Artificial Intelligence: An Armchair Philosopher's Perspective

Mark Maloof

Department of Computer Science
Georgetown University
Washington, DC 20057-1232
http://www.cs.georgetown.edu/~maloof

Philosophy and Star Trek (PHIL-180)

20 October 2015

Objective

In this talk, I argue that computers

- have overcome Lady Lovelace's objection
- ground symbols
- have intentional states

And if you act now, you get the Ginsu knife for free!

Outline

- Lower your expectations!
  - (although it is a multimedia presentation, with color even!)
- Out on a limb: Sí, se puede!
- Approaches to AI
- Computation and Turing machines
- Hypercomputation (and pseudo-hyper computation!)
- Philosophy bric-à-brac
- Stanley: A reason to be optimistic
- Bring it on home

The Pitfalls of AI Talks

- All talks on the philosophy of AI fail
- Mine will too
- Why?
  - We haven't succeeded
  - We don't know when or if we'll succeed
  - We don't really even know what *success* means
  - We know how computers work; we don't know how brains work
  - One has to know math, computer science, physics, psychology, neuroscience, and philosophy
- And of course in the end, the robots always rise to kill and enslave their human creators

Sí, se puede!

McCarthy et al., 1955

- "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

Haugeland, 1985

- "The exciting new effort to make computers think...machines with minds, in the full and literal sense."

Charniak and McDermott, 1985

- "...the study of mental faculties through the use of computational models."

Rich and Knight, 1992, 2009

- "The study of how to make computers do things at which, at the moment, people are better."

Nilsson, 1998

- "Artificial intelligence, broadly (and somewhat circularly) defined, is concerned with intelligent behavior in artifacts. Intelligent behavior, in turn, involves perception, reasoning, learning, communicating, and acting in complex environments."

Russell and Norvig's Four Approaches

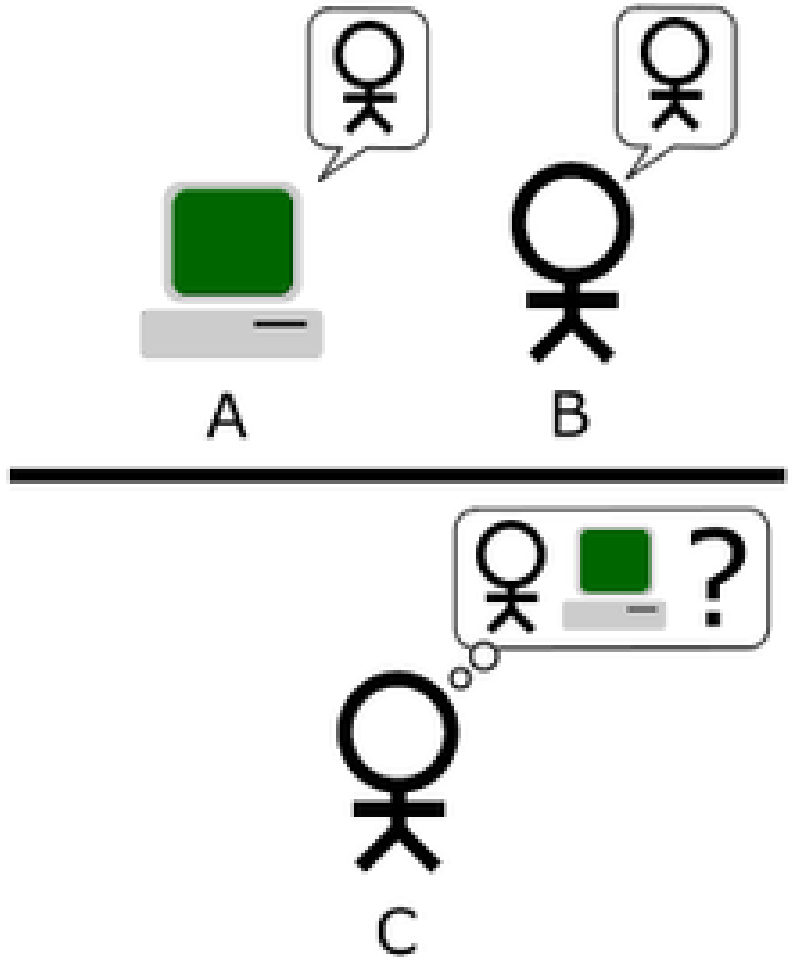
1. Think like a human
2. Act like a human
3. Think rationally
4. Act rationally

Think Like A Human

- "...machines with minds, in the full and literal sense"
- Put simply, program computers to do what the brain does
- How do humans think?
- What is thinking, intelligence, consciousness?
- If we knew, can computers do it, think like humans?
- Does the substrate matter, silicon versus meat?
- Computers and brains have completely different architectures
- Is the brain carrying out computation?
- If not, then what is it?
- Can we know ourselves well enough to produce intelligent computers?

Act Like A Human

Turing Test



Source: http://en.wikipedia.org/wiki/Turing_test

Obligatory xkcd Comic

Source: http://xkcd.com/329/

The Brilliance of the Turing Test

- Sidesteps the hard questions:
  - What is intelligence?
  - What is thinking?
  - What is consciousness?
- If humans can't tell the difference between human intelligence and artificial intelligence, then that's it
- Proposed in 1950, Turing's Imitation Game is still relevant

Think Rationally

- Think rationally? Think logic!
- Put simply, write computer programs that carry out logical reasoning
  - Logic: propositional, first-order, modal, temporal, ...
  - Reasoning: deduction, induction, abduction, ...
- Possible problem: Humans don't really think logically
- Do we care? Strong versus weak AI
- One problem: often difficult to establish the truth or falsity of premises
- Another: conclusions aren't strictly true or false

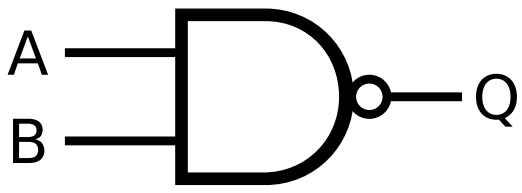
Act Rationally

- Act rationally? Think probability and decision theory!
- "A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome" (Russell and Norvig, 2010, p. 4)
- <jab>"when there is uncertainty"</jab>
- When *isn't* there uncertainty?
- Predominant approach to AI (for now)

Computation

- Everything in a computer is binary: 0 or 1
- Start with one wire and two voltage levels:
  - 0–2 volts $\Rightarrow$ 0
  - 3–5 volts $\Rightarrow$ 1
- Take one wire, one binary digit, or one bit
- What can you do?
  - change 0 to 1
  - change 1 to 0
- Not very interesting, but wait! There's more!
- This state change is computation at its most basic level

Computation: Beautiful NAND



inputs		output
A	B	Q
0	0	1
0	1	1
1	0	1
1	1	0

NAND: What's the big deal?

- It is functionally complete
- Meaning: Anything computable can be computed using only NAND gates
- This is not controversial
- It's descriptive, but it's not constructive
  - Tells you *that*, but not *how*
- So is the brain carrying out computation?
- That's the difficult question
- You can't just answer no
- You have to explain that not-computation process
- That's even more difficult

Computation: It's a bit more complicated

$$\text{NANDs} \equiv \text{Computers} \equiv \text{Programs} \approx \text{Algorithms} \equiv \text{Turing Machines}$$

$$\text{Formal-Symbol Systems} \equiv \text{Physical-Symbol Systems} \equiv \text{Turing Machines}$$

Hypercomputation

- "The new field of hypercomputation studies models of computation that can compute more than the Turing machine and addresses their implications" (Ord, 2002)
- Computers $\approx$ Turing machines $<$ Hypercomputers

- On the other hand, "...there is no such discipline as hypercomputation" (Davis, 2006)
- Furthermore, Turing was not an idiot

Hypercomputation in a Nutshell

- Computers and Turing machines are digital (i.e., binary)
- The brain is analog (i.e., continuous)
  - what about spike trains?
- Digital is only an approximation to analog
  - yeah, but, sampling theorems!
- *Approximation* matters for some people
  - are we watching reality or just a movie?
- For some *approximation* means Turing machines can't be minds
- Perhaps a device carrying out hypercomputation could
- But there are not yet any sufficiently powerful hypercomputers
- ...except, of course, the brain
- That is, brains perform hypercomputation; Turing machines can not; therefore, Turing machines can not be minds

The Chinese Room

- Searle argues that formal systems are not minds
- Takeaway: The Chinese symbols have no meaning to the person in the room
- "Hey! Chinese Room! How many questions have I asked?"
  - can the Room count?
  - counting rules must be in English
  - what would Searle understand?
  - if the Room can not count, then it's not a Turing machine
- Don't we also have to argue that minds are not formal systems?
- Where is the meaning in
  - a release of γ-aminobutyric acid?
  - a neuron?
  - a synapse?
  - a spike train?

Lady Lovelace's Objection

- Lady Ada Lovelace worked with Charles Babbage on his Difference Engine, a mechanical computer
- Worked also on the Analytical Engine, a mechanical computer that was never built
- Regarded as the first programmer
- (October 14 was Ada Lovelace Day)
- She remarked that the machine "has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths"
- Known as Lady Lovelace's objection to artificial intelligence (Turing, 1950)

Intentional States

- "Intentionality is the power of minds to be about, to represent, or to stand for, things, properties and states of affairs" (Pierre, 2014)
- the power of minds. . . to represent things. . .
- Can computers or robots form representations of things in the external world?

Symbol-Grounding Problem

In direct response to the Physical Symbol System Hypothesis (Newell and Simon, 1976), Harnad (1990) asks:

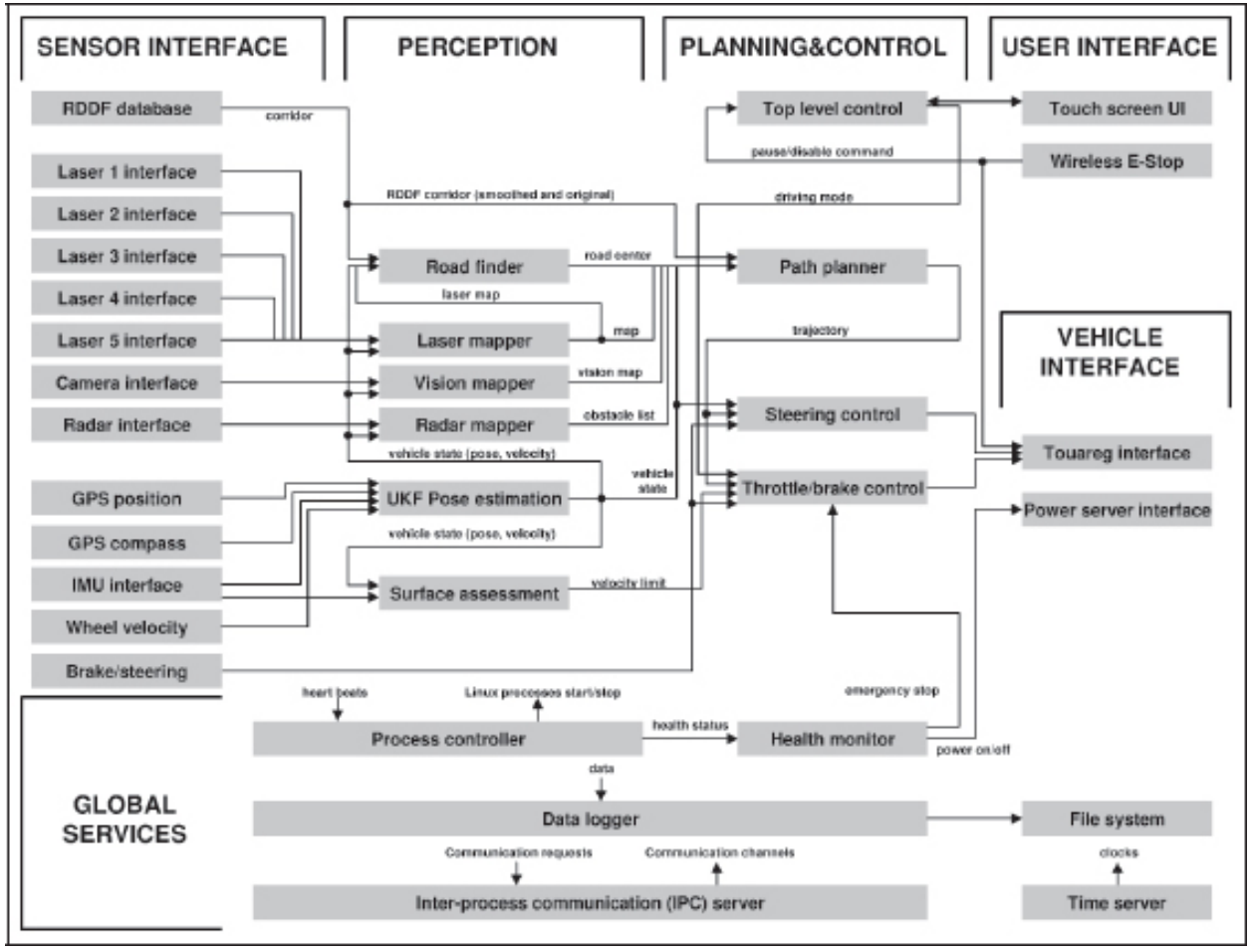
- "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?"
- "How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?"
- Computers $\approx$ FSSs $\equiv$ PSSs $\equiv$ Turing Machines
- Again, is there is meaning everywhere in the brain?
- By the way, Steels (2008) claims the SGP is solved

Stanley: A Reason to be Optimistic

- A self-driving car, a precursor to Google's self-driving car
- In 2005, drove a 175-mile course in the Mojave Desert
- Unaided by humans, who had only two-hours prior notice of the route
- Stanley used terrain maps to plan its overall route
- As it drove, it relied on its own analysis of "analytical relations and truths" to anticipate what lay ahead, by navigating the road itself, assessing its condition, and avoiding obstacles

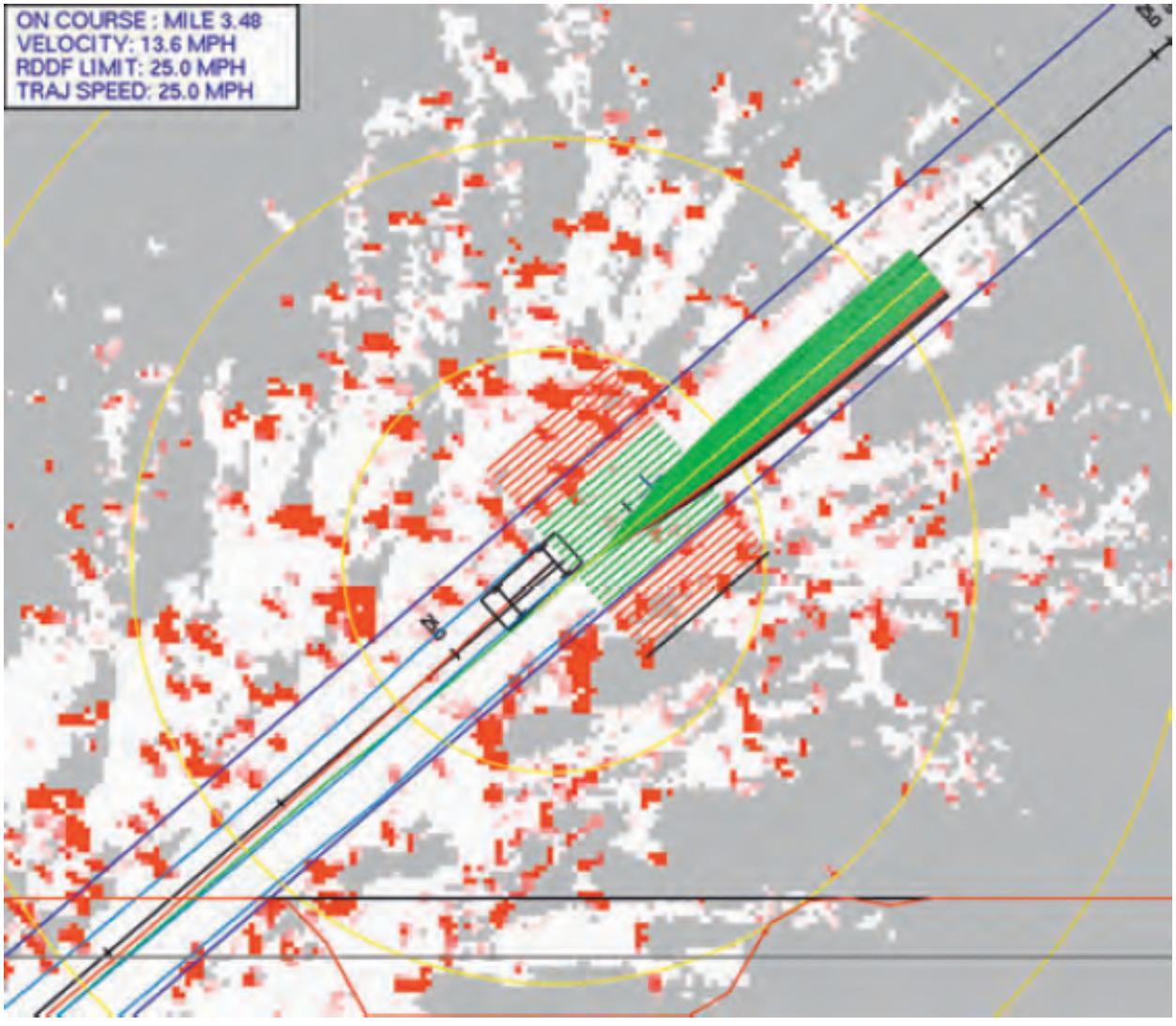
Video: The Great Robot Race

Stanley



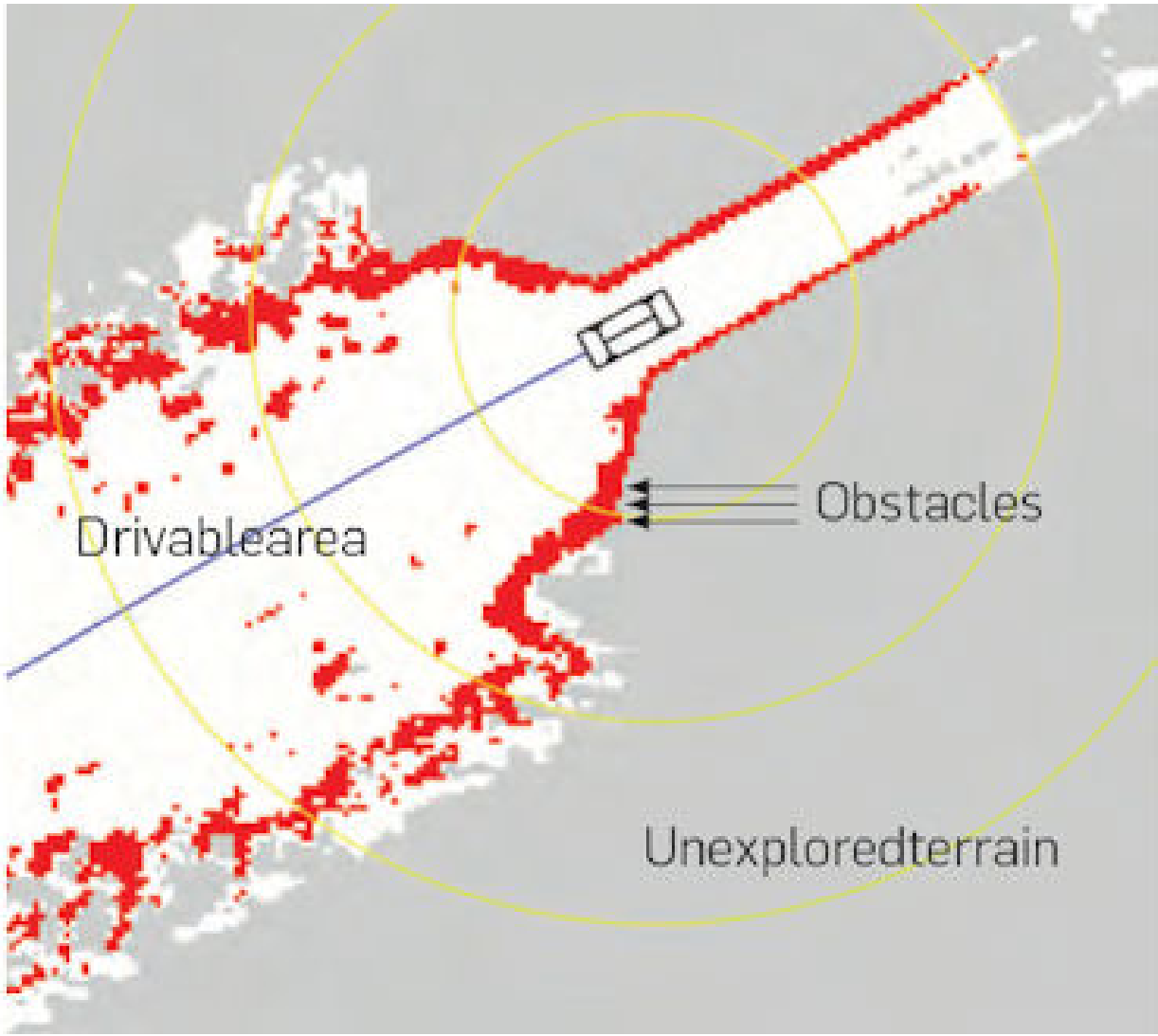
Source: Thrun (2010, Figure 2)

Stanley



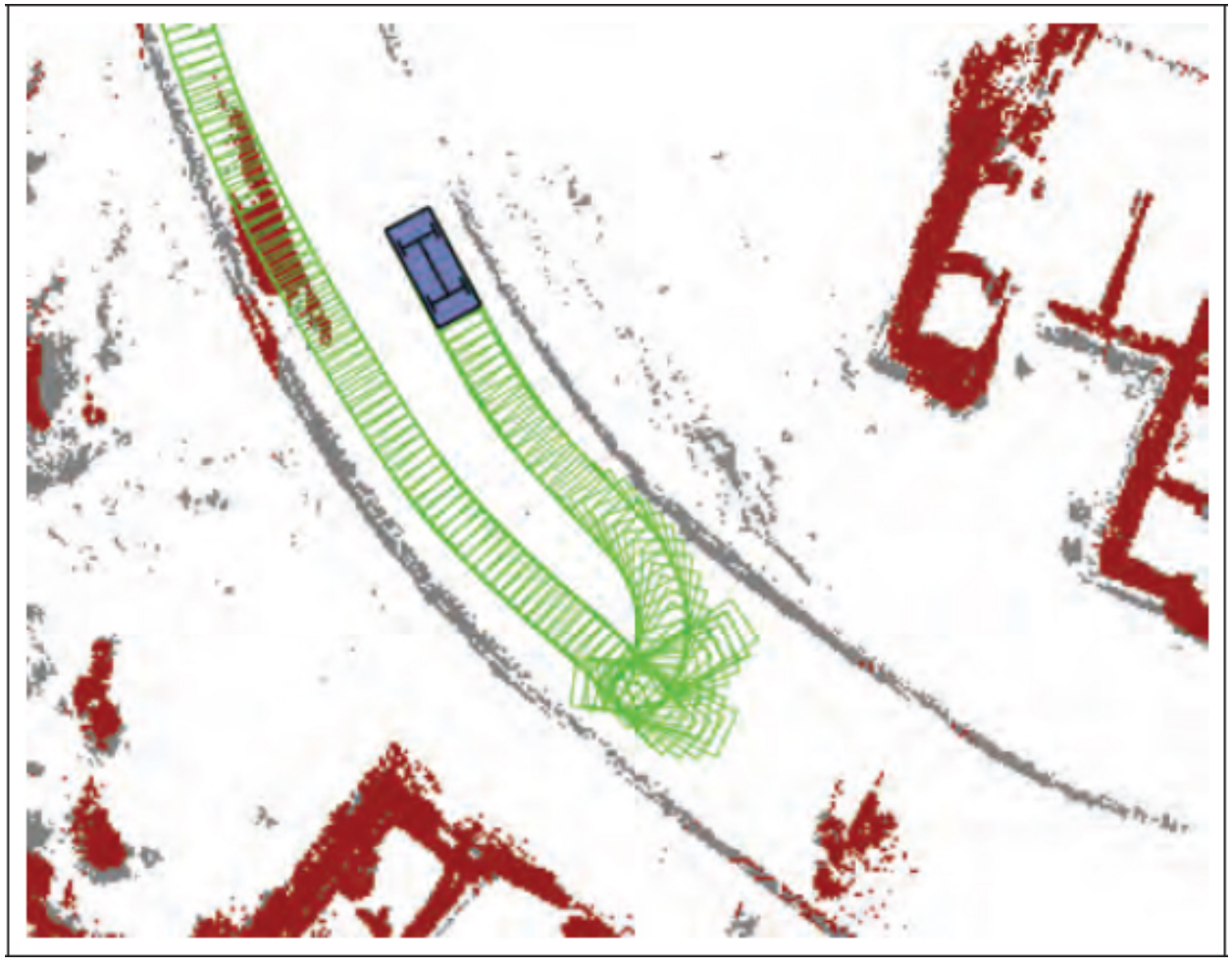
Source: Thrun (2010, Figure 7)

Stanley



Source: Thrun (2010, Figure 9a)

Stanley



Source: Thrun (2010, Figure 13)

Bring it on Home

- Sí, se puede!
- Stanley refutes Lady Lovelace's objection
  - no one programmed it to avoid *that* obstacle in the desert
- Stanley grounds symbols
  - it associates semantic representations with objects in the external world
- Stanley has intentional states
  - it has beliefs about objects in the external world
- Does Stanley know that it knows about obstacles?

What I Told You

- I'm doomed to fail!
- Went on a limb: Sí, se puede!
- Approaches to AI
- Computation and Turing machines
- Hypercomputation (and pseudo-hyper computation!)
- Philosophy bric-à-brac
- Stanley: A reason to be optimistic
- Brought it on home

A Parting Shot: Tesler's Theorem

- "Intelligence is whatever machines haven't done yet."
- Commonly quoted as "AI is whatever hasn't been done yet."

Questions?

Artificial Intelligence: An Armchair Philosopher's Perspective

Mark Maloof

Department of Computer Science
Georgetown University
Washington, DC 20057-1232
http://www.cs.georgetown.edu/~maloof

Philosophy and Star Trek (PHIL-180)

20 October 2015

References I

E. Charniak and D. McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA, 1985.

M. Davis. Why there is no such discipline as hypercomputation. *Applied Mathematics and Computation*, 178(1): 4–7, 2006. doi: http://dx.doi.org/10.1016%2Fj.amc.2005.09.066.

S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

J. Haugeland. *Artificial intelligence: The very idea*. MIT Press, Cambridge, MA, 1985.

J. McCarthy, M. I. Minsky, N. Rochester, and C. E. Shannon. A proposal for the Dartmouth summer research project on artificial intelligence, 1955. URL `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`. [Online; accessed 7 August 2014].

A. Newell and H. A. Simon. Computer science as empirical enquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976.

N. J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco, CA, 1998.

T. Ord. Hypercomputation: Computing more than the Turing machine. Technical Report arXiv:math/0209332 [math.LO], arXiv, 2002. URL `http://arxiv.org/abs/math/0209332`. [Online; accessed 8 October 2014].

J. Pierre. Intentionality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, winter 2014 edition, 2014.

E. Rich and K. Knight. *Artificial intelligence*. McGraw-Hill, New York, NY, 2nd edition, 2009.

E. Rich, K. Knight, and S. B. Nair. *Artificial intelligence*. Tata McGraw-Hill, New Delhi, 3rd edition, 2009.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2010.

L. Steels. The symbol grounding problem has been solved. So what's next? In M. de Vega, A. Glenberg, and A. Graesser, editors, *Symbols and embodiment: Debates on meaning and cognition*. Oxford University Press, Oxford, 2008. URL `http://www.csl.sony.fr/downloads/papers/2008/steels-08d.pdf`.

S. Thrun. Toward robotic cars. *Communications of the ACM*, 53(4):99–106, 2010. URL `http://cacm.acm.org/magazines/2010/4/81485-toward-robotic-cars/`.

A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. URL `http://mind.oxfordjournals.org/content/LIX/236/433`.