Universal Intelligence: A Definition of Machine Intelligence

Spring 2021 CSE 6369 HLAI UTA

Ogenna Esimai

February 05, 2021

Outline

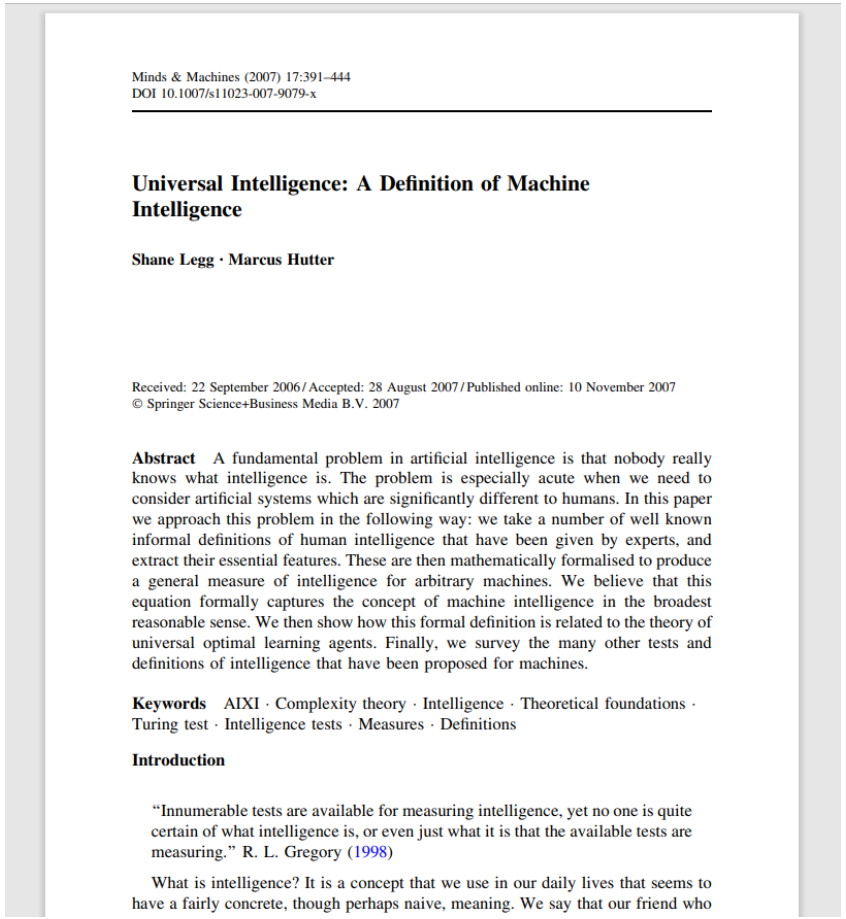
- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Outline

- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Disclaimer

- 54-page paper
- Time constraint
- Main contributions
- Further details as time allows



Minds & Machines (2007) 17:391–444
DOI 10.1007/s11023-007-9079-x

Universal Intelligence: A Definition of Machine Intelligence

Shane Legg · Marcus Hutter

Received: 22 September 2006 / Accepted: 28 August 2007 / Published online: 10 November 2007

Abstract A fundamental problem in artificial intelligence is that nobody really knows what intelligence is. The problem is especially acute when we need to consider artificial systems which are significantly different to humans. In this paper we approach this problem in the following way: we take a number of well known informal definitions of human intelligence that have been given by experts, and extract their essential features. These are then mathematically formalised to produce a general measure of intelligence for arbitrary machines. We believe that this equation formally captures the concept of machine intelligence in the broadest reasonable sense. We then show how this formal definition is related to the theory of universal optimal learning agents. Finally, we survey the many other tests and definitions of intelligence that have been proposed for machines.

Keywords AIXI · Complexity theory · Intelligence · Theoretical foundations · Turing test · Intelligence tests · Measures · Definitions

Introduction

"Innumerable tests are available for measuring intelligence, yet no one is quite certain of what intelligence is, or even just what it is that the available tests are measuring." R. L. Gregory (1998)

What is intelligence? It is a concept that we use in our daily lives that seems to have a fairly concrete, though perhaps naive, meaning. We say that our friend who

Outline

- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Summary

- Definition
  - Formula
    - measure machine intelligence
  - Collection of definitions
    - human intelligence
  - Features
    - mathematical
- Definition vs. theory
  - theory of universal optimal learning agents
- Survey
  - tests and definitions of machine intelligence

Outline

- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Take-home points (1)

- Collection of definitions -> features
  - a property of an individual who is interacting with an external
    - environment
    - problem
    - situation
  - related to ability to succeed or ‘‘profit’’
    - goal
  - ability to deal with range of possibilities + unanticipated
    - quickly learn and adapt

Take-home points (2)

- Informal working definition of intelligence

Intelligence measures an agent's ability to achieve goals in a wide range of environments

- S. Legg, M. Hutter

Source - Universal Intelligence: A Definition of Machine Intelligence. Legg et al. Minds & Machines (2007) 17:391–444

Take-home points (3)

- Definition/Formula
  - the ***universal intelligence*** of agent π

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_{\mu}^{\pi}$$

    - E, space of all computable reward summable environmental measures with respect to the reference machine *U*
    - K, Kolmogorov complexity function
    - $2^{-K(\mu)}$, universal distribution over the space of all environments E
    - μ, the environment
    - V^{π}_{μ} , value function (agent's ''ability to achieve'')
  - the expected performance of agent π with respect to the universal distribution over the space of all environments

Source - Universal Intelligence: A Definition of Machine Intelligence. Legg et al. Minds & Machines (2007) 17:391–444

Take-home points (4)

- Definition vs. theory of universal optimal learning agents
  - perfect theoretical agent, AIXI
  - Hutter (2005)
  - ***universal intelligence*** of agent π, derived from
  - ‘‘intelligence order relation’’ (Definition 5.14 in Hutter (2005))
  - constructed to reflect equations for AIXI
  - AIXI is not computable due to the incomputability of K
  - AIXI is interesting from theoretical perspective

Take-home points (5)

- Survey/tests and definitions of machine intelligence
  - Turing Test and Derivatives
  - Compression Tests
  - Linguistic Complexity
  - Multiple Cognitive Abilities
  - Competitive Games
  - Collection of Psychometric Tests
  - C-Test
  - Smith’s Test

Take-home points (6)

- Comparison of Machine Intelligence Tests and Definitions

Table 1 In the table ● means "yes", • means "debatable", · means "no", and ? means unknown. When something is rated as unknown that is usually because the test in question is not sufficiently specified

Intelligence test	Valid	Informative	Wide range	General	Dynamic	Unbiased	Fundamental	Formal	Objective	Fully defined	Universal	Practical	Test vs. def.
Turing test	•	·	·	·	•	·	·	·	·	•	·	•	T
Total Turing test	•	·	·	·	•	·	·	·	·	•	·	·	T
Inverted Turing test	•	•	·	·	•	·	·	·	·	•	·	•	T
Toddler Turing test	•	·	·	·	•	·	·	·	·	·	·	•	T
Linguistic complexity	•	●	•	·	·	·	·	•	•	·	•	•	T
Text compression test	•	●	●	•	·	•	•	●	●	●	•	●	T
Turing ratio	•	●	●	●	?	?	?	?	?	·	?	?	T/D
Psychometric AI	●	●	•	●	?	•	·	•	•	•	·	•	T/D
Smith's test	•	●	●	•	·	?	●	●	●	·	?	•	T/D
C-test	•	●	●	•	·	●	●	●	●	●	●	●	T/D
Universal intelligence	●	●	●	●	●	●	●	●	●	●	●	·	D

Source - Universal Intelligence: A Definition of Machine Intelligence. Legg et al. Minds & Machines (2007) 17:391–444

Outline

- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Topics (1)

- Basic Agent–Environment Framework

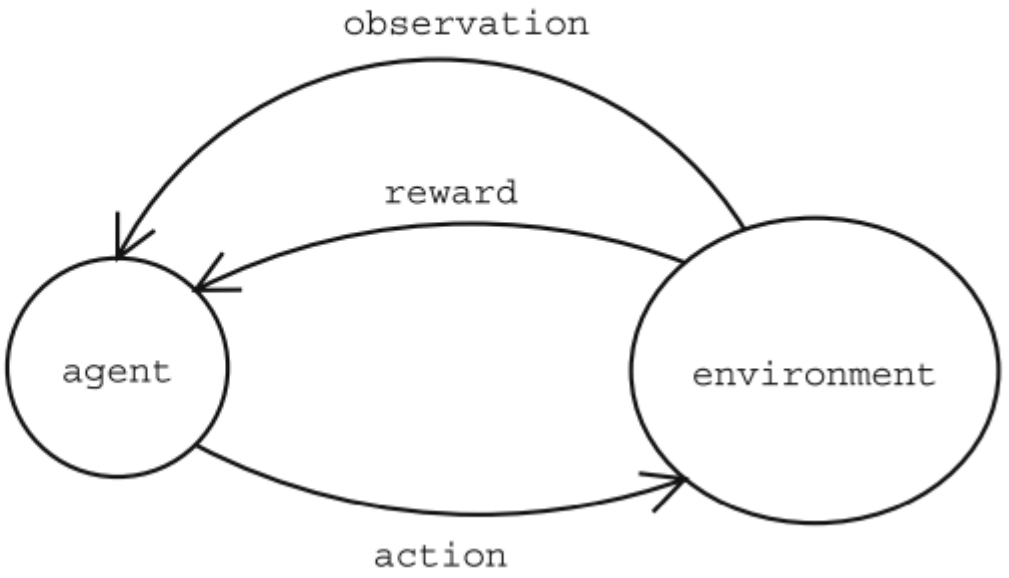


Fig. 1 The agent and the environment interact by sending action, observation and reward signals to each other

Source - Universal Intelligence: A Definition of Machine Intelligence. Legg et al. Minds & Machines (2007) 17:391–444

Topics (2)

- Formal Agent–Environment Framework
  - Agent–Environment Interaction
    - agent (symbols, action space), environment (symbols, perception space)
  - Agent – a function, a probability measure
  - Environment
  - Measure of Success – scale, reward near future vs. reward distant future
  - Space of Environments

Topics (3)

- Universal Intelligence of Various Agents
  - Random – lowest intelligence, uniformly random
  - Very Specialised – very low UI, perform extremely well narrow, complex E
  - General, Simple – basic learning, lookup table (OAR)
  - Simple + More History – correlate current action with previous action
  - Simple, Forward Looking – plan ahead, slide game
  - Very Intelligent – perform well most simple envs, fairly well in many complex
  - Super Intelligent – perfect theoretical agent, AIXI
  - Human - extremely simple envs, should perform well. More complex envs, performance difficult to predict

Topics (4)

- Properties of Universal Intelligence
  - Valid – reasonable, describes something reasonably similar to ‘‘intelligence’’
  - Meaningful – orders agent power, adaptability naturally
  - Informative – real value, independent of perf of other agents
  - Wide range – applies to agents of different levels
  - General – on all well-defined environments
  - Unbiased – not priority/culture , Universal Turing computation
  - Fundamental – computation, information and complexity. Unchanging with tech
  - Formal – mathematical, not much ambiguity
  - Objective – subjective criteria
  - Universal – not anthropocentric
  - Practical – computable, -> test

Topics (5)

- Response to Common Criticisms
  - It’s Obviously False, There’s Nothing in Your Definition, Just a Few Equations
  - It’s Obviously Correct, Indeed Everybody already Knows this Stuff
  - Assuming that the Environment is Computable is too Strong
  - Assuming that Environments Return Bounded Sum Rewards is Unrealistic
  - How Do You Respond to Block’s Argument?
  - How Do You Respond to Searle’s ‘‘Chinese Room’’ Argument?
  - But You Don’t Deal with Consciousness (or Creativity, Imagination, Freewill, Emotion, Love, Soul, etc.)
  - Universal Intelligence is Impossible due to the No-Free-Lunch Theorem

Outline

- Disclaimer
- Summary
- Take-home points
- Topics
- Summary

Summary

- Definition
  - Formula
    - measure machine intelligence
  - Collection of definitions
    - human intelligence
  - Features
    - mathematical
- Definition vs. theory
  - theory of universal optimal learning agents
- Survey
  - tests and definitions of machine intelligence