Uncontrollability of Artificial Intelligence

Dr. Roman.Yampolskiy@louisville.edu

Computer Engineering and Computer Science
University of Louisville - ceecs.louisville.edu/ry
Director – CyberSecurity Lab

@romanyam

/roman.yampolskiy
AI Control Problem

How can humanity remain safely in control while benefiting from a superior form of intelligence?

Is the AI control problem:
  Solvable?
  Partially Solvable?
  Unsolvable?
  Undecidable?
Types of Control

• **Explicit** control – Commands are interpreted nearly literally. This is what we have today with many AI assistants such as SIRI and other narrow AIs.

• **Implicit** control – AI has some common sense, but still tries to follow commands.

• **Aligned** control – AI relies on its model of the human to understand intentions behind the command and uses common sense interpretation of the command to do what human probably hopes will happen.

• **Delegated** control – A superintelligent and human-friendly system which knows better what should happen to make the human happy and keep them safe, AI is in control.
Uncontrollability

Figure 1: Control and Autonomy curves as Capabilities of the system increase.