Contestable AI by Design: Towards a Framework

Kars Alfrink
TU Delft

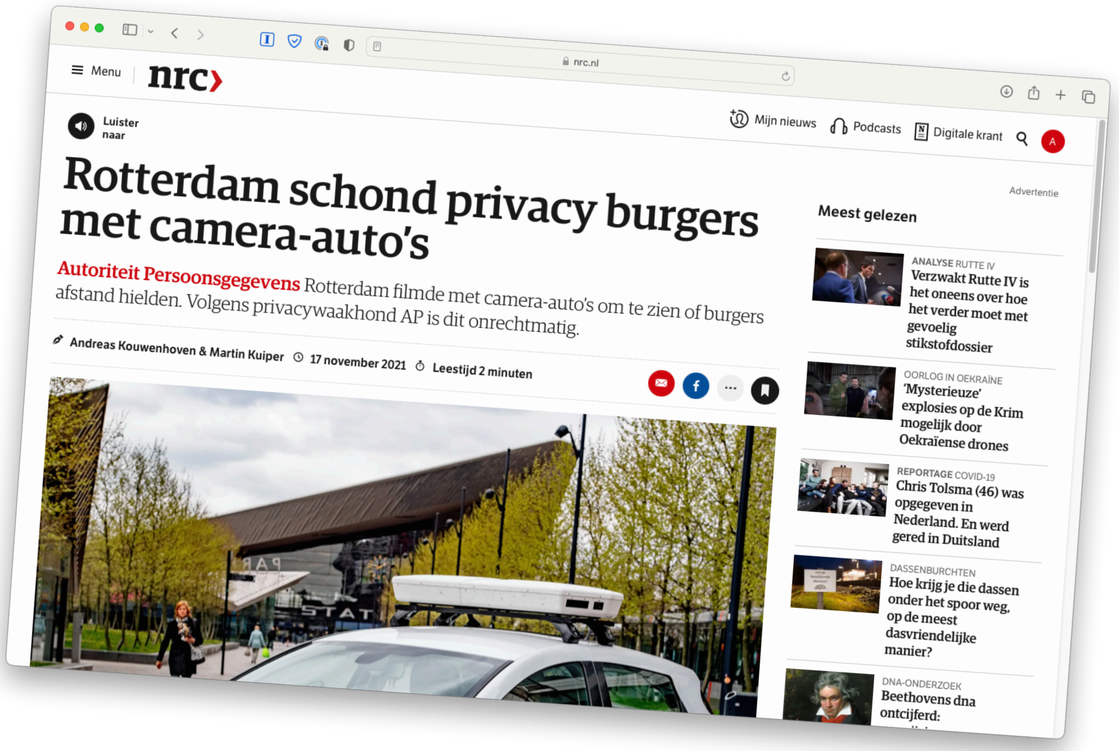
ICT.OPEN
20 April 2023

www.contestable.ai

TU Delft | Knowledge & Intelligence Design | chi nl

Credit: AlgorithmWatch

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



Contestable AI

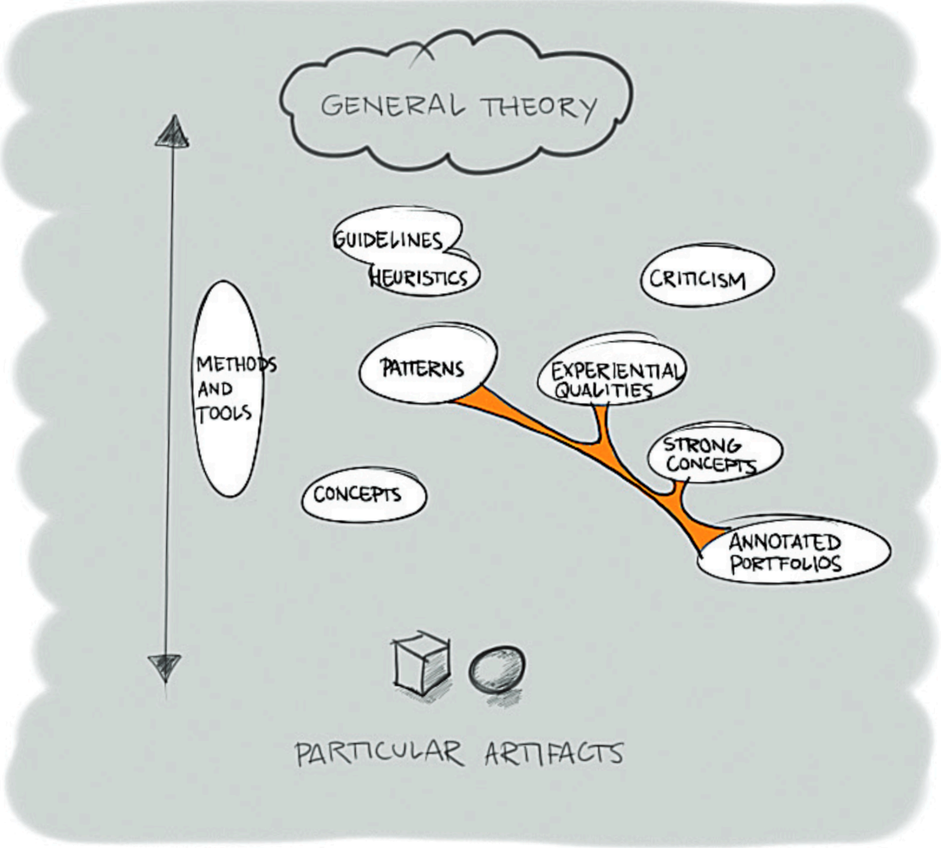
AI that is open and responsive to dispute, throughout the system lifecycle, establishing a dialectical relationship between decision subjects and system operators.

- Humans challenging machine predictions (Hirsch et al., 2017)
- Deep system property (Vaccaro et al., 2019)
- Human intervention, contestability by design (Almada, 2019)
- Procedural relationship (Sarra, 2020)
- Justifications (Henin & Le Métayer, 2021)

From principles to practices
(Morley et al., 2019)

Intermediate-level generative design knowledge
(Höök & Löwgren, 2012; Löwgren et al., 2013)

Design frameworks
(Obrenovic, 2011)

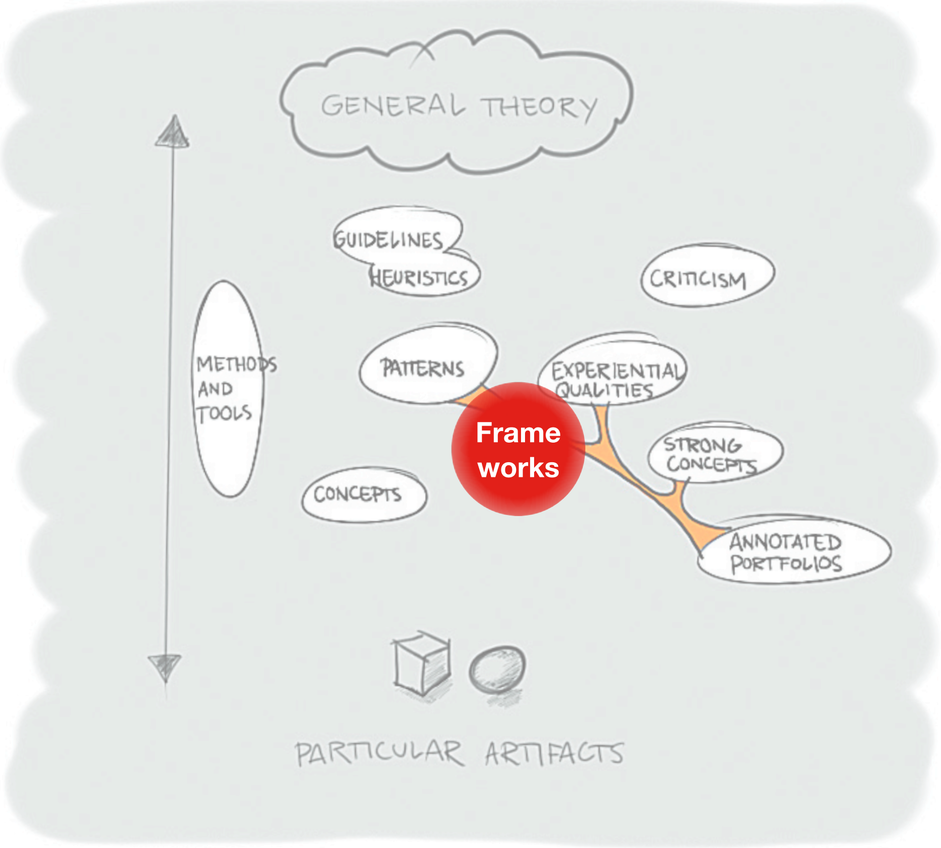


Examples of approaches to intermediate-level interaction design knowledge (Löwgren et al., 2013).

From principles to practices
(Morley et al., 2019)

Intermediate-level generative design knowledge
(Höök & Löwgren, 2012; Löwgren et al., 2013)

Design frameworks
(Obrenovic, 2011)



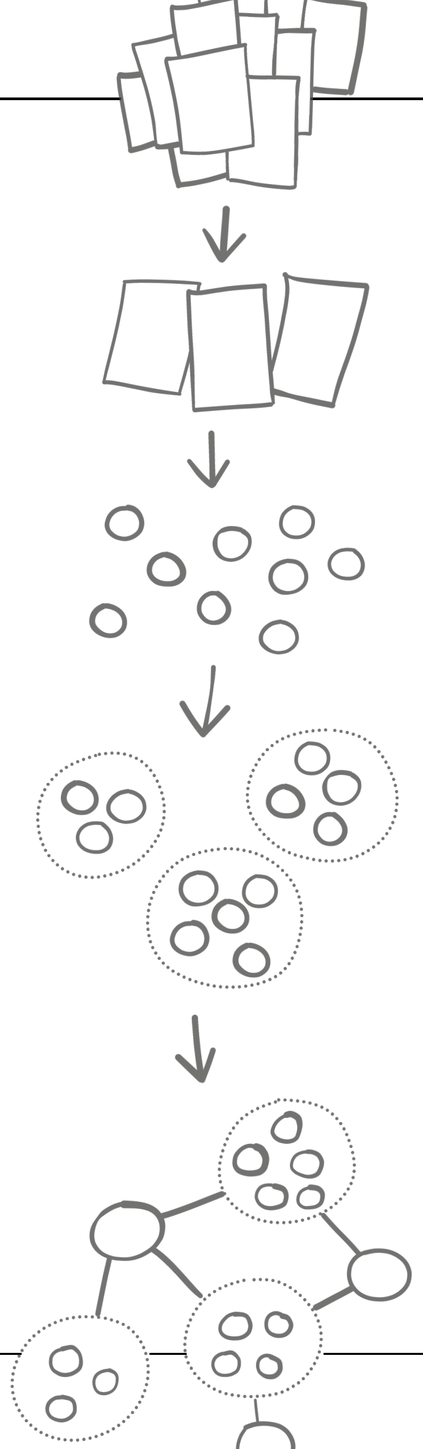
Examples of approaches to intermediate-level interaction design knowledge (Löwgren et al., 2013).

Method

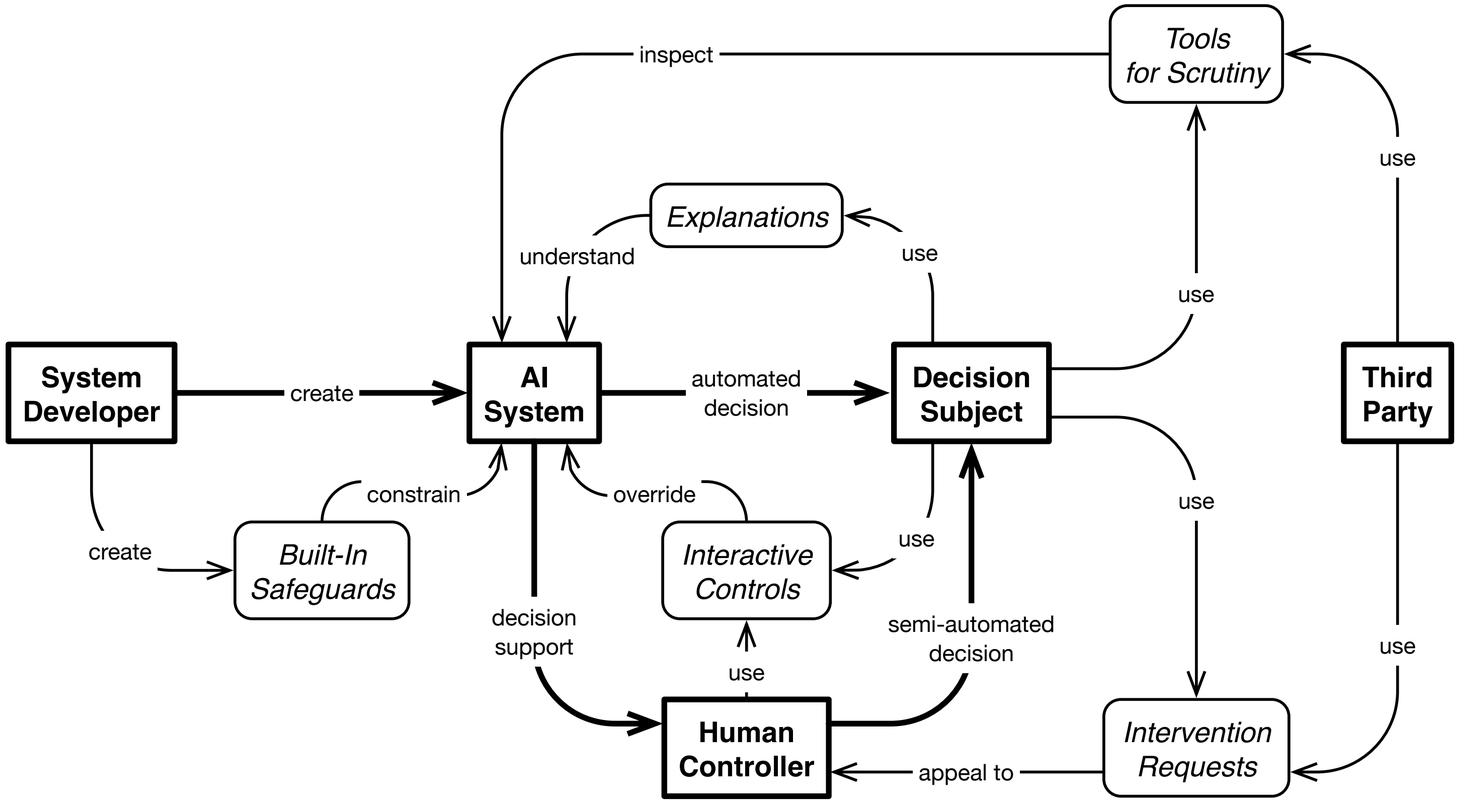
Systematic literature review:
AI, design
& contestability.

Reflexive thematic analysis:
“active ingredients.”

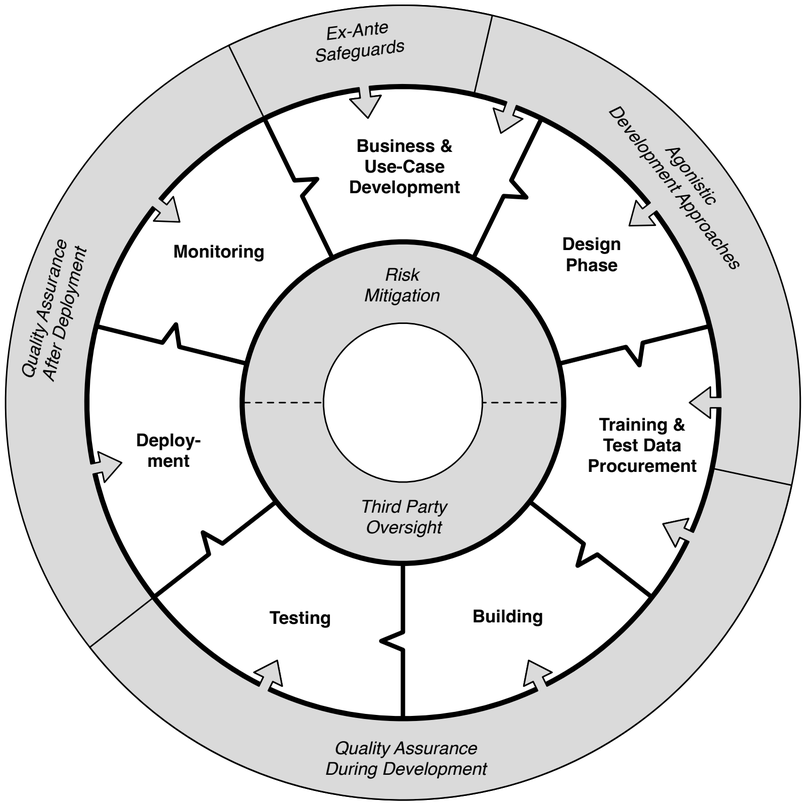
Visual mapping techniques:
lifecycle stages & actors.



Features



Practices



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

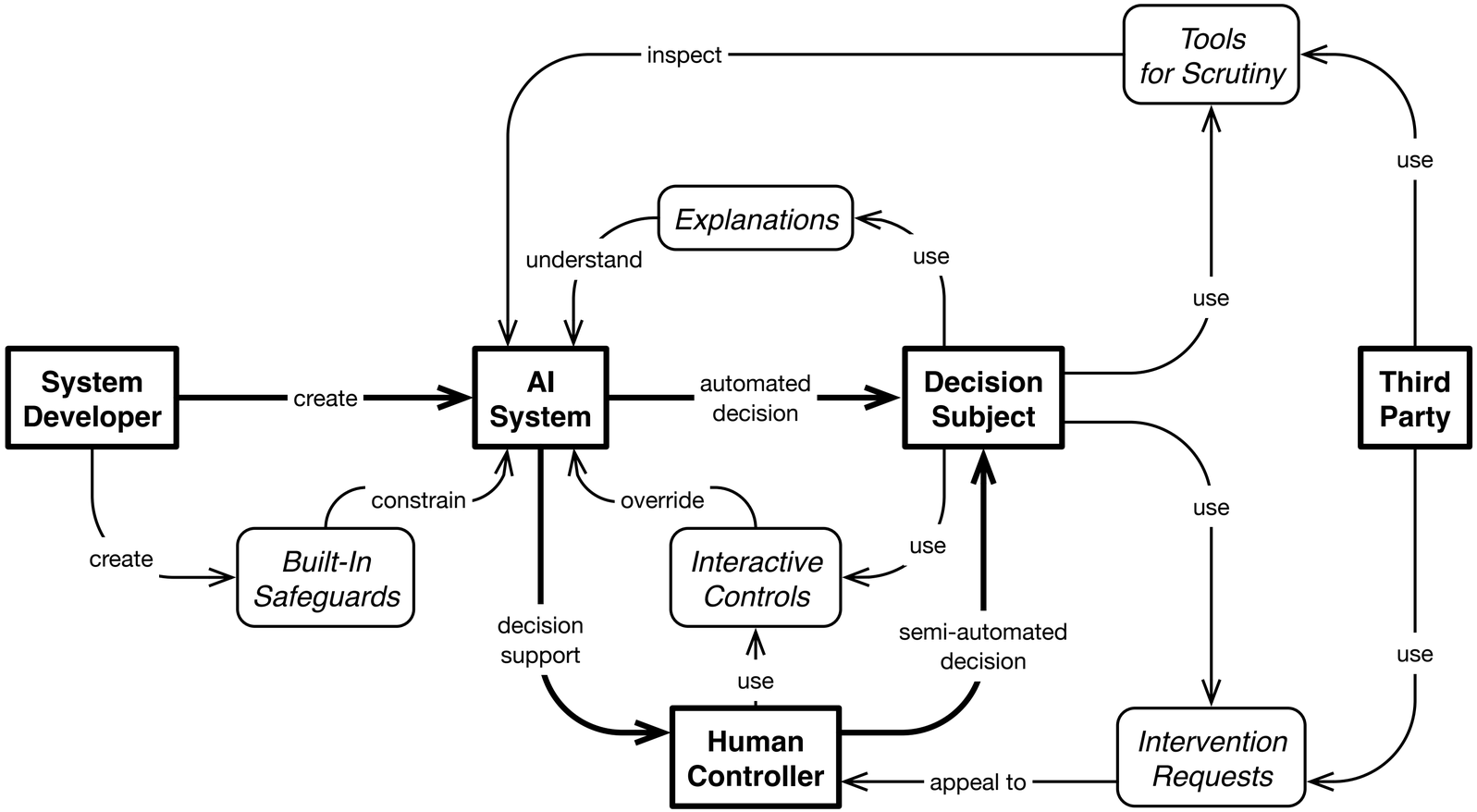
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

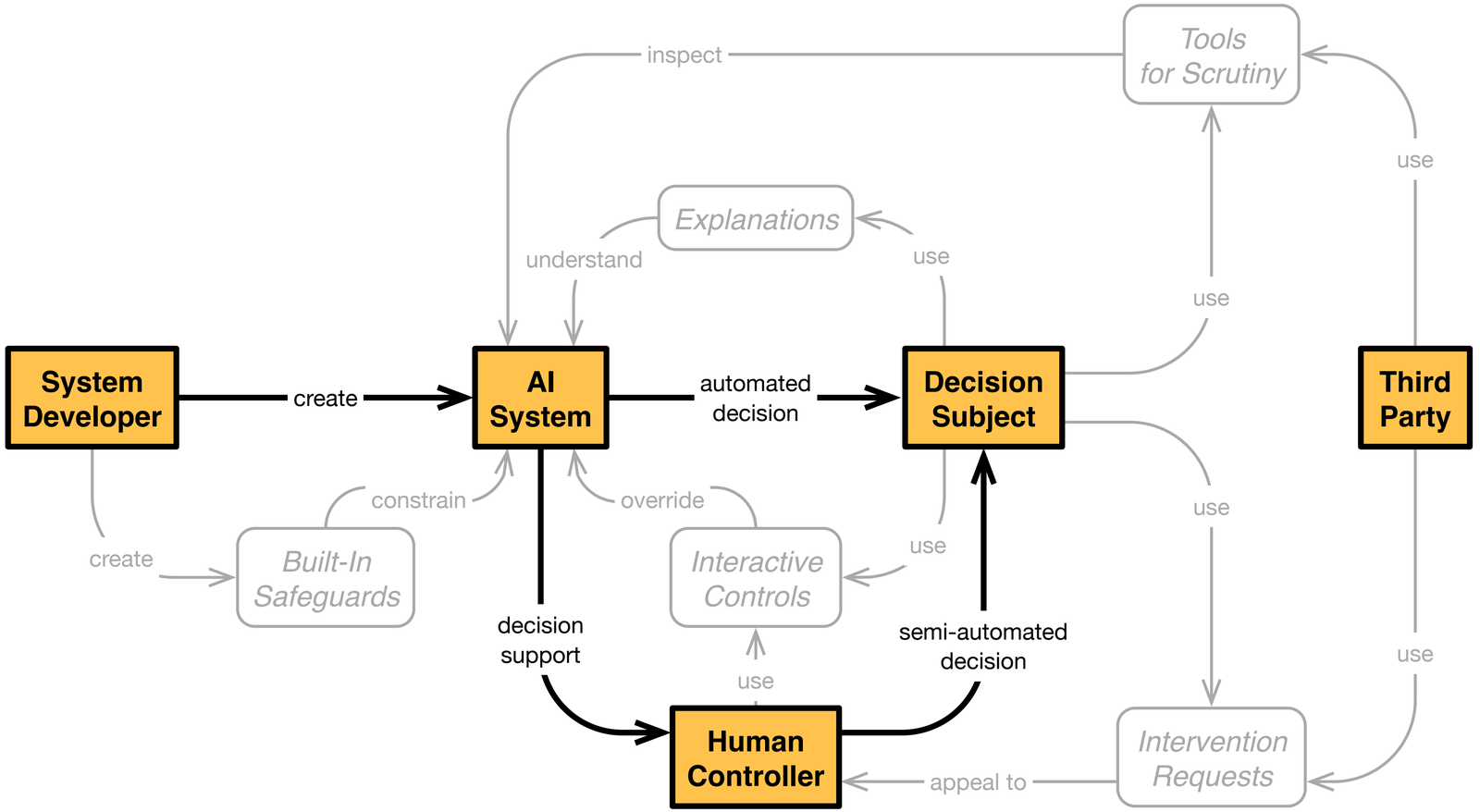
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

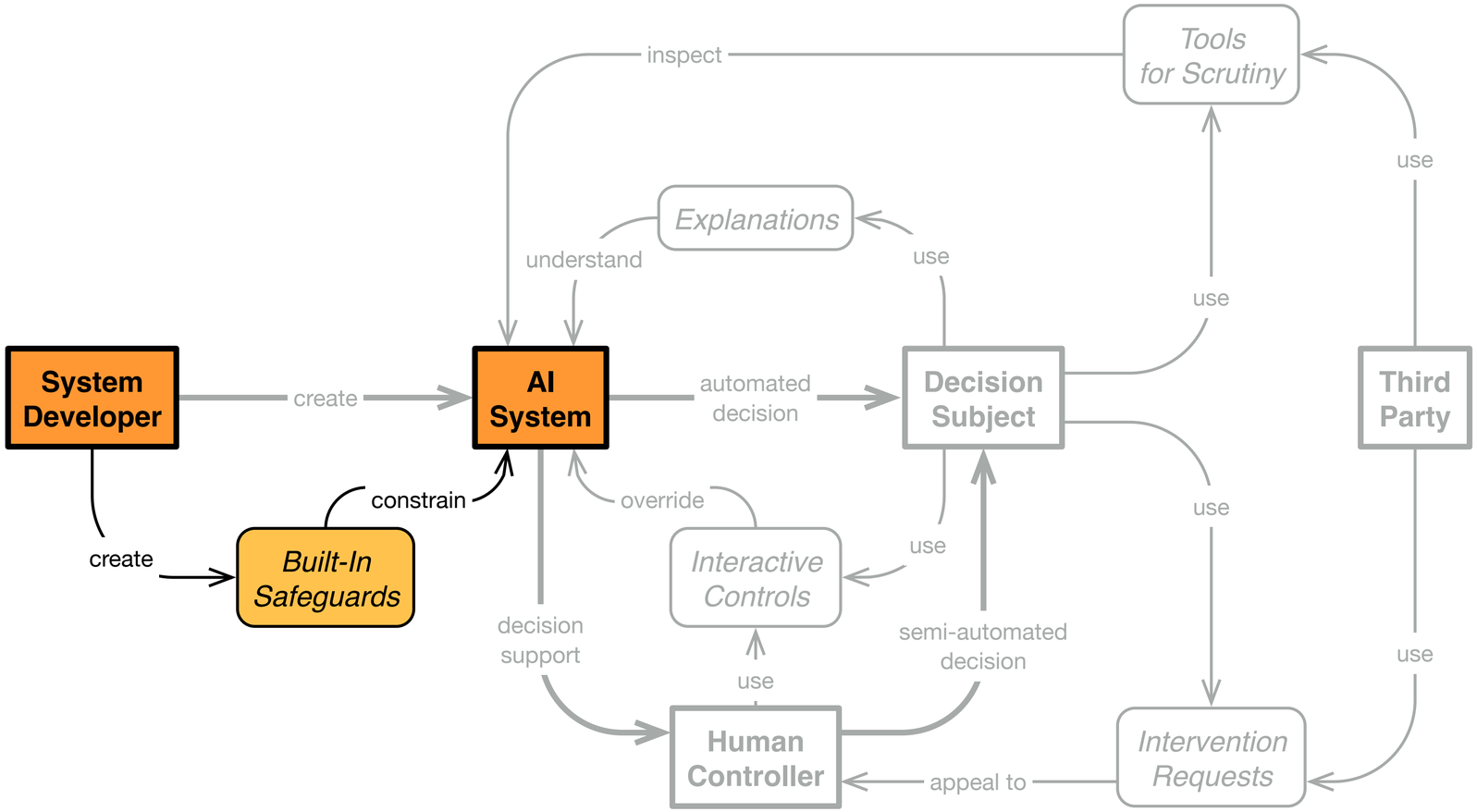
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

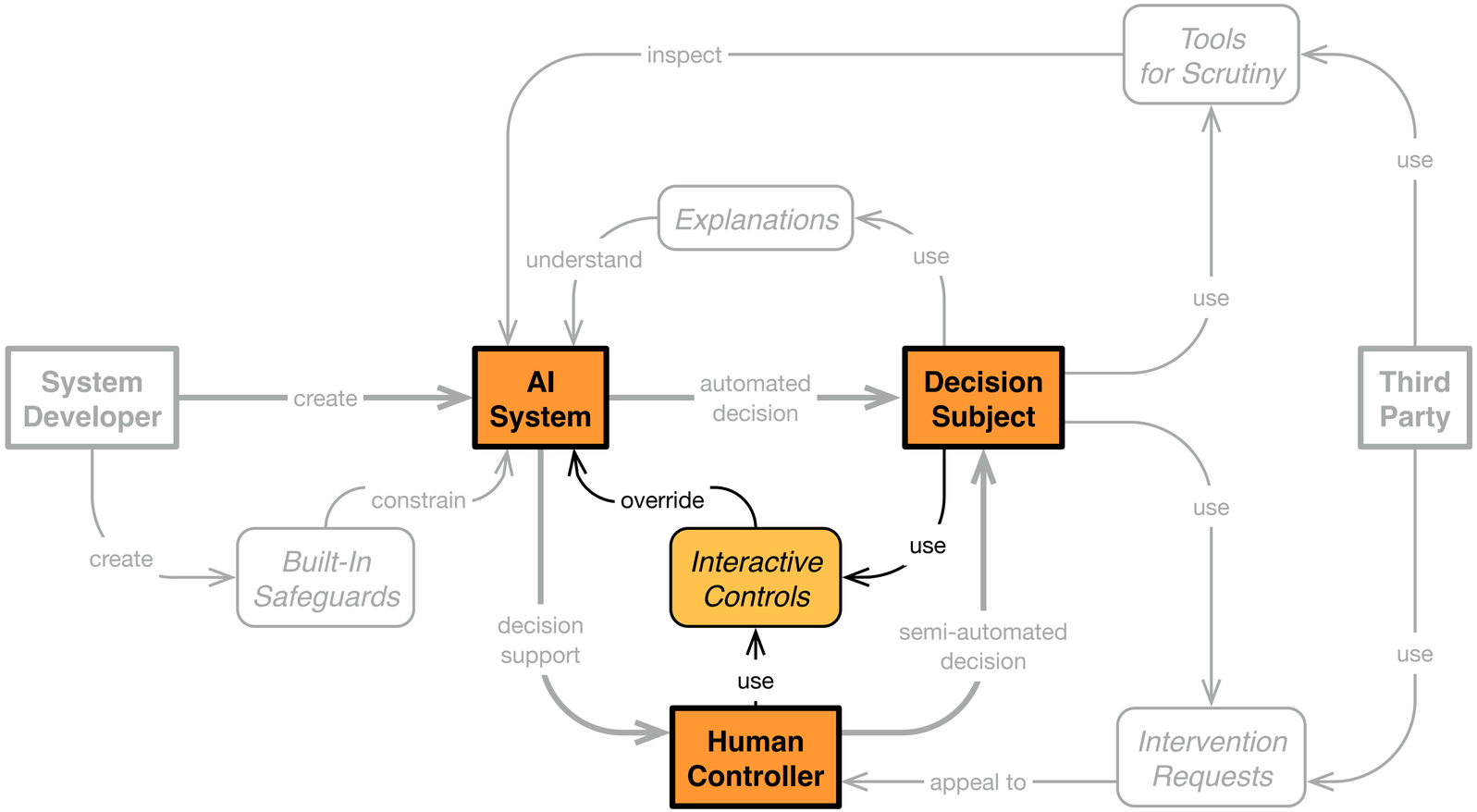
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

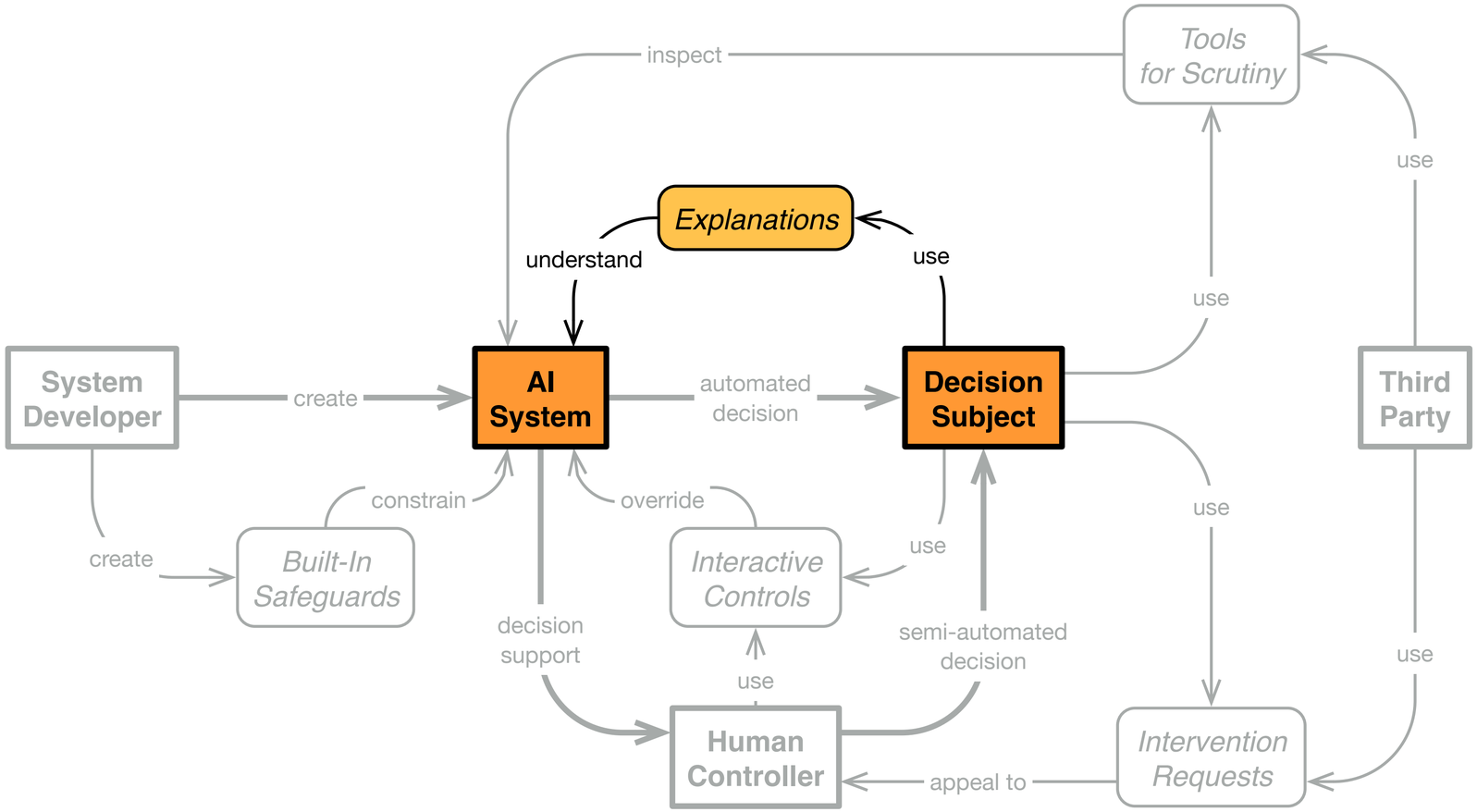
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

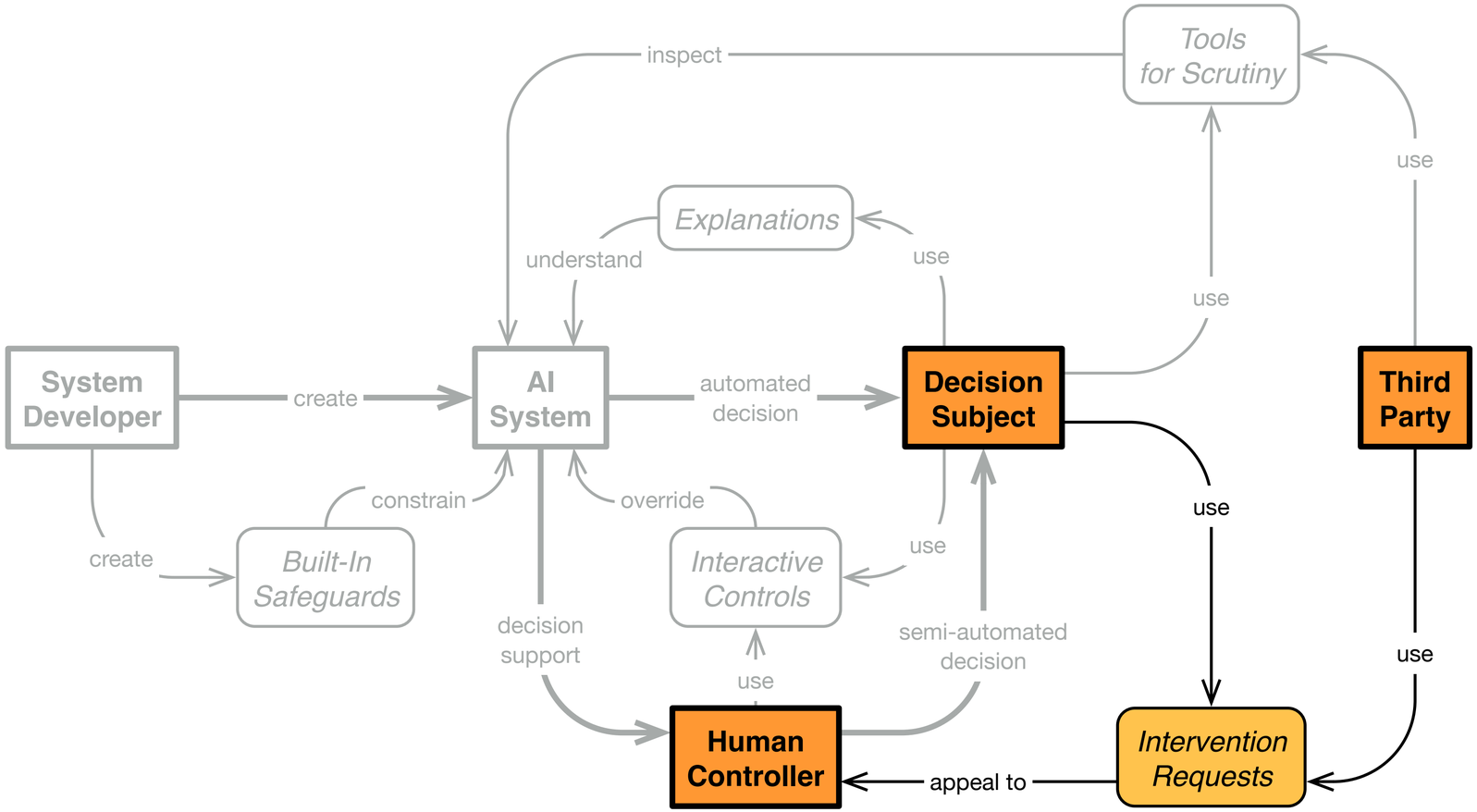
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Features

Built-in safeguards

External adversarial system • Formal constraints

Interactive controls

Negotiate, correct, or override machine decision • Feedback loop back to training • Supplement local contextual data

Explanations

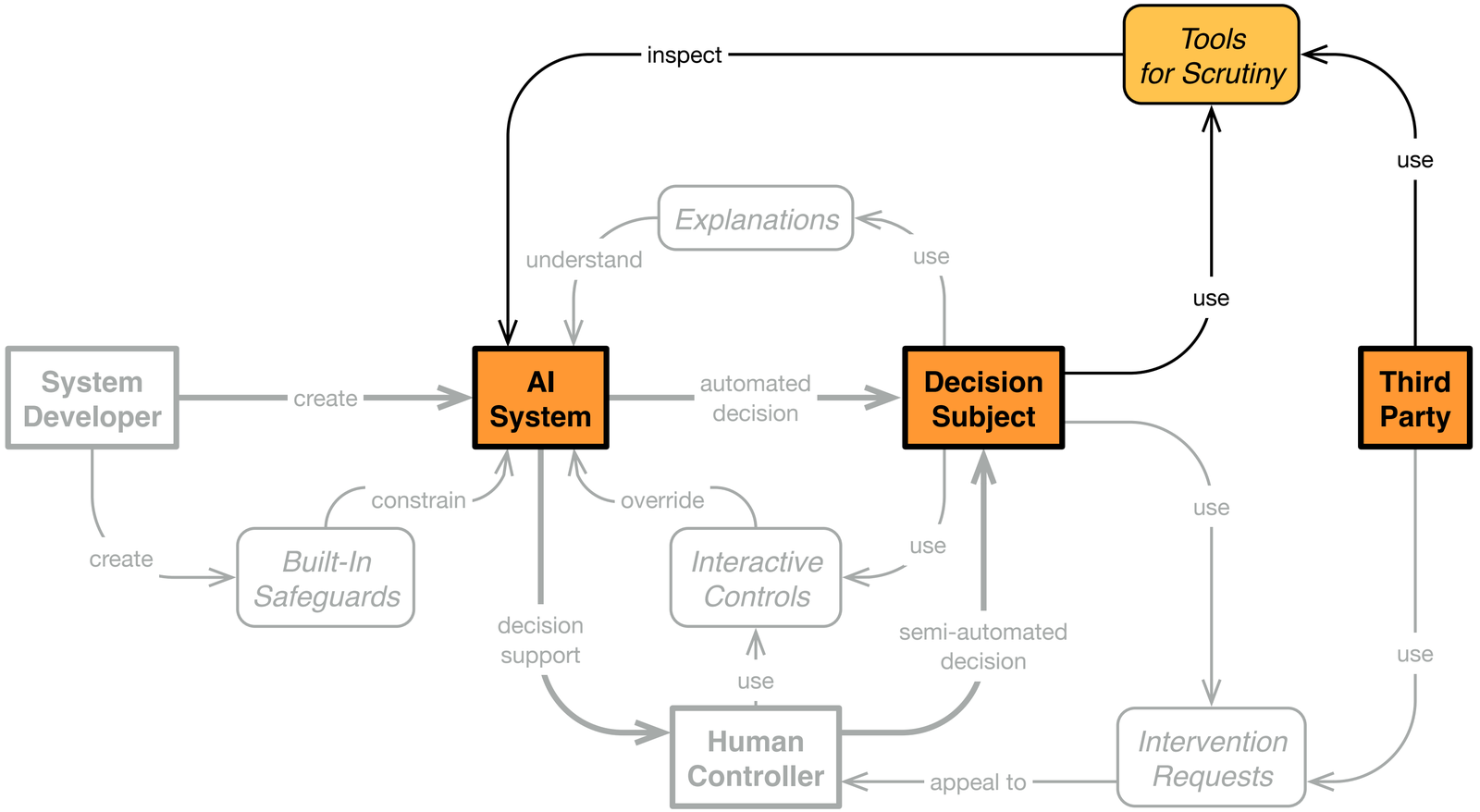
Traceable decision chains • Behavioral explanations • Sandboxing • Local approximations • Justifications

Intervention requests

Human review • Supportive, synchronous channels • Third party representation • Collective action • Dialectical exchange

Tools for scrutiny

Norms linked to implementation • Documentation • Formal proofs • Comparative measures • Opaque assurances



Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

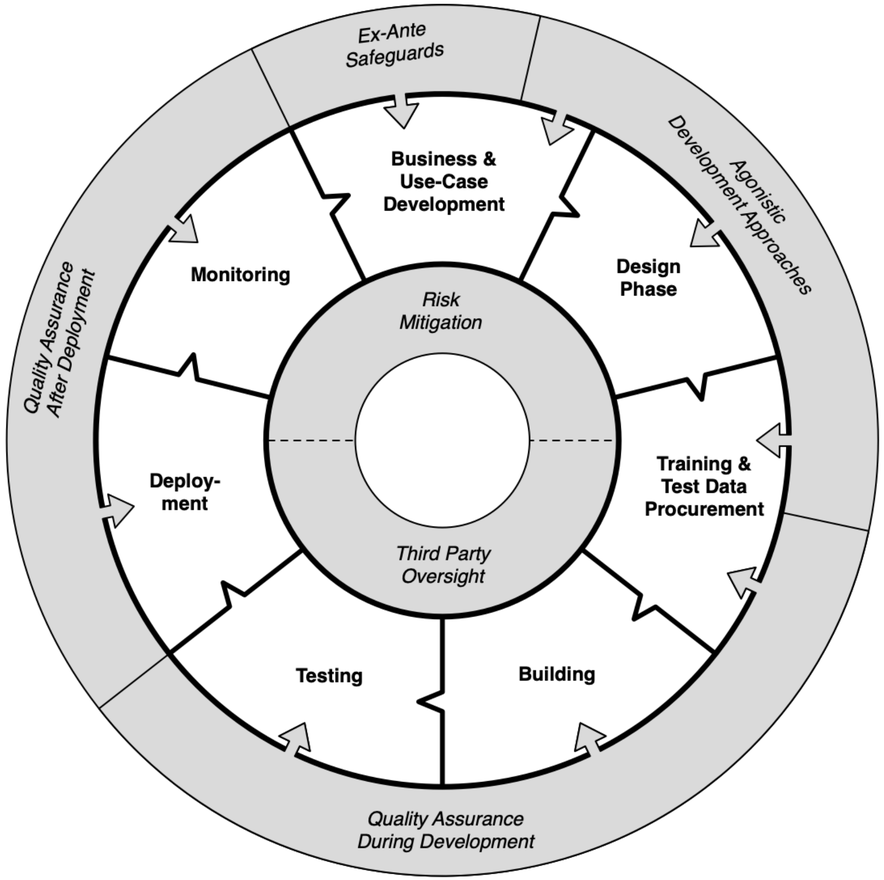
Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments



Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Practices

Ex-ante safeguards

Anticipating impacts • Acceptance criteria • Certification

Agonistic dev approaches

Co-construct decision-making process • Ongoing adversarial dialogue

QA measures during dev

Stakeholder needs guiding development • Bias prevention • Living labs • Stakeholder feedback

QA measures after deploy

Procedural integrity • Monitoring for bias, misuse • Feedback from corrections, appeals and additional contextual info

Risk mitigation

User education • Environmental limits

Third party oversight

Model-centric tools for auditing • Trusted intermediaries • Secure environments

Limitations

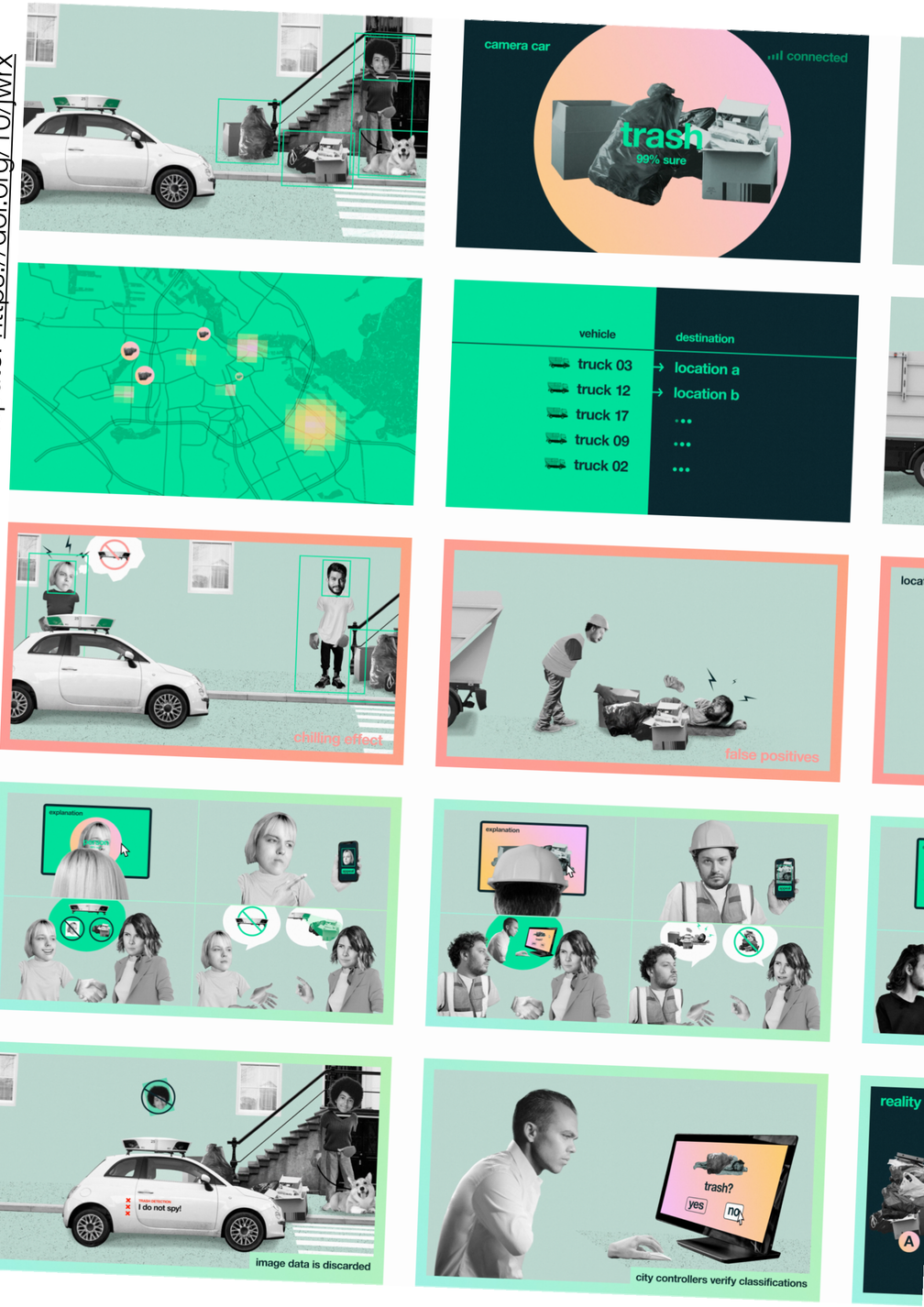
- **Lack of context**
- **Majority theoretical**
- **Validation pending**

Transferability

- **Level of impact**
- **Time-sensitivity**

Future work

- **Directions for use**
- **Example outcomes**



Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2023). Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. https://doi.org/10/jwrx

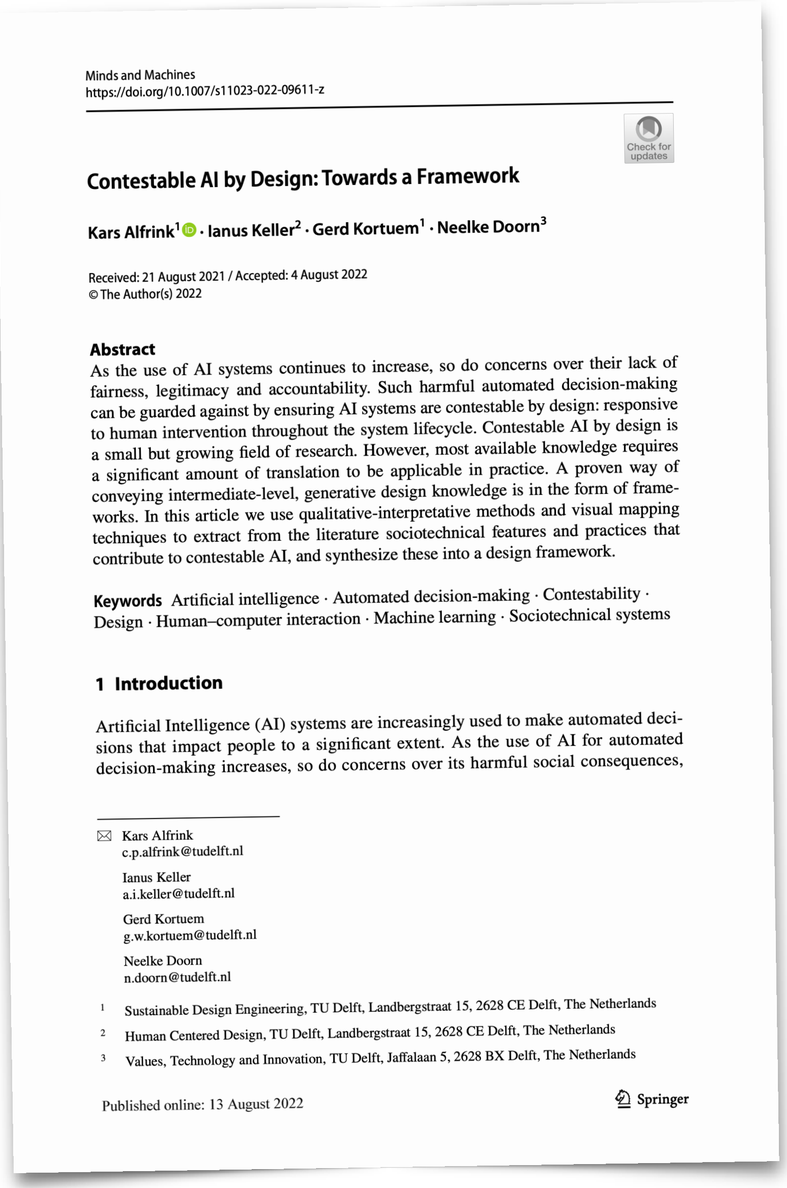
Avoid resolving disputes up-front at all costs

Controversy is at times inevitable

Agree on procedures for disagreement resolution

“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.



Minds and Machines
https://doi.org/10.1007/s11023-022-09611-z

Check for updates

Contestable AI by Design: Towards a Framework

Kars Alfrink[1] · Ianus Keller[2] · Gerd Kortuem[1] · Neelke Doorn[3]

Received: 21 August 2021 / Accepted: 4 August 2022
© The Author(s) 2022

Abstract
As the use of AI systems continues to increase, so do concerns over their lack of fairness, legitimacy and accountability. Such harmful automated decision-making can be guarded against by ensuring AI systems are contestable by design: responsive to human intervention throughout the system lifecycle. Contestable AI by design is a small but growing field of research. However, most available knowledge requires a significant amount of translation to be applicable in practice. A proven way of conveying intermediate-level, generative design knowledge is in the form of frameworks. In this article we use qualitative-interpretative methods and visual mapping techniques to extract from the literature sociotechnical features and practices that contribute to contestable AI, and synthesize these into a design framework.

Keywords Artificial intelligence · Automated decision-making · Contestability · Design · Human–computer interaction · Machine learning · Sociotechnical systems

1 Introduction

Artificial Intelligence (AI) systems are increasingly used to make automated decisions that impact people to a significant extent. As the use of AI for automated decision-making increases, so do concerns over its harmful social consequences,

✉ Kars Alfrink
c.p.alfrink@tudelft.nl

Ianus Keller
a.i.keller@tudelft.nl

Gerd Kortuem
g.w.kortuem@tudelft.nl

Neelke Doorn
n.doorn@tudelft.nl

[1] Sustainable Design Engineering, TU Delft, Landbergstraat 15, 2628 CE Delft, The Netherlands

[2] Human Centered Design, TU Delft, Landbergstraat 15, 2628 CE Delft, The Netherlands

[3] Values, Technology and Innovation, TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 13 August 2022

Springer

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by Design: Towards a Framework. Minds and Machines. https://doi.org/10/gqnjcs

edu.nl/963n7

Contestable AI by Design: Towards a Framework

Kars Alfrink
TU Delft

ICT.OPEN
19-20 April 2023

www.contestable.ai