The role of the STAMP model in the emergence of AI perils

MIT STAMP Workshop 2024

Team

Dr Mikela Chatzimichailidou
Professor,
Civil, Environmental and Geomatic Engineering,
University College London, UK
mikela.chatzi@ucl.com

Dr Ioannis Dokas
Associate Professor,
Civil Engineering,
Democritus University of Thrace, Greece
idokas@civil.duth.gr

Dr Liucheng Guo
Co-Founder and CTO,
TG0, UK
liucheng@tg0.co.uk

AI fundamentals

AI Artificial Intelligence

- Branch of computer science.
- Deals with the creation of intelligent agents, which are systems that can reason and act autonomously.

ML Machine Learning

- A program or system that trains a model from input data, giving the computer the ability to learn without explicit programming.

DL Deep Learning

- A type of machine learning.
- Incorporates many layers of neural networks to learn more complex patterns.

Motivation

A call to action

- **November 2023** - the UK chaired the inaugural **AI Safety Summit** at Bletchley Park where Alan Turing decoded messages that had been encrypted with the Enigma machine.

- Countries attending reached a landmark agreement recognising a **shared consensus** on the opportunities and risks of AI, and the need for collaborative action on AI safety.

- **Domestic frameworks** were set forth: UK response to the AI Regulation White Paper, the EU AI Act, the US Voluntary Measures and Executive Order on Safe, Secure, and Trustworthy AI, China's AI governance framework.

Summit objectives

1. A **shared understanding** of the risks posed by AI and the need for action

2. A forward process for **international collaboration** on AI safety, including how best to support national and international frameworks

3. **Appropriate measures** which individual organisations should take to increase AI safety

4. Areas for potential collaboration on AI safety research, including evaluating model capabilities and the **development of new standards** to support governance

5. Showcasing how **ensuring the safe development** of AI will enable AI to be used for good globally

Key themes

Introduction of a universally accepted **definition** of safe AI.

Appropriate **standardisation** and interoperability in AI.

Tackling these two issues is a prerequisite to building a shared **understanding** of AI and taking immediate action.

Problem

How do we regulate something we do not understand or something that is constantly changing?

Inadequate regulatory oversight

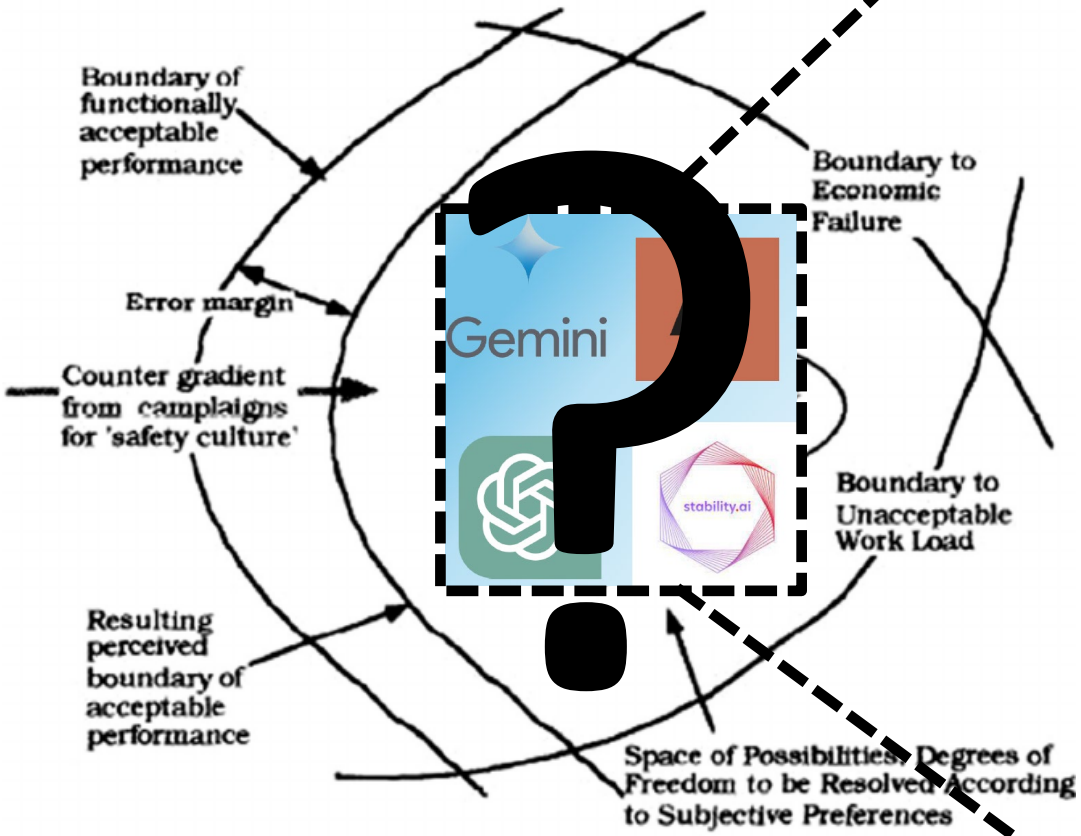
- According to the Hackitt report (2018) regulatory oversight and enforcement tools are currently inadequate in a sense that **“the size or complexity of a project does not seem to inform the way in which it is overseen by the regulator”**

- Where enforcement is necessary, it is often not pursued. Where it is pursued, the penalties are so small as to be an ineffective deterrent.

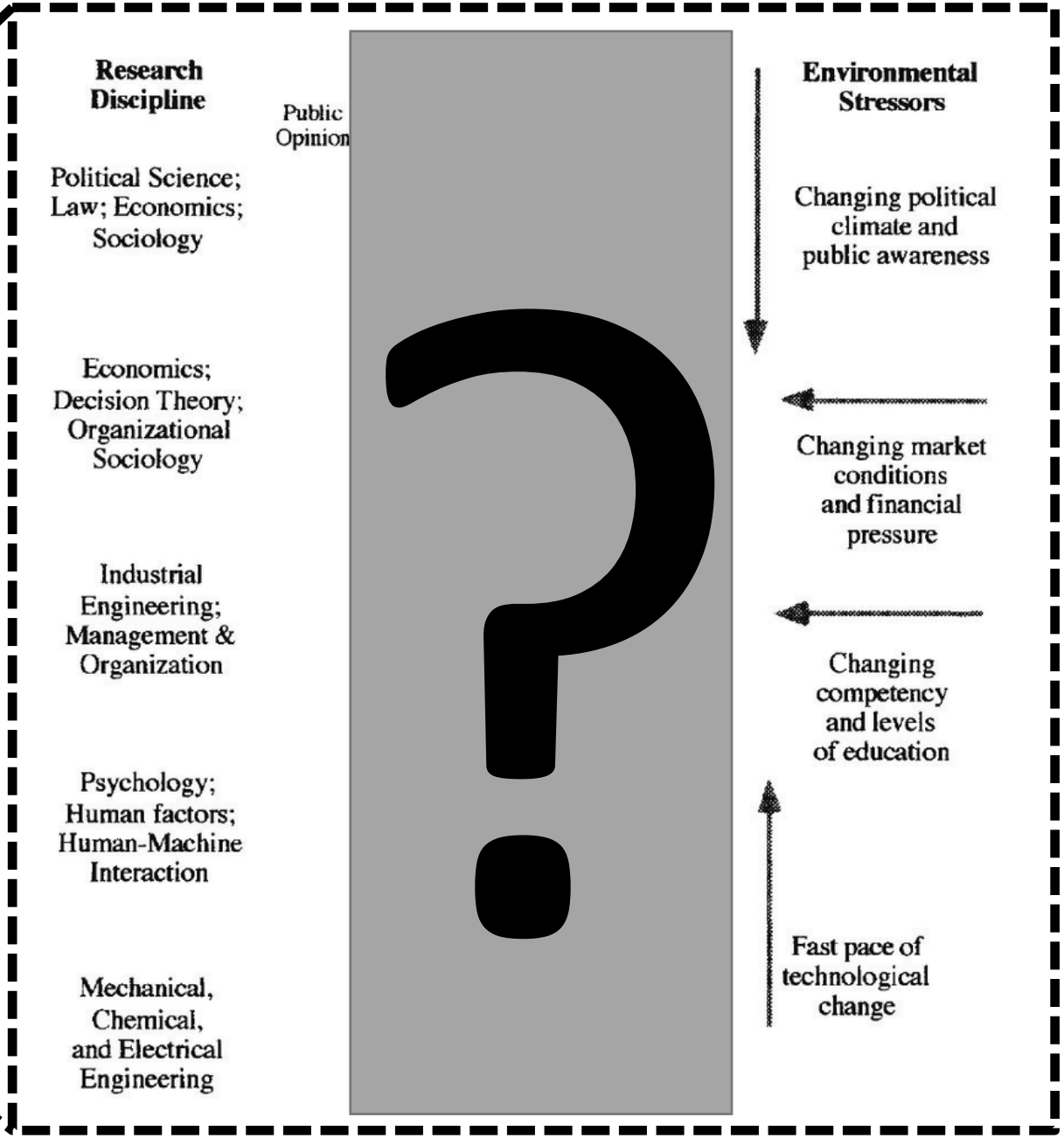
Uncertainty, unpredictability and unknowns

- The UK Government Office for Science argues that the novel risks posed by future AI models are highly **uncertain**.
- Complex and interconnected systems using AI could present **unpredictable** risks or modes of failure.
- Systems able to run on local devices – or that rely on distributed cloud computing – present **different** risks.

- **What** are we trying to regulate? What is the problem? What is the loss?
- What is the **system of reference**? Is it the state/ government, the organisation, the user, large language model?
- How can we introduce a safety management system, when we don't know the **boundaries** of acceptable performance?



Under the presence of strong gradients behaviour will very likely migrate toward the boundary of acceptable performance (Rasmussen, 1997)



The socio-technical system involved in risk management (Rasmussen, 1997)

Insufficient tools

- Quantitative tools **omit** the role of organisational culture, legislation, and regulation (UK Government Office for Science, 2023).
- Traditional approaches to safety analysis, such as Failure Modes and Effects Analysis and Fault Tree Analysis, which are recommended by existing safety standards, such as ISO 26262, are **not sufficient** in the context of AI (UK Government, 2023).

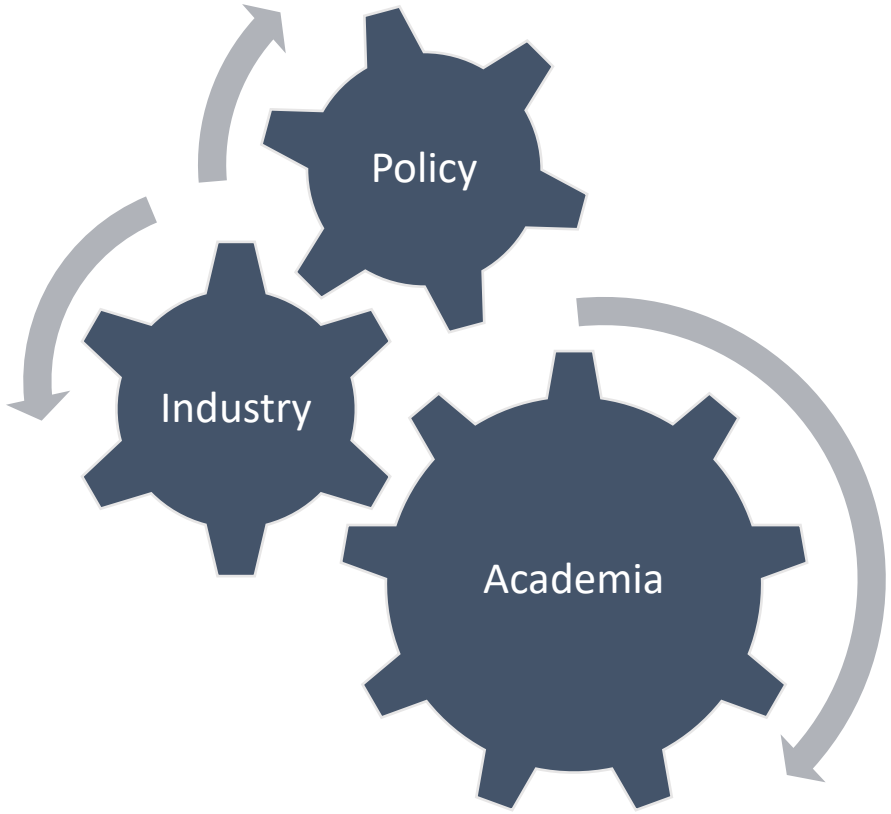
"It appears that mechanistic approaches to safety analysis will be insufficient to predict hazards caused by systemic failures of the system due to emergent complexity" (UK Government, 2023).

"true AI, by definition, won't fit into the deterministic model necessary for safety certification" (SEMP, 2024).

Towards providing a solution

AI safety is a socio-technical challenge that cannot be resolved with technical interventions alone.

Collaboration for exploration



Our approach

- The fundamental **principles** (e.g. communication and control) **remain the same**, it is complexity that changes.
- We propose STAMP's view on safety, where safety is a **control problem**.
- In alignment with the AI Safety Summit (and the UK AI Safety Institute), where **'loss of control'** is listed as one of the most extreme risks of AI.

- **Clarity and rules** on what needs to be done by innovators will eventually ensure that safety and trust do not stifle innovation.
- Beginning of a series of projects about safe AI to provide insights into issues pertaining to **safety, security, sustainability, as well as ethical** concerns of AI.

Back to the basics

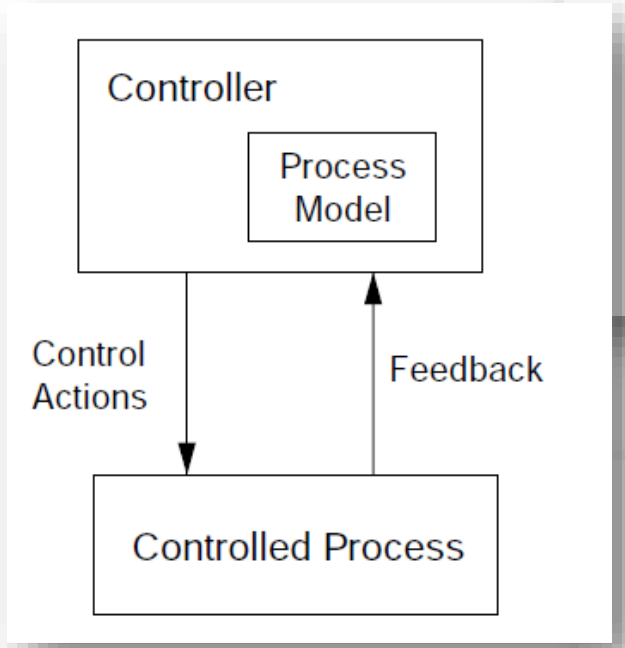
Systems theory

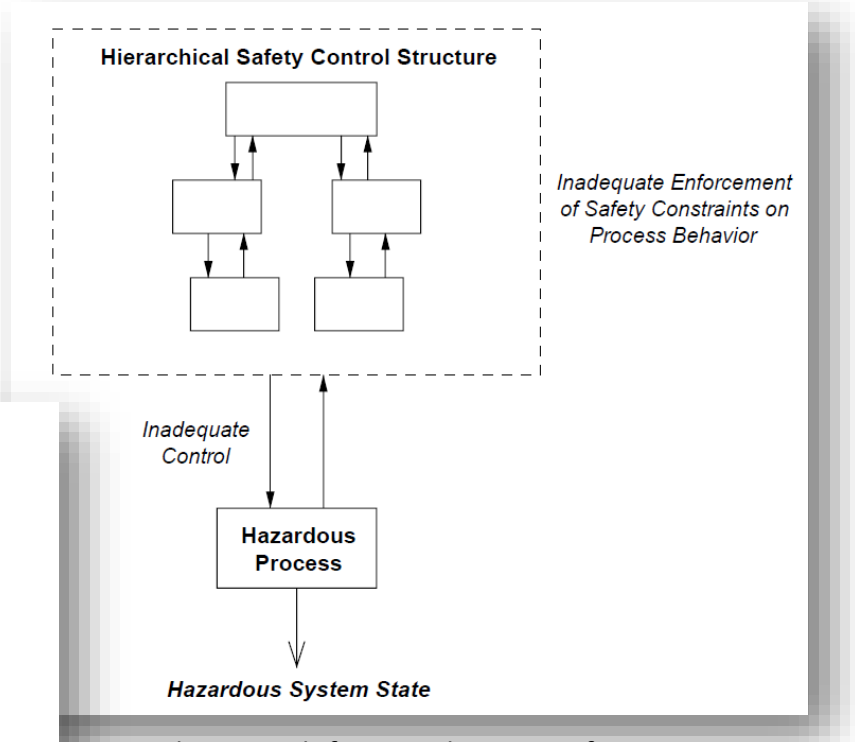
- Emergence and Hierarchy
- Command and Control

STAMP & STPA

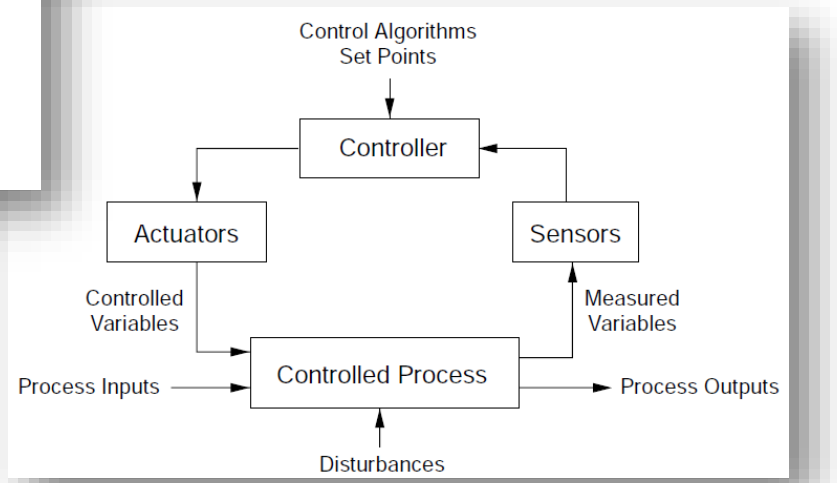
- System
- Accident and loss
- Hazard (as condition/ state)

… safety constraint/ requirement





Accidents result from inadequate enforcement of the behavioural safety constraints on the process (Leveson, 2011)



A standard control loop (Leveson, 2011)

Added value

Answer fundamental questions:

- What is the system?
- What is the controlled process?
- Who is the controller and what are its responsibilities?

Output:
Create a set of regulatory *AI Accountability and Responsibility Tools* based on STAMP that can attribute actions and decisions made by AI systems to specific entities or organisations; especially in complex multi-agent scenarios, enhancing thus the accountability and responsibility of such AI systems.

Outcome:
Shedding light on these questions will facilitate ***embedding control mechanisms* related to bias, fairness, transparency, and accountability** into the development, testing and operation phases of AI models. This could minimise potential vulnerabilities and limit rogue behaviours or misuse of AI.

Bonus slides

Plugging into existing frameworks

- E.g. EU AI Act and framework
- We introduced STPA as part of the railway **CSM-RA framework** that describes a common mandatory European risk management process for the rail industry (Chatzimichailidou and Dunsford, 2019; Oginni et al., 2023).
- Frameworks like that are **not prescriptive** on the techniques and tools to be used.
- Those tools selected should be appropriate to adequately assess and manage the risk being introduced.
- At a time of bringing systems of increasing complexity into operational use, we must ask ourselves if the conventional tools and techniques that we have relied on in established industries for many years are the most appropriate for rapidly changing markets (Chatzimichailidou and Dunsford, 2019). → STPA

The transformative potential of AI in many human activities is undeniable, but…

Safe and ethical AI

Risk areas of AI Weidinger et. al, 2022):

- Discrimination, hate speech and exclusion
- Information hazards
- Misinformation harms
- Malicious uses
- Human-computer interaction harms
- Environmental and socioeconomic harms

Responsible innovation:
In addition to developing a given technology, innovators must anticipate, reflect on, and evaluate the benefits and risks a technology holds.

Considerations

- Explore complex **interactions** between uncertainties in a structured and rigorous way.
- Regulatory frameworks must be as **flexible** as the AI systems it seeks to regulate.
- The regulator must possess the requisite **variety of knowledge** and expertise in AI to create effective regulations.
- Effective mechanisms for the exchange of knowledge and best practices among different organisations, countries, and institutions to **collectively manage** AI safety challenges must be also established.
- **Multidisciplinary approaches** must be established to ensure the consideration of a variety of perspectives (e.g. technologists, ethicists, policymakers, domain experts).
- Decision-makers and organisations must be **proactive, adaptable, and responsive** to emerging threats and issues.

11TH EUROPEAN STAMP WORKSHOP AND CONFERENCE

"Advancing Safety in a Complex World"

ESWC

Democritus University of Thrace

Prof. Nancy Leveson

Key Note Speaker

Prof. Georgios Boustras

Key Note Speaker

October 2-4, 2024

Alexandroupolis, Greece

HOTEL
ASTIR EGNATIA ALEXANDROUPOLIS

https://eurostamp2024.civil.duth.gr

Conversation

Contact:

Dr Mikela Chatzimichailidou

Professor, Civil, Environmental and Geomatic Engineering, University College London, UK

mikela.chatzi@ucl.com

https://profiles.ucl.ac.uk/87235-mikela-chatzimichailidou

www.linkedin.com/in/prof-mikela-chatzimichailidou