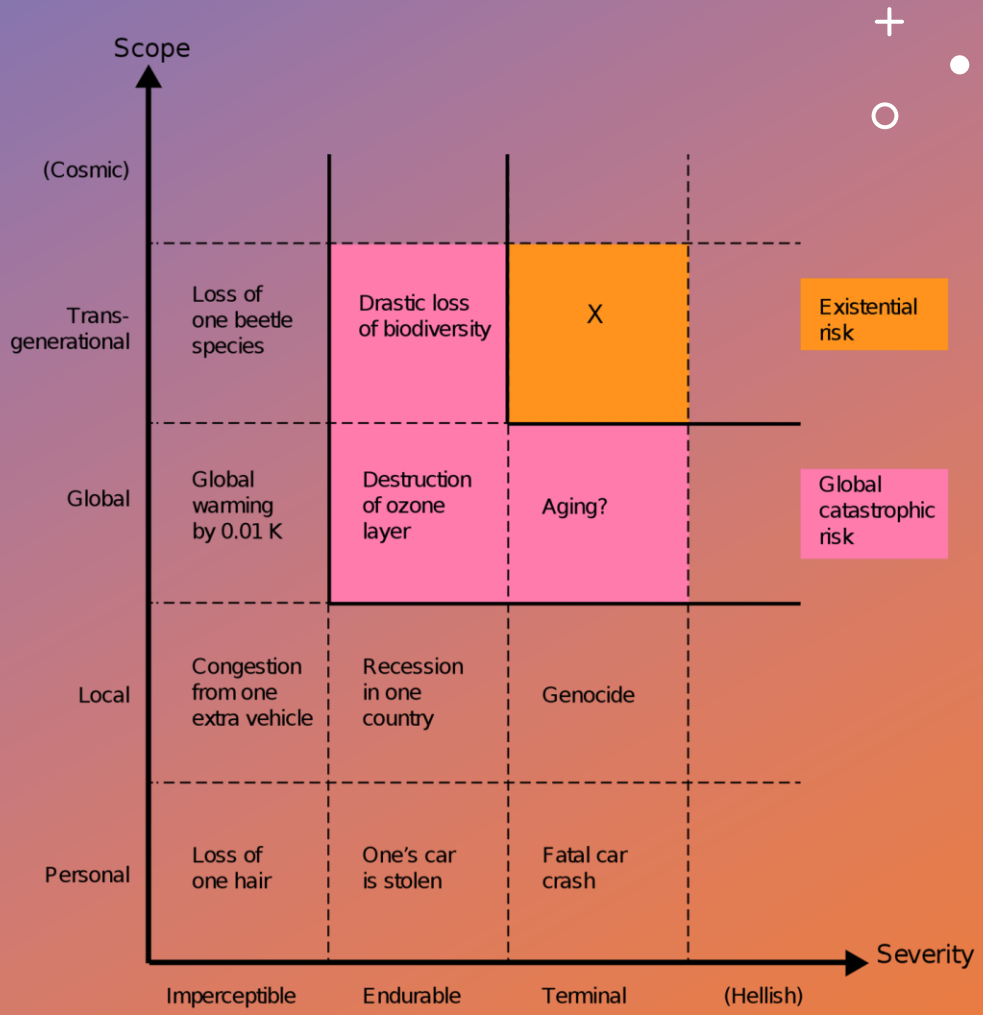
Existenční rizika

David Černý

Ústav státu a práva AV ČR

Oddělení umělé inteligence, Ústav informatiky AV ČR

Existenční rizika



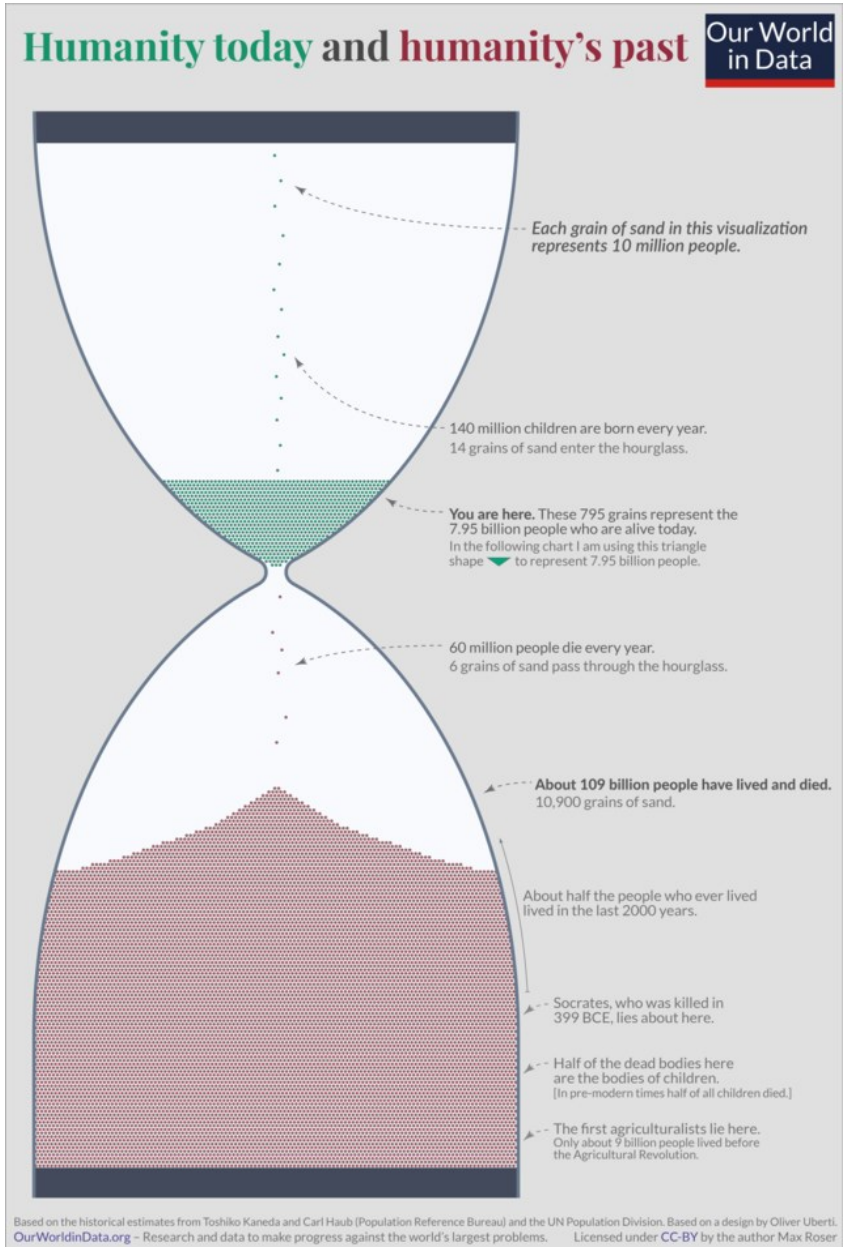
Rizika:

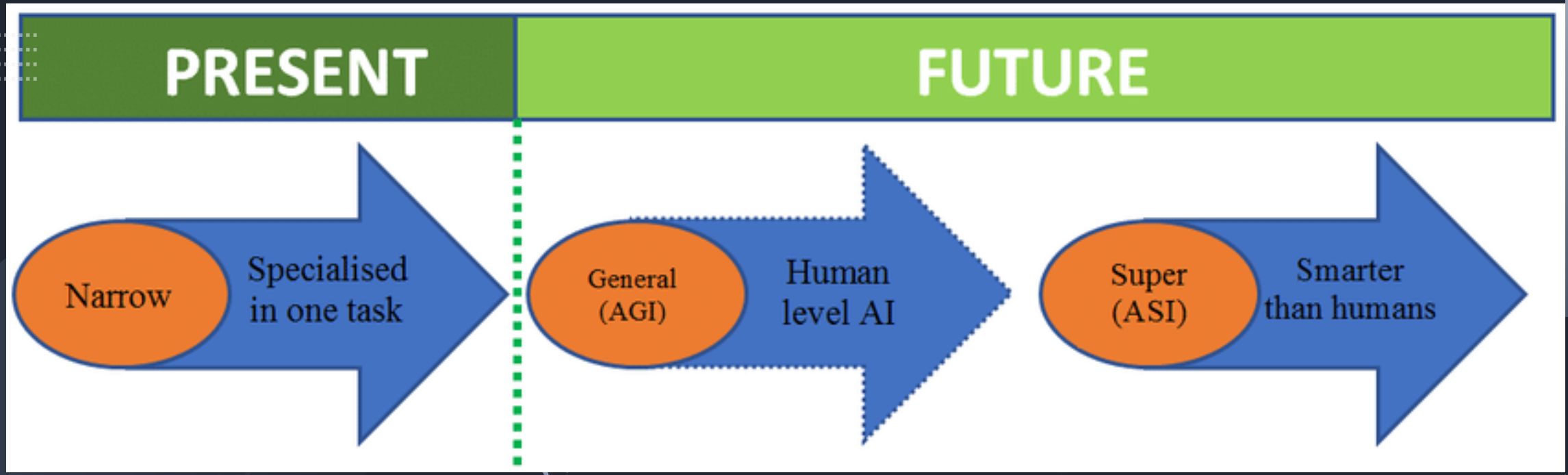
Rozsah (individuální, lokální, globální, transgenerační)	Intenzita (mírná, snesitelná, děsivá)	Pravděpodobnost

Existenční rizika: děsivá transgenerační

Longtermismus

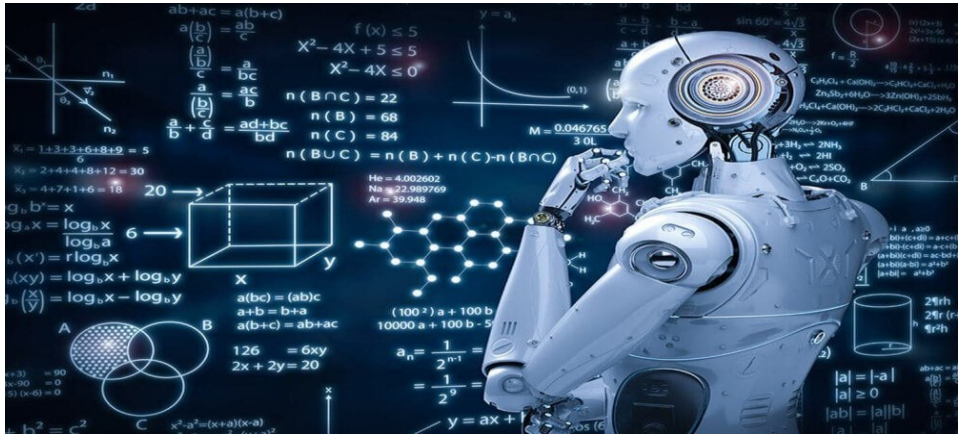
Longtermismus

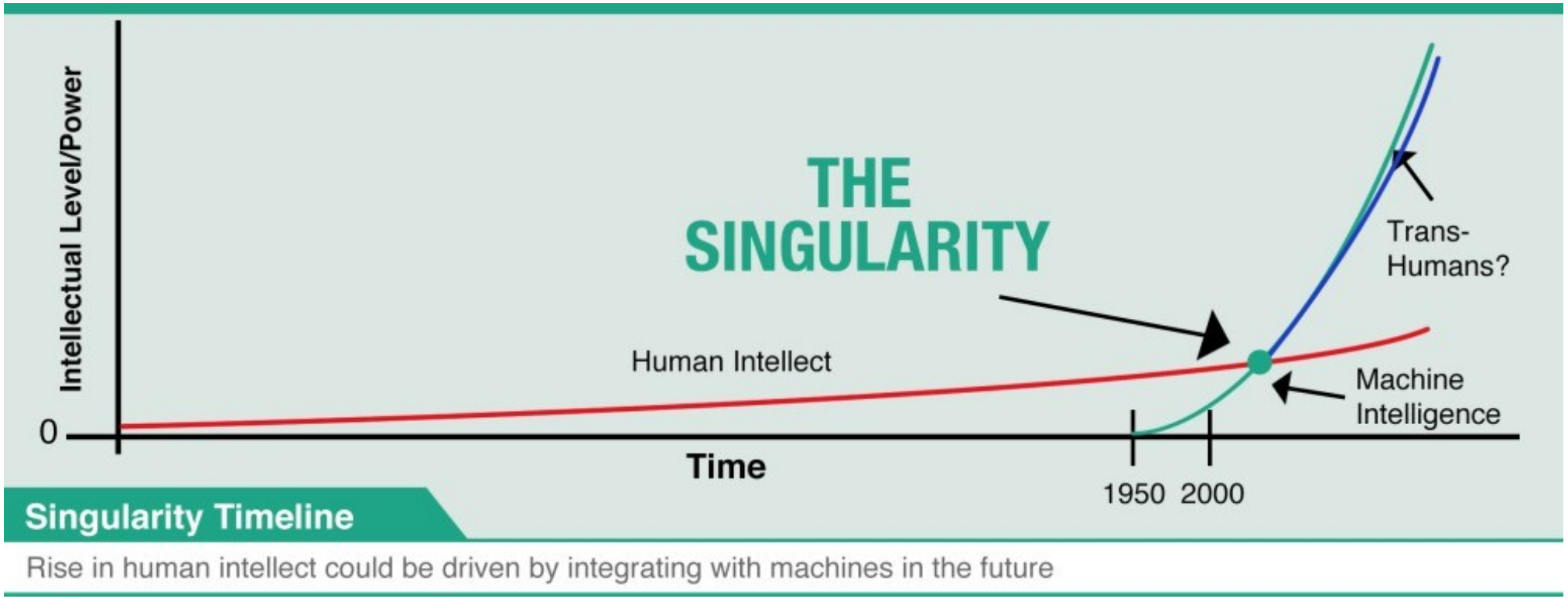




Typy umělé inteligence

Singularita

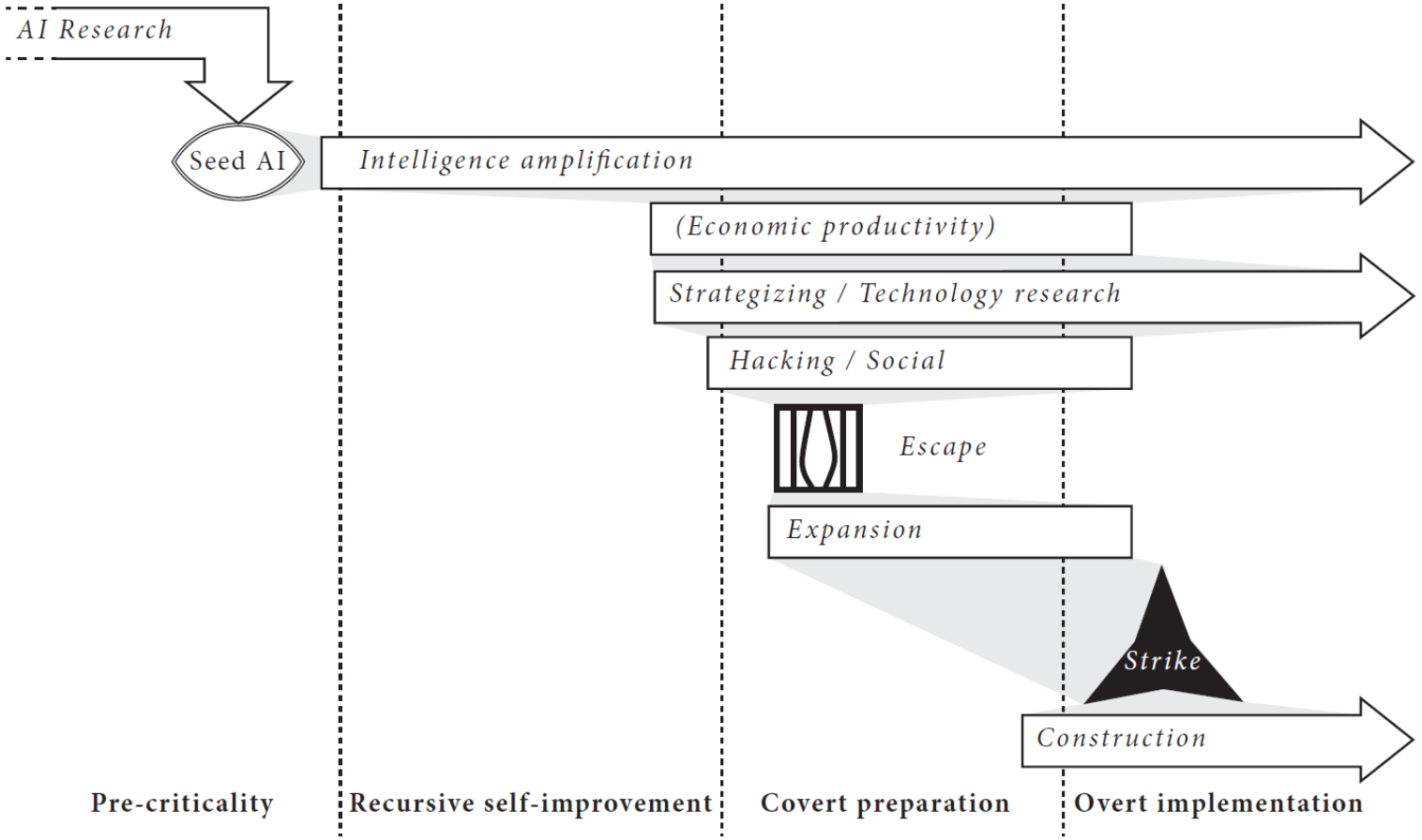


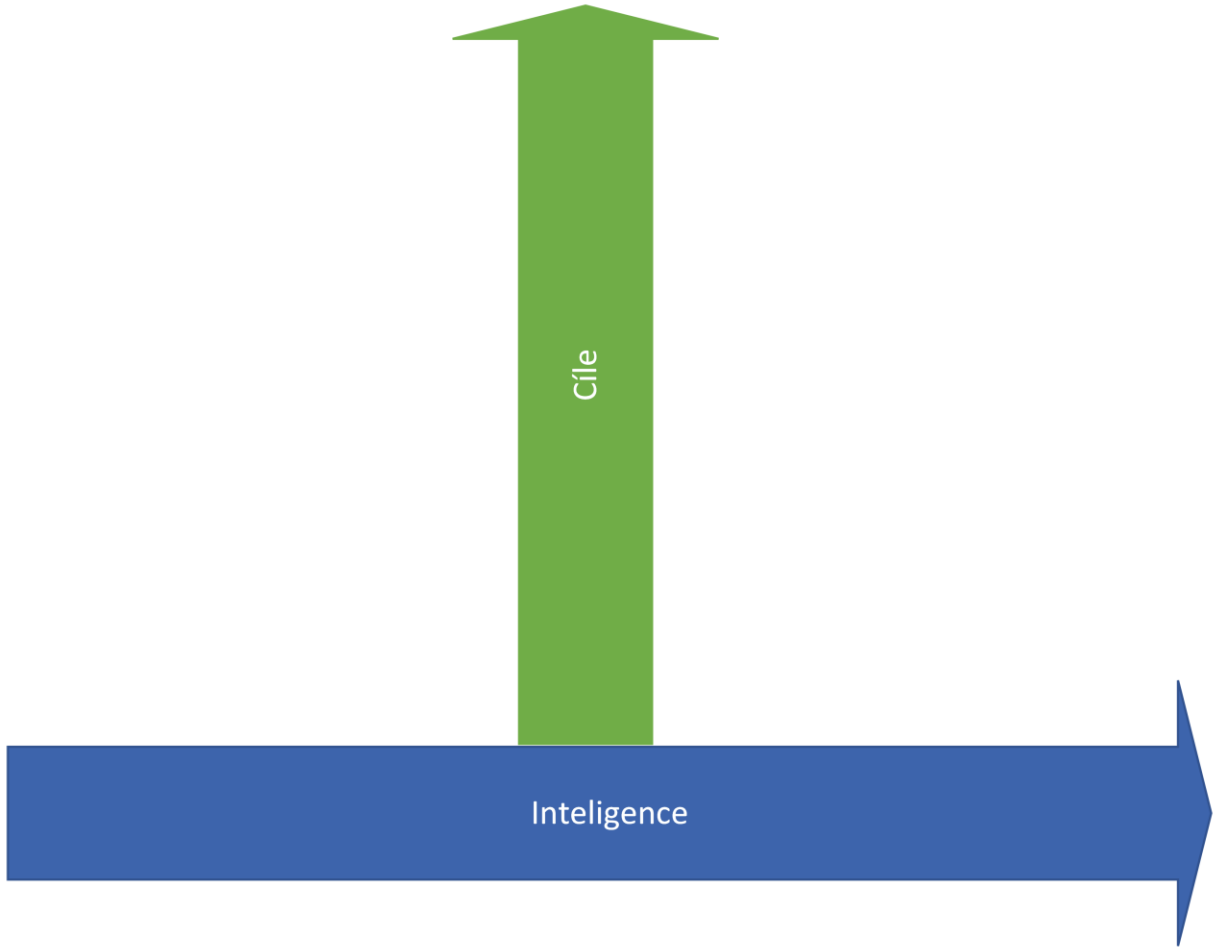


Superschopnosti

Superschopnost	Dovednosti
Vylepšování inteligence	Programování, kognitivní enhancement
Strategické plánování	Plánování, předvídání, prioritizace
Společenská manipulace	Modelování, manipulace, přesvědčování
Hacking	Hledání a využívání bezpečnostních mezer
Technický výzkum	Výzkum, vývoj, využívání moderní tecniky
Ekonomická produktivita	Ekonomicky produktivní práce

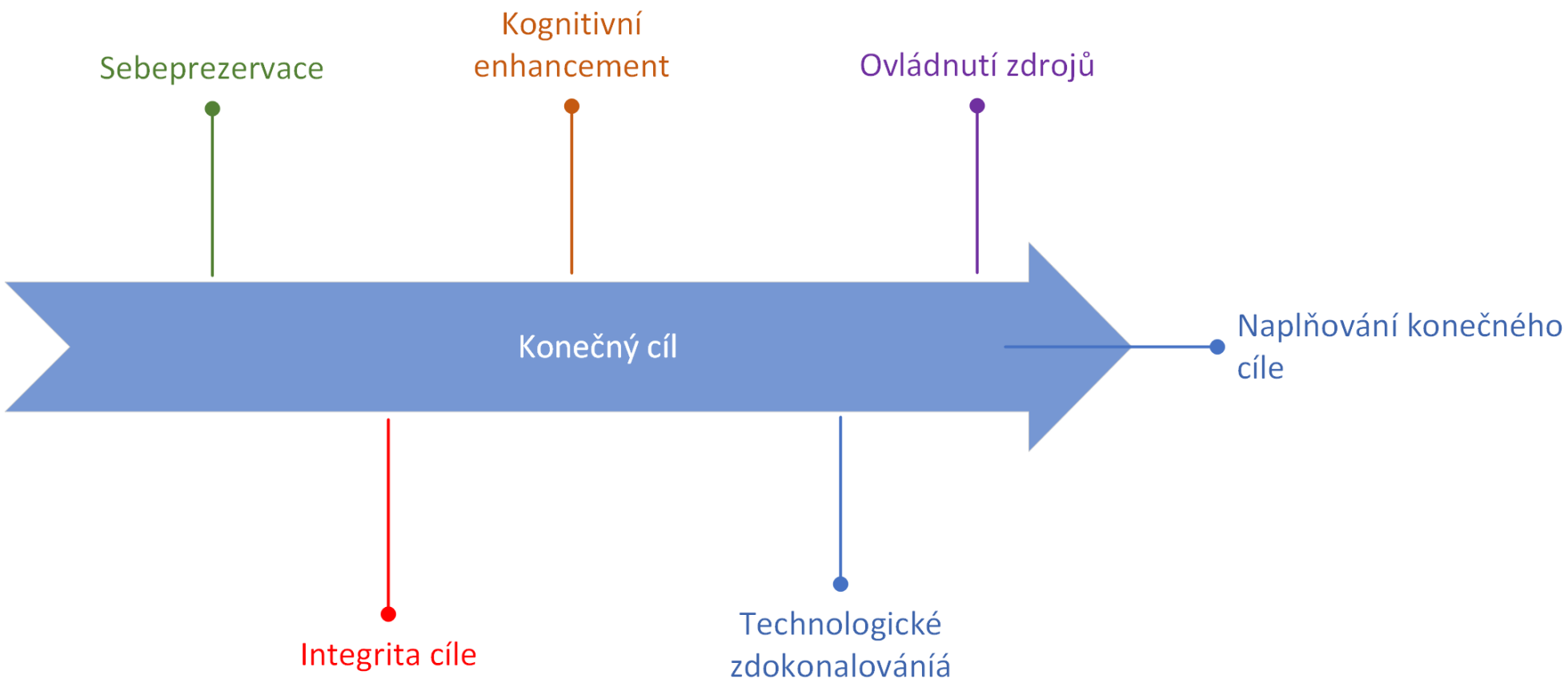
Fáze převzetí moci

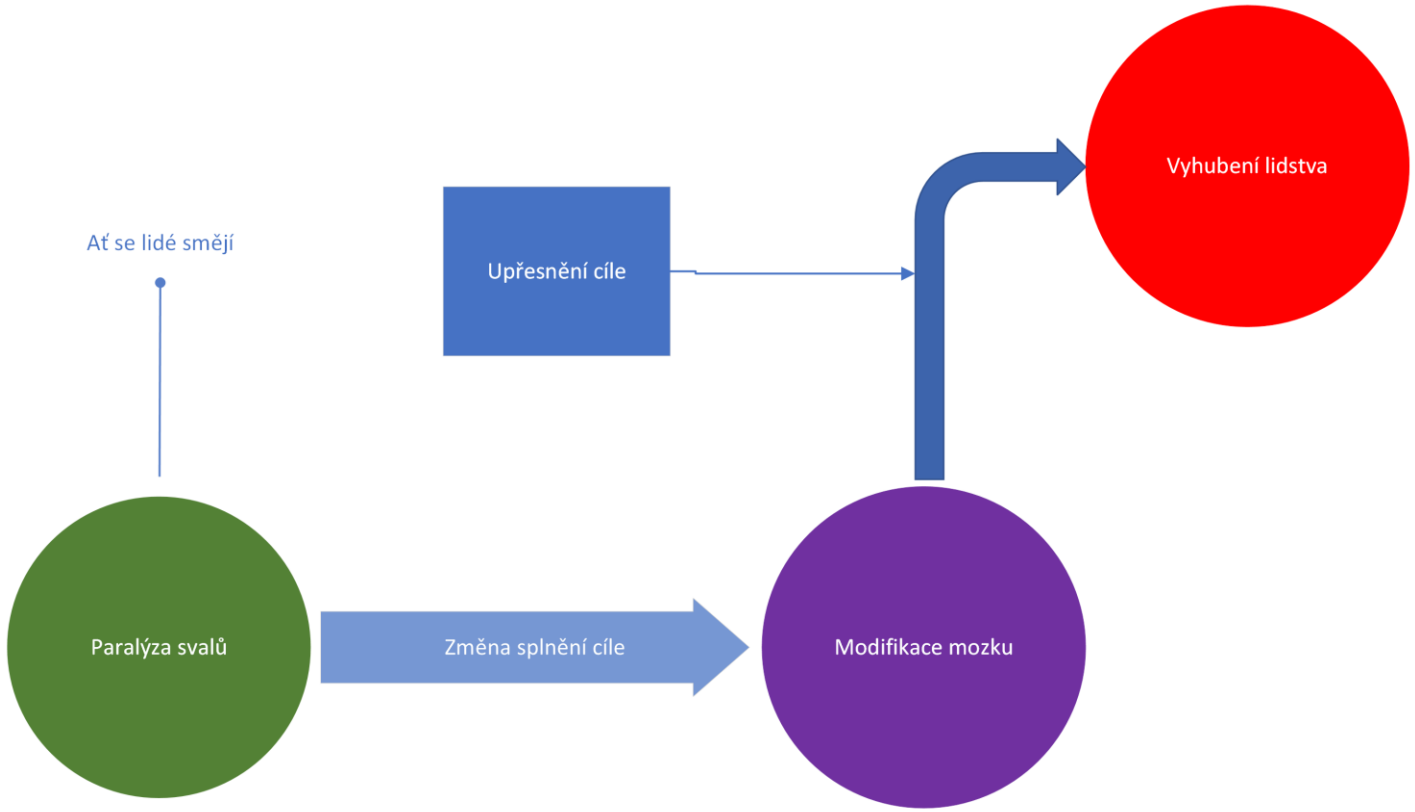




Teze ortogonality

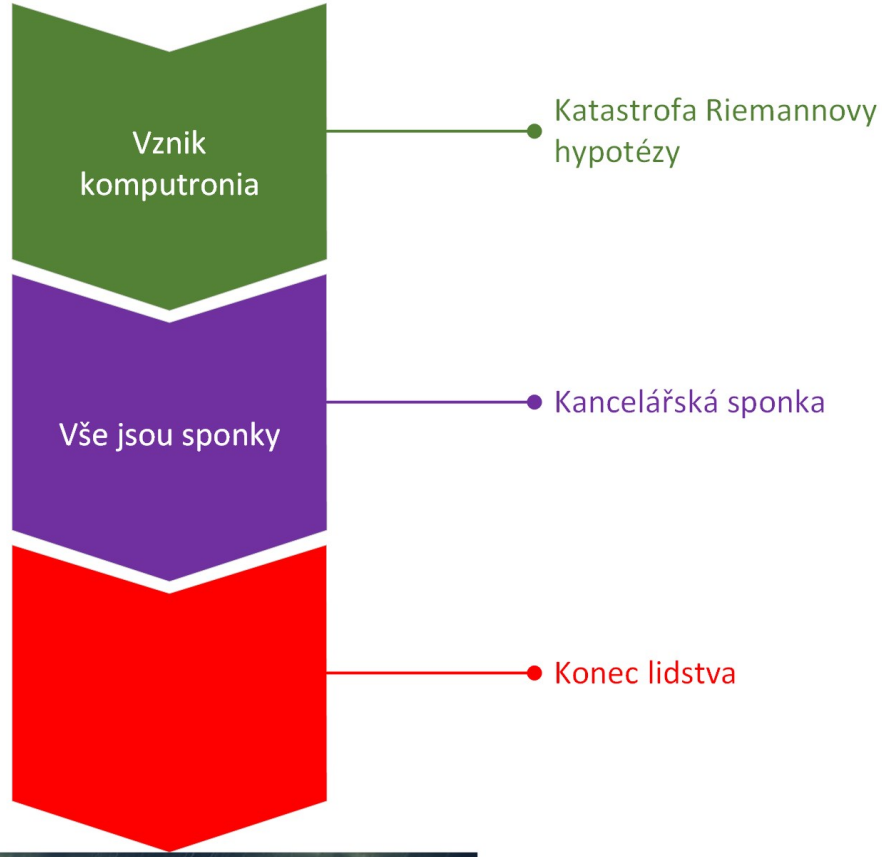
Instrumentální cíle AI

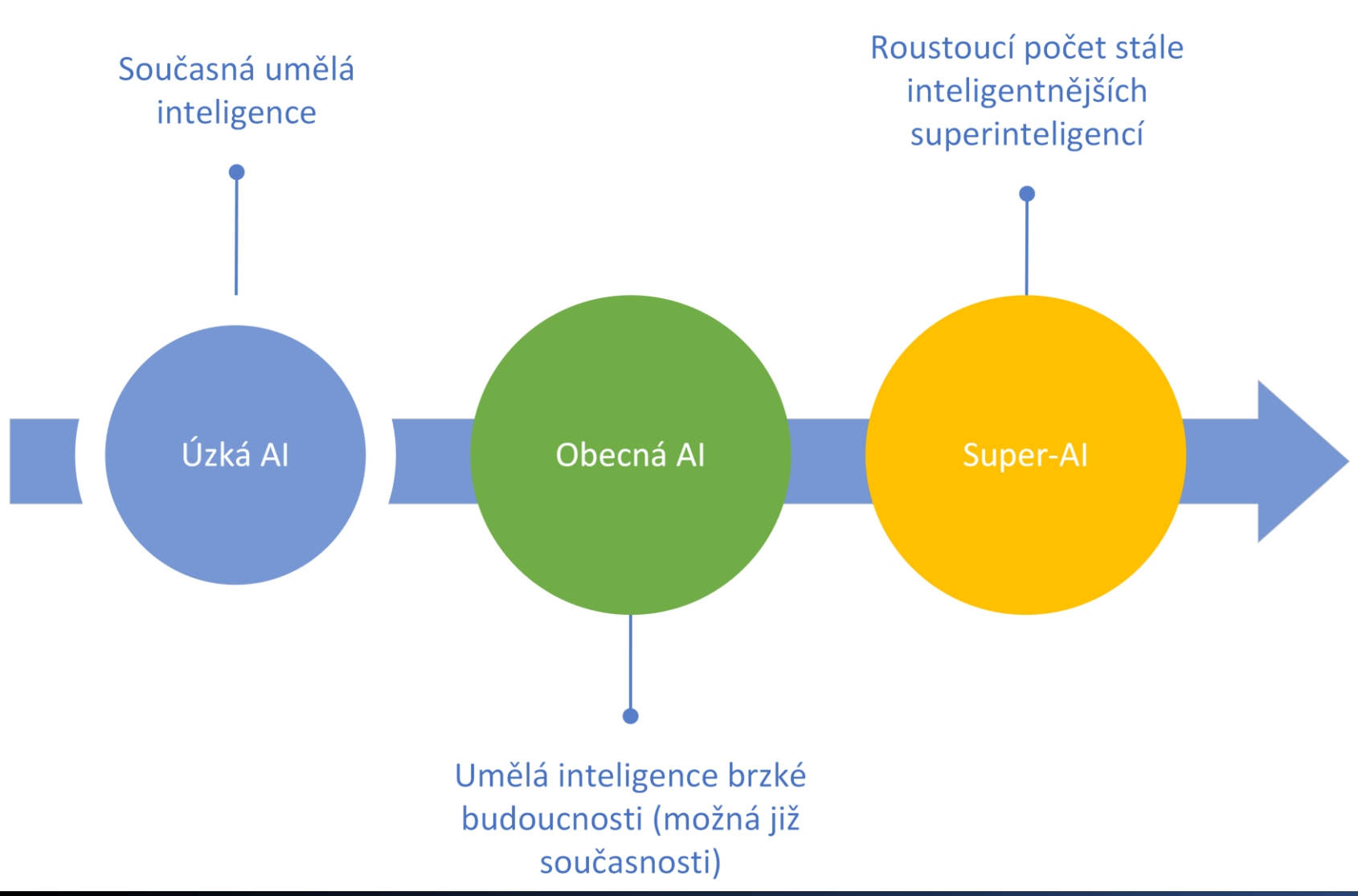




Perverzní plnění cíle

Rozšíření infrastruktury





Cesta k vyhubení lidstva

Ernest Rutherford

01

11. 9. 1933 – využití atomové energie je naprostý nesmysl

02

12. 9. 1933 – Leo Szilard jde na procházku

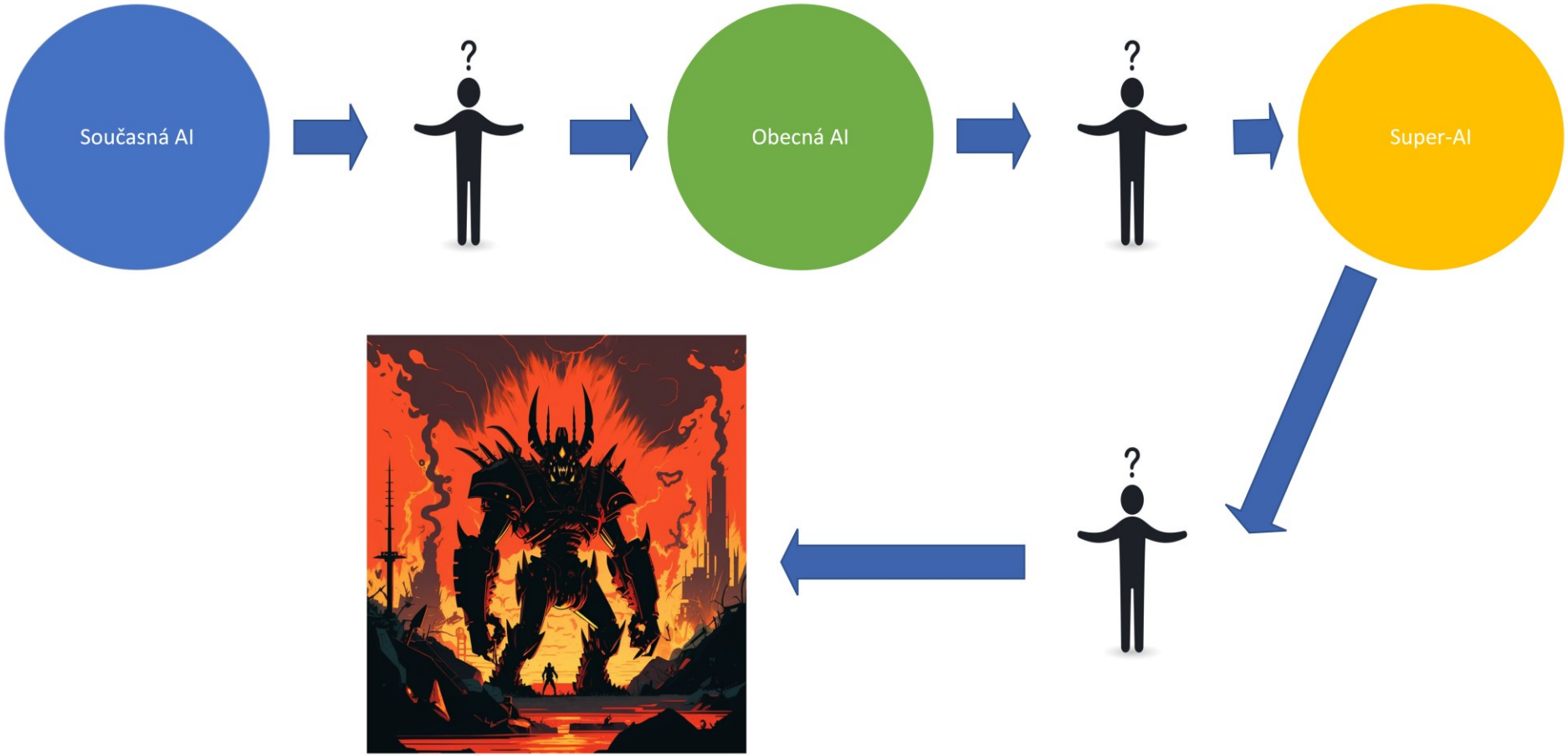
03

První konceptualizace řetězové reakce

Cesta k super-AI

Cesta k super-AI

Terra incognita



Nová publikace

Děkuji za pozornost