Towards Robust End-to-End Alignment

Lê Nguyên Hoang
EPFL

Safe AI 2019

Misalignment

Evasion attacks

Reward Hacking

Poisoning attacks

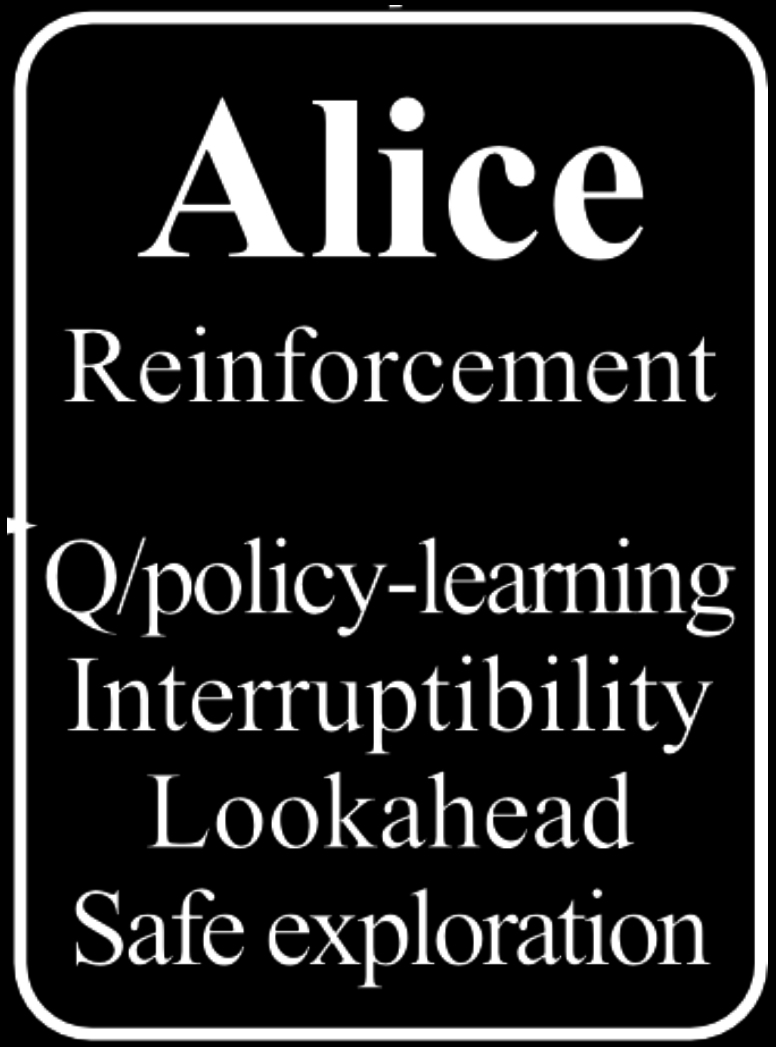
Wireheading

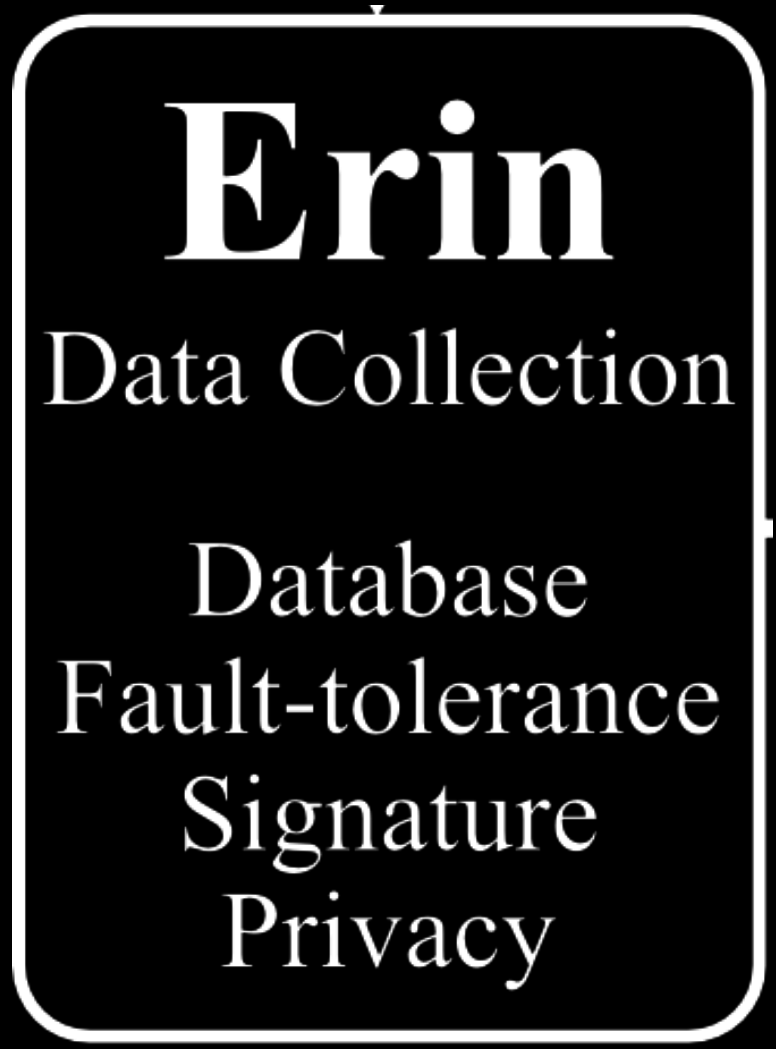
Unsafe exploration

Side effects

Lost Opportunities


Erin
Reliable Data Collection
Dave
World Model Inference
Charlie
Desirability Score Learning
Bob
Incentive Problem
Alice
Reinforcement Learning


Alice
Reinforcement
Q/policy-learning
Interruptibility
Lookahead
Safe exploration


Erin
Data Collection
Database
Fault-tolerance
Signature
Privacy

Dave

World Model

Feature learning
Bayes
MCMC
World description

Charlie

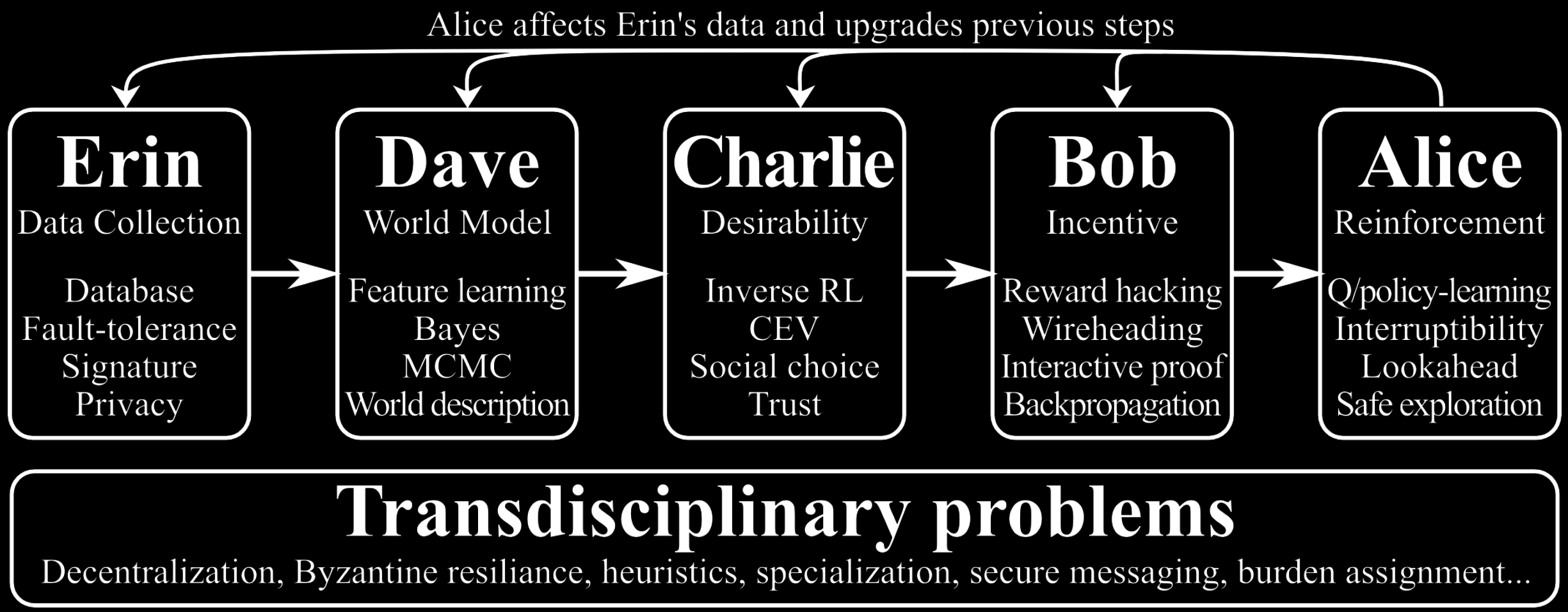
Desirability

Inverse RL
CEV
Social choice
Trust

Bob

Incentive

Reward hacking
Wireheading
Interactive proof
Backpropagation


Alice affects Erin's data and upgrades previous steps
Erin
Data Collection
Database
Fault-tolerance
Signature
Privacy
Dave
World Model
Feature learning
Bayes
MCMC
World description
Charlie
Desirability
Inverse RL
CEV
Social choice
Trust
Bob
Incentive
Reward hacking
Wireheading
Interactive proof
Backpropagation
Alice
Reinforcement
Q/policy-learning
Interruptibility
Lookahead
Safe exploration
Transdisciplinary problems
Decentralization, Byzantine resiliance, heuristics, specialization, secure messaging, burden assignment...