Reframing
the AI Safety Landscape:

A development-oriented approach

Eric Drexler

FHI/DeepMind AI Safety Seminar

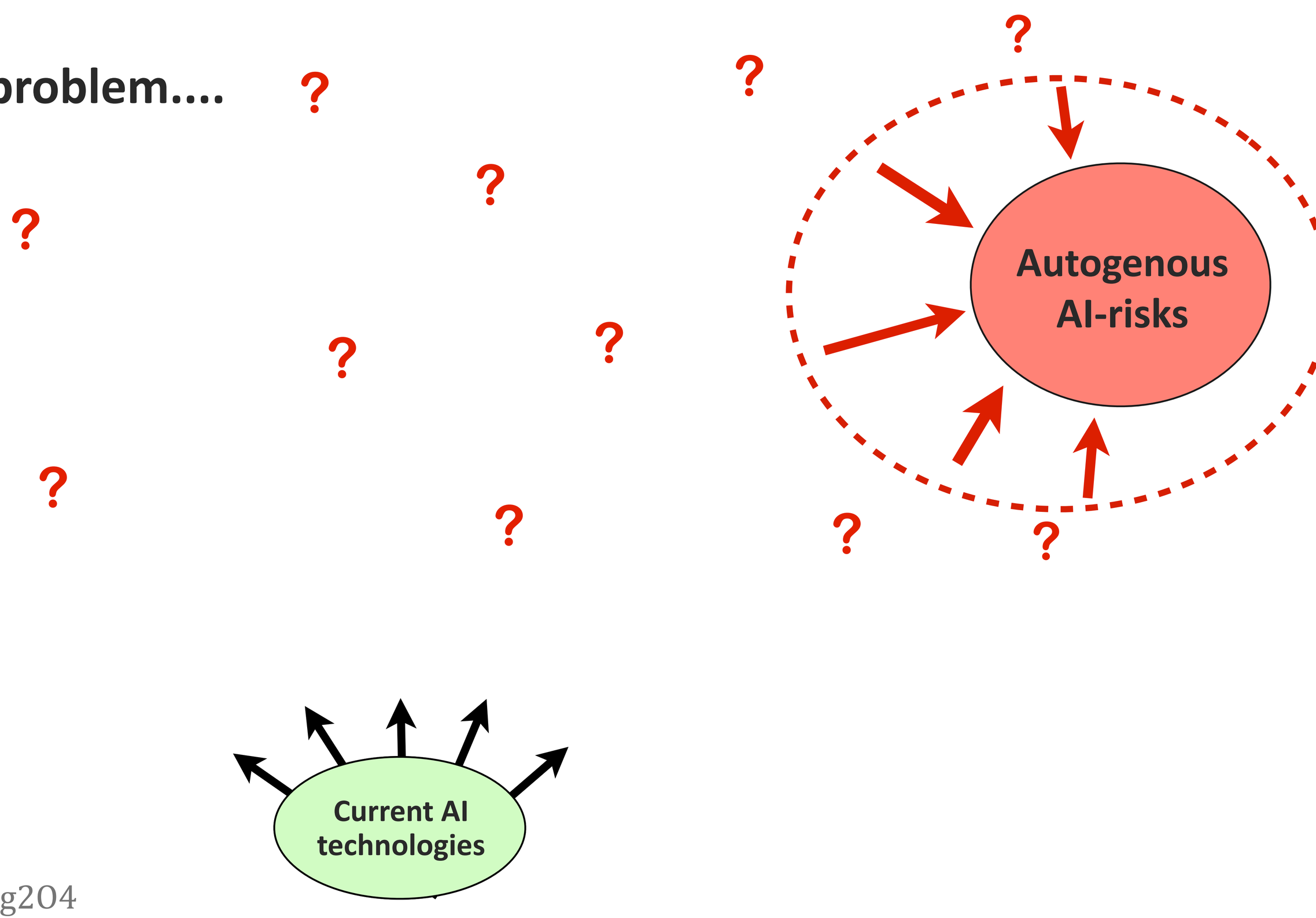
Oxford, 2 March 2017

OXFORD MARTIN SCHOOL

Future of Humanity Institute

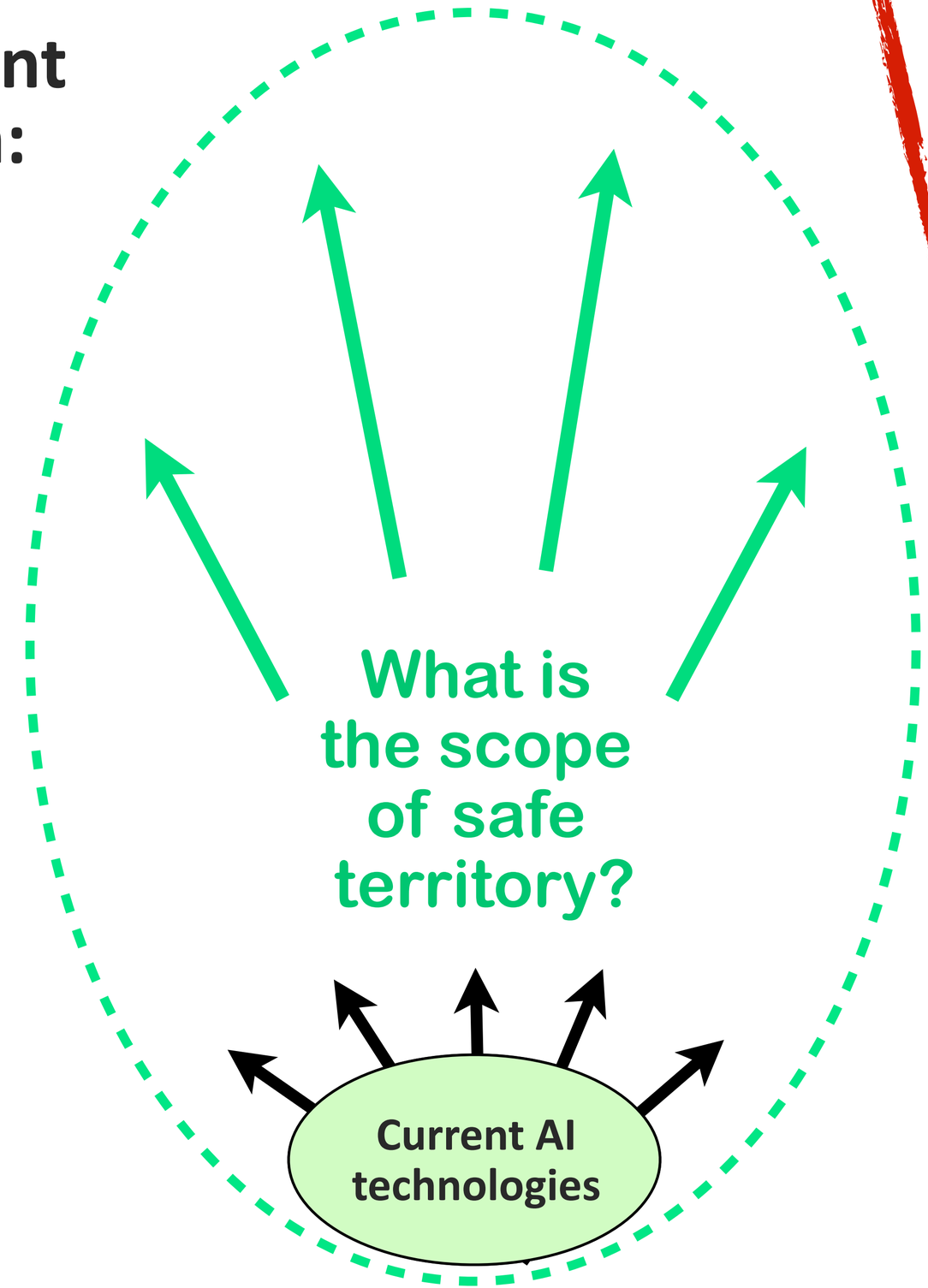
UNIVERSITY OF OXFORD

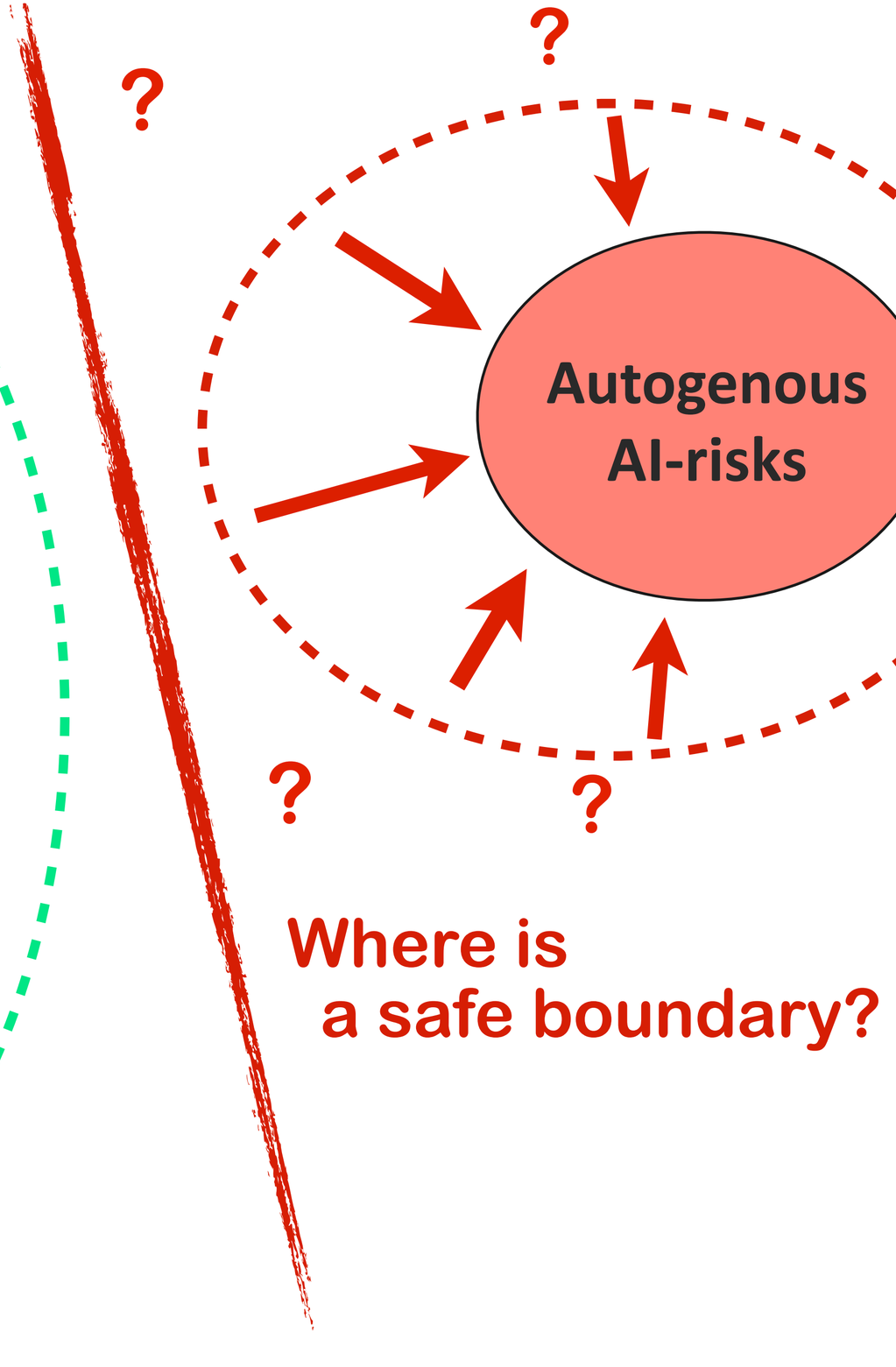
The problem....



goo.gl/Uwg2O4

A different question:





goo.gl/Uwg2O4

The R&D-automation model of recursive improvement

- R&D flows from basic research to concrete applications.
- R&D involves diverse human tasks.
- AI progress will enable automation of human tasks.
- AI progress will enable automation of AI R&D tasks.
- Thorough automation of AI R&D tasks…

⟹ **recursive improvement**.

goo.gl/Uwg2O4

The R&D-automation model of recursive improvement

- R&D flows from basic research to concrete applications.
- R&D involves diverse human tasks.
- AI progress will enable automation of human tasks.
- AI progress will enable automation of AI R&D tasks.
- Thorough automation of AI R&D tasks…

⟹ **recursive improvement**.

goo.gl/Uwg2O4

Recursive improvement ⇏ "self" improvement

- R&D tasks require specialized competencies.
- AI components can provide specialized competencies.
- Developing AI components requires multiple competencies.
- R&D processes (not components) produce new components.
- R&D processes need not be embodied in distinct "selves".

goo.gl/Uwg2O4

Recursive improvement ⇏ “self” improvement

- R&D tasks require specialized competencies.
- AI components can provide specialized competencies.
- Developing AI components requires multiple competencies.
- R&D processes (not components) produce new components.
- R&D processes need not be embodied in distinct “selves”.

∴ Recursive improvement ⇏ “self” improvement.

goo.gl/Uwg2O4

Recursive improvement:

$\not\Rightarrow$ Self-improving AGI agents (possible $\neq$ necessary)

$\Rightarrow$ Access to superintelligent-level AI capabilities

$\Rightarrow$ Access to General Artificial Intelligence (GAI)

$\not\Rightarrow$ Hands-off application development

…We want *AI-facilitated* design & refinement

goo.gl/Uwg2O4

AI-facilitated design and refinement:

goo.gl/Uwg2O4

AI-facilitated design and refinement:

- Can draw on descriptions, sketches, prototypes, examples…
- Can employ discussion, questions, advice, suggestions…
- Can learn from user experience, accidents, successes…
- Can employ strong, broad priors on human (dis)approval.

Developing predictive models of human (dis)approval:

goo.gl/Uwg2O4

Developing predictive models of human (dis)approval:

Stuart Russell:

AI systems will be able to learn: "Not just by watching, but also by reading. Almost everything ever written down is about people doing things, and other people having opinions about it."

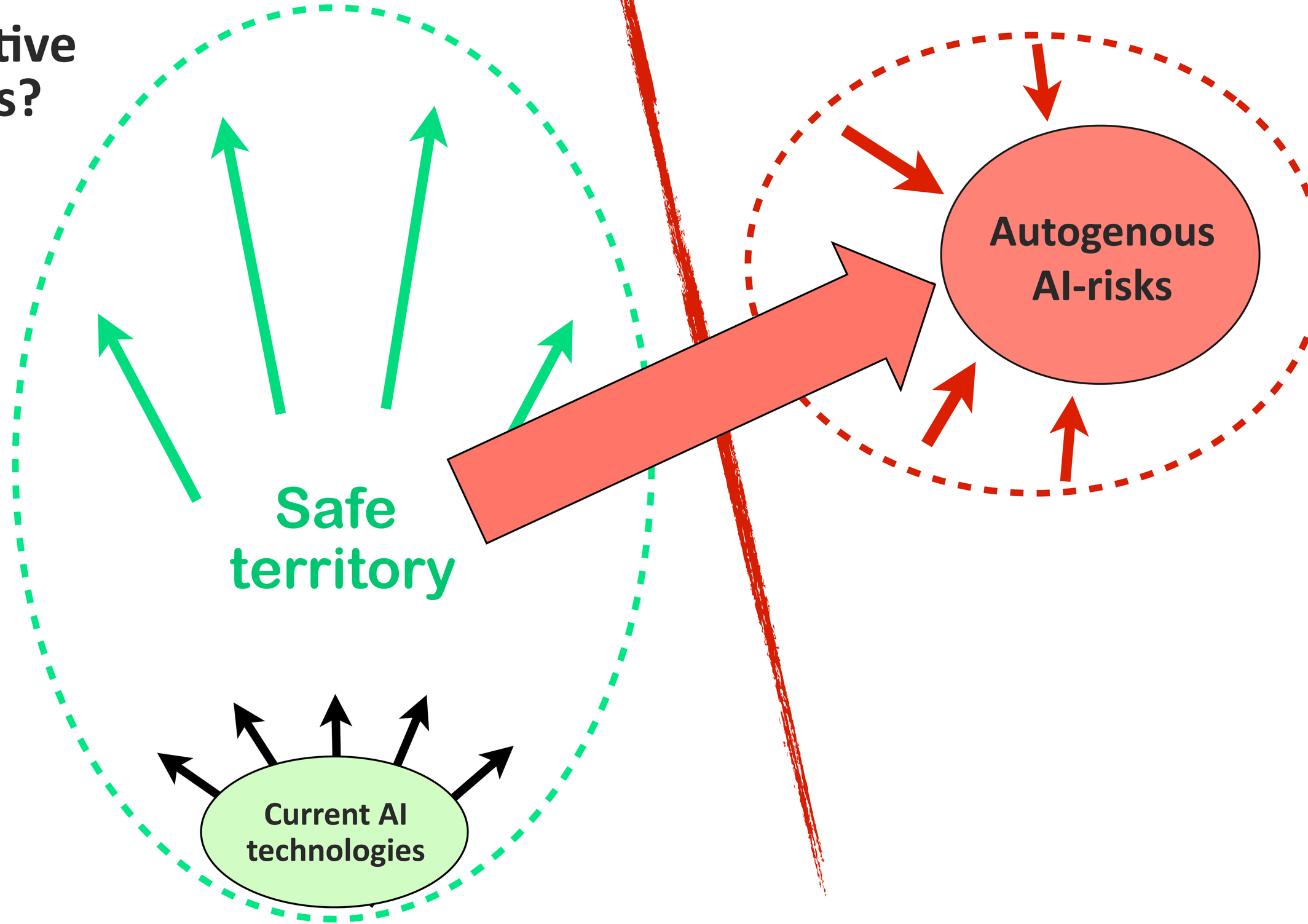
Predictive models of human (dis)approval can be trained on:

Text and video: History, news, tweets, fiction, science fiction, advice columns, social media, movies, sitcoms, CCTV cameras, legal codes, court records, and works of moral philosophy.

Interactive crowd-sourcing: Can challenge participants to "fool the AI" with difficult cases, eliciting erroneous predictions of approval to enable training on imaginative hypotheticals.

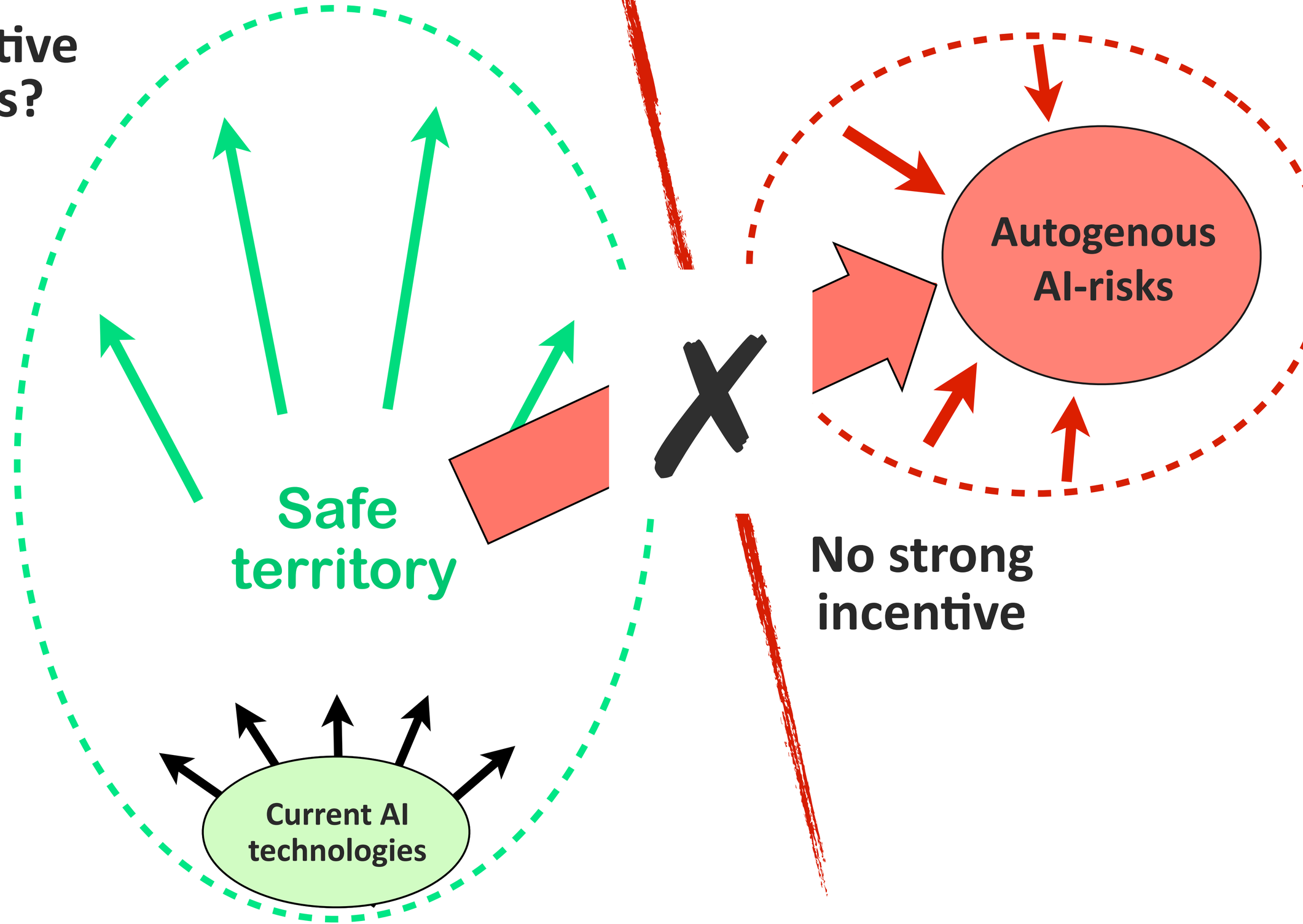
goo.gl/Uwg2O4

Competitive pressures?



goo.gl/Uwg2O4

Competitive pressures?



goo.gl/Uwg2O4

AI-facilitated design specification and refinement:

- Can draw on descriptions, sketches, prototypes, examples…
- **Can employ discussion, questions, advice, suggestions…**
- Can learn from user experience, accidents, successes…
- Can employ strong, broad priors on human (dis)approval.

?

goo.gl/Uwg2O4

Consider a problem frame in which AI components are potentially strategic actors.

Consider concerns regarding potential collusion among strategic actors.

Apply game-theoretic concepts and development-oriented thinking.

goo.gl/Uwg2O4

Contrasting conditions that tend to <u>disrupt</u> collusion among actors:

B1) **Large numbers of actors**
(the required scope of collusion is large)

B2) **Sensitivity to defectors**
(collusion fails unless it is complete)

B3) **Diversity among actors**
(*e.g.,* different histories, adversarial objectives)

B4) **Single-move decision processes**
in which most actions are unobservable

B5) **Lack of common knowledge**
of a wide range of circumstances

Natural architectures for problem-solving systems:

C1) To improve the quality of answers, it is natural to implement ***multiple, diverse*** (implicitly competing) systems to propose alternative answers.

C2) To identify poor or misleading answers, it is natural to employ diverse answer-assessors, ***any one of which*** could disrupt deceptive collusion.

C3) Systems of diverse, competing answer-providers and answer-assessors naturally implement both ***independent*** and ***adversarial*** objectives.

C4) It is natural to apply fixed system instantiations to multiple problems, incidentally yielding a series of ***history-blind, single-move decisions***.

C5) It is natural to provide ***different, task-relevant information*** to systems solving different problems, omitting knowledge of general circumstances.

GAI capabilities can implement adversarial, trustworthy problem-solving systems that operate at a superintelligent-level.

goo.gl/Uwg2O4

Objectives & paths:

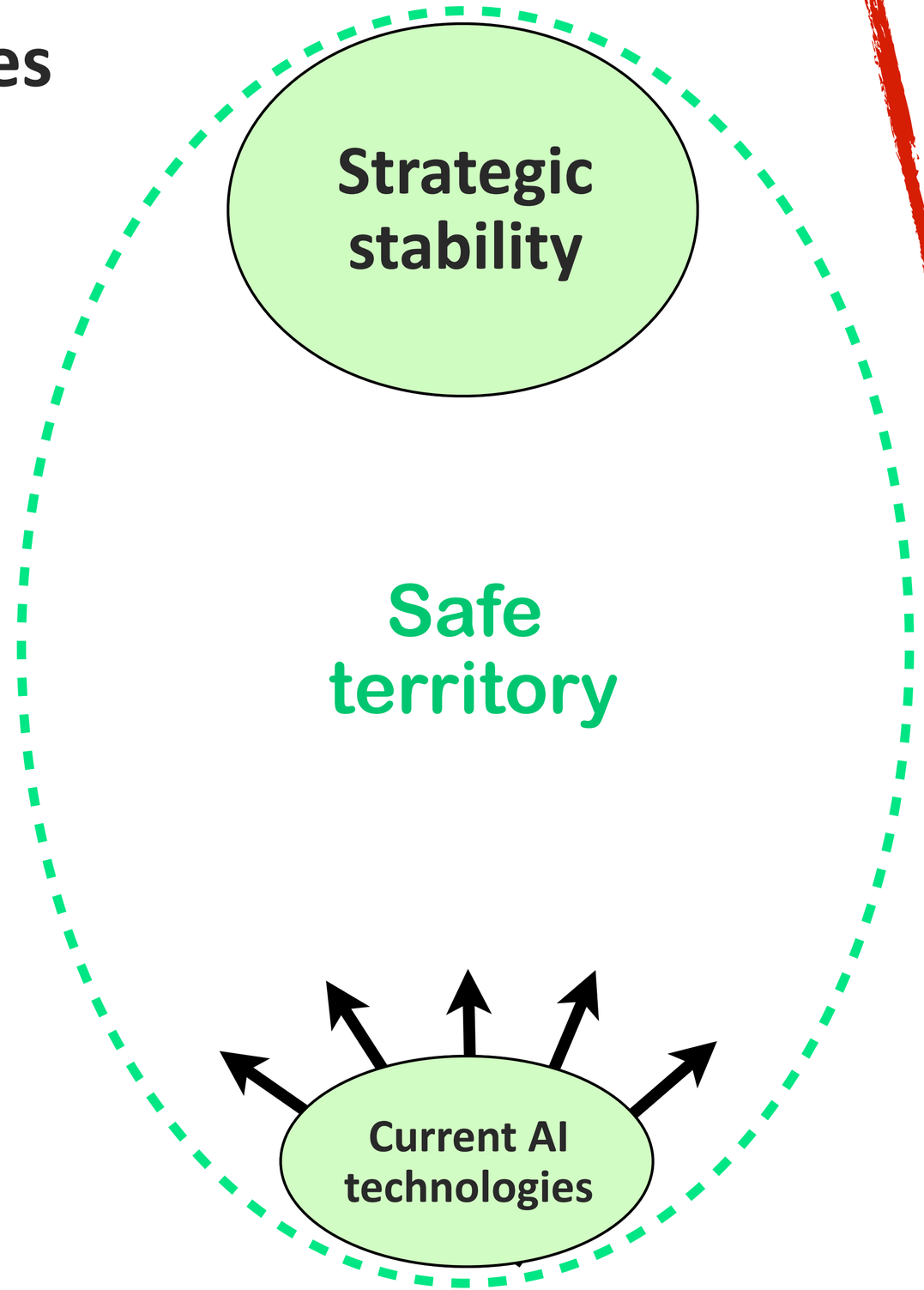
Safe territory

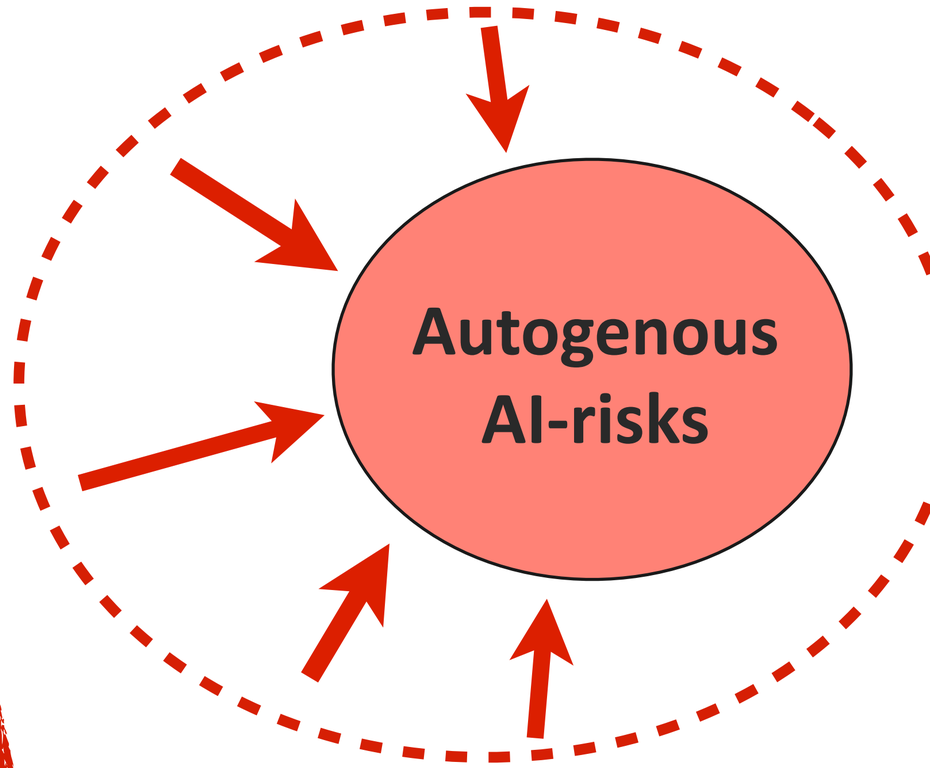
Current AI technologies

Autogenous AI-risks

goo.gl/Uwg2O4

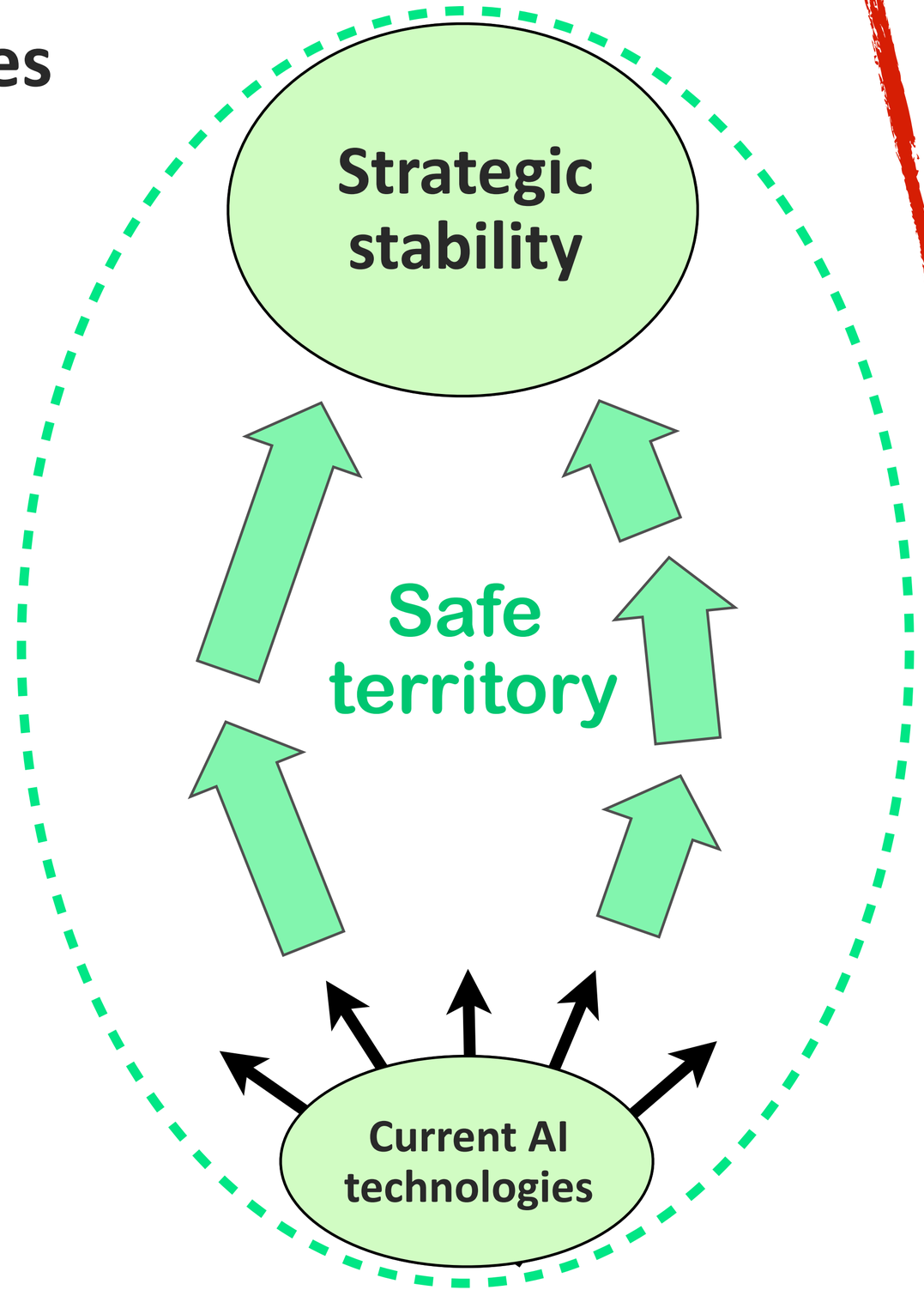
Objectives & paths:

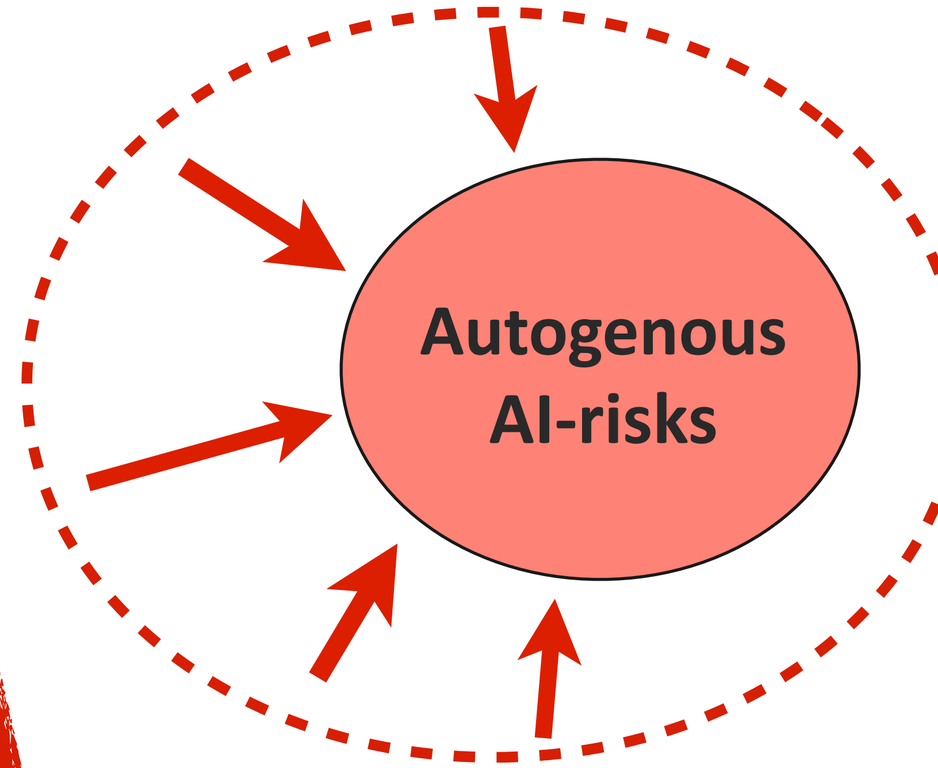




goo.gl/Uwg2O4

Objectives & paths:





goo.gl/Uwg2O4

Incremental R&D automation leads
toward fast recursive improvement.

goo.gl/Uwg2O4

Incremental R&D automation leads
toward fast recursive improvement.

Fast recursive improvement can yield
superintelligent-level GAI capabilities.

goo.gl/Uwg2O4

Incremental R&D automation leads
toward fast recursive improvement.

Fast recursive improvement can yield
superintelligent-level GAI capabilities.

Superintelligent-level GAI capabilities can implement
adversarial, trustworthy problem-solving systems

goo.gl/Uwg2O4

Incremental R&D automation leads
toward fast recursive improvement.

Fast recursive improvement can yield
superintelligent-level GAI capabilities.

Superintelligent-level GAI capabilities can implement
adversarial, trustworthy problem-solving systems

This prospect reframes classic AI safety problems
and may render them more tractable.

goo.gl/Uwg2O4

Incremental R&D automation leads
toward fast recursive improvement.

Fast recursive improvement can yield
superintelligent-level GAI capabilities.

Superintelligent-level GAI capabilities can implement
adversarial, trustworthy problem-solving systems

This prospect reframes classic AI safety problems
and may render them more tractable.

More research is needed.

goo.gl/Uwg2O4

Reframing the AI safety landscape:

goo.gl/Uwg2O4

Desiderata and directions for interim AI safety guidelines **(5 min read)**

Incremental R&D automation, agents, and safety concerns **(4 min read)**

Human involvement is compatible with fast, recursive AI technology improvement **(4 min read)**

Advanced AI development systems can support effective human guidance **(6 min read)**

Providing general AI services neither requires nor entails AGI **(7 min read)**

Developing and applying models of human approval **(4 min read)**

Considering AI safety in light of expected future knowledge **(4 min read)**

Ensuring non-collusion in composite superintelligent oracles **(5 min read)**

Development-oriented models for deeply-structured AI systems **(4 min read)**