The Alan Turing Institute

Queen Mary University of London

AI Ethics and Governance in Practice Programme

Professor David Leslie

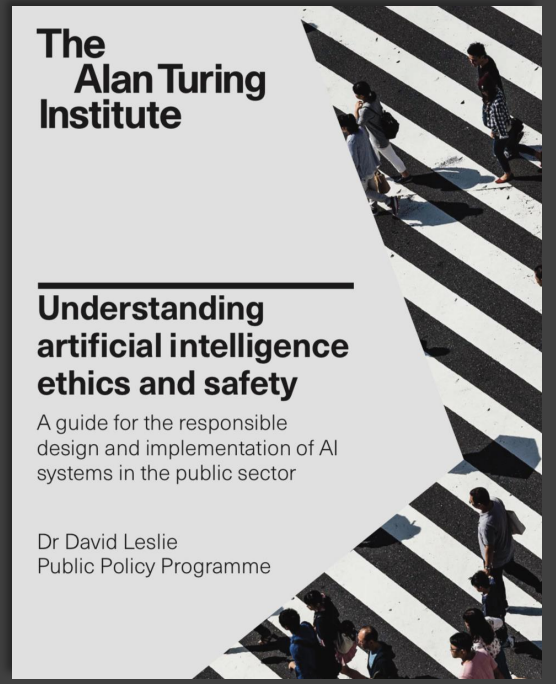
Director of Ethics and Responsible Innovation Research

Professor of Ethics, Technology and Society (QMUL)

July 2024

UK National Guidance on AI ethics and safety in the public sector

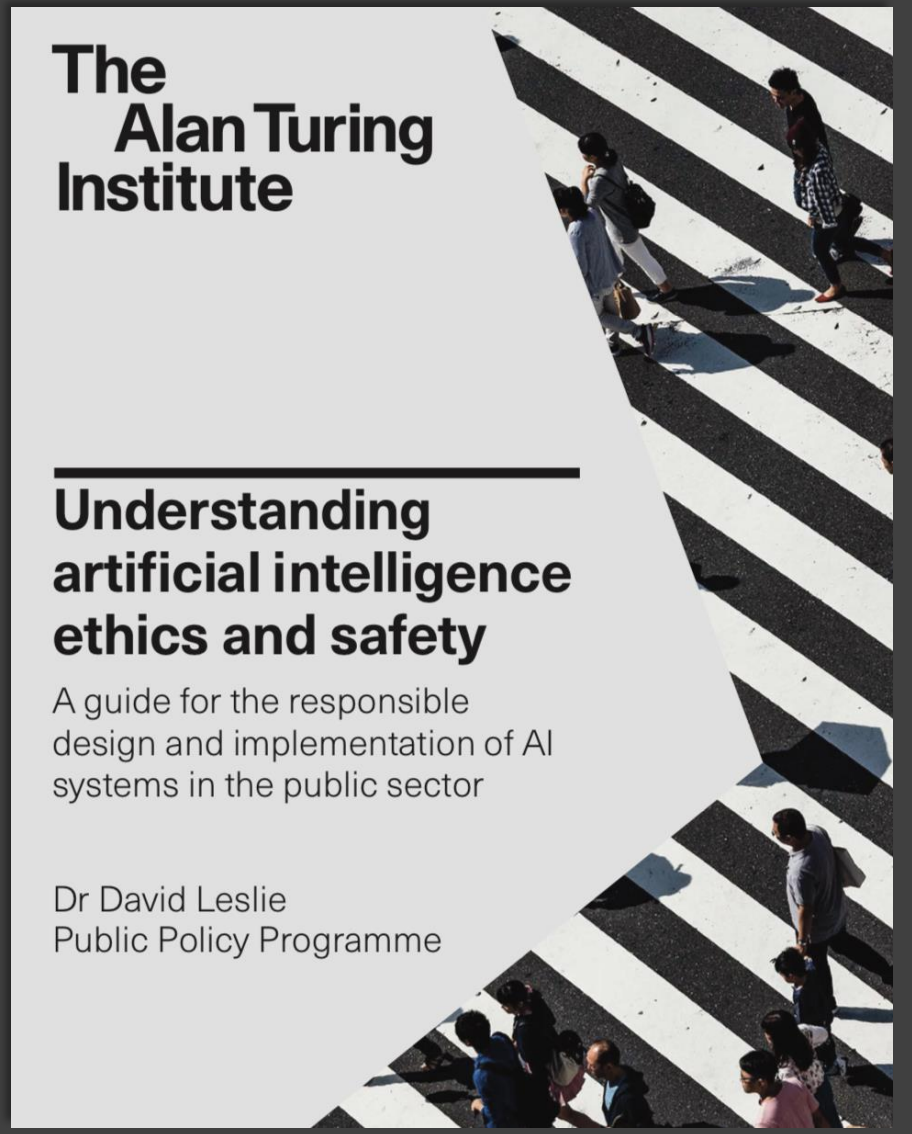
- Funded by EPSRC in 2018 to do fieldwork with Ministry of Justice to develop ethics framework which formed the normative foundation of the guidance
- Published June 2019 in collaboration with the Office for AI and GDS as part of the government's official guide to using AI in the public sector with Ministerial approval
- Now the **most accessed and cited public sector AI ethics guidance in the world**.
- Credited with initiating the move from "principles to practice in the field of AI policy and governance



UK National Guidance on AI ethics and safety in the public sector

Guidance has since been put into use by:

- Ministry of Justice
- Ministry of Defence
- Department for Environment Food & Rural Affairs
- Centre for Connected & Autonomous Vehicles
- Home Office
- ico. Information Commissioner's Office
- NHSX
- FCA Financial Conduct Authority
- GCHQ
- Geospatial Commission
- Department for Digital, Culture, Media & Sport
- Government Digital Service
- Office for Artificial Intelligence

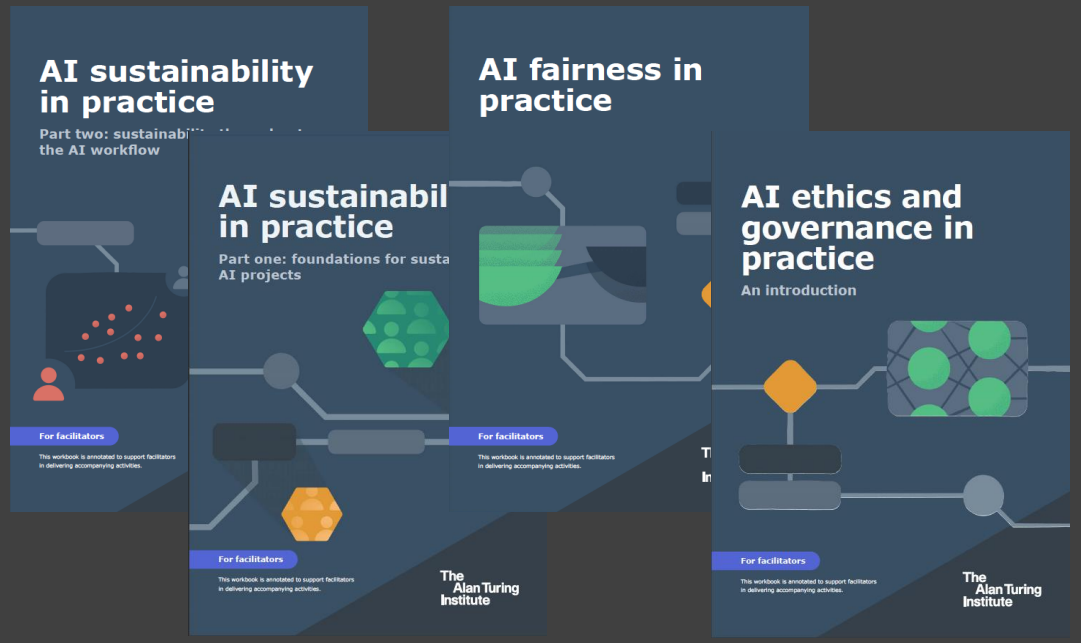


AI Ethics and Governance in Practice Programme

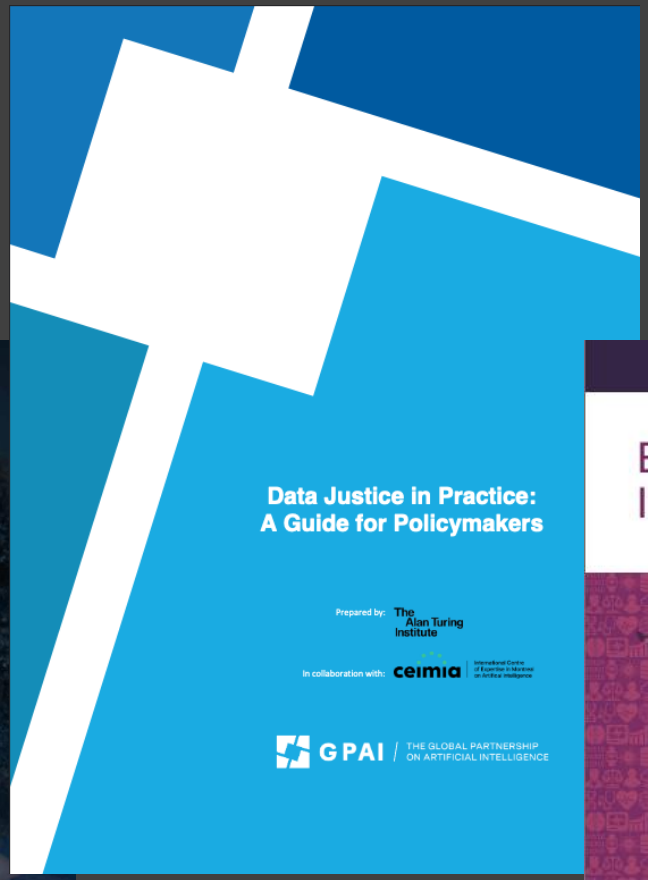
- 2021 UK National AI Strategy made expansion of public sector guidance a priority

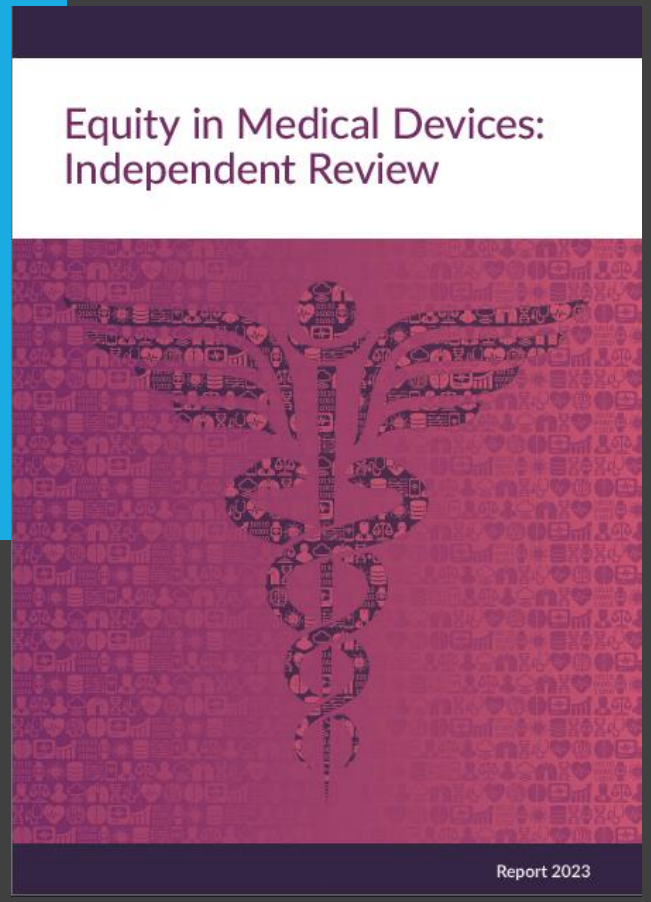
- Supported by the Office for AI and EPSRC funding to construct a series of 8 practice-based workbooks—co-designed and piloted with civil servants

- This is now complimented by an interactive digital platform to facilitate accessibility, uptake, and participation across government



AI Ethics and Governance in Practice Programme





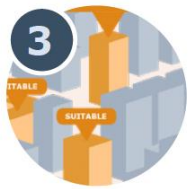
Programme Curriculum

AI Ethics and Governance in Practice: An Introduction

Multiple Domains

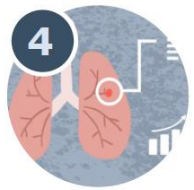
AI Sustainability in Practice Part One

AI in Urban Planning



AI Sustainability in Practice Part Two

AI in Urban Planning



AI Fairness in Practice

AI in Healthcare

Responsible Data Stewardship in Practice

AI in Policing and Criminal Justice

AI Safety in Practice

AI in Transport

AI Transparency and Explainability in Practice

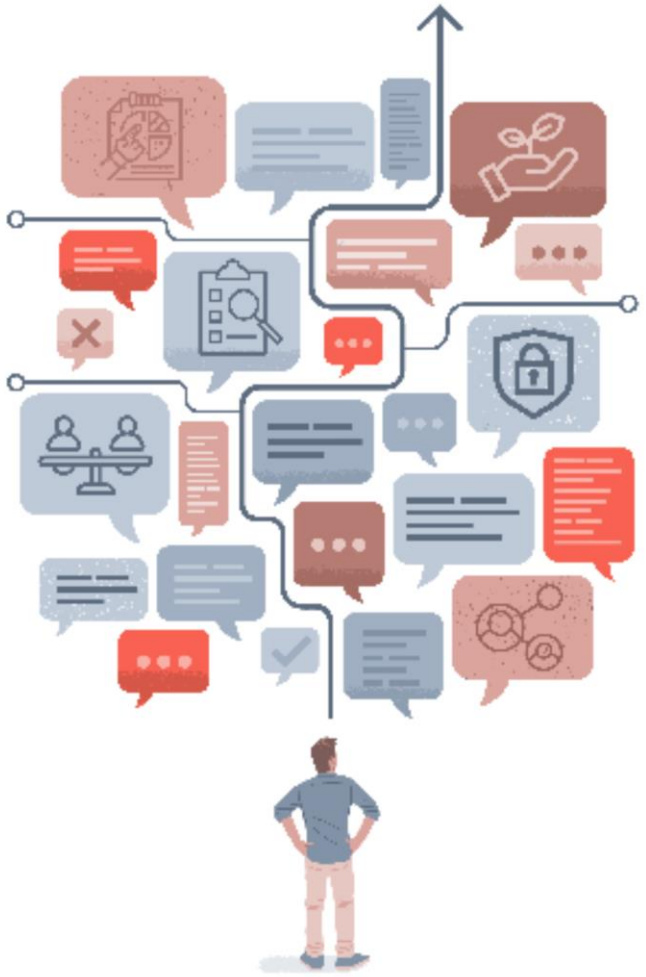
AI in Social Care

AI Accountability in Practice

AI in Education

The CARE and Act Framework

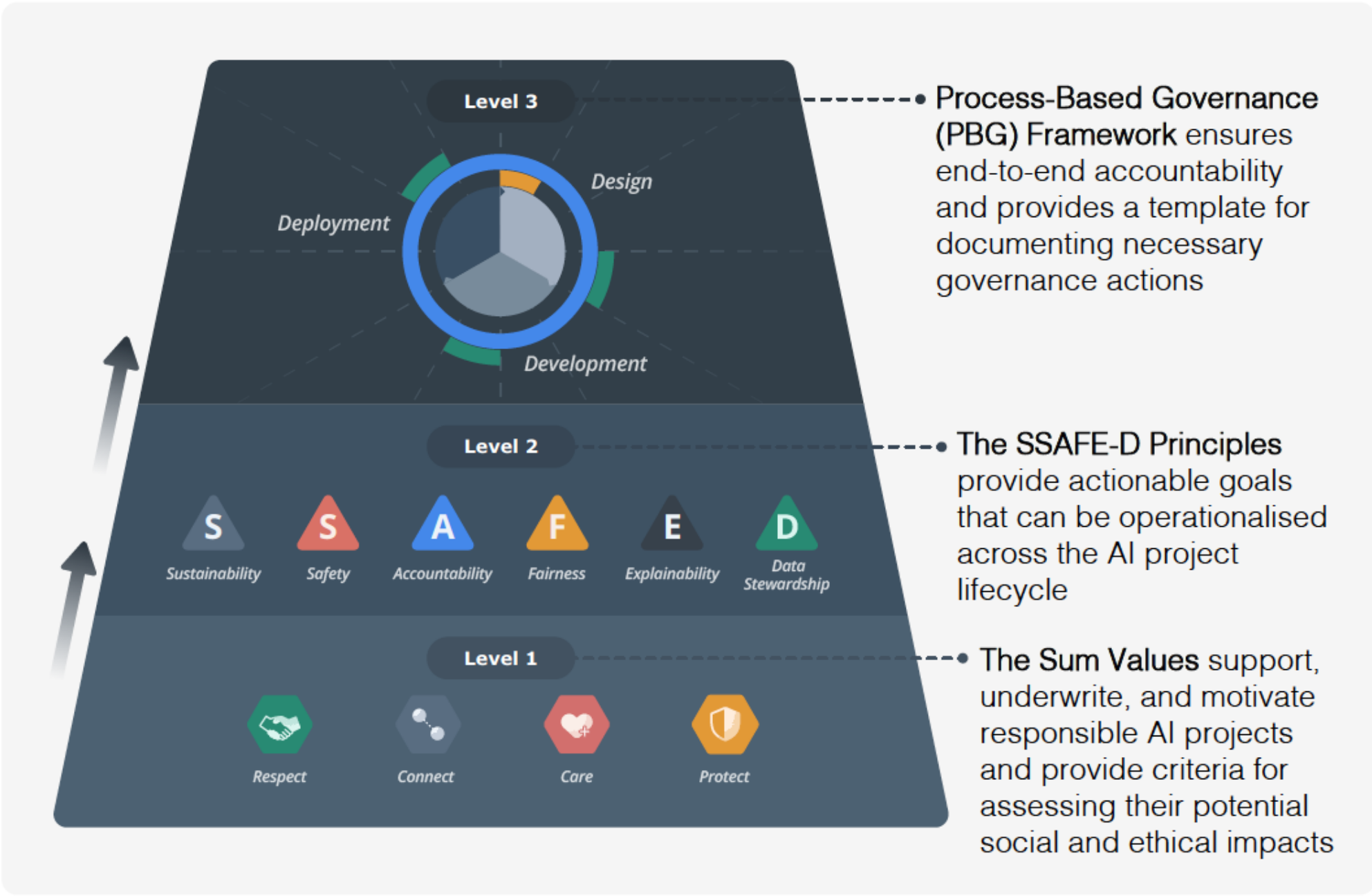
- The CARE and Act framework serves as a preliminary tool for critical self-reflection about the values that steer AI projects, as well as projects' real-world implications.
- Focused on building an organisational culture of responsible research and innovation
- Considering the CARE and Act maxims at each stage of the AI lifecycle can help establish and sustain habits of responsible research and innovation.



The CARE and Act Framework

- **C onsider Context**
  - Think about the conditions and circumstances surrounding your research project.
- **A nticipate impacts**
  - Describe and analyse the impacts, intended or not, that might arise from your project.
- **R eflect on purposes**, positionality, and power
  - Reflect on the goals of, motivations for and potential implications of the research and engage in reflexive practices that scrutinise the way potential perspectival limitations and power imbalances can exercise influence on the equity and integrity of research projects and on the associated uncertainties, areas of ignorance, assumptions, framings, and questions.
- **E ngage inclusively**
  - Open up such visions, impacts, and questioning to broader deliberation, dialogue, engagement, and debate in an inclusive way.
- **Act responsibly** and transparently
  - Use these processes to influence the direction and trajectory of the research and innovation process itself. Produce research that is both scientifically and ethically justifiable. (EPSRC, 2013, expanded) and provide transparent documentation, following best governance, self-assessment, and reporting practices.

Three building blocks of a responsible and trustworthy AI project lifecycle



SUM Values: Responding to the ethical concerns raised by real-world hazards and risks

Risks that Emerge From the Use of AI/ML Technologies	Related Ethical Implications and Concerns	SUM Values
 Loss of autonomy	**Agency** Human agency, dignity, and individual flourishing	**Respect** the dignity of individual persons
 Loss of interpersonal connection and empathy	**Interaction** Solidarity, communication, and integrity of social interaction	**Connect** with each other sincerely, openly, and inclusively
 Poor quality or hazardous outcomes	**Wellbeing** Individual, communal, and biospheric wellbeing	**Care** for the wellbeing of each and all
 Bias, injustice, inequality, and discrimination	**Justice** Justice, equity, and the common good	**Protect** the priorities of social values, justice, and the public interest

SSAFE-D Principles: Operationalising top-level normative goals

Achieving this goal requires assuring AI projects being developed with continuous sensitivity to real-world impacts.

Achieving this goal requires an AI system to be technically accurate, reliable, secure, and robust.

Achieving this goal requires assuring the project's end-to-end answerability and auditability.

Achieving this goal requires assuring a minimum threshold of discriminatory non-harm and bias mitigation.

Achieving this goal requires the ability to explain and justify AI project processes and AI-supported outcomes.

Achieving this goal requires data quality, integrity, protection, and privacy to be assured.

Process Based Governance (PBG) Framework: Evidencing and documenting necessary governance actions

Sustainability

SEP (Stakeholder Engagement Process)

Process facilitating the uptake of proportionate stakeholder engagement and input throughout the AI lifecycle. The SEP enables a contextually informed understanding of the social environment and human factors that may be impacted by, or may impact, individual AI projects.[58]

SIA (Stakeholder Impact Assessment)

Process facilitating the iterative evaluation of the social impact and sustainability of individual AI projects, as well as the corroboration of these potential impacts in dialogue with stakeholders, when appropriate.

Accountability

PBG Log

Live document outlining governance actions, relevant team members and roles involved in each action, timeframes for follow-up actions, and logging protocols, for individual AI projects.[59]

Explainability

EAM (Explainability Assurance Management)

Iterative process aimed to facilitate the implementation and evaluation of transparency and explainability assurance activities across the project lifecycle and assist in providing clarification of AI system outputs to a range of impacted stakeholders.

Safety

SSA & RM (Safety Self-Assessment and Risk Management)

Process facilitating the evaluation of how AI projects align with safety objectives through the iterative identification and documentation of potential safety risks across the lifecycle and assurance actions implemented to address these.

Fairness

BSA, BRM & FPS (Bias Self-Assessment, Bias Risk Management & Fairness Position Statement)

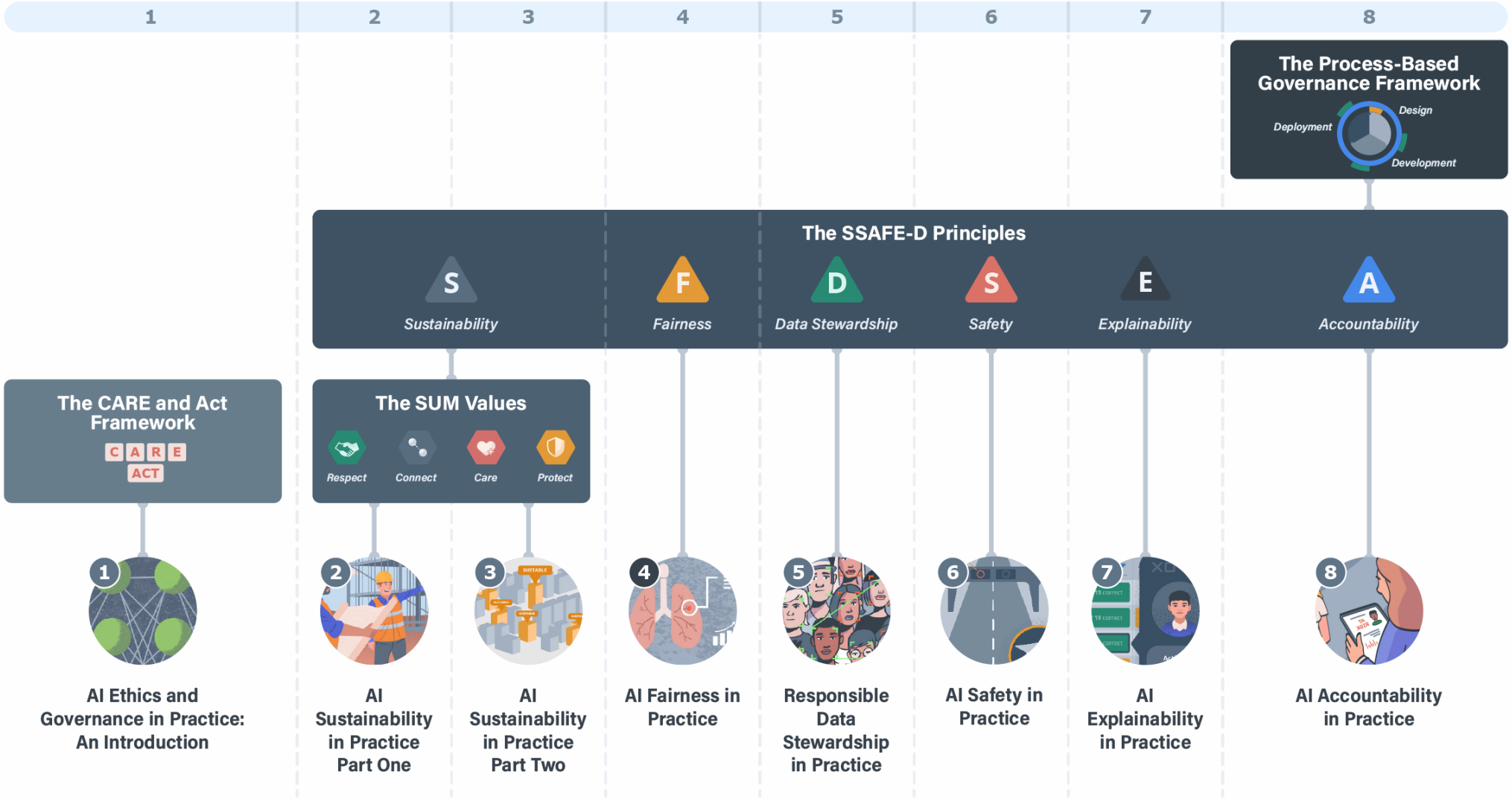
Process facilitating the evaluation of how AI projects align with the principle of fairness through the iterative identification and documentation of risks of bias across the lifecycle and assurance actions implemented to address these. The FPS is a document establishing the metric-based fairness criteria for individual AI projects, providing an explanation in plain and nontechnical language.

Data Stewardship

Data Factsheet

Live document facilitating the uptake of best practices for responsible data management and stewardship across the AI project workflow. The document consists of a comprehensive record of the data lineage and iterative assessments of data integrity, quality, protection, and privacy.

Programme Roadmap



Guidance Briefs:

Scan for more information:

The Alan Turing Institute

- **Project team: David Leslie, Ann Borda, Antonella Maia Perini, Smera Jayadeva**

  Previous contributors: Cami Rincon, Morgan Briggs

- **Graphics: Conor Rigby**