who are we explaining to?
  - Expectations and attitudes
  - Capabilities
  - Mental models

Tasks

- What decisions/ recommendations/actions are we trying to explain?
  - High risk versus low risk
  - Level of automation
  - Situational context

Physical systems

- How does it work?
  - Models
  - Interfaces
  - Interactions

Five take-aways

- Design with humans in mind
- Know why you are explaining
- Understand the intended users
- Analyse the task and the situational context
- Think about what you want to optimise