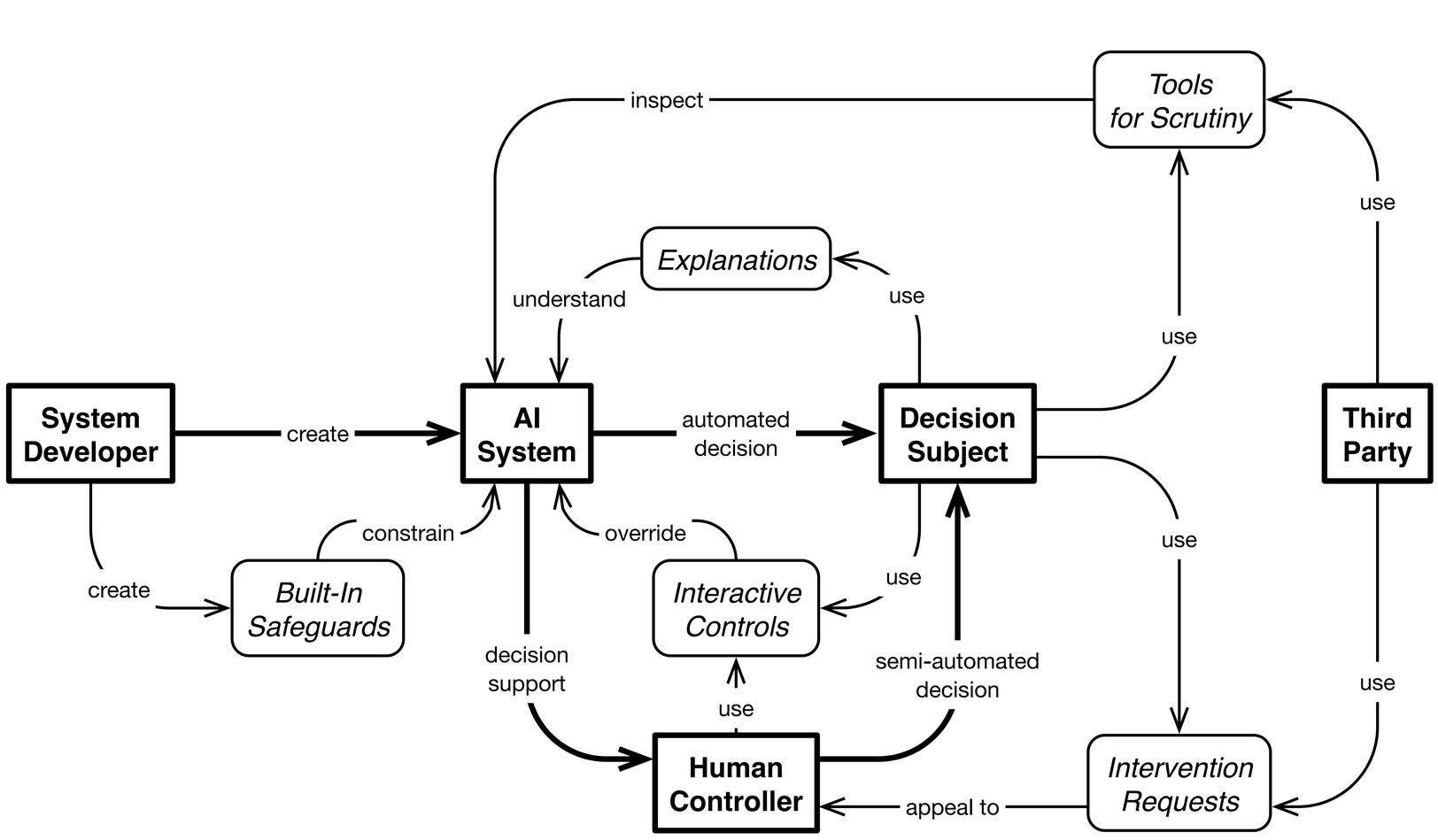
Contestable AI by Design

To ensure public artificial intelligence systems are responsive to human values, they must be **contestable by design**.

– open and responsive to human intervention
– throughout the whole system lifecycle
– a procedural relationship between decision subjects and system controllers
– leveraging disagreement for continuous improvement

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

Features

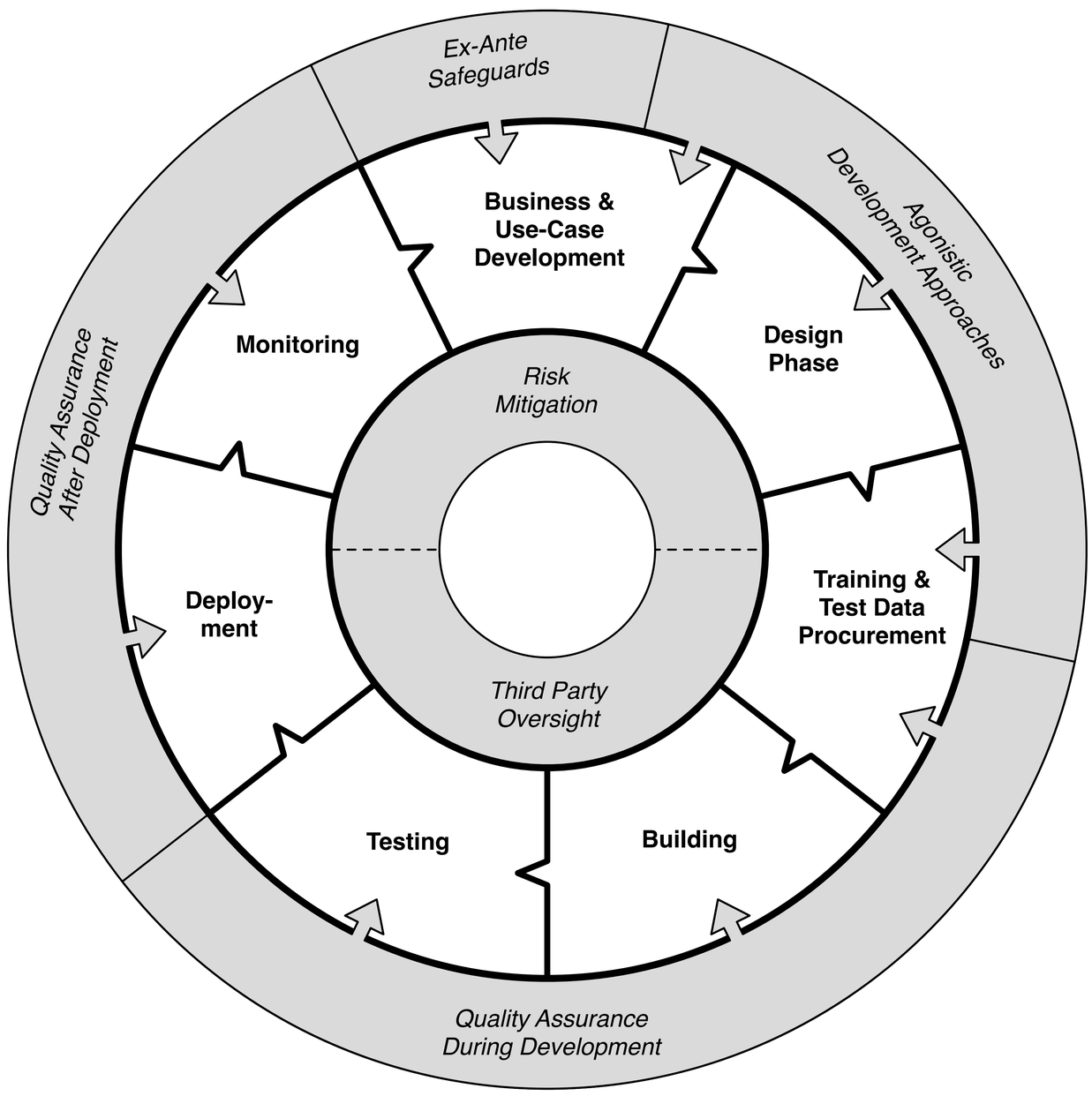


System developers create *built-in safeguards* to constrain the behavior of AI systems.
Human controllers use *interactive controls* to correct or override AI system decisions.
Decision subjects use *interactive controls*, *explanations*, *intervention requests*, and *tools for scrutiny* to contest AI system decisions.
Third parties also use *tools for scrutiny* and *intervention requests* for oversight and contestation on behalf of individuals and groups.

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

During **business and use-case development**, *ex-ante safeguards* protect against potential harms.
During **design** and **training and test data procurement**, *agonistic development approaches* enable stakeholder participation, making room for and leveraging conflict towards continuous improvement.
During **building** and **testing**, *quality-assurance measures* ensure stakeholder interests are centered, and progress towards shared goals is tracked.
Finally, during **deployment** and **monitoring**, further *quality assurance measures* enable tracking of system performance on an ongoing basis, and the feedback loop with future development of the system is closed.
Throughout, *risk mitigation* intervenes in the system context to reduce the odds of failure, and *third party oversight* strengthens the role of external reviewers to enable ongoing outside scrutiny.



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

AI system lifecycle adapted from: Binns, R., & Gallo, V. (2019, March 26). An overview of the Auditing Framework for Artificial Intelligence and its core components. *Information Commissioner's Office (ICO)*. https://ico.org.uk/about-the-ico/media-centre/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/