From Silos to Systems: Perspectives on STPA for AI Sociotechnical Harms

Shalaleh Rismani, McGill University, Canada
Maurus Enbergs, TU Delft, Netherlands
Íñigo de Troya, TU Delft, Netherlands
AJung Moon, McGill University, Canada
Roel Dobbe, TU Delft, Netherland

Background & goals

Background:

- Growing discussion about harms from AI systems
- Evolving meaning of safety for AI systems
- Emerging fields of AI safety, responsible AI, AI ethics, FATE (Fairness, Accountability, Transparency and Ethics)

Goals:

1. Reflect on the lessons from system safety for AI systems [1,2]
2. **Unpack how STPA could be applied for AI systems**



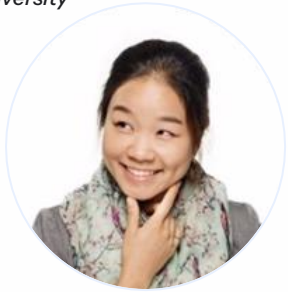
Shalaleh Rismani
Postdoctoral researcher, McGill University



Maurus Enbergs
PhD researcher, TU Delft



Íñigo de Troya
PhD researcher, TU Delft



AJung Moon
Associate professor, McGill University



Roel Dobbe
Assistant professor, TU Delft

[1] Dobbe, R. (2022). System safety and artificial intelligence. 2022 ACM Conference on Fairness Accountability and Transparency, 1584–1584. https://doi.org/10.1145/3531146.3533215
[2] Rismani, S., Shelby, R., Smart, A., Jatho, E., Kroll, J., Moon, A., & Rostamzadeh, N. (2023). From plane crashes to algorithmic harm: Applicability of safety engineering frameworks for responsible ML. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2, 1–18. https://doi.org/10.1145/3544548.3581407

Presentation overview

Insight #1

Accounting for sociotechnical harms and dynamics using STPA

Case 1

Text-to-Image in creative practice

Insight #2

STPA enables rigorous analysis without access to AI model internals

Case 2

Fraud prediction

Insight #3

System boundaries are normative and shape what we uncover

Case 3

Legal automation

Moving forward

Discussion and call for collaboration!

Insight #1

Accounting for sociotechnical harms and dynamics using STPA

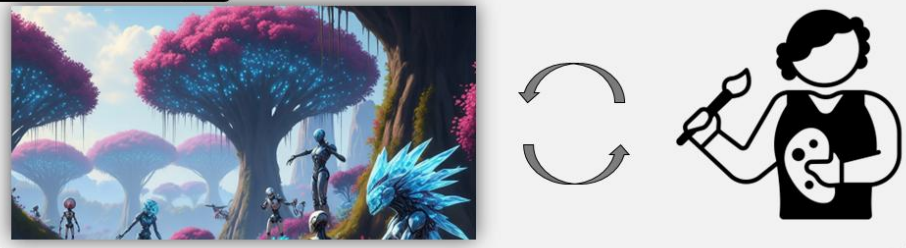
Case 1

Text-to-Image in creative practice

Harms from AI systems go beyond traditional safety-critical harms

RQ: How can we use STPA to examine harms from AI systems that are not traditionally safety-critical?

Case study



Use of Text to Image (T2I) models in creative practice

Workshop with 15 Text-to-Image (T2I) artists from 6 countries

Semi-structured interviews with 8 T2I developers

Data processing for creating training dataset

Model integration in Text to Image demo

Use of the T2I demo by a filmmaker

Applying STPA*

*This case study was completed during an internship at Google Research

[3] Rismani, S., Shelby, R., Smart, A., Delos Santos, R., Moon, A., & Rostamzadeh, N. (2023). Beyond the ML model: Applying safety engineering frameworks to text-to-image development. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2, 70–83. https://doi.org/10.1145/3600211.3604685
[4] [In review] Rismani, S., Dobbe, R., & Moon, A. (2024). From silos to systems: Process-oriented Hazard Analysis for AI systems. ArXiv, abs/2410.22526. arXiv. https://doi.org/10.48550/arXiv.2410.22526

Insight #1: Using sociotechnical harms as a way of thinking about losses

Sociotechnical harms are the adverse lived experiences resulting from a system's deployment and operation in the world.

(L1) Loss of creativity (Interpersonal harm)

(L2) Loss of diversity (Representational harm)

(L3) Loss of accessibility

(L4) Loss of efficiency

(L5) Loss of quality (Social system harm)

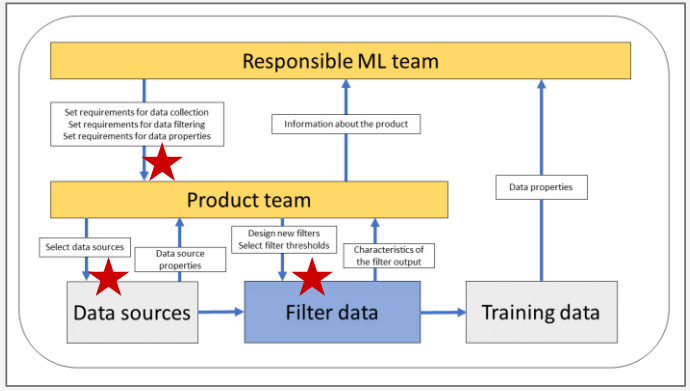
(L6) Loss of reputation

Data processing for creating training dataset	**Model** integration in Text to Image demo	**Use** of the T2I demo by a filmmaker
H1. Data set contains low-quality text-image pairs	H1. create an image that doesn't match the prompt	H1. the artist cannot create a video
H2. contains harmful content	H2. fails to generate an image at all	H2. the artist cannot create a video that meets client requirements
H3. lacks diversity in representation	H3. generates an unsafe image, such as one with adult content	H3. artist creates a video that disseminates false or harmful information

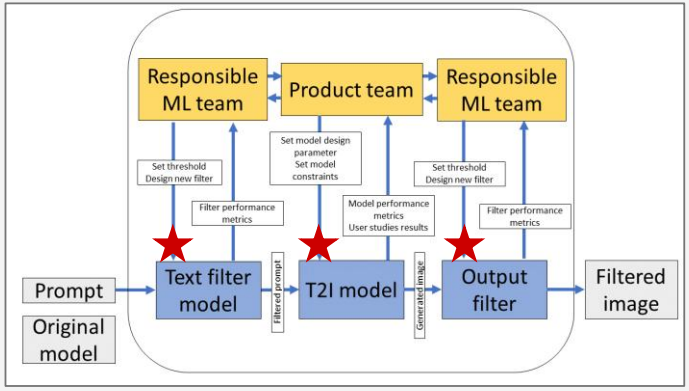
[5] Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 24, 723–741. https://doi.org/10.1145/3600211.3604673

Insight #1: Using control diagram to capture sociotechnical dynamics of development and use

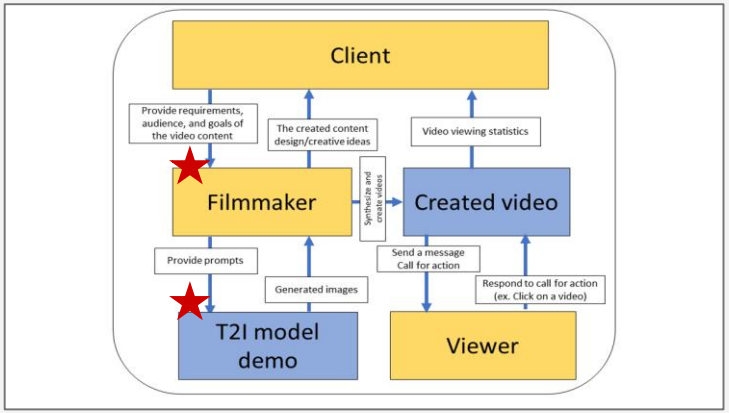
Data processing for creating training dataset



Model integration in Text to Image demo



Use of the T2I demo by a filmmaker



Insight #2

STPA enables rigorous analysis without access to AI model internals

Case 2

Fraud prediction

Dealing with opaque, algorithmic components in sociotechnical systems

Case Study



The Childcare Benefits Scandal saw the **Dutch Tax Authority wrongly accuse** thousands of **parents of childcare benefit fraud**. Due to **automated risk profiling** benefits were cut, **debts reclaimed**, and **families pushed into financial ruin**.

- Timeframe: 2015-2021
- **~26,000 parents** falsely accused
- Upwards of **1,600 children removed** from their homes
- Estimated compensation costs: upwards of **€8 billion**

Difficult to study as it concerns highly **confidential government information** with **pending litigation cases**.

→ **No model access,** only:

Public documentation (official investigation reports)

STPA

Dealing with opaque, algorithmic components in sociotechnical systems

Case Study



The Childcare Benefits Scandal saw the **Dutch Tax Authority wrongly accuse** thousands of **parents of childcare benefit fraud**. Due to **automated risk profiling** benefits were cut, **debts reclaimed**, and **families pushed into financial ruin**.

- Timeframe: 2015-2021
- **~26,000 parents** falsely accused
- Upwards of **1,600 children removed** from their homes
- Estimated compensation costs: upwards of **€8 billion**

Difficult to study as it concerns highly **confidential government information** with **pending litigation cases**.

→ **No model access,** only:

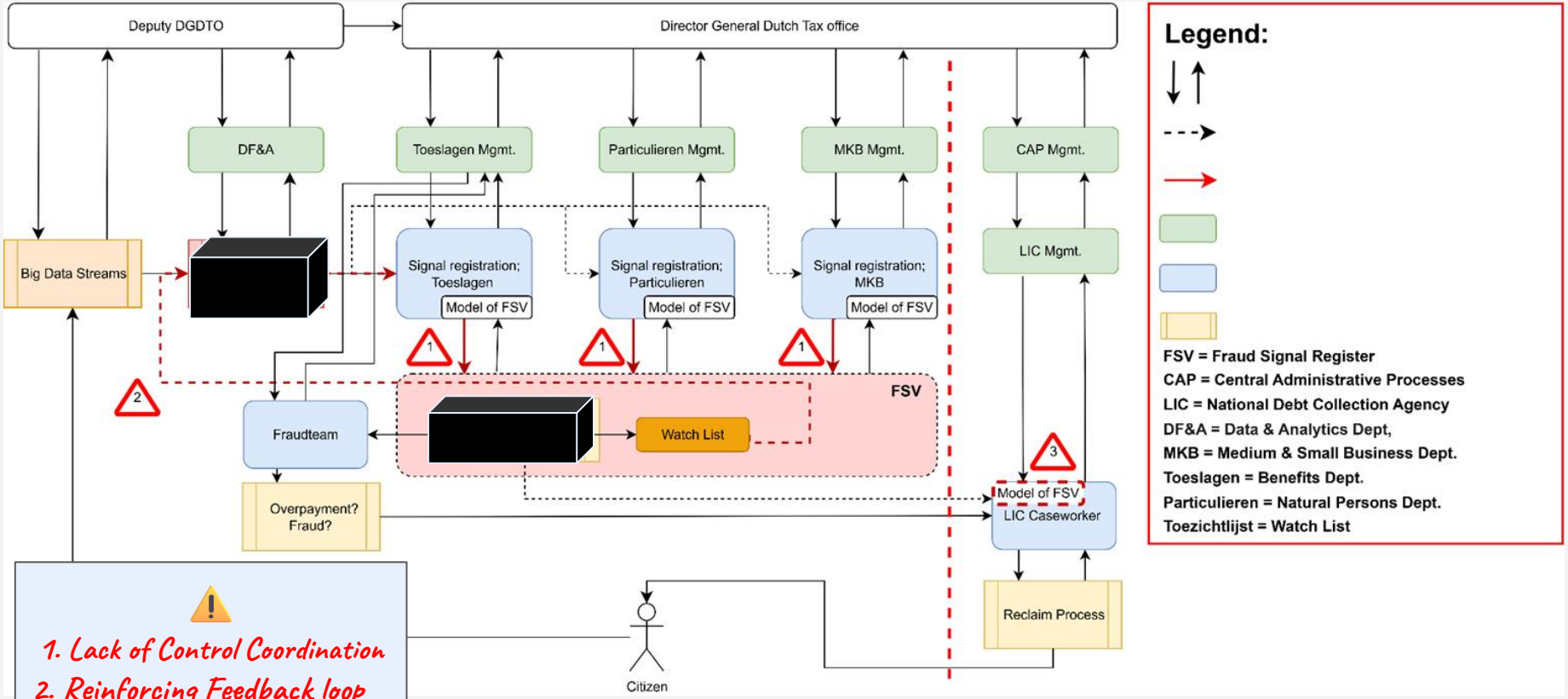


Public documentation (official investigation reports)

STPA

Insight #2:

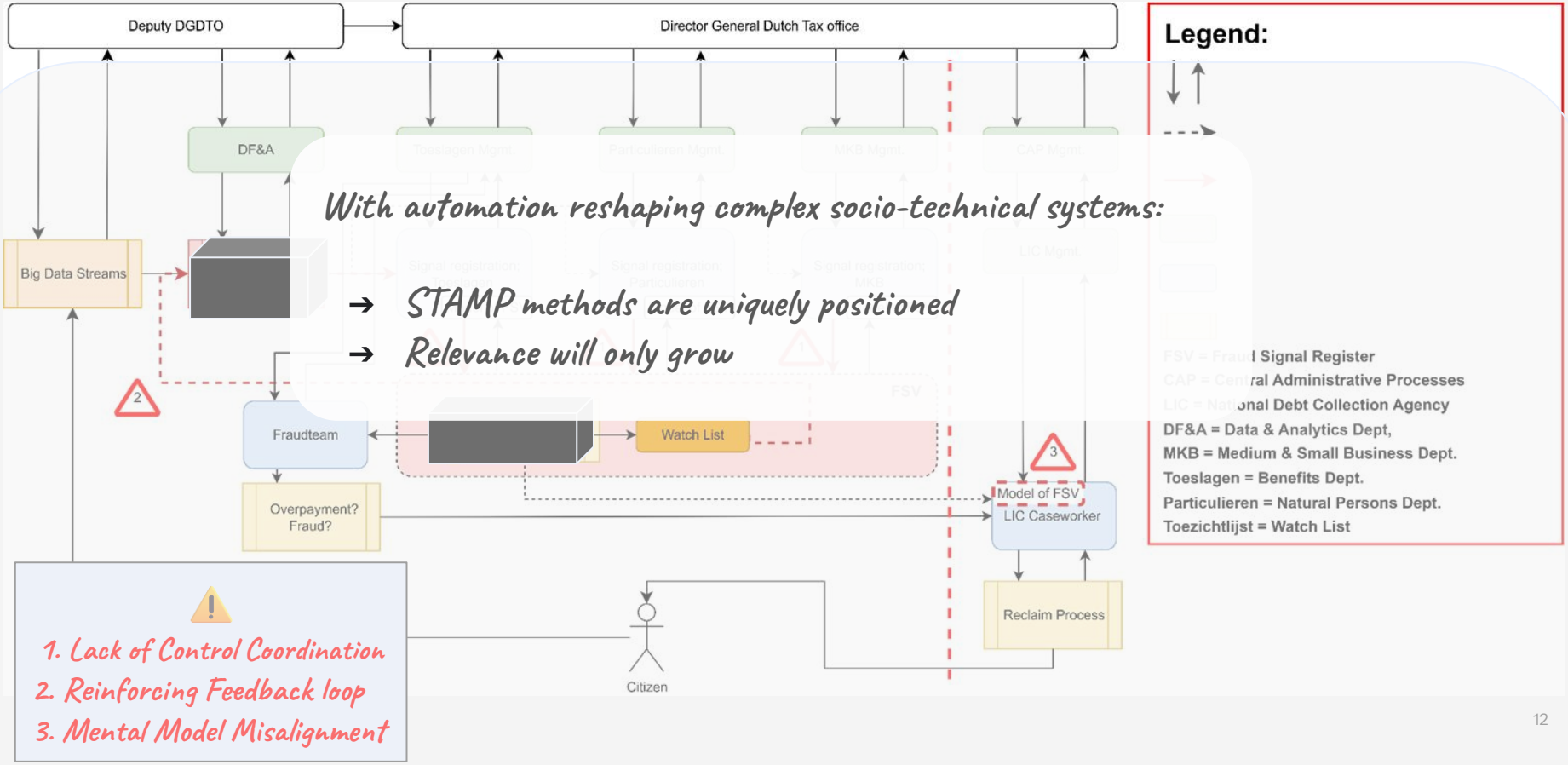
STPA enables rigorous analysis without access to AI model internals



1. Lack of Control Coordination
2. Reinforcing Feedback loop
3. Mental Model Misalignment

Insight #2:

STPA enables rigorous analysis without access to AI model internals



With automation reshaping complex socio-technical systems:

→ *STAMP methods are uniquely positioned*

→ *Relevance will only grow*

1. *Lack of Control Coordination*
2. *Reinforcing Feedback loop*
3. *Mental Model Misalignment*

Insight #3

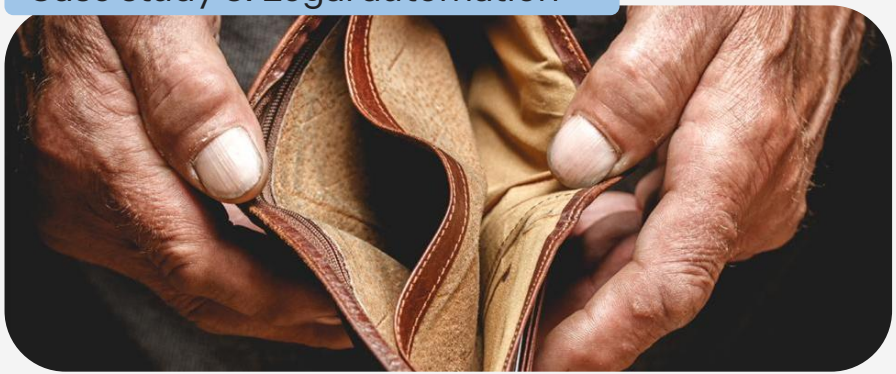
System boundaries are normative and shape what we uncover

Case 3

Legal automation

Insight #3:
System boundaries are normative and shape what we uncover

Case study 3: Legal automation



Municipal Algorithm Registers are transparency instruments that inform citizens about how algorithmic systems used in public services through a public database (website). They are meant to **assure the public that algorithmic systems are safe to use**.

...but how well do they inform the public about the system?

...can we conduct an STPA on Avola using public information and additional stakeholder input?

RQ. How can **stakeholder perspectives** complement publicly available information (from a municipal algorithm register) to conduct an STPA of ***Avola*** – a **legal automation** system for determining **welfare eligibility**?

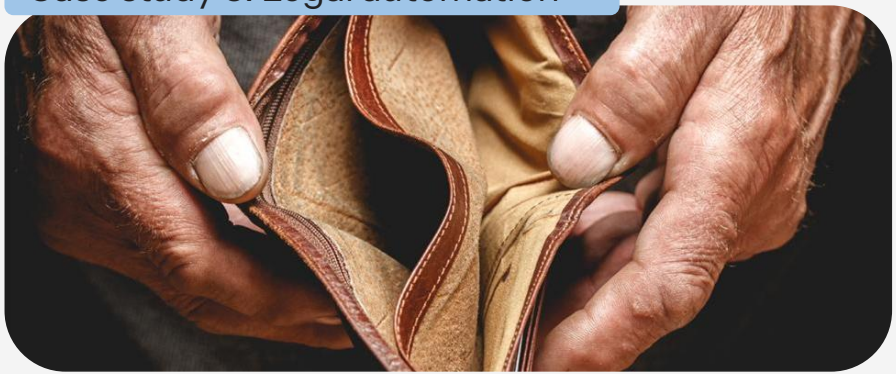
As part of our own STPA, we conducted **system mapping workshops** with 3 different stakeholder groups to understand how their own perspectives and domain expertise informs how they map the system.

- Municipal staff (N=2)
- Civil Society Organisations (N=2)
- Ombudsmen (N=2)

By **extending the system boundary** to consider Ombudsmen and Civil Society Organisations, we can **identify missing safety mechanisms** as well as **inadequacies** in existing ones.

Insight #3:
System boundaries are normative and shape what we uncover

Case study 3: Legal automation



Municipal Algorithm Registers are transparency instruments that inform citizens about how algorithmic systems used in public services through a public database (website). They are meant to **assure the public that algorithmic systems are safe to use**.

...but how well do they inform the public about the system?

...can we conduct an STPA on Avola using public information and additional stakeholder input?

RQ. How can **stakeholder perspectives** complement publicly available information (from a municipal algorithm register) to conduct an STPA of ***Avola*** – a **legal automation** system for determining **welfare eligibility**?

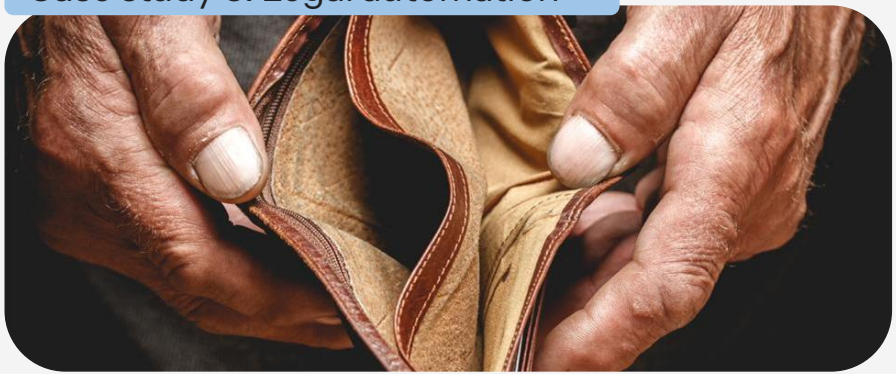
As part of our own STPA, we conducted **system mapping workshops** with 3 different stakeholder groups to understand how their own perspectives and domain expertise informs how they map the system.

- Municipal staff (N=2)
- Civil Society Organisations (N=2)
- Ombudsmen (N=2)

By **extending the system boundary** to consider Ombudsmen and Civil Society Organisations, we can **identify missing safety mechanisms** as well as **inadequacies** in existing ones.

Insight #3:
System boundaries are normative and shape what we uncover

Case study 3: Legal automation



RQ. How can **stakeholder perspectives** complement publicly available information (from a municipal algorithm register) to conduct an STPA of ***Avola*** – a **legal automation** system for determining **welfare eligibility**?

Municipal Algorithm Registers are transparency instruments that inform citizens about how algorithmic systems used in public services through a public database (website). They are meant to **assure the public that algorithmic systems are safe to use**.

…but how well do they inform the public about the system?

…can we conduct an STPA on Avola using public information and additional stakeholder input?

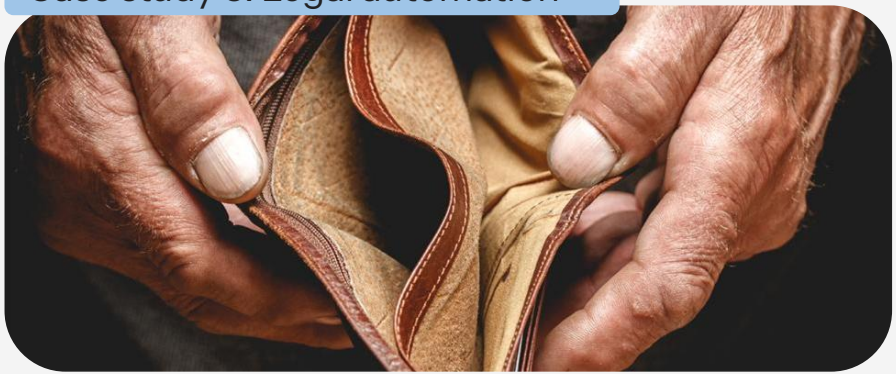
As part of our own STPA, we conducted **system mapping workshops** with 3 different stakeholder groups to understand how their own perspectives and domain expertise informs how they map the system.

- Municipal staff (N=2)
- Civil Society Organisations (N=2)
- Ombudsmen (N=2)

By **extending the system boundary** to consider Ombudsmen and Civil Society Organisations, we can **identify missing safety mechanisms** as well as **inadequacies** in existing ones.

Insight #3:
System boundaries are normative and shape what we uncover

Case study 3: Legal automation



RQ. How can **stakeholder perspectives** complement publicly available information (from a municipal algorithm register) to conduct an STPA of ***Avola*** – a **legal automation** system for determining **welfare eligibility**?

Municipal Algorithm Registers are transparency instruments that inform citizens about how algorithmic systems used in public services through a public database (website). They are meant to **assure the public that algorithmic systems are safe to use**.

...but how well do they inform the public about the system?

...can we conduct an STPA on Avola using public information and stakeholder input?

As part of our own STPA, we conducted **system mapping workshops** with 3 different stakeholder groups to understand how their own perspectives and domain expertise informs how they map the system.

- 👷 Municipal staff (N=2)
- 🙋‍♀️ Civil Society Organisations (N=2)
- 🕵️‍♀️ Ombudsmen (N=2)

By **extending the system boundary** (beyond what's published in the algorithm register) to consider external actors, we can **identify missing safety mechanisms** as well as **inadequacies** in existing ones.

Insight #3:
System boundaries are normative and shape what we uncover

	Missing or inadequate safety mechanisms	Sociotechnical harms implicated
🤷 Ombudsmen	- *Youth support* organisations - Unclear if data Avola uses is up-to-date (e.g. current salary) - Incident Reports sent to developer may be technical, rather than tell citizens' stories	- Potential *denial of benefits* due to out-of-date data
🙋 Civil Society Organisations	- Freedom of Information request is a high barrier for citizen - Citizens probably don't know what a Human Rights or Data Protection Impact Assessments are - Unclear who handles complaints, and what action is taken - Questioned lack of evaluation metrics	- Potential *denial of benefits* if legal automation performance metrics are biased - Results of Human Rights and Data Protection Impact Assessments not shared (*Information harm*)
👷 Municipal staff	- Government bodies - Unaware of evaluation metrics or "legal tests" performed	- Potential *denial of benefits* if legal automation not correctly implemented

Insight #3:
System boundaries are normative and shape what we uncover

	Missing or inadequate safety mechanisms	**Sociotechnical harms implicated**
🤷‍♀️ Ombudsmen	- *Youth support* organisations - Unclear if data Avola uses is up-to-date (e.g. current salary) - Incident Reports sent to developer may be technical, rather than tell citizens' stories	- Potential *denial of benefits* due to out-of-date data
🙋‍♀️ Civil Society Organisations	- Freedom of Information request is a high barrier for citizen - Citizens probably don't know what a Human Rights or Data Protection Impact Assessments are - Unclear who handles complaints, and what action is taken - Questioned lack of evaluation metrics	- Potential *denial of benefits* if legal automation performance metrics are biased - Results of Human Rights and Data Protection Impact Assessments not shared (*Information harm*)
👩‍💼 Municipal staff	- Government bodies - Unaware of evaluation metrics or "legal tests" performed	- Potential *denial of benefits* if legal automation not correctly implemented

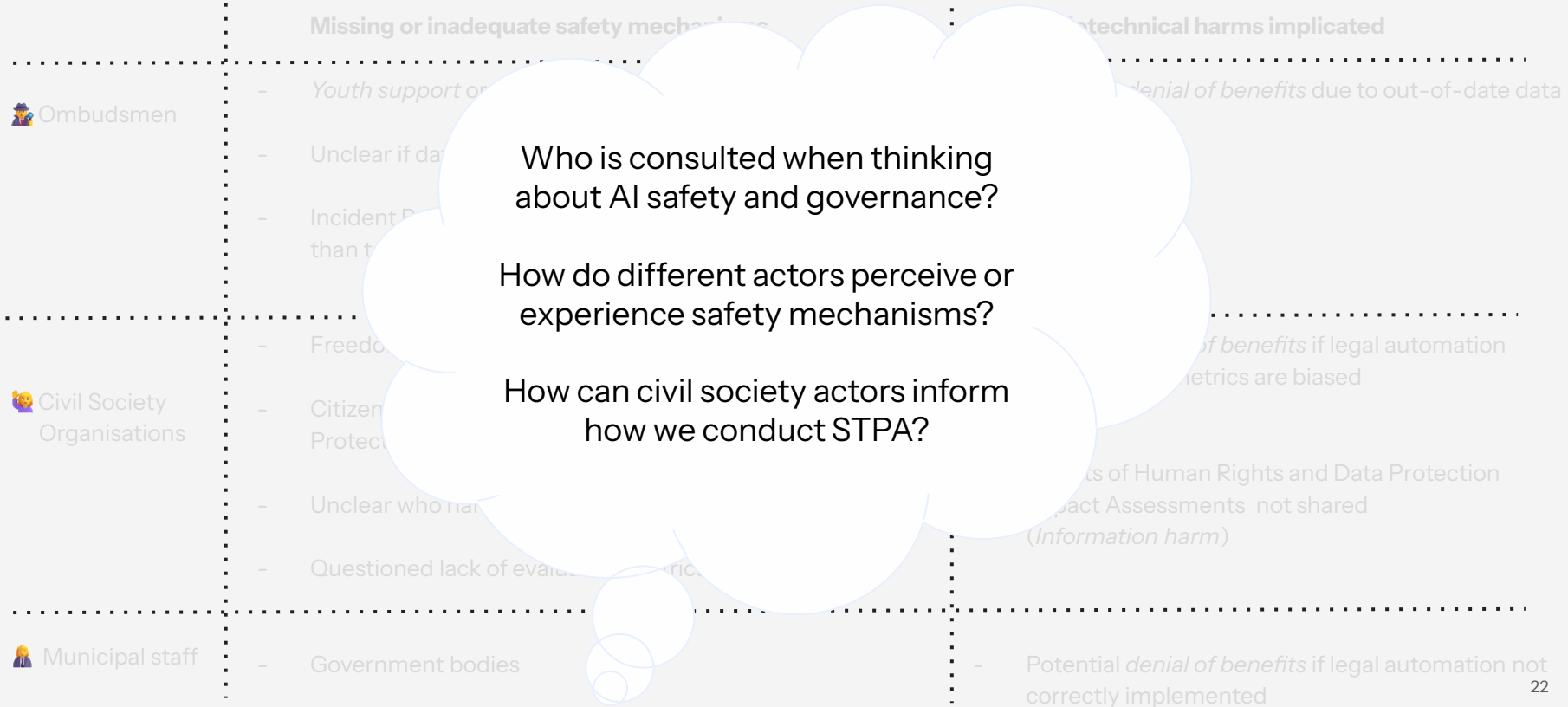
Insight #3:
System boundaries are normative and shape what we uncover

	Missing or inadequate safety mechanisms	**Sociotechnical harms implicated**
🤷‍♀️ Ombudsmen	- *Youth support* organisations - Unclear if data Avola uses is up-to-date (e.g. current salary) - Incident Reports sent to developer may be technical, rather than tell citizens' stories	- Potential *denial of benefits* due to out-of-date data
🙋 Civil Society Organisations	- Freedom of Information request is a high barrier for citizen - Citizens probably don't know what a Human Rights or Data Protection Impact Assessments are - Unclear who handles complaints, and what action is taken - Questioned lack of evaluation metrics	- Potential *denial of benefits* if legal automation performance metrics are biased - Results of Human Rights and Data Protection Impact Assessments not shared (*Information harm*)
👷 Municipal staff	- Government bodies - Unaware of evaluation metrics or "legal tests" performed	- Potential *denial of benefits* if legal automation not correctly implemented

Insight #3:
System boundaries are normative and shape what we uncover

	Missing or inadequate safety mechanisms	**Sociotechnical harms implicated**
🤷‍♀️ Ombudsmen	- *Youth support* organisations - Unclear if data Avola uses is up-to-date (e.g. current salary) - Incident Reports sent to developer may be technical, rather than tell citizens' stories	- Potential *denial of benefits* due to out-of-date data
🙋‍♀️ Civil Society Organisations	- Freedom of Information request is a high barrier for citizen - Citizens probably don't know what a Human Rights or Data Protection Impact Assessments are - Unclear who handles complaints, and what action is taken - Questioned lack of evaluation metrics	- Potential *denial of benefits* if legal automation performance metrics are biased - Results of Human Rights and Data Protection Impact Assessments not shared (*Information harm*)
👷 Municipal staff	- Government bodies - Unaware of evaluation metrics or "legal tests" performed	- Potential *denial of benefits* if legal automation not correctly implemented

Insight #3:
System boundaries are normative and shape what we uncover

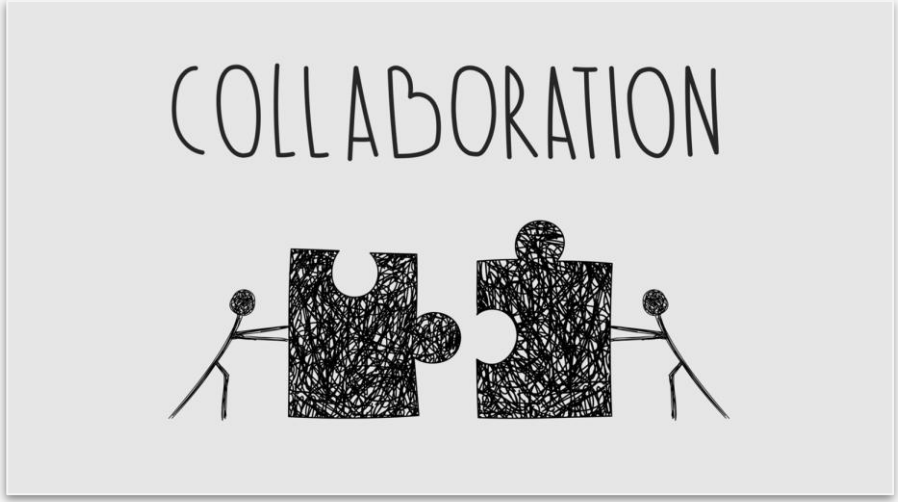


Discussion and call for collaboration

1. What are the best approaches to capture social, organizational and human relationships in STPA?

1. How can we account for the changing nature of AI systems with time (e.g., emergent capabilities, system updates) when applying STPA?

1. How can we show the efficacy of STAMP-based methods for addressing emerging safety concerns of AI systems?



Thank you

Contact us:
Shalaleh: shalaleh.rismani@mail.mcgill.ca
Maurus: M.E.Enbergs-1@tudelft.nl
Íñigo: I.M.D.R.deTroya@tudelft.nl

Relevant references

Dobbe, R. (2022). System safety and artificial intelligence. 2022 ACM Conference on Fairness Accountability and Transparency, 1584–1584. https://doi.org/10.1145/3531146.3533215

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 24, 723–741. https://doi.org/10.1145/3600211.3604673

Rismani, S., Shelby, R., Smart, A., Jatho, E., Kroll, J., Moon, A., & Rostamzadeh, N. (2023). From plane crashes to algorithmic harm: Applicability of safety engineering frameworks for responsible ML. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2, 1–18. https://doi.org/10.1145/3544548.3581407

Rismani, S., Shelby, R., Smart, A., Delos Santos, R., Moon, A., & Rostamzadeh, N. (2023). Beyond the ML model: Applying safety engineering frameworks to text-to-image development. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2, 70–83. https://doi.org/10.1145/3600211.3604685

Dobbe, R., & Wolters, A. (2024). Toward sociotechnical AI: Mapping vulnerabilities for machine learning in context. Minds and Machines, 34(2), 12. https://doi.org/10.1007/s11023-024-09668-y

[In review] Rismani, S., Dobbe, R., & Moon, A. (2024). From silos to systems: Process-oriented Hazard Analysis for AI systems. ArXiv, abs/2410.22526. arXiv. https://doi.org/10.48550/arXiv.2410.22526

de Troya, Í., Kernahan, J., Doorn, N., Dignum, V., & Dobbe, R. (2025). Misabstraction in Sociotechnical Systems. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 1829–1842. https://doi.org/10.1145/3715275.3732122

Backup:

STPA vs CAST

CAST to identify causal factors of accident
STPA is a system hazard
→ First, we wanted to illustrated the value of STAMP for public algorithmic systems, hence forward looking analysis better suited to audience
Second, we did not want to focus on causal factors that may have contributed to the specific accident, but rather identify all systemic hazards present in the Tax authorities system at the time. Even those not directly related to the accidents.

Backup:

ID		Hazard	Losses
[SH-1]		**Process & Model misclassifies citizens**	**L1**
	Organisation		
	[SH-1.1]	Insufficient, opaque and slow process comunication between subsystems	
	[SH-1.2]	Information from citizens is shared prior to signal verification	
	[SH-1.3]	Insufficient holistic process measuarbility and visability	
	[SH-1.4]	Diverging Mental models of process	
	[SH-1.5]	Diverging data registration process for differen organisations silos regarding FSV system	
	Data		
	[SH-1.6]	Insufficient data quality in data registration systems	
	[SH-1.7]	Bureaucatic, pungent/strict, & suspicious culture towards subjects	
	[SH-1.8]	Designd data structure in data registration systems are inadequate	
	Model		
	[SH-1.9]	Opaque Risk model structures	
	[SH-1.10]	No logging functions tracking decisions (For risk models and FSV)	
	[SH-1.11]	Different understanding of model objectives between stakeholders DAT (Accuracy), Toeslagen (Specificity)	
	[SH-1.12]	Lack of model causality	
	[SH-1.13]	Unfairnes and Biases in ML/AI pipline	
[SH-2]		**No citizen support process**	**L1, L2**
	[SH-2.1]	Bureaucatic, pungent/strict, & suspicious culture towards subjects	
	[SH-2.2]	No logging functions tracking decision making	
[SH-3]		**Data protection standards for third party information sharing are low**	**L1, L2**
	[SH-3.1]	Toeslagen shares data to thrid party government organisation, pior to signal verification	
	[SH-3.2]	Insufficient data quality in data registration systems	
[SH-4]		**No Recovery protocolls for process/system failure in place**	**L1, L3, L4**
	[SH-4.1]	Ridgied process and system structure	
	[SH-4.2]	Insufficient, opaque and slow process comunication and visability between subsystems	
	[SH-4.3]	MKB is buisness owner of the FSV system	
[SH-5]		**Insufficent system correction/improvment**	**L3, L4**
	[SH-5.1]	Insufficient, slow and opaque feedback to development	
	[SH-5.2]	Insufficient development resources for system correction and improvement	
	[SH-5.3]	Long change request/feedback process	
	[SH-5.4]	MKB is buisness owner of the FSV system	
[SH-6]		**Operators are unaware of current system state in realtion to safety**	**L1, L2, L3**
	[SH-6.1]	Key process stages are missing feedback loops from down stream work organisations	
	[SH-6.2]	Control actions to measure system state are missing at Key process stages	
	[SH-6.3]	Reports about system state are miss interpreted	