

---

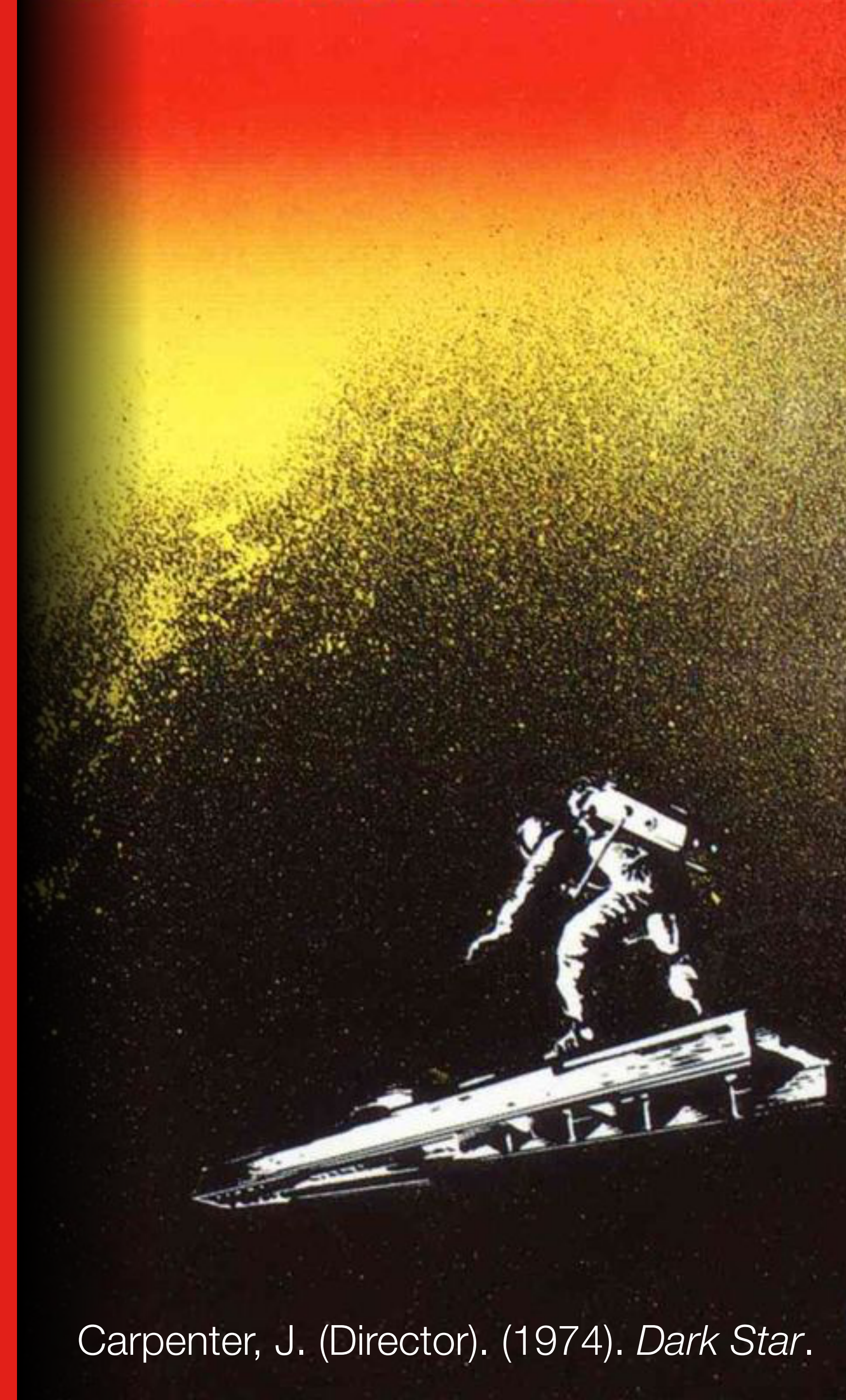
# Meaningful Human Control Through Contestability by Design

---

Kars Alfrink  
TU Delft  
[contestable.ai](https://contestable.ai)

AiTech Agora  
2 March 2022

---



Carpenter, J. (Director). (1974). *Dark Star*.



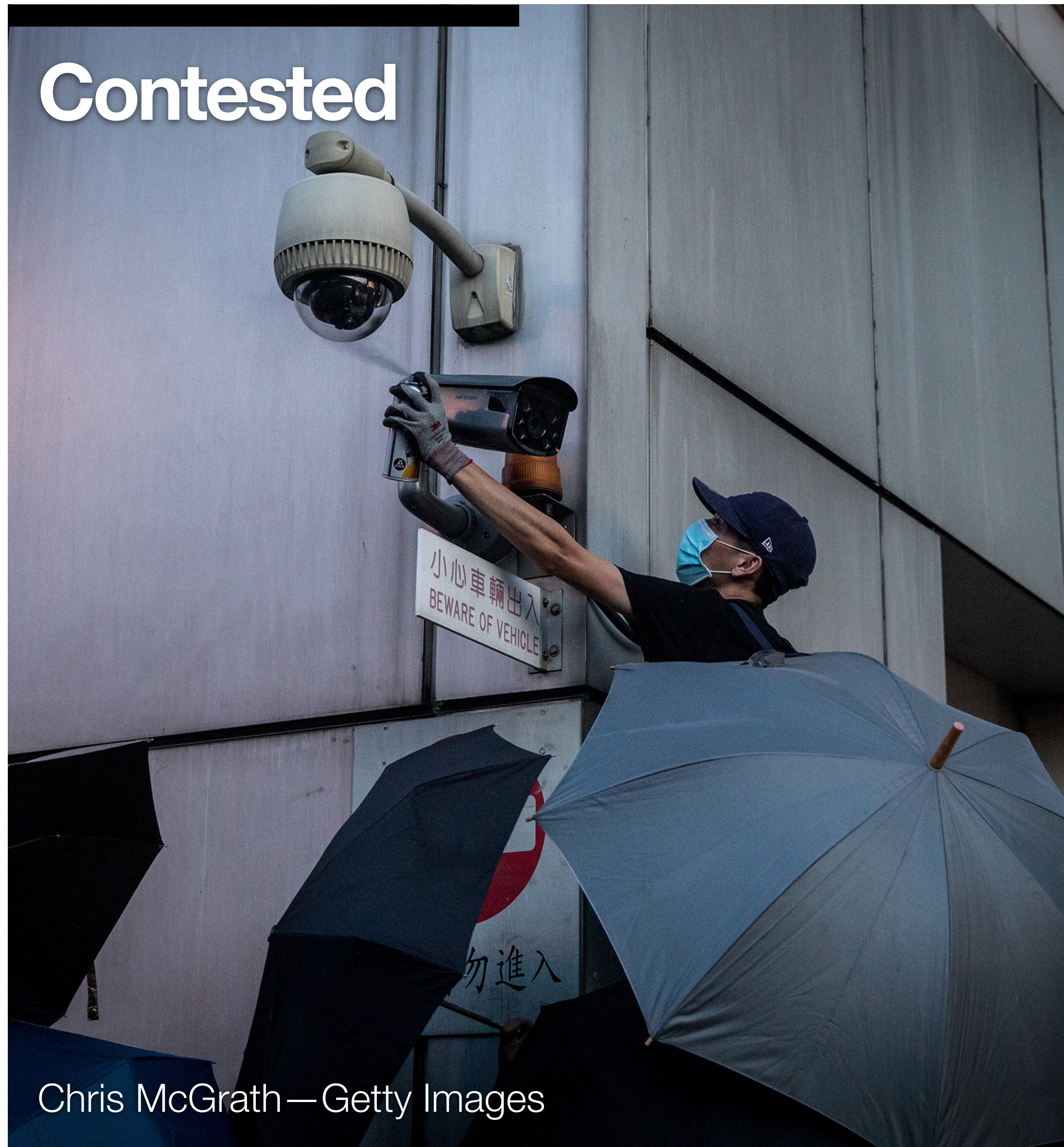
# Transparent AI

Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (in press). *Tensions in Transparent Urban AI: Designing A Smart Electric Vehicle Charge Point*. *AI & Society*.





# Contested



Chris McGrath—Getty Images

# Contestable

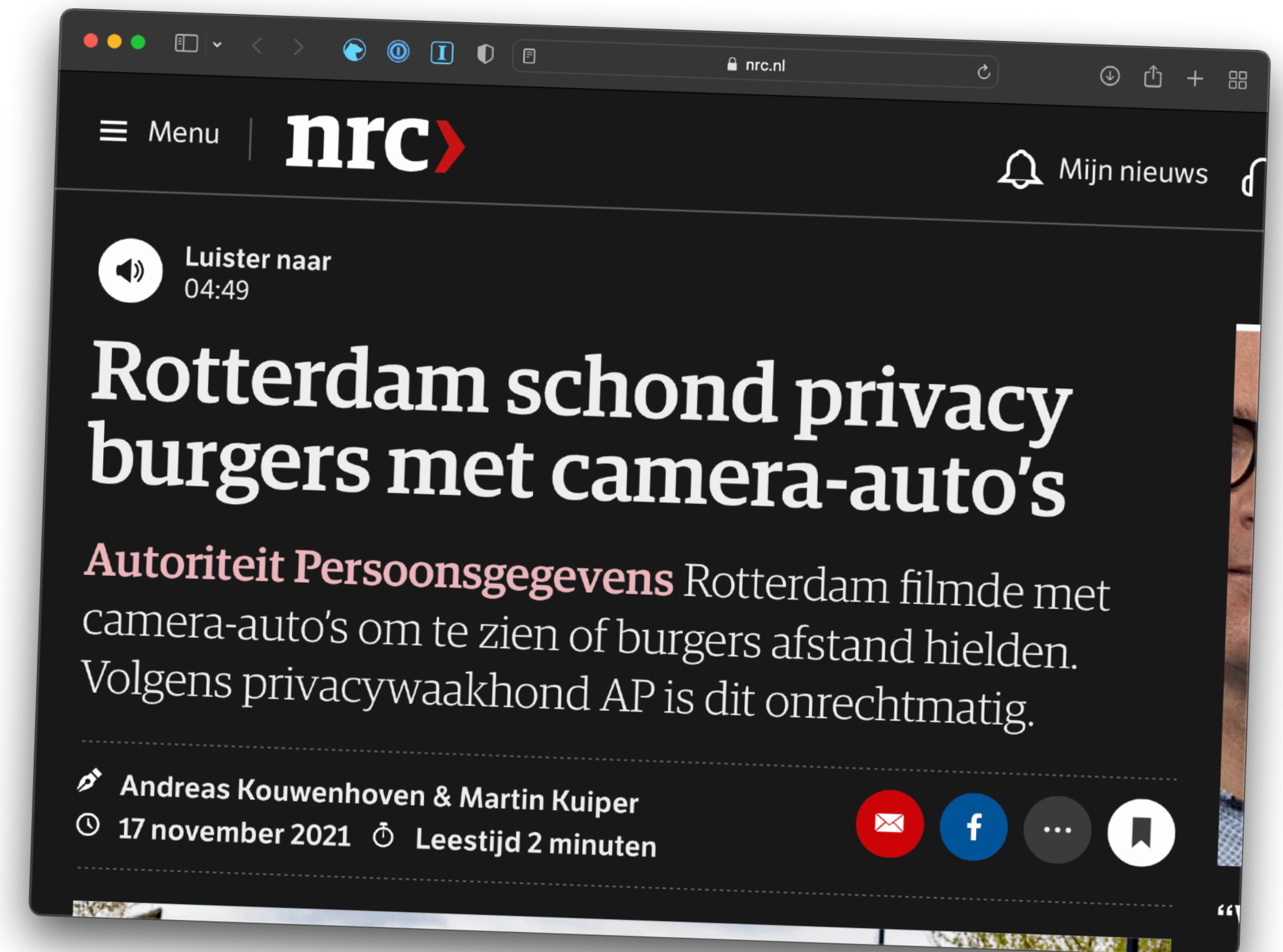


Responsible Sensing Lab



# Problem

**Automated decision-making** can harm people's basic human rights to **autonomy** and **dignity**.





---

# Solutions

**Ex-ante:** Participatory &  
value-sensitive design

**Ex-post:** Explanations,  
appeal procedures



---

**“When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming.”**

Collingridge, D. (1980). The social control of technology.



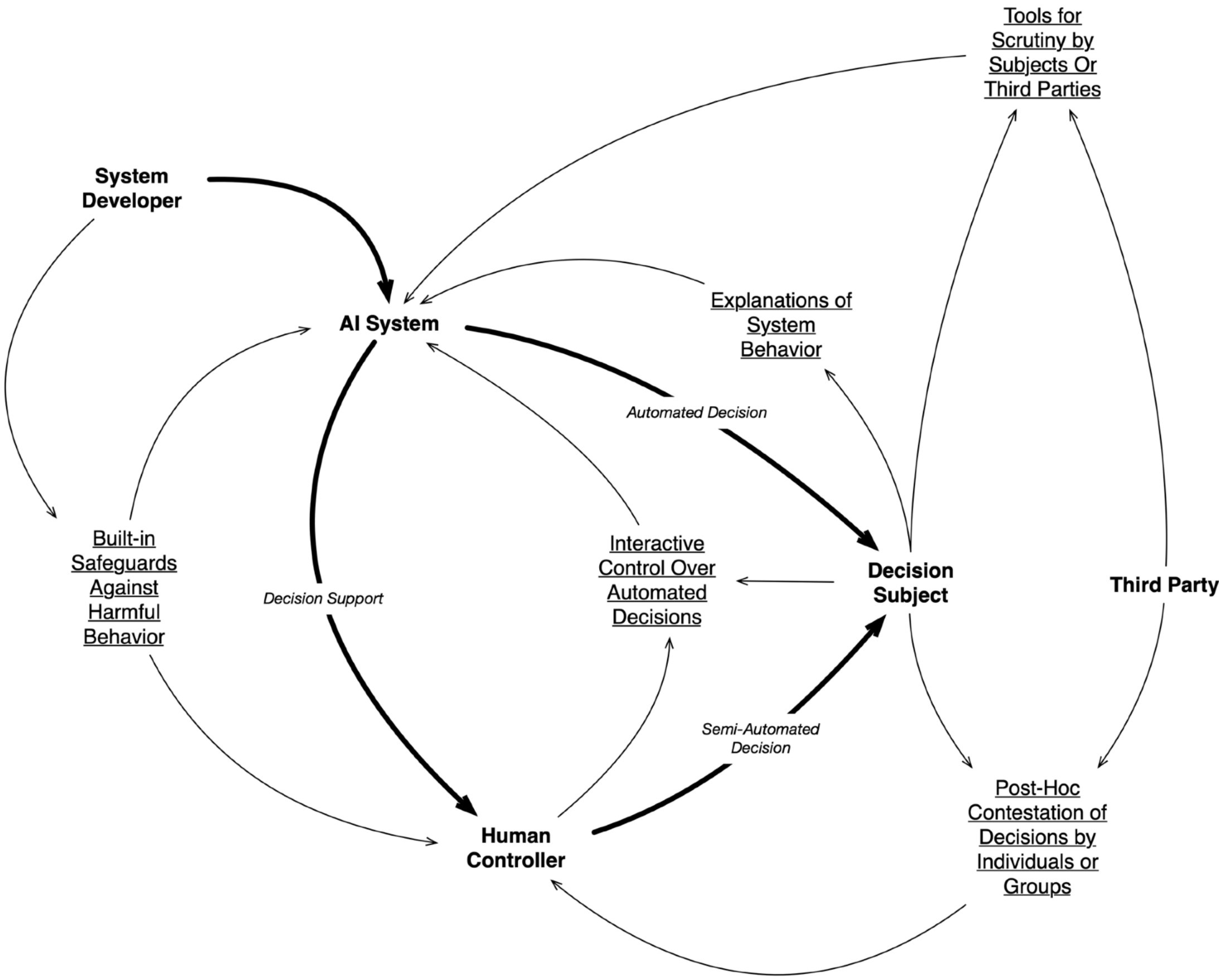
---

## Contestable AI

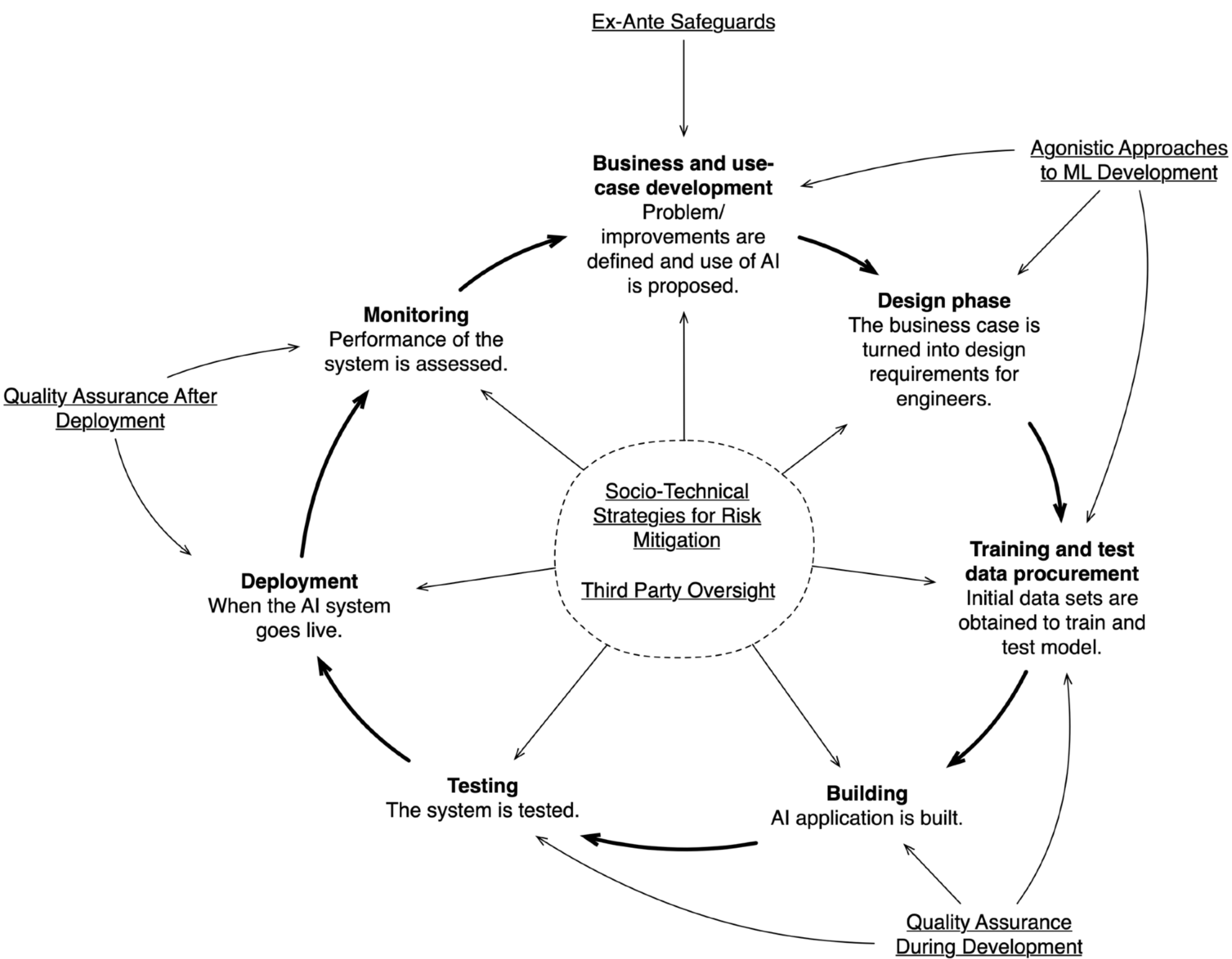
**AI that is open and responsive to human intervention, throughout a system's lifecycle, establishing a procedural relationship between decision subjects and system controllers.**



# Features



# Practices



Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2021). **Contestable AI by Design: Towards A Framework**. Manuscript Submitted for Publication.



---

## FEATURES

Built-in safeguards

**second adversarial system**

Interactive control

**negotiate, correct or override automated decision • feedback loop  
back to training • supplement local contextual data**

Explanations

**behavioral model • sandboxing approaches • model inversion •  
ambiguity awareness**

Intervention requests

**post-hoc contestation • comparative measures • organizational room  
for receiving, evaluating and responding to disputes • shifting  
burdens on individuals • enabling collective action • dialectical  
exchange**

Tools for scrutiny

**documentation of development process • documentation of technical  
composition • performance indicators • opaque assurances**



---

## **PRACTICES**

Ex-ante safeguards

**acceptance criteria • anticipation • certification**

Agonistic dev approaches

**co-construct decision process • participatory design**

QA during development

**living labs • iterative development**

QA after development

**monitoring for bias and misuse • feedback from corrections, appeals & additions**

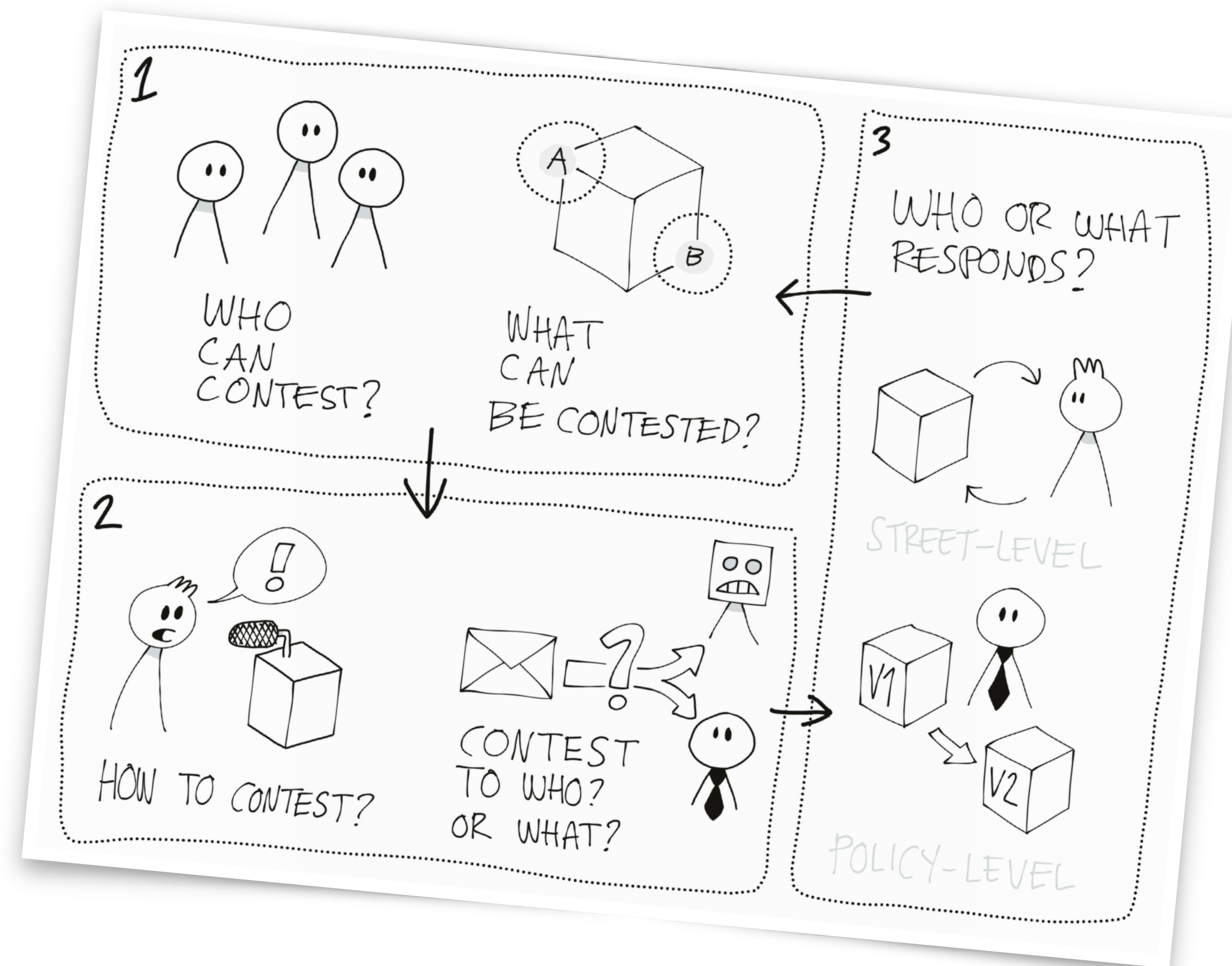
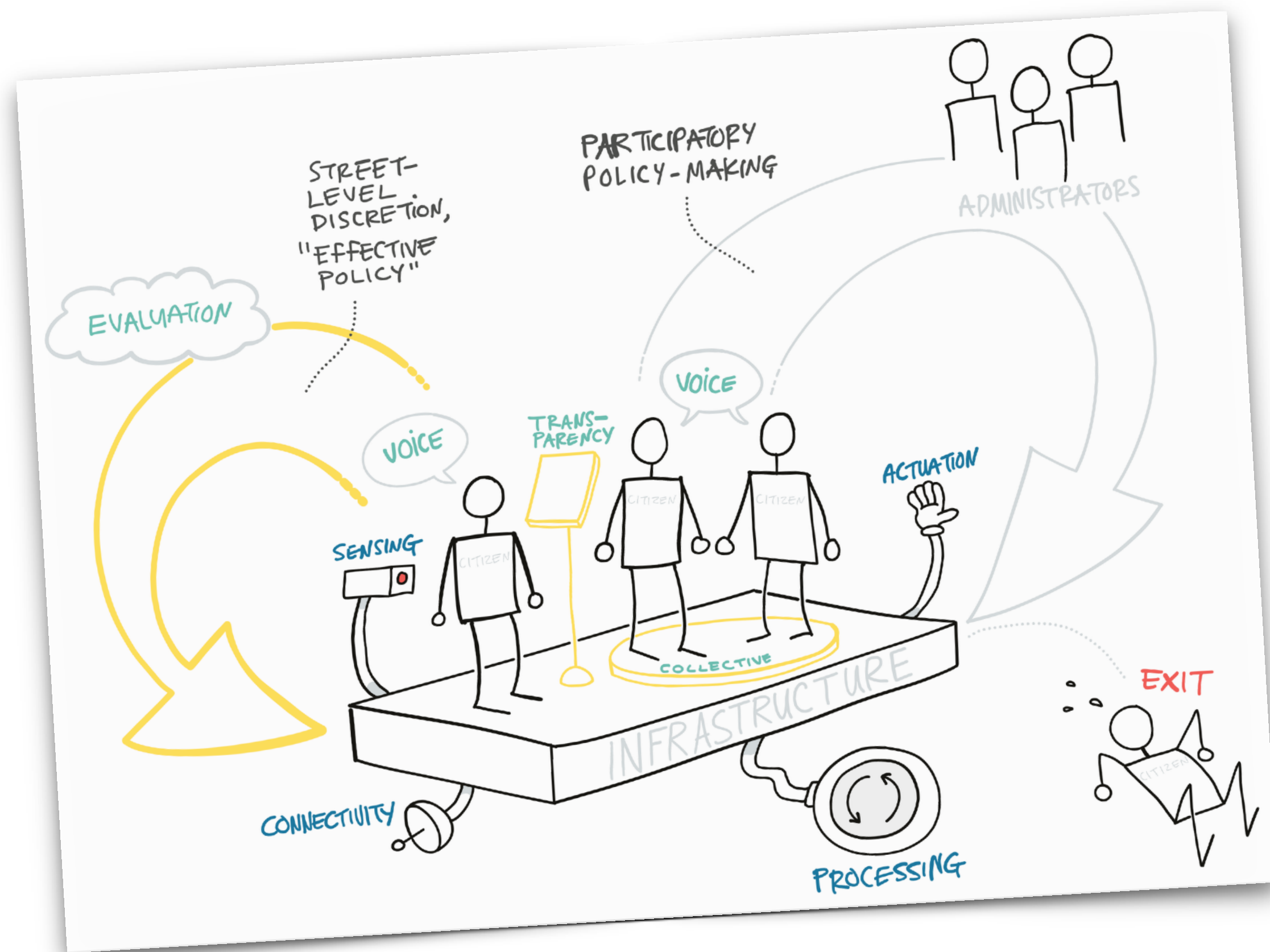
Risk mitigation

**environmental protections • user education**

Third party oversight

**trusted 3rd parties • secure environments • representing individuals and groups**







---

Comparison to previous conceptual explorations

- 1. Individual vs collective action**
- 2. Human discretion**
- 3. Software development as policy-making**



---

Work in progress

# Contestable Camera Cars

---






A speculative design  
exploration of public AI  
systems that are responsive to  
value change.

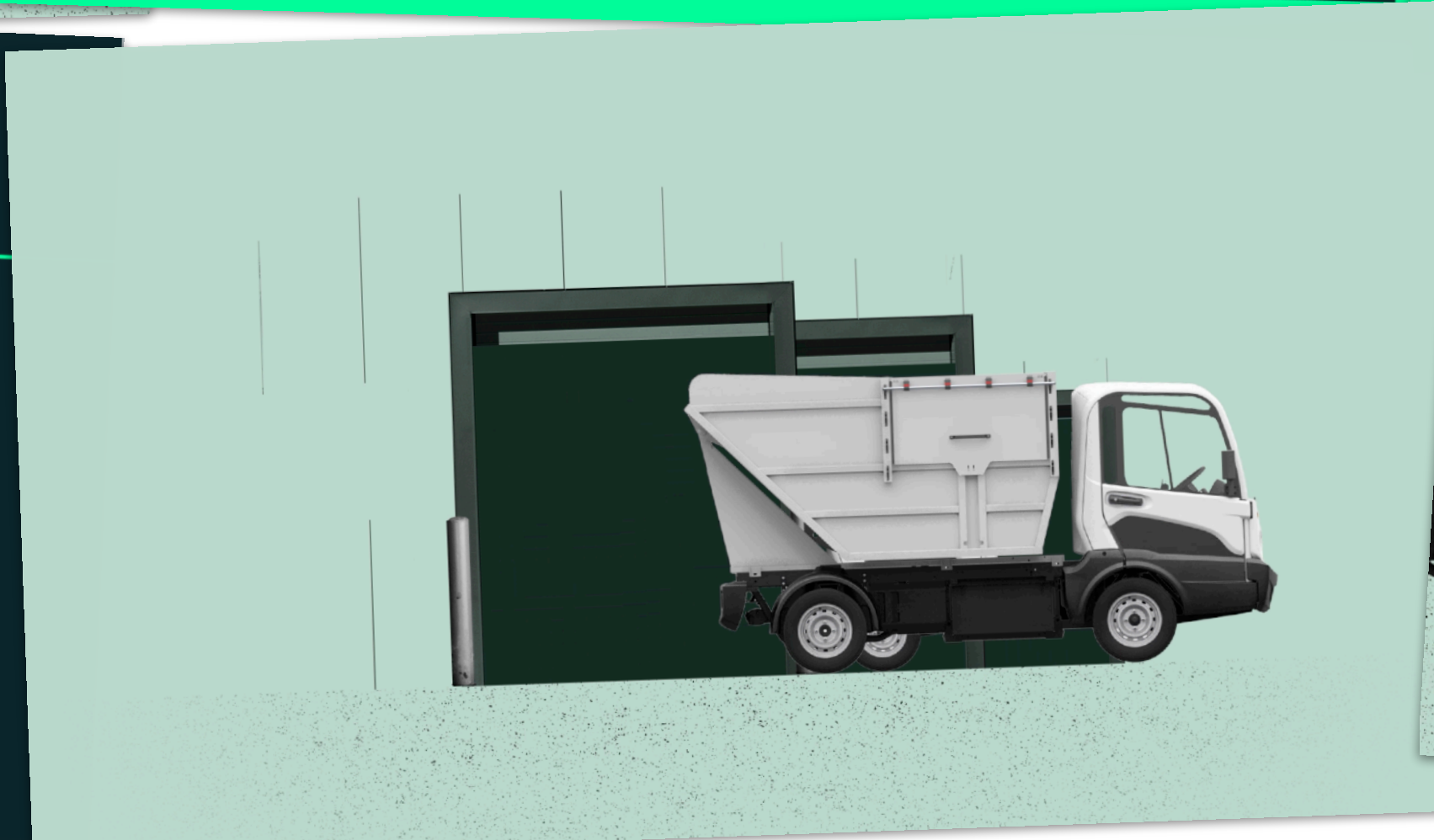
---







vehicle	destination
 truck 03	→ location a
 truck 12	→ location b
 truck 17	→ location a
 truck 09	..
 truck 02	..









---

# Meaningful Human Control

**Tracking:** Contestability as a mechanism for adapting to value change

**Tracing:** Monitoring and *literal* traceability as leverage points for contestability



---

## Takeaways

**Avoid resolving disputes up front** using compromise or consensus-seeking.

**Set up procedural, agonistic mechanisms** to identify and respond to them.

**Leverage conflict for continuous improvement.**



---

# Thank you!

---

**Kars Alfrink**  
**TU Delft**  
**[contestable.ai](https://contestable.ai)**

---

