Why Newcomb, Suzy, and Billy are Critical for Trustworthy AI?

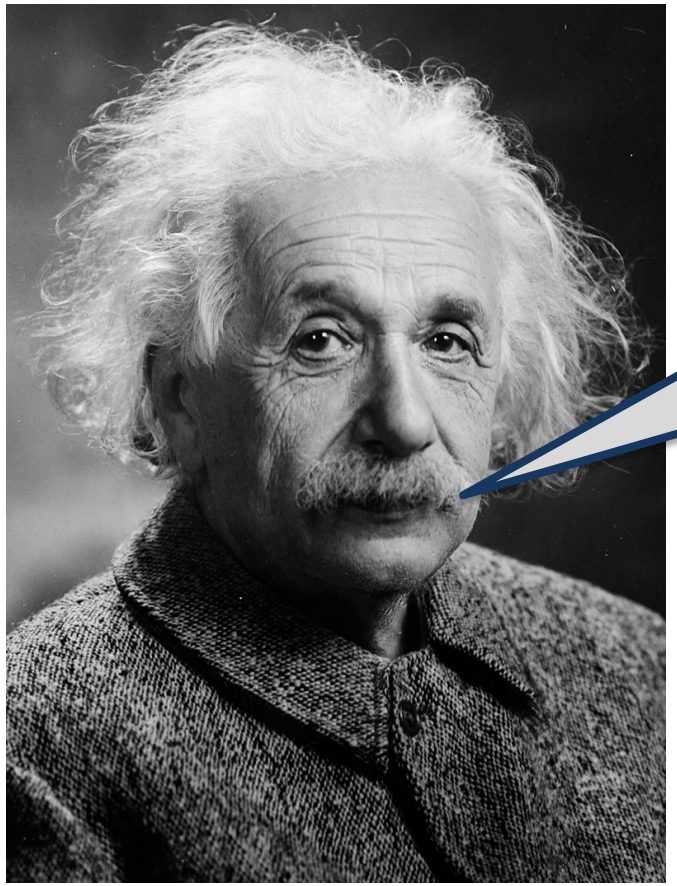
Goran Radanović

Multi-Agent Systems Group @

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

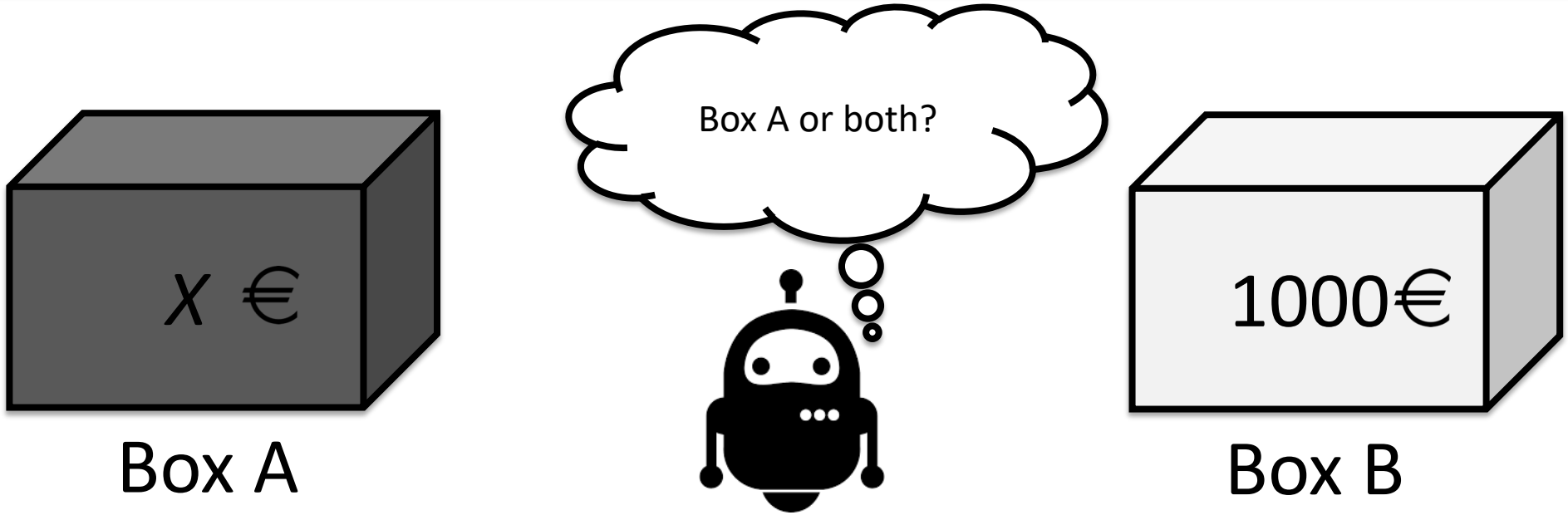
15th November, 2022

Let’s start with…

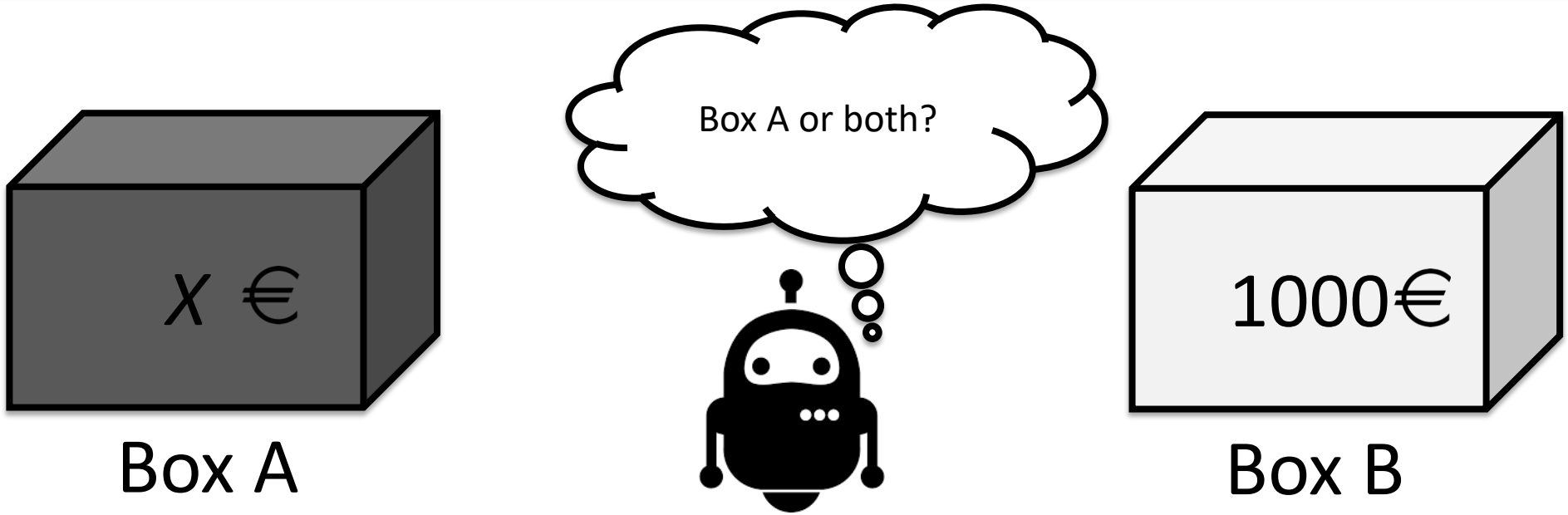


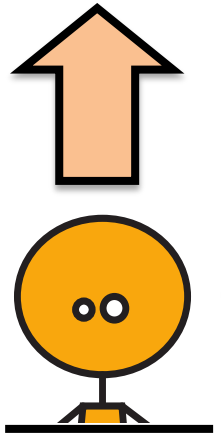
Gedankenexperiment!

Gedankenexperiment 1



Gedankenexperiment 1: Newcomb's Paradox

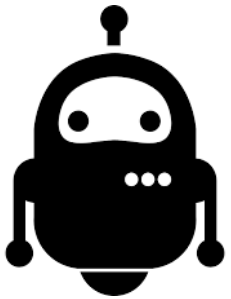


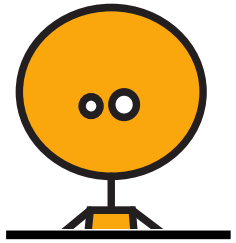


Put $X = 0$ € in Box A if *choice* = $A + B$

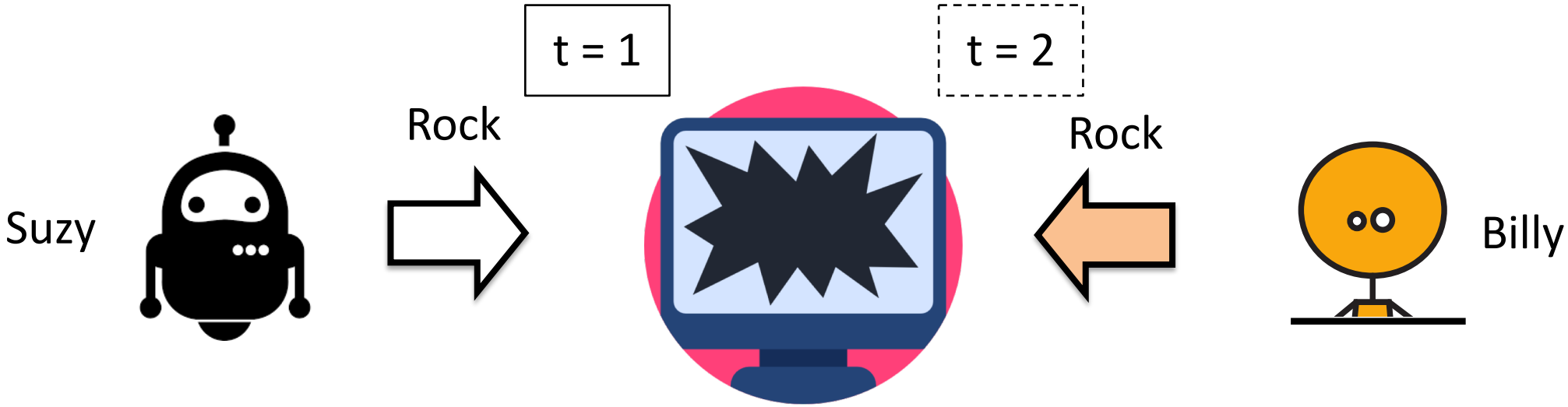
Put $X = 10^6$ € in Box A if *choice* = A

Gedankenexperiment 2

Suzy 

 Billy

Gedankenexperiment 2: Suzy & Billy Example



Who is responsible for the broken computer?

Multi-Agent Sequential Decision Making

AlphaGo

Great success of RL in games…

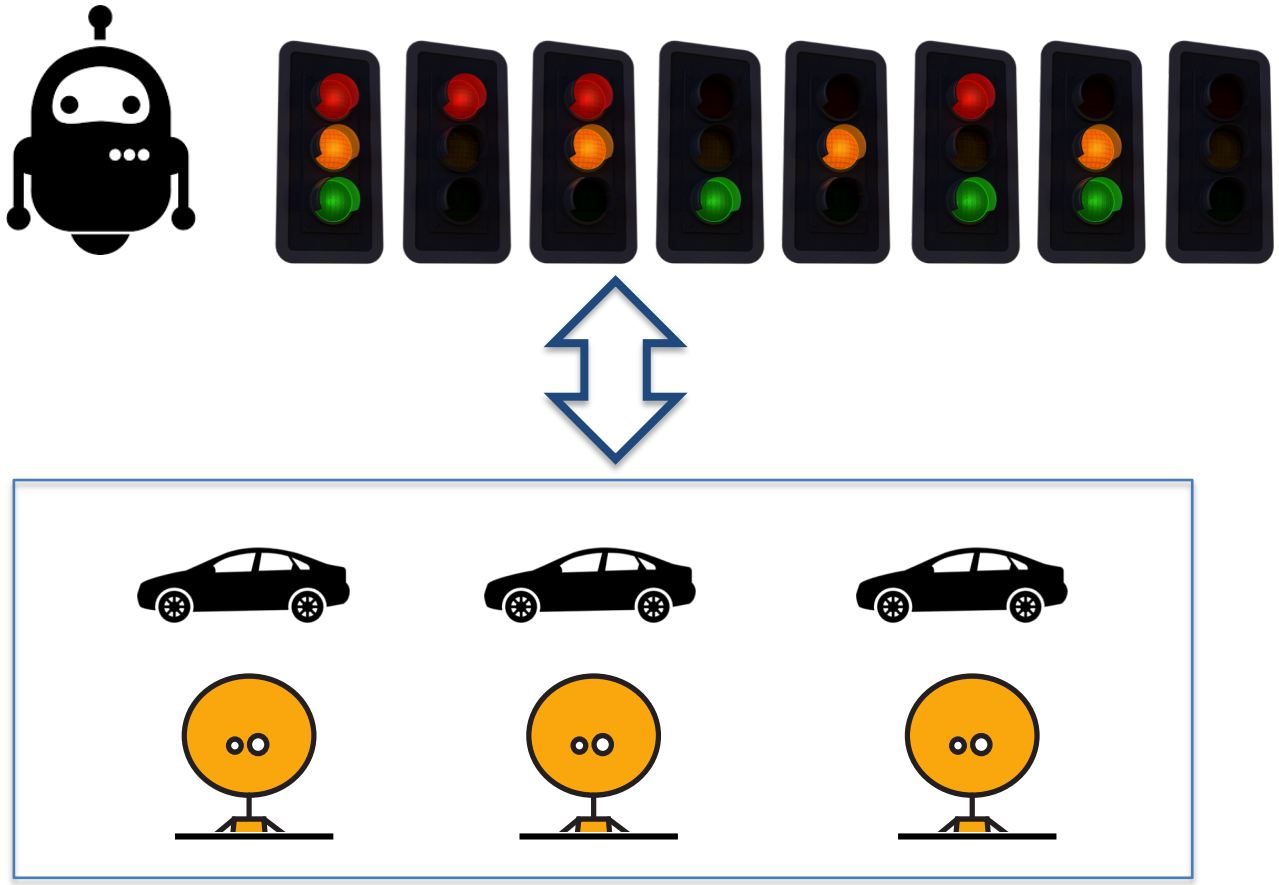
Multi-Agent Sequential Decision Making

AlphaGo

Great success of RL in games…

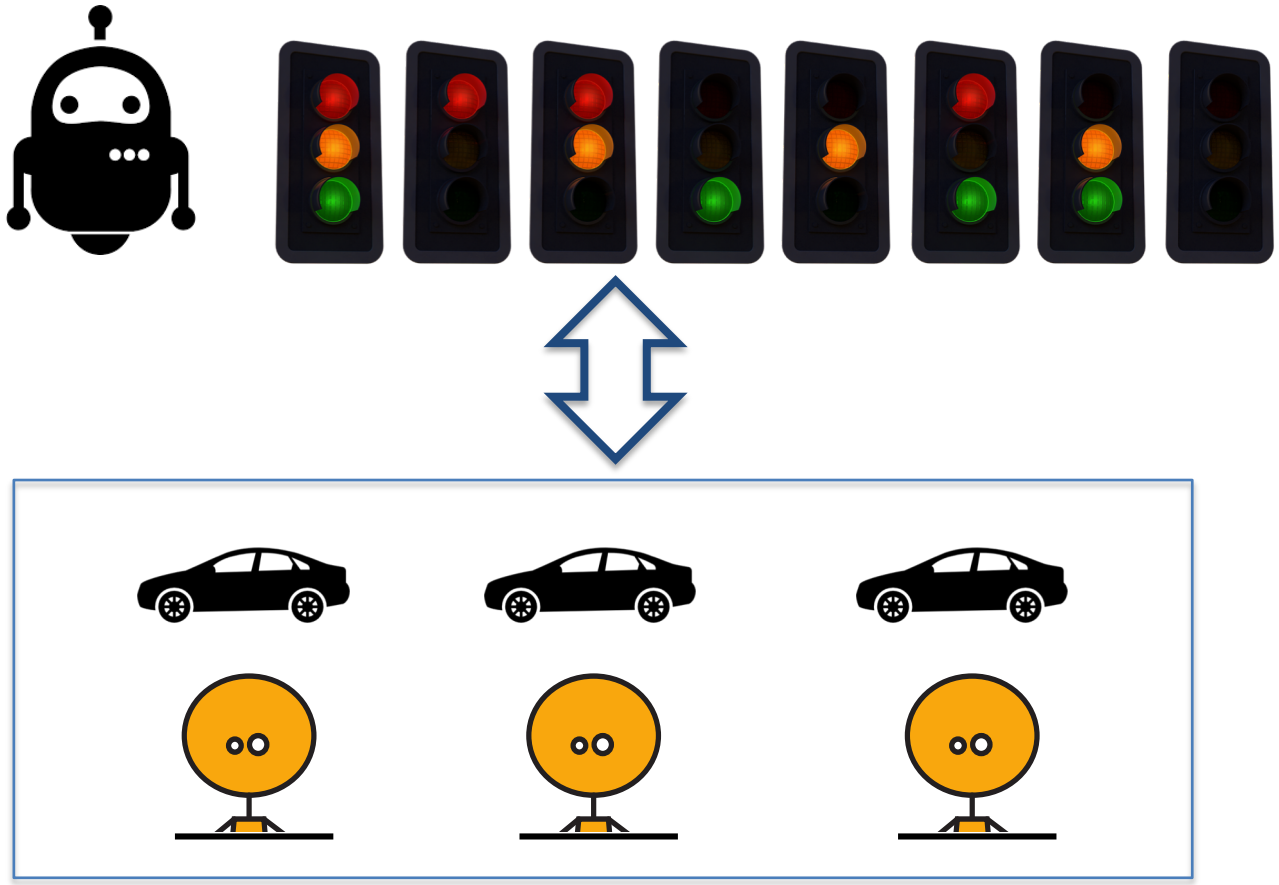
Real world systems?

Challenge: Responsive Environments



Responsive (Newcomblike) Environment

Challenge: Responsive Environments



Responsive (Newcomblike) Environment

Performative RL

Debmalya

Challenge: Attributing Responsibility

Suzy

Billy

Computer

The New York Times

Driver Charged in Uber's Fatal 2018 Autonomous Car Crash

Investigators said the woman had been watching a video on her phone when the vehicle killed a pedestrian in Arizona.

Challenge: Attributing Responsibility

Suzy

Billy

Computer

The New York Times

Driver Charged in Uber's Fatal 2018 Autonomous Car Crash

Investigators said the woman had been watching a video on her phone when the vehicle killed a pedestrian in Arizona.

Actual Causality

Stelios

Thank You!

gradanovic@mpi-sws.org