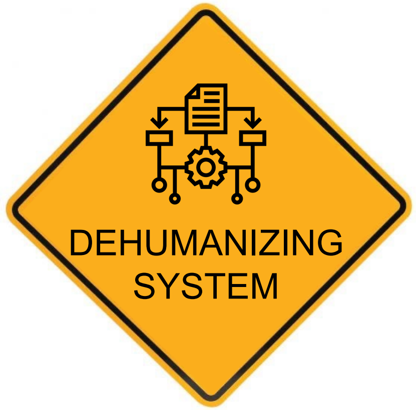


A System Safety Perspective for Developing and Governing Artificial Intelligence

Roel Dobbe

Assistant Professor
Delft University of Technology
Technology, Policy and Management – Engineering Systems and Services

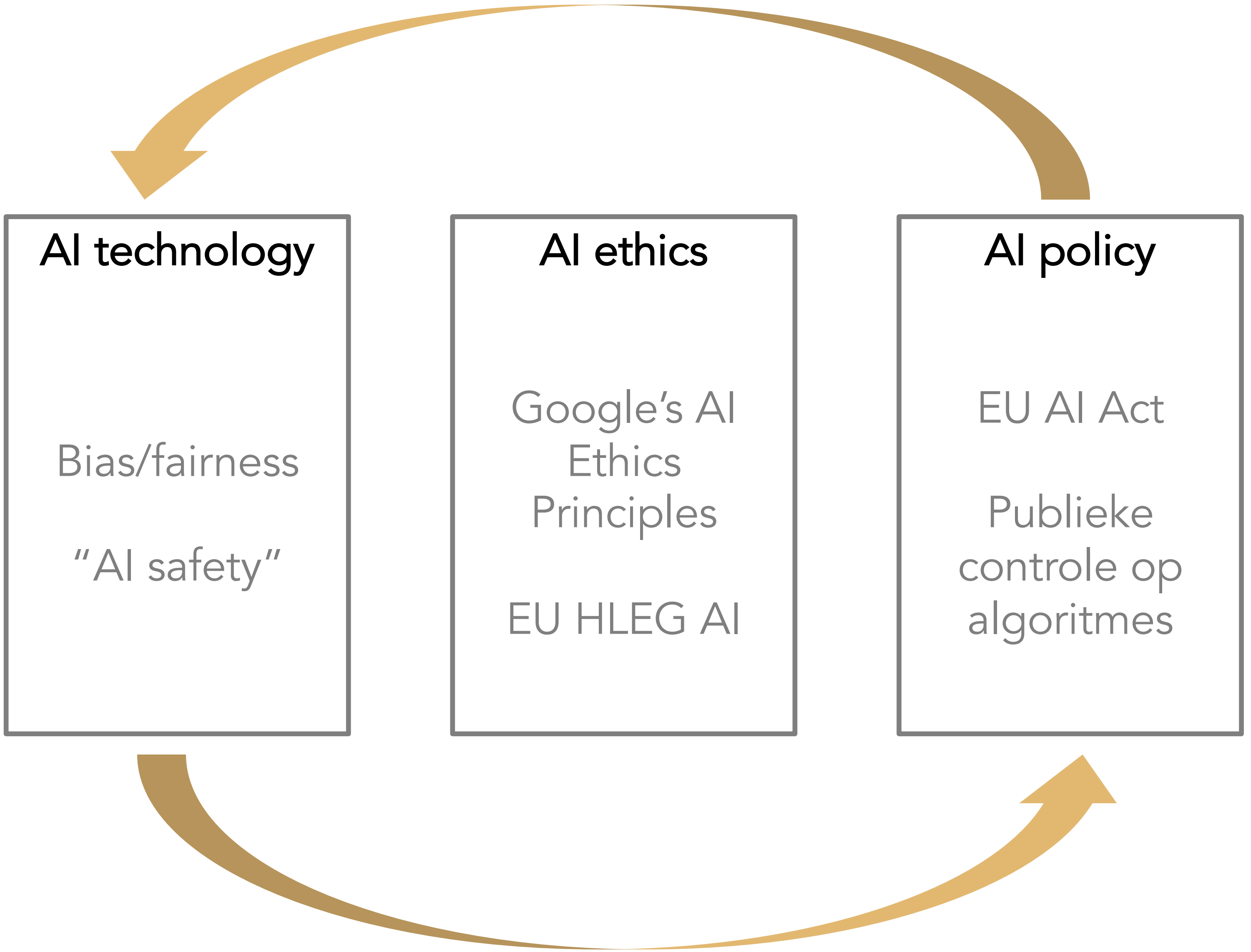
SafeAI 2022 – March 1st, 2022


DEHUMANIZING
SYSTEM

Responses to new harms

AI technology	AI ethics	AI policy
Bias/fairness “AI safety”	EU HLEG AI Google’s AI Ethics Principles	EU AI Act Public control on algoritmes

Measures run risk missing actual harms



Reflexes widen gaps – efforts may miss risks & opportunities

AI tech reflex

Bias/fairness

"AI safety"

AI ethics reflex

Google's AI Ethics Principles

EU HLEG AI

AI policy reflex

EU AI Act

Public control over algorithms

Socio-technical gap

Technical fixes may widen gap between what is socially desired and technically possible. Ethical, legal and social issues follow a push for technological solutions.

Accountability gap

Ethics washing/shopping/ etc widens gap between those who develop/profit from AI and those most likely to suffer the consequences of negative effects.

Policy implementation gap

Push for policy instruments may increase bureaucracy and put disproportionate power in hands of developers/tech, while missing the actual risks for citizens.

Alternative: A Systems Perspective

Socio-technical gap	Accountability gap	Policy implementation gap
Technical fixes widen gap between what is socially desired and technically possible. Ethical, legal and social issues follow a push for technological solutions.	*Ethics washing widens gap between those who develop and profit from AI and those most likely to suffer the consequences of negative effects.*	*Large push for policy instruments increases bureaucracy and puts too much onus for ethical, legal and social implications on the developer.*

	2 Vagueness	3 Infrastructure
	Safety is understood, formalized and experienced differently by different people, requiring socio-technical specification and validation.	*AI systems affect safety by reshaping public infra-structure, requiring democratic checks/balances and citizen engagement for just anticipation of and response to risks.*

ELSEVIER

Artificial Intelligence

Available online 14 July 2021, 103555

In Press, Journal Pre-proof

Hard choices in artificial intelligence

Roel Dobbe [a], Thomas Krendl Gilbert [b], Yonatan Mintz [c, 1]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.artint.2021.103555

Alternative: A Systems Perspective

Socio-technical gap	Accountability gap	Policy implementation gap
Technical fixes widen gap between what is socially desired and technically possible. Ethical, legal and social issues follow a push for technological solutions.	*Ethics washing widens gap between those who develop and profit from AI and those most likely to suffer the consequences of negative effects.*	*Large push for policy instruments increases bureaucracy and puts too much onus for ethical, legal and social implications on the developer.*

1 Emergence	2 Vagueness	3 Infrastructure
Safety is an emergent properties. They are <u>controlled</u> for across integral/iterative design of technical AI artefacts and their institutional context.	*Safety is understood, formalized and experienced differently by different people, requiring <u>socio-technical specification and validation</u>.*	*AI systems affect safety by reshaping public infra-structure, requiring <u>demo-cratic checks/balances and citizen engagement</u> for just anticipation of and response to risks.*

Enter System Safety

Charles Otto
Miller

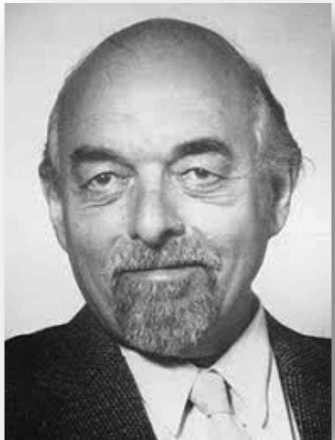
Jens
Rasmussen

Nancy
Leveson

What did system safety respond to?

Charles Otto Miller

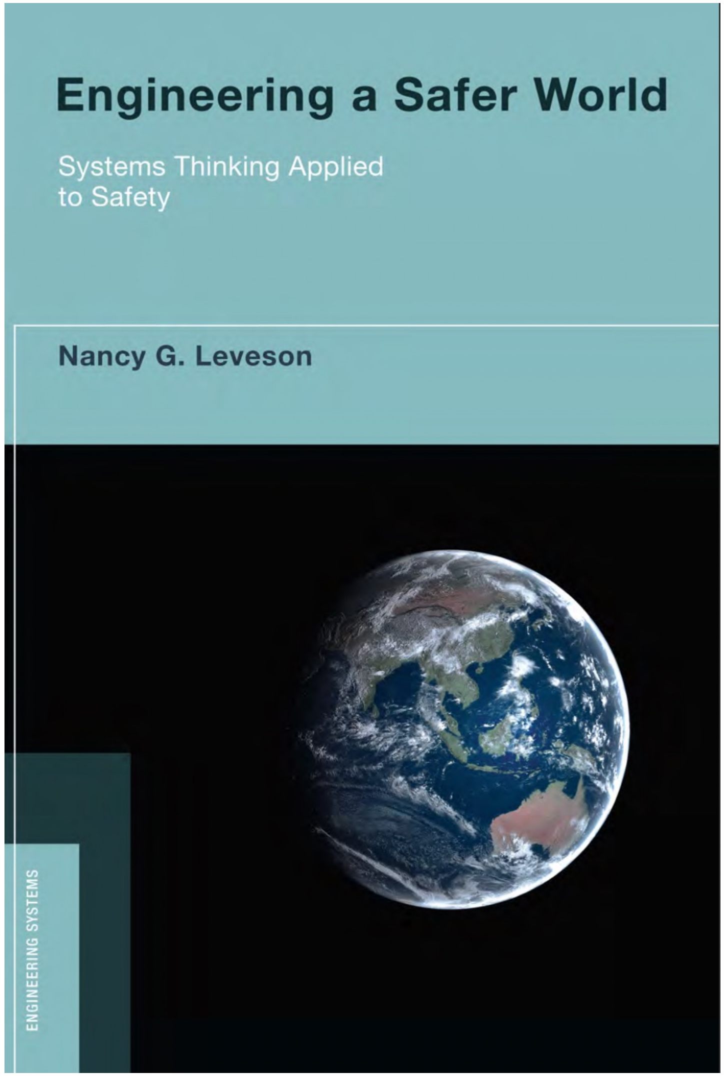
To cope with the increasing complexity of aerospace systems. Many of the ideas have been lost or displaced by more mainstream practices in reliability engineering.



Jens Rasmussen

Applying systems thinking to safety and human factors engineering. Prolific academic who put forward concepts such as boundaries of safe operation, ecological interfaces and methods such as cognitive work analysis.

Professor Nancy Leveson



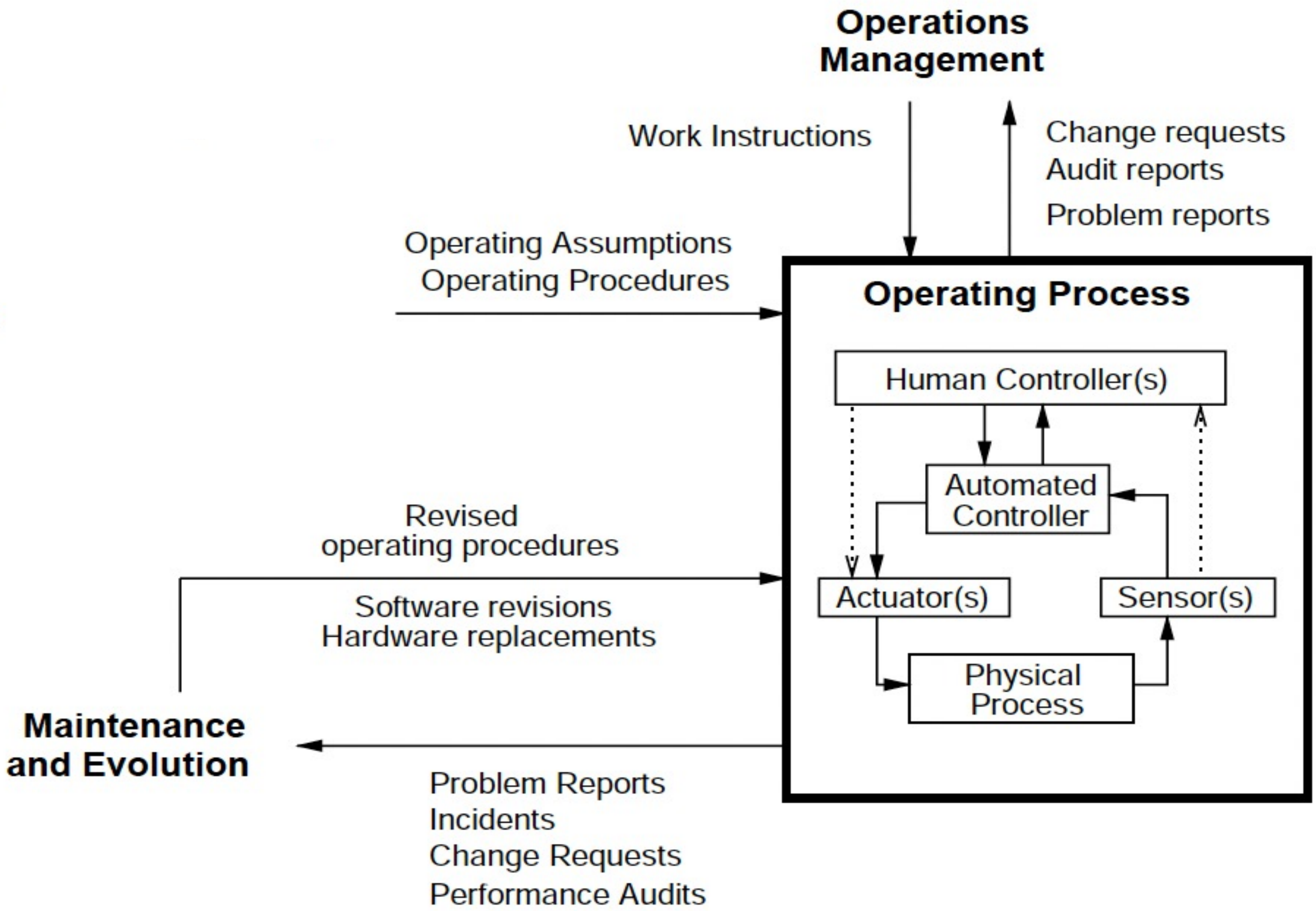
<u>Zeroth</u> assumption: safety is emergent

In systems theory, emergent properties, such as safety, arise from the interactions among the system components. The emergent properties are controlled by imposing constraints on the behavior of and interactions among the components.

Safety then becomes a ***<u>control</u>*** *problem where the goal of the control is to enforce the safety constraints. Accidents result from inadequate control or enforcement of safety-related constraints on the development, design, and operation of the system.*

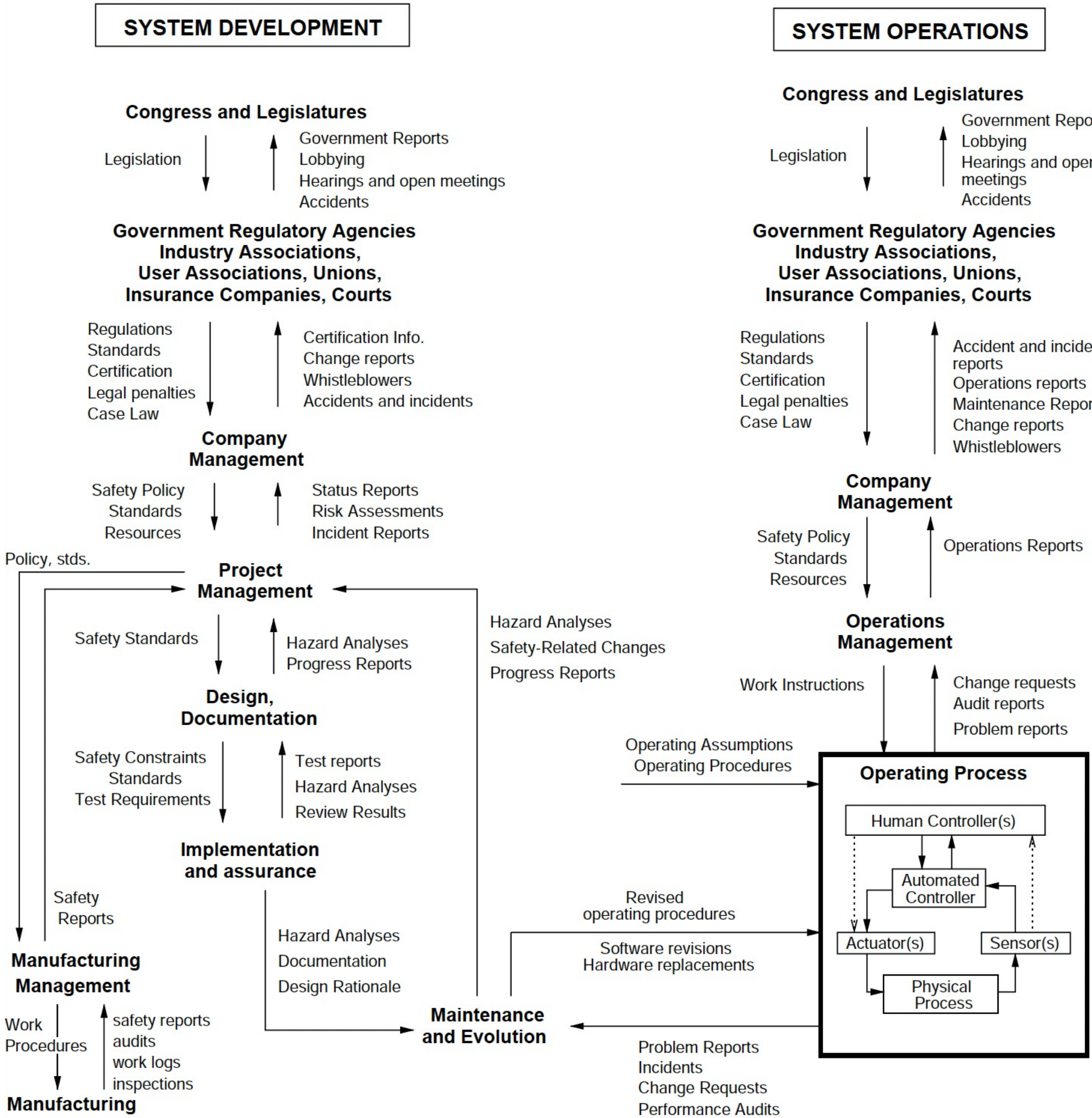
Leveson, Nancy G.. *Engineering a Safer World : Systems Thinking Applied to Safety*, MIT Press, 2012.

What does sociotechnical control look like?



Source: "Engineering a Safer World: applying system thinking to safety", Leveson (2012)

What does sociotechnical control look like?

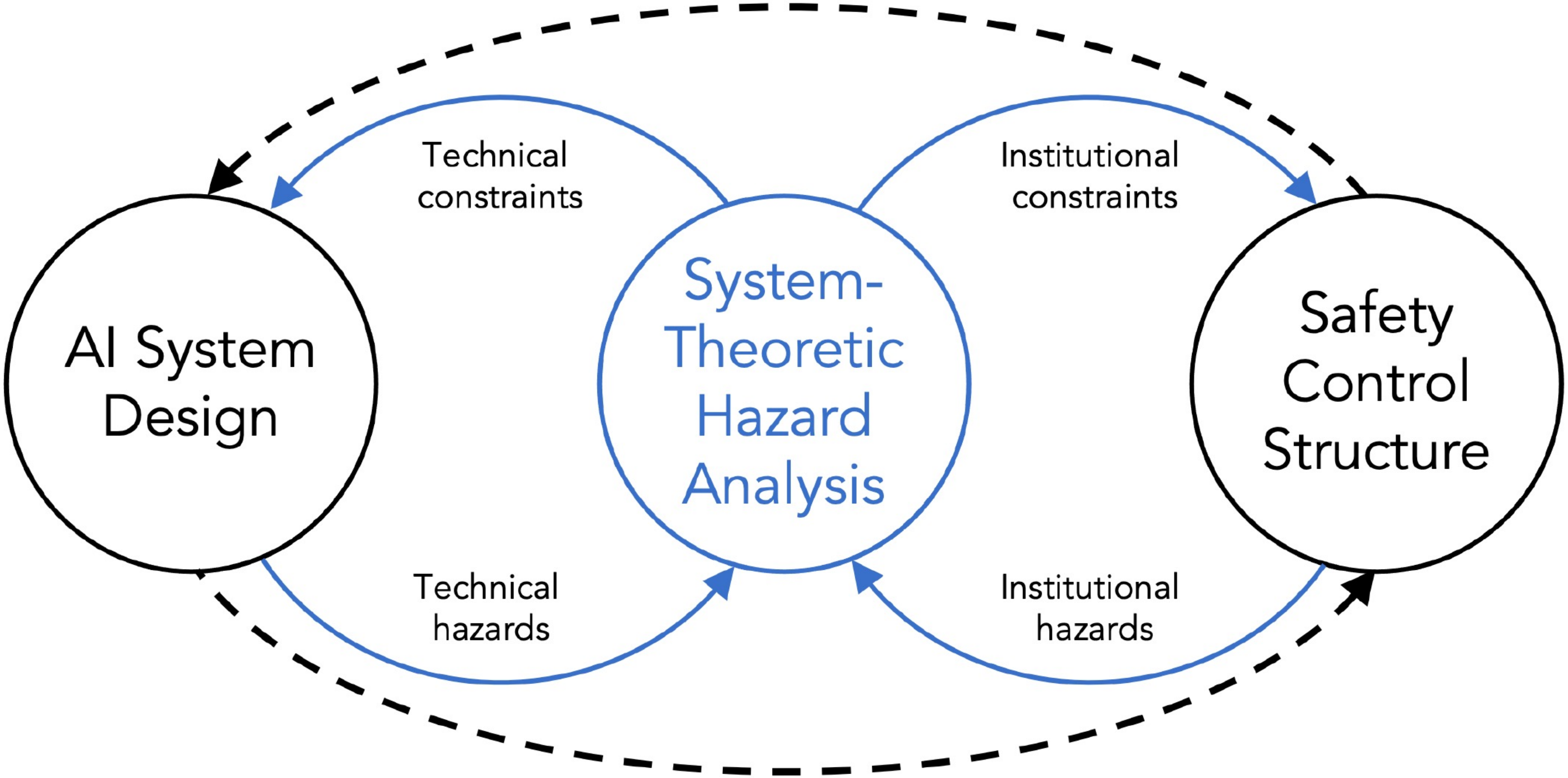


Source: "Engineering a Safer World: applying system thinking to safety", Leveson (2012)

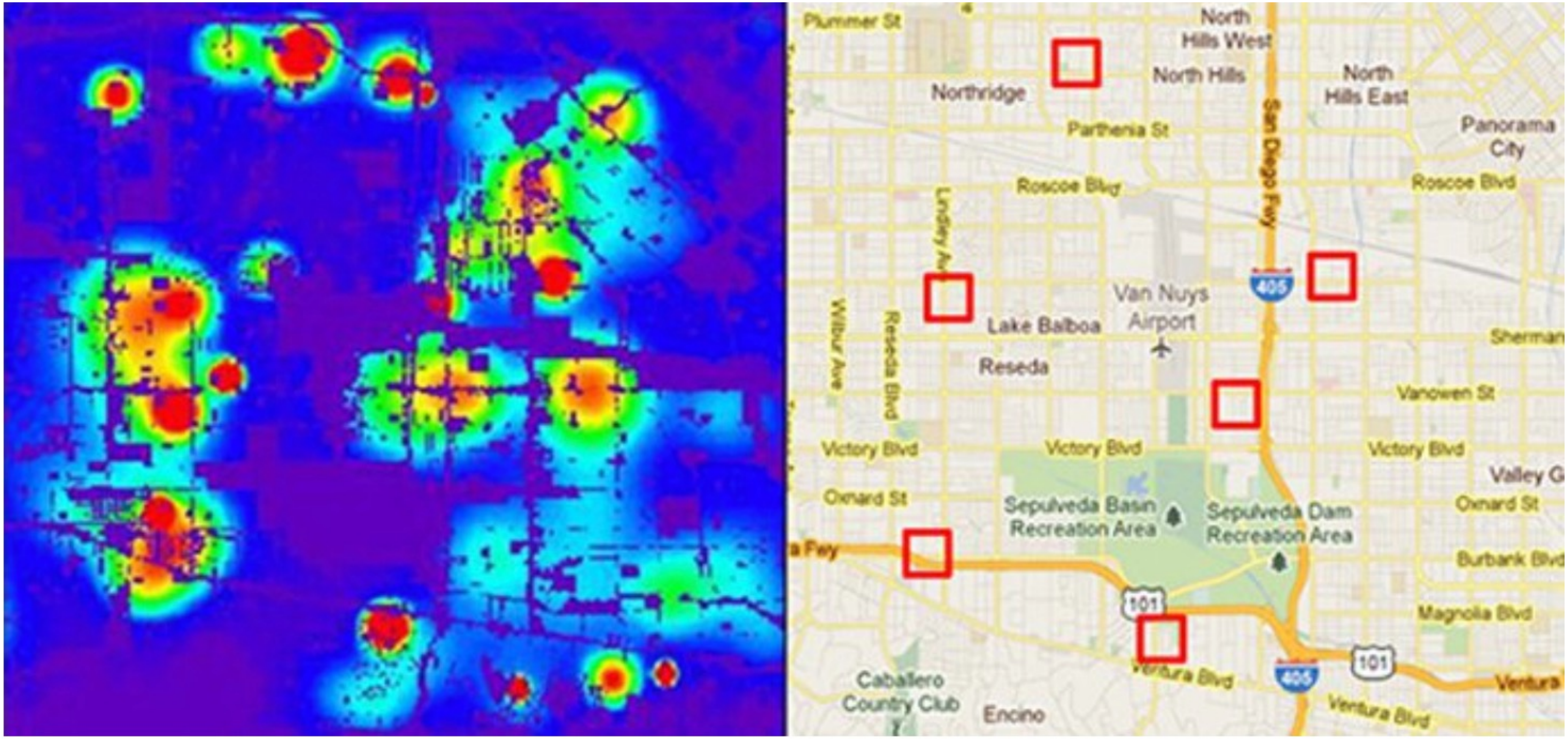
1

Shift Focus from AI Component Reliability to AI System Hazard Elimination

Leveson Lesson 1: *High reliability is neither necessary nor sufficient for safety.*


AI System Design
Technical constraints
Technical hazards
System-Theoretic Hazard Analysis
Institutional constraints
Institutional hazards
Safety Control Structure

Predictive Policing



Predictive policing is built around algorithms that identify potential crime hotspots.. PredPol

Source: Smithsonian Magazine (2018)

2 Shift from Event-based to Constraint-based Accident Models for AI Systems

Leveson Lesson 2: *Accidents are complex processes involving the entire sociotechnical system. Traditional event-chain models cannot describe this process adequately.*

Tesla Crash Williston, Florida, 2016

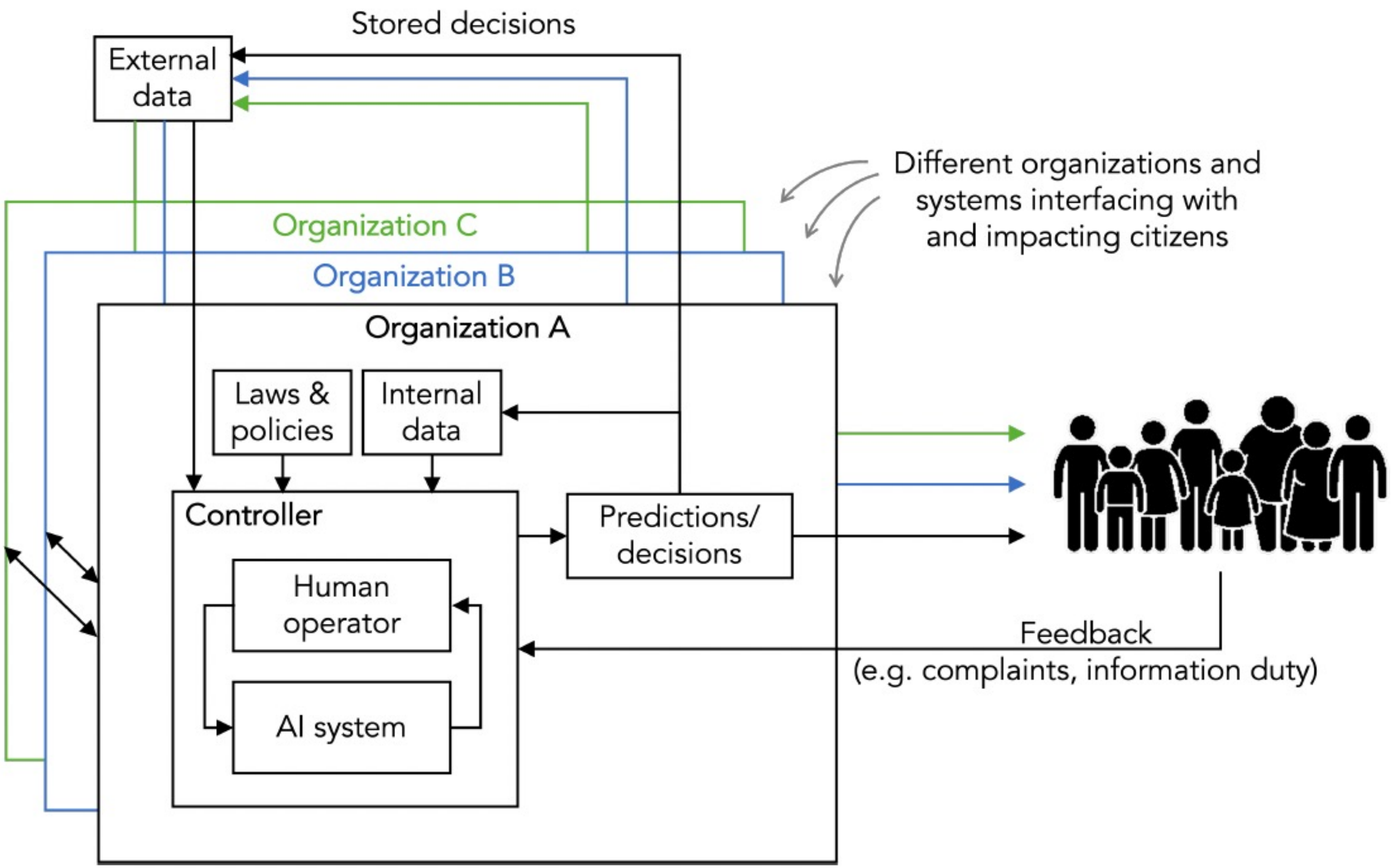
3 Shift from a Probabilistic to a System-theoretic Safety Perspective for AI

Leveson Lesson 3: *Risk and safety may be best understood and communicated in ways other than probabilistic risk analysis.*

Process Model

1. *The goal:* the objectives and safety constraints that must be met and enforced by the controller;
2. *The action condition:* the controller must be able to affect the state of the system;
3. *The observability condition:* the controller must be able to ascertain the state of the system, through feedback, observations and measurements;
4. *The model condition:* the controller must be or contain a model of the process. A human controller should also have a model of the behavior of the AI techniques used for control and decision-making.

The Dutch System Risk Indication System

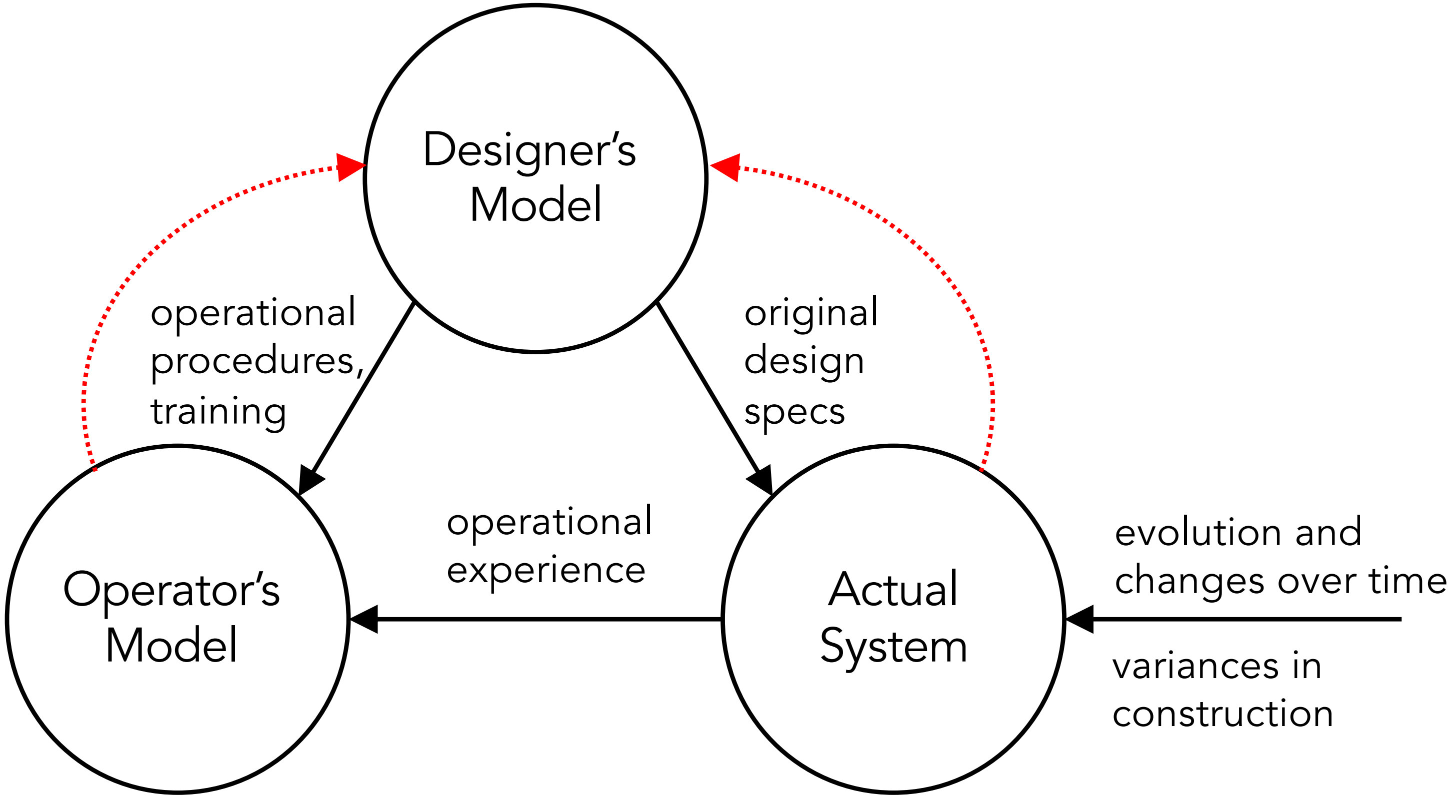


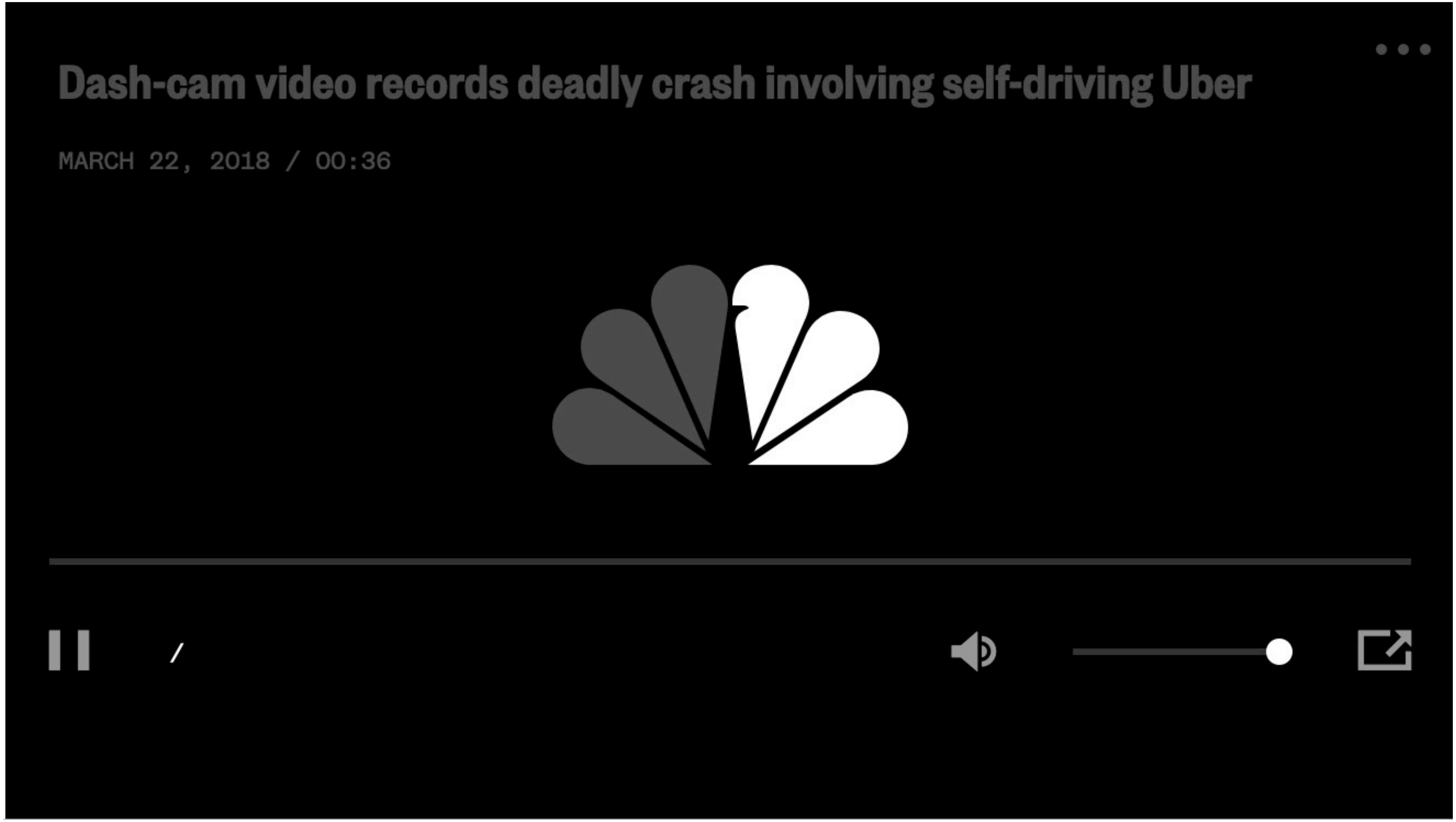
4

Shift from Siloed Design and Operation of AI Systems to Aligning Mental Models

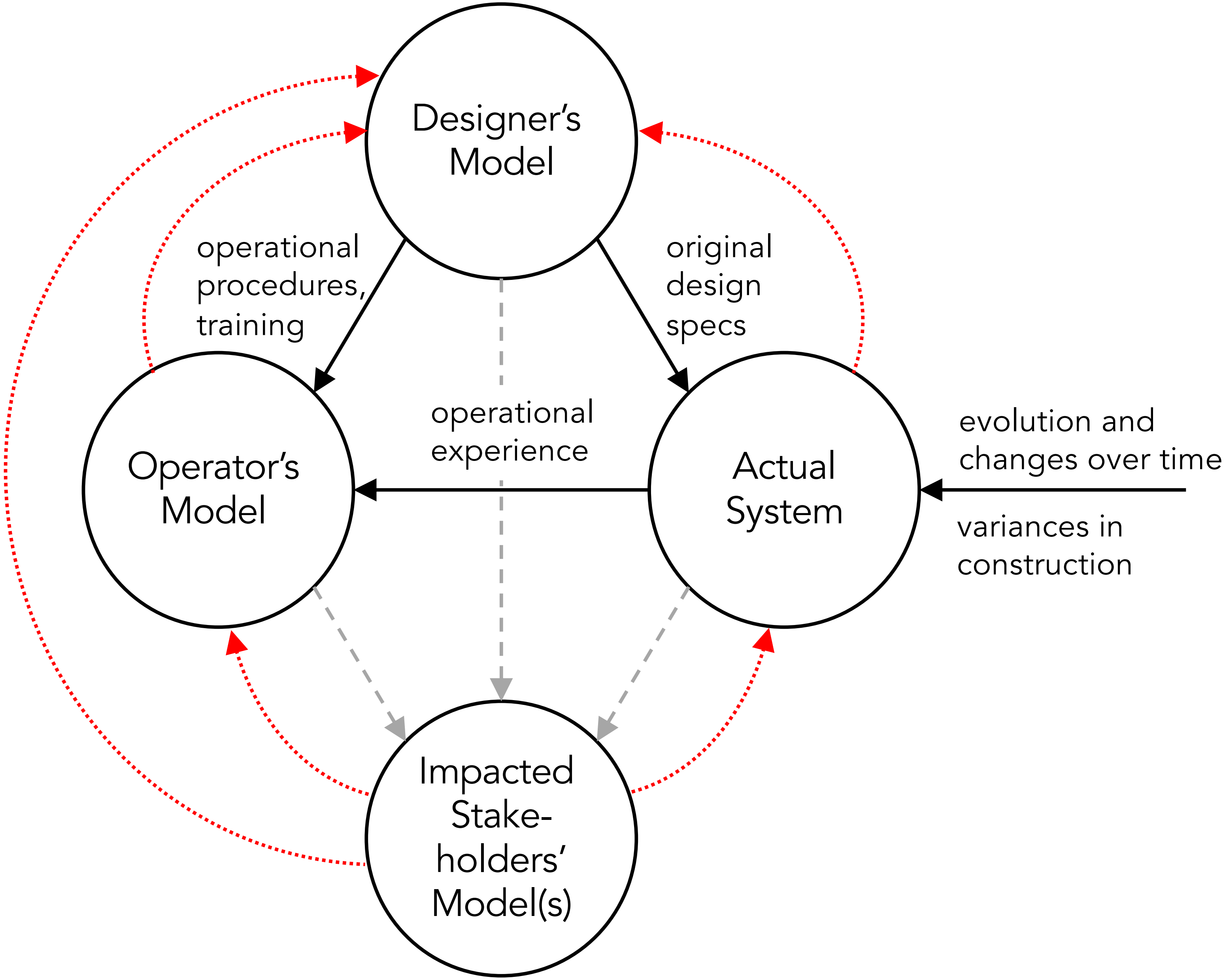
Leveson Lesson 4: *Operator error is a product of the environment in which it occurs. To reduce operator "error" we must change the environment in which the operator works.*

Aligning Mental Models

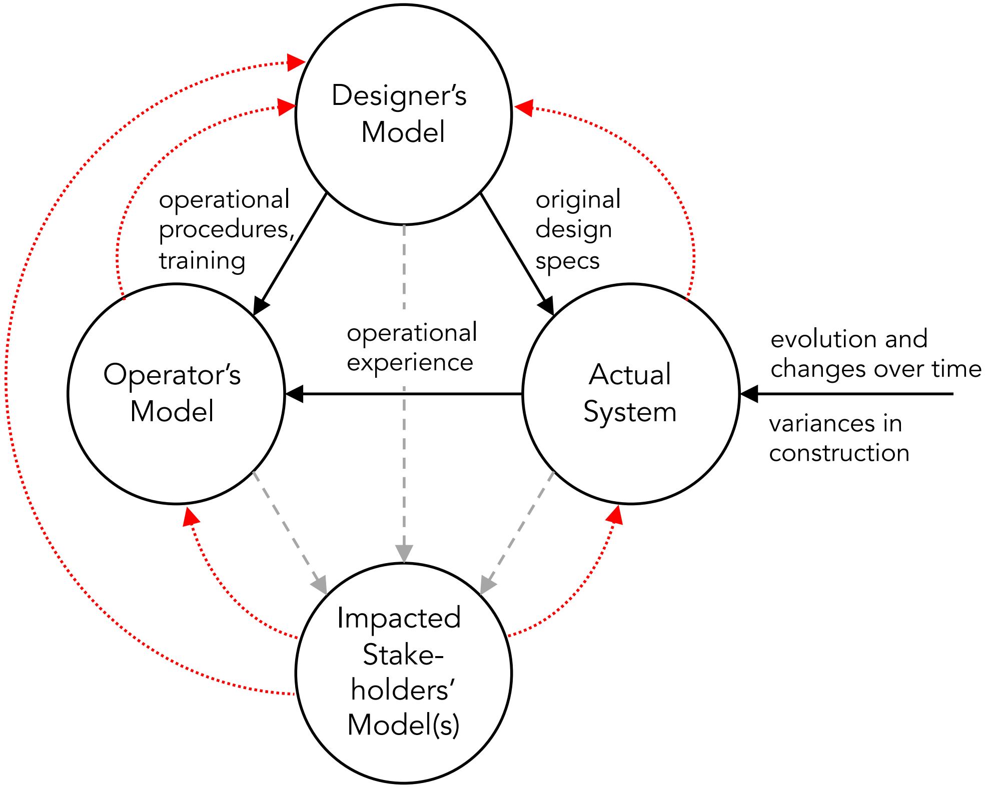



Dash-cam video records deadly crash involving self-driving Uber
MARCH 22, 2018 / 00:36

Aligning Mental Models



AV Fleets as Public Infrastructure



5

Curb the Curse of Flexibility in AI Software Development

Leveson Lesson 5: *Highly reliable software is not necessarily safe. Increasing software reliability or reducing implementation errors will have little impact on safety.*

5. Curse of flexibility 1/2

"Many software requirements problems arise from what could be called the *curse of flexibility*.

The computer is so powerful and so useful because it has eliminated many of the physical constraints of previous machines. [..]

With software, the limits of what is possible to accomplish are different than the limits of what can be accomplished successfully and safely – the limiting factors change from the structural integrity and physical constraints of our materials to limits on our intellectual capabilities."

Source: "Engineering a Safer World", Leveson (2012)

5. Curse of flexibility 2/2

“Nearly all the serious accidents in which software has been involved in the past twenty years can be traced to requirements flaws, not coding errors. [..]

The most serious problems arise, however, when nobody understands what the software should do or even what it should not do. We need better techniques to assist in determining these requirements.”

Source: ”Engineering a Safer World", Leveson (2012)

Parameters in LLMs since 2012

AlexNet (2012)	60,000,000
ELMo (2018)	94,000,000
Megatron-Turing NLG (2021)	530,000,000,000
GPT-4 (expected 2023)	~100,000,000,000,000

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender[1*], Timnit Gebru[2*],
Angelina McMillan-Major[1], Shmargaret Shmitchell[3]

[1] University of Washington [2] Black in AI [3] The Aether
*These authors contributed equally.

Computational Infrastructure

FEATURESDIALOG

XXVIII.6 November - December 2021
Page: 50
Digital Citation

THE STEEP COST OF CAPTURE

Authors:
Meredith Whittaker

This is a perilous moment. Private computational systems marketed as artificial intelligence (AI) are threading through our public life and institutions, concentrating industrial power, compounding marginalization, and quietly shaping access to resources and information.

Programmable Infrastructures

The term "programmable infrastructures" refers to the political, economic and technological vision that advocates for the introduction of computational infrastructure onto our common infrastructures.

If common infrastructures come with extensive planning and expensive updates, the promise of programmability is that by adding a digital layer, the plans and policies of common infrastructures can be abstracted from their underlying physical constraints.

This, it is claimed, will make them easy to reconfigure just like digital systems. In other words, legacy physical infrastructures can be further freed from their physical constraints and can ostensibly be made as programmable as native computational systems.

Source: "Programmable Infrastructures", Gürses, Poon and Dobbe (2020)

6 Translate Safety Constraints to the Design and Operation of the AI System

Leveson Lesson 6: *Systems will tend to migrate toward states of higher risk. Such migration is predictable and can be prevented by appropriate system design or detected during operations using leading indicators of increasing risk.*

7

Build an Organization and Culture that is Open to Understanding and Learning

Leveson Lesson 7: *Blame is the enemy of safety. Focus should be on understanding how the system behavior as a whole contributed to the loss and not on who or what to blame for it.*

Importance of management and culture

"The key to effectively accomplishing
any of the goals described in [the system safety discipline] lies in management.

Most people want to run safe organizations, but they may misunderstand the tradeoffs required and how to accomplish the goals."

Source: "Engineering a Safer World: applying system thinking to safety", Leveson (2012)

A ‘Just Culture’ balances safety and accountability

“Only responding to calls for accountability is not likely to lead you to justice or to improved safety.

People will feel unfairly singled out, and disclosure of safety problems will suffer.”

Source: ”Just Culture: balancing safety and accountability", Dekker (2016)

A ‘Just Culture’ balances safety and accountability

“A just culture, then, also pays attention to safety, so that people feel comfortable to

(1) bring out information about what should be improved to levels or groups that can do something about it; and

(2) allow the organization to invest resources in improvements that have a safety dividend,
rather than deflecting resources into legal protection and limiting liability."

Source: ”Just Culture: balancing safety and accountability", Dekker (2016)

Seven lessons for AI Design & Governance

	Leveson Lesson	AI System Safety Implication	Example System Safety Strategy
1	Component reliability is insufficient for safety	Identify and eliminate hazards at system level	System hazard-informed system design and safety control structure
2	Causal event models cannot capture system complexity	Understand safety through socio-technical constraints	System-theoretic accident models: integrating safety constraints, the process model and the safety control structure
3	Probabilistic methods don't provide safety guarantees	Capture safety conditions and requirements in a system-theoretic way	Process model: AI system goals, actions, observation and model of controlled process and automation
4	Operator error is a product of the environment	Align mental models across design, operation and affectedstakeholders	Leveson's design principles for shared human-AI controller design: redundancy, incremental control and error tolerance
5	Reliable software is not necessarily safe	Include (AI) software in hazard analysis	System-theoretic process analysis
6	Systems migrate to states of higher risk	Ensure operational safety	Feedback mechanisms (audits, investigations and reporting systems)
7	Blame is the enemy of safety	Build an organization and culture that is open to understanding and learning	Just Culture

Source: "System Safety and Artificial Intelligence", Dobbe (forthcoming)

Thank you!

- r.i.j.dobbe@tudelft.nl

Main references:

- Roel Dobbe, “System Safety and Artificial Intelligence,” in *Oxford Handbook on AI Governance (forthcoming)*, Oxford, 2022. https://arxiv.org/abs/2202.09292
- R. Dobbe, T. Krendl Gilbert, and Y. Mintz, “Hard choices in artificial intelligence,” *Artificial Intelligence*, vol. 300, p. 103555, Nov. 2021, doi: 10.1016/j.artint.2021.103555.
- N. G. Leveson and J. Moses, Engineering a Safer World: Systems Thinking Applied to Safety. Cambridge, UNITED STATES: MIT Press, 2012.
- S. Dekker, *Just culture: Balancing safety and accountability*. CRC Press, 2016.