Abstract for MIT 2024 STAMP Workshop.

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

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

Ioannis Dokas, Assistant Professor, Civil Engineering, Democritus University of Thrace, Greece

Liucheng Guo, Co-founder and CTO, Tangi0 LTD, UK

The transformative potential of Artificial Intelligence (AI) in many human activities is undeniable. Today's AI algorithms, trained on vast amounts of data of different forms, such as text, videos, images and voice, are capable of generating new data in response to a specific prompt. Organisations worldwide are acknowledging the impact of these AI innovations. AI however, does not come without its challenges. Weidinger et. al. [1] provide twenty-one risks, draw on expertise and literature from computer science, linguistics, and the social sciences, and situate these risks in a taxonomy of six risk areas:

i. Discrimination, hate speech and exclusion
ii. Information hazards
iii. Misinformation harms
iv. Malicious uses
v. Human-computer interaction harms
vi. Environmental and socioeconomic harms

These risk areas are stressing the need for what it is known as responsible innovation, which entails that in addition to developing a given technology, innovators must anticipate, reflect on, and evaluate the benefits and risks a technology holds. This may include engaging multiple perspectives and communities and then acting upon these insights [2]. The taxonomy shows that the potential for an AI system to cause harm is driven by a mix of technical and non-technical factors. AI safety is therefore a socio-technical challenge that cannot be resolved with technical interventions alone [3]. The safe development of AI has therefore become a priority for most governments and organisations.

In November 2023, the United Kingdom chaired the inaugural AI Safety Summit at Bletchley Park where Alan Turing decoded messages that had been encrypted with the Enigma and Tunny machines. Countries attending the AI Safety Summit agreed to the Bletchley Declaration on AI safety, a landmark agreement recognising a shared consensus on the opportunities and risks of AI, and the need for collaborative action on AI safety [4]. Many participants set forth views on their own domestic frameworks including the UK's anticipated response to the AI Regulation White Paper, the EU AI Act, the US Voluntary Measures and Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, and China's AI governance framework.

Ultimately, the five objectives of the Summit were [5]:

- a shared understanding of the risks posed by AI and the need for action
- a forward process for international collaboration on AI safety, including how best to support national and international frameworks
- appropriate measures which individual organisations should take to increase AI safety
- areas for potential collaboration on AI safety research, including evaluating model capabilities and the development of new standards to support governance
- showcasing how ensuring the safe development of AI will enable AI to be used for good globally

After reviewing relevant documents published by governments around the world, the authors concluded that the key themes are:

1. The value of introducing a universally accepted definition of safe AI.
2. The value of appropriate standardisation and interoperability in AI.

Answering these questions is a prerequisite to building a shared understanding of AI and taking immediate action.

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. Similarly, systems able to run on local devices – or that rely on distributed cloud computing – present different risks. It is, therefore, an imperative to explore complex interactions between uncertainties in a structured and rigorous way. The Office also contend that quantitative tools omit the role of organisational culture, legislation, and regulation [3]. Along the same lines, representatives of the UK private sector explain that "true AI, by definition, won't fit into the deterministic model necessary for safety certification" [6].

Complexity gives rise to a certain 'attitude' to approaching a **complex** world. According to the Hackitt report [7] 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" [7]. For example, where enforcement is necessary, it is often not pursued. Where it is pursued, the penalties are so small as to be an ineffective deterrent [7]. Although the report describes how the regulatory system covering complex buildings is not fit for purpose, the criticism is valid in the case of regulating AI and beyond. Therefore, the authors suggest the use of the law of requisite variety to analyse the problem of understanding and regulating AI. In short, in order to effectively control or manage an AI system, the controller (or decision-maker) – and in turn the control mechanism – must have at least as much variety (or flexibility) as the system it is trying to control. Thus, it is important for the regulator to possess the requisite **variety** of knowledge and expertise in AI to create effective regulations. In essence, the regulatory framework must be as flexible as the AI systems it seeks to regulate. Decision-makers and organisations must be proactive, adaptable, and responsive to emerging threats and issues. Furthermore, a multidisciplinary approach must be established to ensure the consideration of a variety of perspectives and the effective collaboration among various stakeholders, including technologists, ethicists, policymakers, and domain experts, together with effective mechanisms for the exchange of knowledge and best practices among different organisations, countries, and institutions to collectively manage AI safety challenges.

"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" [8]. 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 [8]. On these grounds, the authors adopt STAMP's view to safety, according to which safety is a control problem. This is in alignment with the AI Safety Summit's objectives as well as the UK AI Safety Institute [9], where 'loss of control' is listed as one of the most extreme risks of AI.

The authors will use STAMP – and STPA's principles – to help define and understand the meaning of 'safe AI' and lay the foundation and structure towards regulating AI safety (i.e. two key themes identified previously). This includes answering the following fundamental questions:

- What is the system?
- What is the shape/form of the control mechanisms?
- What is the controlled process?
- Who is the controller and what are its responsibilities?

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 lead to minimising potential vulnerabilities and, in turn, limiting rogue behaviours or misuse of AI technologies.

The authors acknowledge that there may be more than one answer, depending on the perspective from which AI safety is being viewed. They believe however, that it is possible to 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.

The results of this applied research can plug into existing frameworks, such as the EU AI framework [10]. Similar contributions where made in the past, where the authors introduced STPA as part of the railway CSM-RA framework that describes a common mandatory European risk management process for the rail industry [11, 12]. Frameworks like that are not prescriptive on the techniques and tools to be used. However, those 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.

The authors, in collaboration with the Health and Safety Executive (HSE, UK) and the Rail Safety and Standards Board (RSSB, UK), are in the early stages of an exploration project where they will revisit the process of generating (and reviewing) standards, laws, and regulations. Usually, industry standards set the benchmark or the threshold value, but this only works when the system of interest is well understood and static. So, the question that arises is: *how do we regulate something we do not understand or something that is constantly changing?* [13] The authors, working cooperatively with key partners from policy and industry, will provide guidance and, where possible, answers to the questions raised above. Clarity and rules on what needs to be done by innovators will eventually ensure that safety and trust do not stifle innovation [14]. It is envisaged that this work is the beginning of a series of publications in the space of safe AI and may provide insights into issues pertaining to security, sustainability, as well as ethical concerns of AI.

References

1. Weidinger, Uesato, Rauh, Griffin, Huang, Mellor, Glaese, Cheng, Balle, Kasirzadeh, Biles, Brown, Kenton, Hawkins, Stepleton, Birhane, Hendricks, Rimell, Isaac, Haas, Legassick, Irving, and Gabriel. 2022. Taxonomy of Risks posed by Language Models. https://doi.org/10.1145/3531146.3533088
2. Stilgoe, Owen, and Macnaghten. 2013. Developing a framework for responsible innovation. https://doi.org/10.1016/j.respol.2013.05.008
3. Government Office for Science. 2023. Future Risks of Frontier AI. https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf

4. UK Government. 2023. Chair's Summary of the AI Safety Summit 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park
5. UK Government. 2023. UK government sets out AI Safety Summit ambitions. https://www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions
6. SEMP. 2024. Certifying the Unpredictable: Ensuring safe AI Implementation in Transportation Systems. https://www.sempltd.com/insights/certifying-the-unpredictable-ensuring-safe-ai-implementation-in-transportation-systems/
7. Hackitt. 2018. Building a Safer World. https://assets.publishing.service.gov.uk/media/5afc50c840f0b622e4844ab4/Building_a_Safer_Future_-_web.pdf
8. UK Government. 2023. Introducing the AI Safety Institute. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute
9. Burton and McDermid. 2023. Closing the gaps: Complexity and uncertainty in the safety assurance and regulation of automated driving. https://publica-rest.fraunhofer.de/server/api/core/bitstreams/c0198205-8061-4fcf-bfa3-02e37bc2780c/content
10. European Union. 2021. Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
11. Oginni, Camelia, Chatzimichailidou, and Ferris. 2023. Applying System-Theoretic Process Analysis (STPA)-based methodology supported by Systems Engineering models to a UK rail project. https://doi.org/10.1016/j.ssci.2023.106275
12. Chatzimichailidou and Dunsford. 2019. Introducing a system theoretic framework for safety in the rail sector: supplementing CSM-RA with STPA. https://doi.org/10.1080/09617353.2019.1709289
13. Leveson. 2019. CAST HANDBOOK: How to Learn More from Incidents and Accidents. http://psas.scripts.mit.edu/home/get_file4.php?name=CAST_handbook.pdf
14. Hurrell. 2024. House of Lords AI report condemns UK government for AI safety focus. https://techmonitor.ai/technology/ai-and-automation/house-of-lords-ai-report-ai-safety-focus-uk