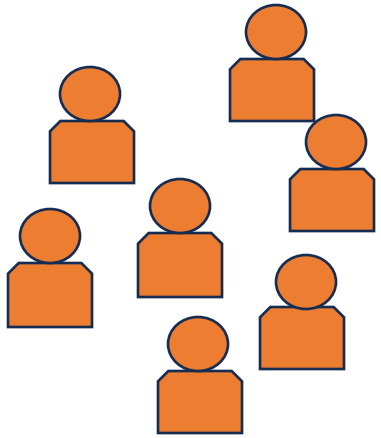
Causality and Auditing

Gourab K. Patro and Koustav Rudra

AI systems deployed at scale

- Deployed AI systems often interact with interact with people, each other, and other ecosystem or environmental parameters.



Other AI Systems

Environmental parameters

AI System

Potential for widespread unforeseen effects

Some examples of unforeseen effects

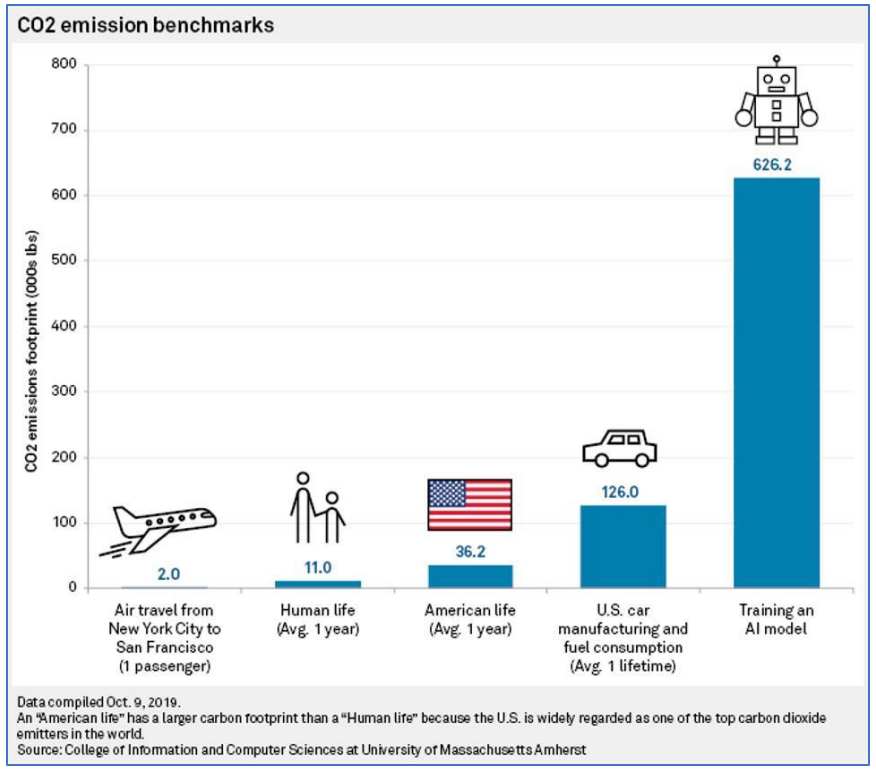
Consumption-centric information retrieval systems powered by ML model

Filter (reinforement) bubbles and echo chambers

Image source: NBC news, Sticky digital, Thecoversation

Some examples of unforeseen effects

Environmental effects of large AI models



Green Intelligence: Why Data And AI Must Become More Sustainable

Forbes

AI can help us fight climate change. But it has an energy problem, too

Horizon
The EU Research & Innovation Magazine

Environmental costs of training AI models are too high

Image source: Forbes

Auditing (Safety and Accountability)

- **What are the risks** for humans after an AI system is deployed in the real world? Is the AI system safe?
- **What are the impacts** of the AI system?
- **Who** or what **is responsible** for the actions (or failures) of the AI system after deployment?

Accountability in AI

- Relates to the expectation that designers, developers, and deployers will comply with standards and legislation
  - to ensure the proper functioning of AIs during their lifecycle

Fjeld J, Achten N, Hilligoss H, Nagy A and Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethicaland rights-based approaches to principles for AI

AI Accountability: Goals

- Compliance:
  - Defines the design, development, and deployment standards to be met throughout the entire lifecycle of an AIs
  - Often translated into preliminary checks by AIs providers

- Report
  - Practices ensuring explanation and justification of AIs' behaviors

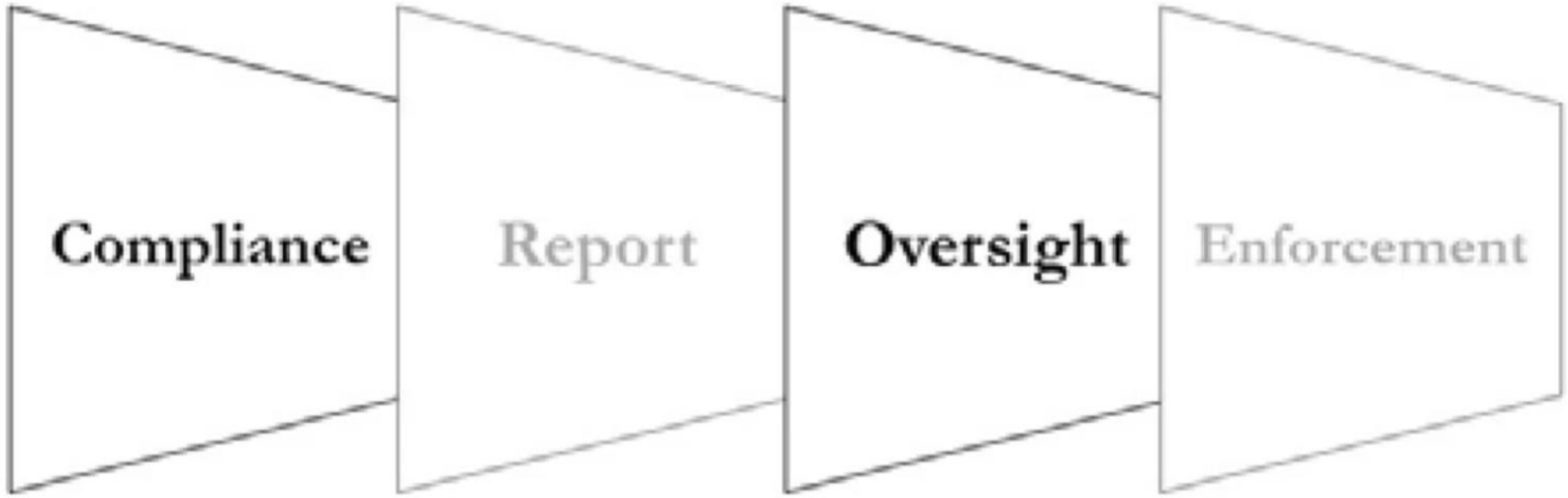
Novelli, C., Taddeo, M. & Floridi, L. Accountability in artificial intelligence: what it is and how it works. *AI & Soc* (2023). https://doi.org/10.1007/s00146-023-01635-y

AI Accountability: Goals

- Oversight:
  - Seeks to find relevant facts or information, and create evidence, to evaluate the life-cycle performance of AIs
- Enforcement
  - Ties the monitoring and evaluation of the performance of AIs to formal or informal consequences

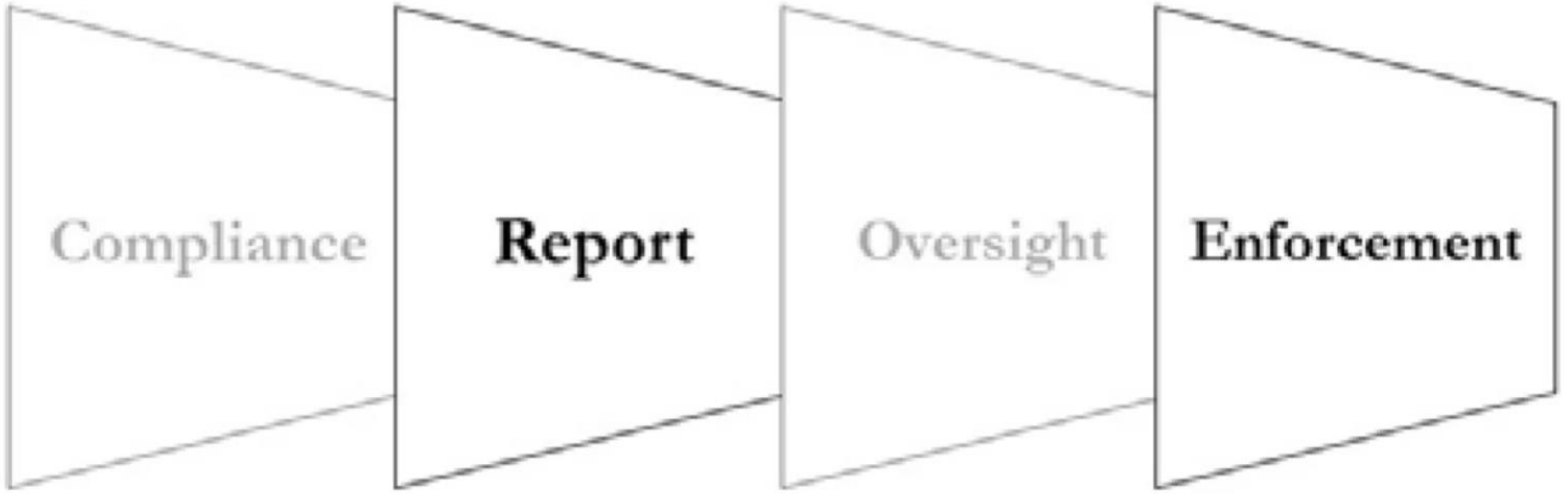
Accountability: How does this work?

- Proactive Accountability
  - Accountability as a virtue
  - Planning purpose
    - Comes before events and aims to prevent failures

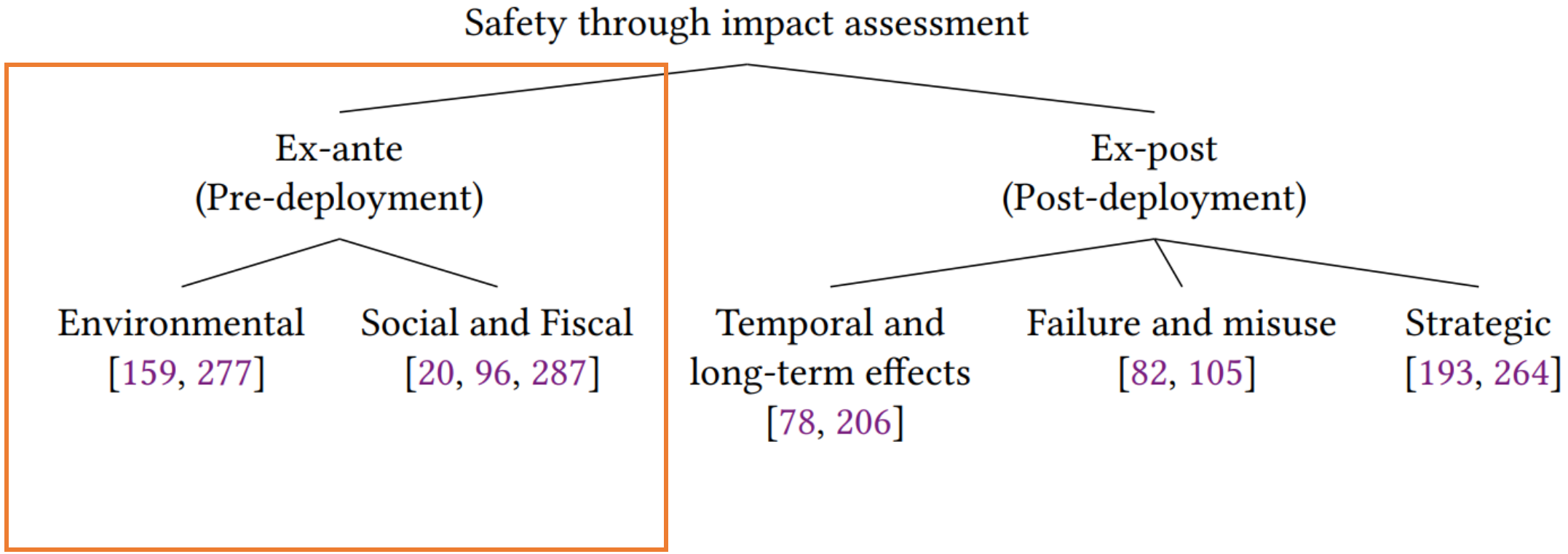


Accountability: How does this work?

- Reactive Accountability
  - Negative sense
  - Responsive purpose
    - Comes after events and aims to address failures



Safety through (causal) impact assessment

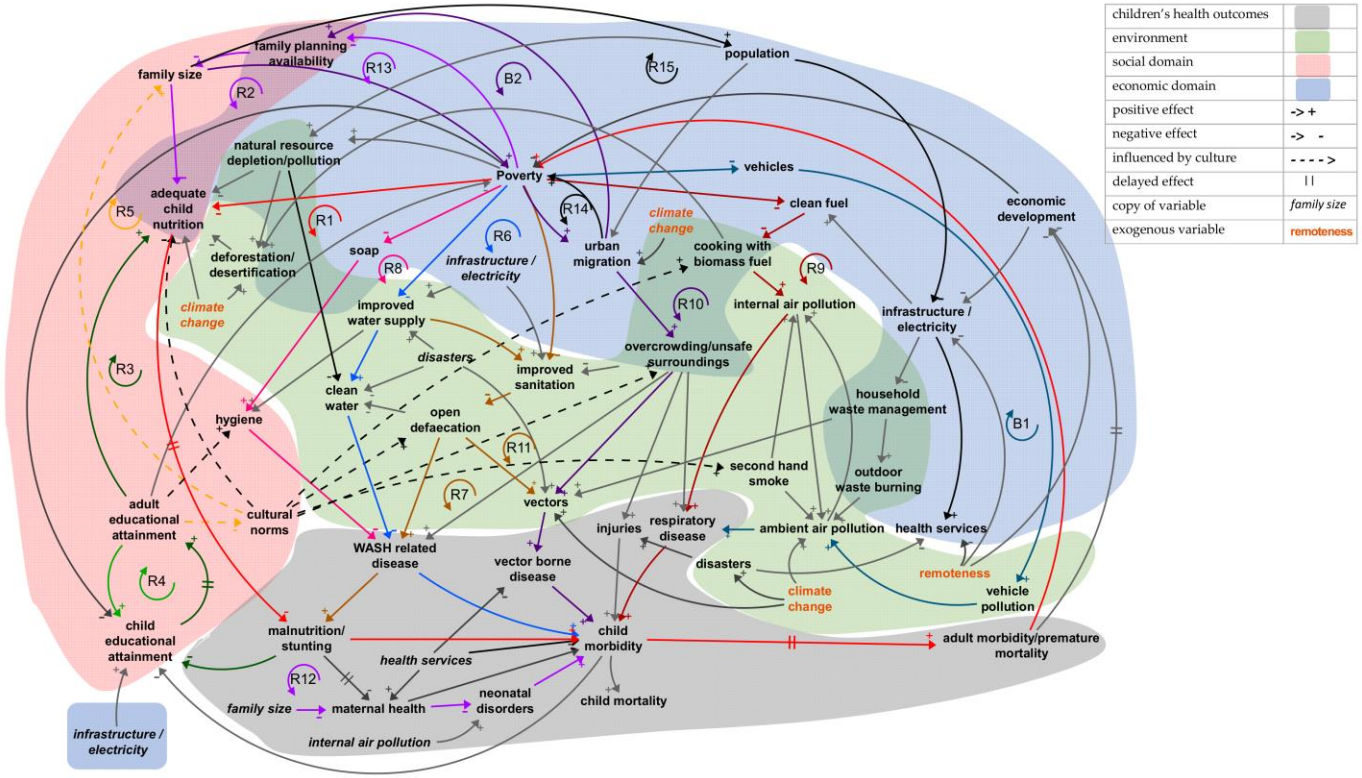


Ex-ante Impact Assessment

- Impacts are assessed before deployment
- Assess potential ramifications of systems, projects, and policies
  - Environmental
  - Financial
  - Social, and human rights
- Grant some measure of control and voice to
  - Designers or developers
  - Affected population
  - Authorities

Ex-ante Impact Assessment: Environmental

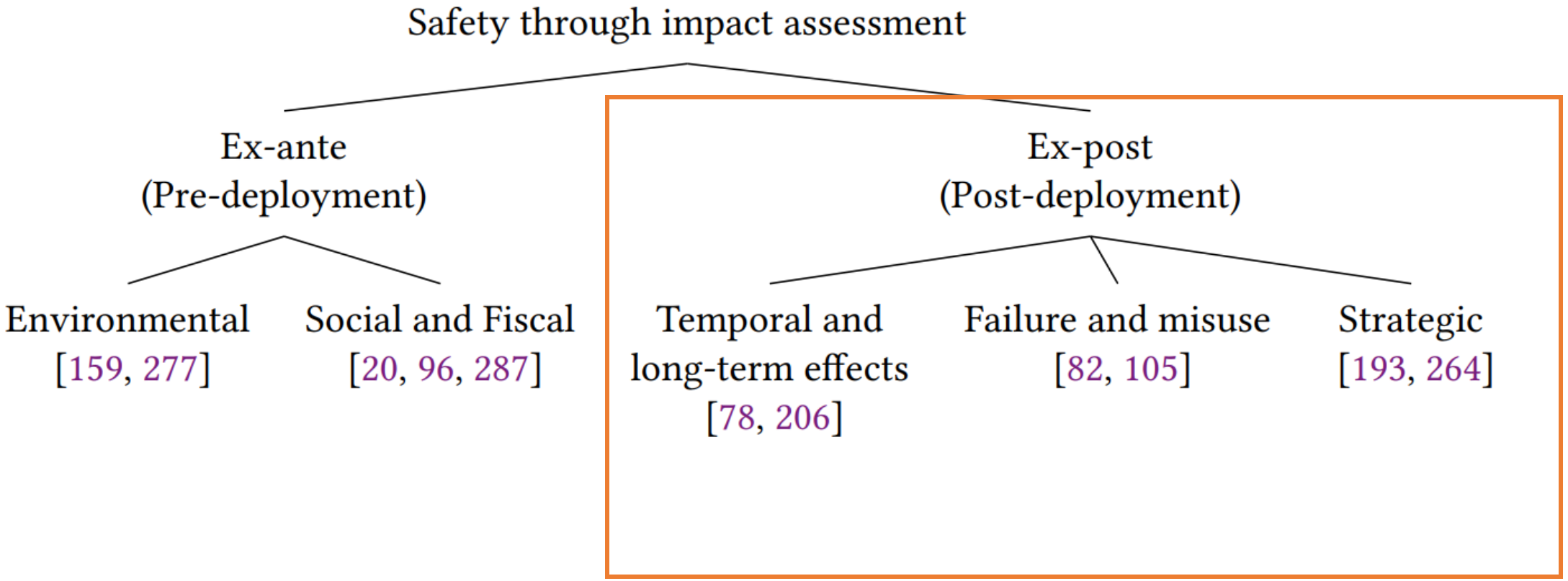
- Causal networks for environmental (causal) impact assessment



Ex-ante Impact Assessment: Social and Fiscal

- Estimate increase in sales, popularity, overall perception of a product, etc.
- In multi-agent settings
  - Modeling of the **preferences** and **choices**
  - Agents along with their ability to infer evaluations and outcomes
  - Understand the effects of introducing new economic policies or changing existing ones

Safety through (causal) impact assessment



Ex-post Impact Assessment

- Impacts are assessed after deployment
  - Often in real-time
  - Using the running record of the system,
  - Real-time audits and evaluations
- Not limited to the available prior knowledge and use cases
- Ex-post assessments are generally broader than ex-ante
  - Since they need to define what constitutes an impact in real-time

Ex-post Impact Assessment: Temporal

- Examples:
  - Online filter bubbles
  - Echo chambers
  - Social network polarization
  - Content homogenization effects

- **Causality** with **behavioral modeling** help assess the effects and find out the responsible elements in design
- But there is a need for simulation either in virtual or real environment

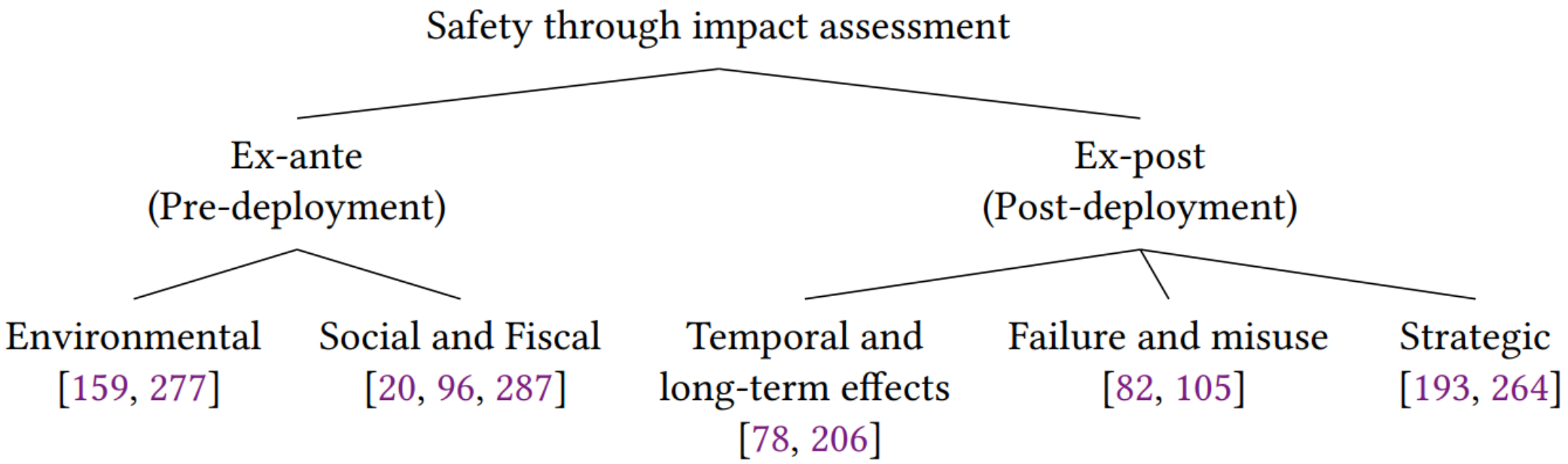
Ex-post Impact Assessment: Failure and Misuse

- Systems are often designed with **some desired criteria** (e.g., accuracy, fairness, robustness, etc.)
- If real-world model **deviates** ==> ***accountability*** measures
- Bottom-up causal approach using goal-specific accountability mechanisms by Ibrahim et al. 2021.
- Goals:
  - Identify the root cause of specific type(s) of events or failures
  - Eliminate the underlying (technical) problem and also to assign blame.
- Misuse like deepfakes: No causal approach has been used yet

Ex-post Impact Assessment: Strategic Risks and Effects

- Shokri et al. 2021: Back-propagation-based **explanations can leak** a significant amount of **sensitive information about individual** training data points.
- Tsirtsis et al. 2020: **Counterfactual explanations can reveal** various **details of decision-making systems**, making them vulnerable to strategic attacks.
- Rational strategic behavior can cause issues of privacy and robustness
- Assessing such issues: **Causality** along with **applied game theory**

Summary


Safety through impact assessment
Ex-ante
(Pre-deployment)
Ex-post
(Post-deployment)
Environmental
[159, 277]
Social and Fiscal
[20, 96, 287]
Temporal and
long-term effects
[78, 206]
Failure and misuse
[82, 105]
Strategic
[193, 264]