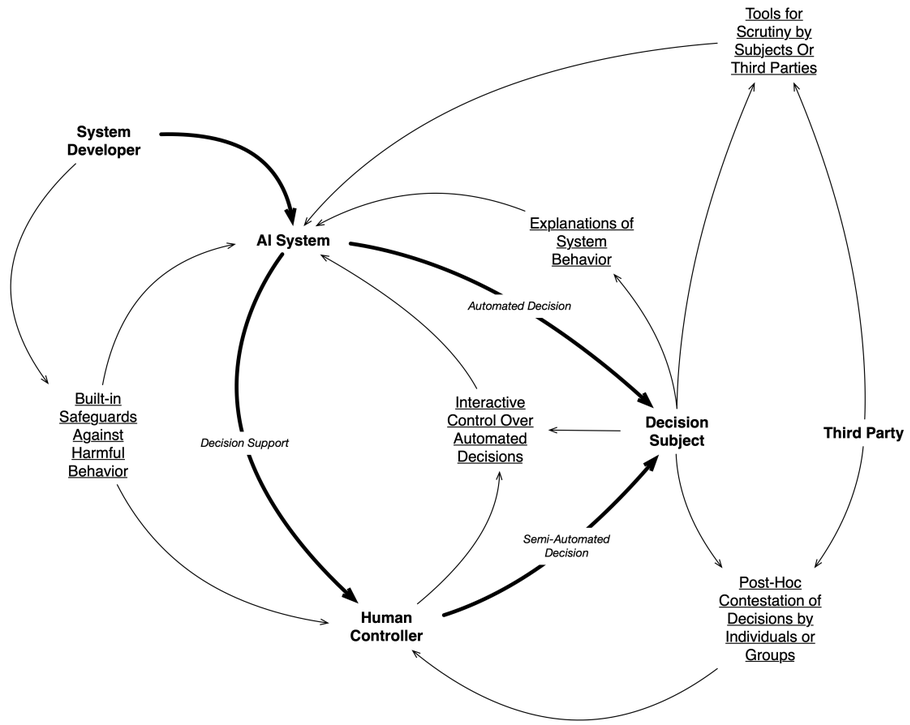
Contestable AI by Design

To ensure public artificial intelligence systems are responsive to value change over time, they must be made **contestable by design**.

Contestable AI

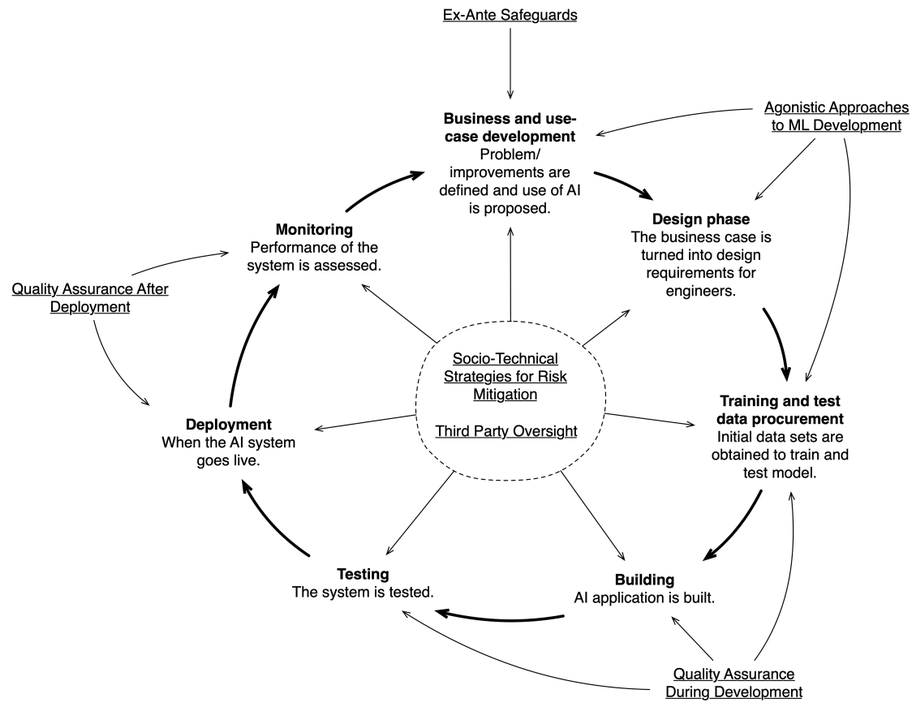
- is open and responsive to human intervention
- encompasses a system's entire lifecycle
- establishes a procedural relationship between decision subjects and system controllers
- leverages disagreement for continuous improvement



Features

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

Built-in safeguards	second adversarial system
Interactive control	negotiate, correct or override automated decision • feedback loop back to (re-)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 • zero-knowledge proofs (opaque assurances)



Practices

During **business and use-case development**, <u>ex-ante safeguards</u> are put in place to protect against potential harms. During the **design, and procurement of training and test data**, <u>agonistic development approaches</u> enable stakeholder participation in a way that makes room for and leverages conflict towards continuous improvement. During **building** and **testing** <u>quality assurance practices</u> are used to ensure stakeholder interests are centered and progress towards shared goals is tracked. Finally, during **deployment** and **monitoring**, further <u>quality assurance</u> measures ensure system performance is tracked on an ongoing basis, and the feedback loop with future development of the system is closed.

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