How Technological Interventions Can Prevent Harms
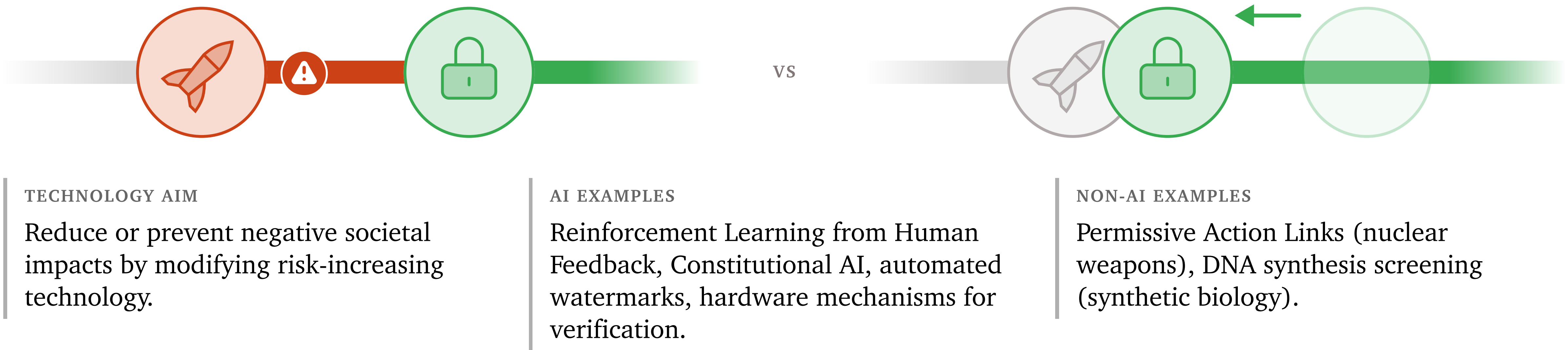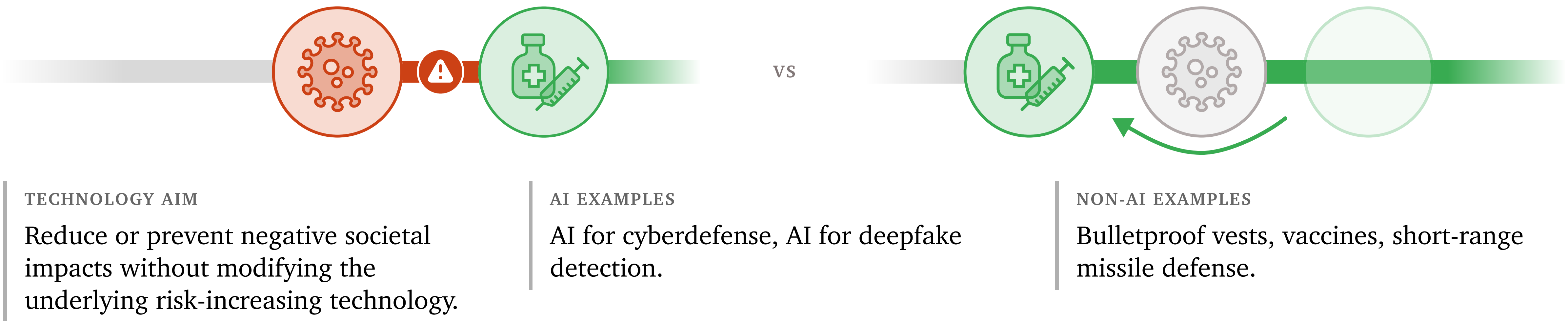
Safety Technologies sooner relative to risk-increasing technologies

vs

TECHNOLOGY AIM
Reduce or prevent negative societal impacts by modifying risk-increasing technology.

AI EXAMPLES
Reinforcement Learning from Human Feedback, Constitutional AI, automated watermarks, hardware mechanisms for verification.

NON-AI EXAMPLES
Permissive Action Links (nuclear weapons), DNA synthesis screening (synthetic biology).

Defensive Technologies before risk-increasing technologies

vs

TECHNOLOGY AIM
Reduce or prevent negative societal impacts without modifying the underlying risk-increasing technology.

AI EXAMPLES
AI for cyberdefense, AI for deepfake detection.

NON-AI EXAMPLES
Bulletproof vests, vaccines, short-range missile defense.

Substitute Technologies instead of risk-increasing technologies

vs

TECHNOLOGY AIM
Create less risky technologies that provide similar benefits to riskier technologies.

AI EXAMPLES
General-purpose AI systems that lack dangerous capabilities, narrow AI systems to substitute for general purpose ones in certain contexts.

NON-AI EXAMPLES
Clean energy (substitute for fossil fuels), hydrofluorocarbons (substitute for chlorofluorocarbons).