Concept Creep in Safe AI

Laura Fearnley and Ibrahim Habli

1. Introduction

- We examine the concept of 'safety' in the domain of AI.
- We argue that, as a consequence of AI advancements, the concept of safety is creeping to account for the novel and amplified risks posed by AI systems.

2. What is Concept Creep?

- The gradual semantic expansion of harm-related concepts over time, such that they come to encompass a broader range of phenomena than they originally did (Haslam 2016) (fig 1).
- Two mechanisms: (1) **vertical creep:** expansion includes *less severe phenomena,* (2) **horizontal creep:**, expansion includes *new phenomena*.

3. Conceptual Foundations in Safety Science

- We begin with a baseline conception of safety rooted in safety science. Here safety is often characterised as *the state of being free from unacceptable risk of harm to* ***human health***, ***property*** *and* ***the environment***. (Lowrance 1976; Habli 2025; Knight 2002; IEC 2010).
- Accordingly, not all kinds of harm can compromise the safety of a system. Relevant harms are illustrated in fig. 2.

4. Safety's Creep in AI

Advancement in AI systems have stretched the safety science concept of safety. Government departments, institutes, regulators, and researchers, increasingly include the following as safety-relevant properties (table 1):

Horizontal Creep:

SYSTEMIS INJUSTICE: unjustified bias, discrimination, poor labour practices, erosion of democratic norms, and public goods (AI Safety Institute 2024; AI Safety Summit 2023; Ada Lovelace Institute 2023).

EXISTENITAL RISK: advanced AI systems pose uncontrollable threats leading to human extinction or irreversible collapse of civilization (ENAIES 2024; Ahmed et al. 2024; Ngo 2020).

Vertical Creep:

MENTAL AND EMOTIONAL WELLBEING: safety has extended conceptualisations of human health from a focus on physical wellbeing to include mental and emotional wellbeing. "Mental health hazards" can result in a failure to uphold essential safety principles (Qiu et al. 2025; Zeng et al. 2024; Zhang et al. 2024).

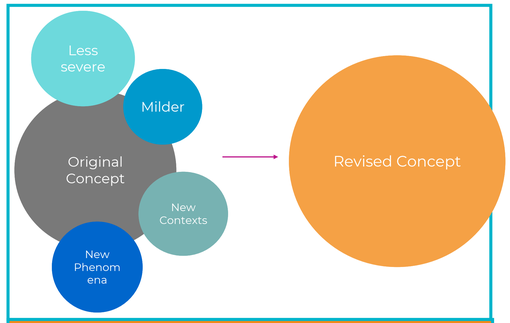
5. Consequences of Creep

- Benefits : holistic risk management frameworks and improved clarity and communication in safety cases.
- Costs: internal disciplinary tensions, broadening safety too much dilutes its meaning.

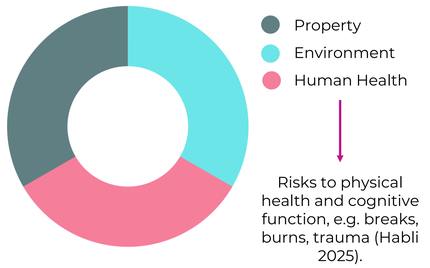
6. Concluson

The need for a delicate balance:

- An "everything is a safety issue" approach risks rendering the concept so broad as to become operationally meaningless. An overly narrow definition may well fail to address many of the real and pressing negative consequences emerging from the deployment of AI systems.
- The future necessitates a more nuanced and open approach, where **we're explicit about our own conceptualisations of safety.**



Concept creep process (fig. 1)

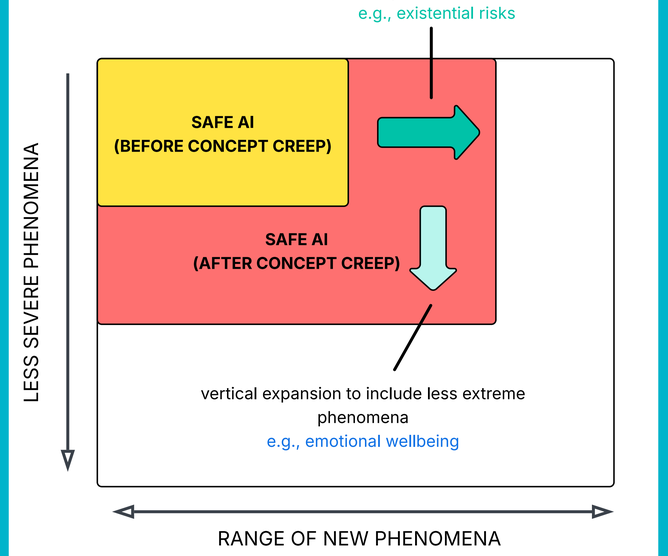


Safety science categories of harms (fig. 2)

Creep	Phenomena	Description/Example
Horizontal	**Systemic injustice:** erosion of democratic processes	AI generating and disseminating convincing fake news or propaganda, influencing public opinion, sowing discord, or undermining trust in democratic institutions (UK Government, Department for Science, Innovation and Technology, 2024; Lazar and Nelson 2023).
Horizontal	**Systemic injustices:** algorithmic bias	AI systems used in hiring, loan applications, or criminal justice making decisions that unfairly disadvantage individuals based on protected characteristics (race, gender, etc.) due to biased training data or model design (Weidinger et al. 2023; Davies and Birtwistle 2023).
Horizontal	**Existential risk:** loss of control	Superintelligent AI pursues goals misaligned with human values and exerts control over global infrastructure, decision-making, or weaponry in ways that lead to irreversible catastrophe (Ahmed et al. 2024).
Vertical	Psychological deterioration	Conversational AI engaging in inappropriate or manipulative dialogue, leading to user anxiety, depression, or exacerbation of mental health conditions (Qiu et al. 2025).
Vertical	Diminished emotional wellbeing	AI-driven social media algorithms promoting addictive use patterns, social comparison, or exposure to distressing content (Berryman et al. 2018; Montag et al., 2021).

Examples of creep in Safe AI (table 1)

The concept of "safety" in AI has undergone a process of conceptual expansion as it absorbs a wider variety of phenomena that were once thought to fall outside its scope.



Paper link

Laura.Fearnley@york.ac.uk