
CURIOUS
ACTORS
A STRONG AI EXPERIMENT
@lemiffe
lemiffe.com

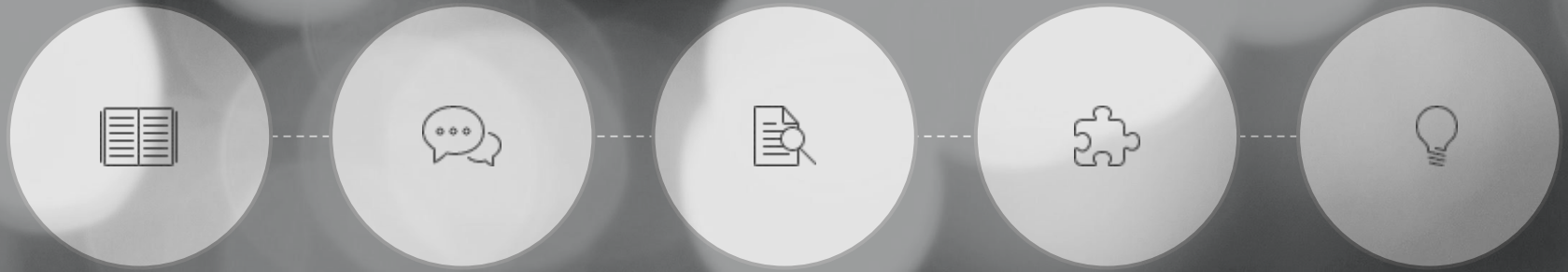

STRONG AI IS STILL FAR,
FAR AWAY

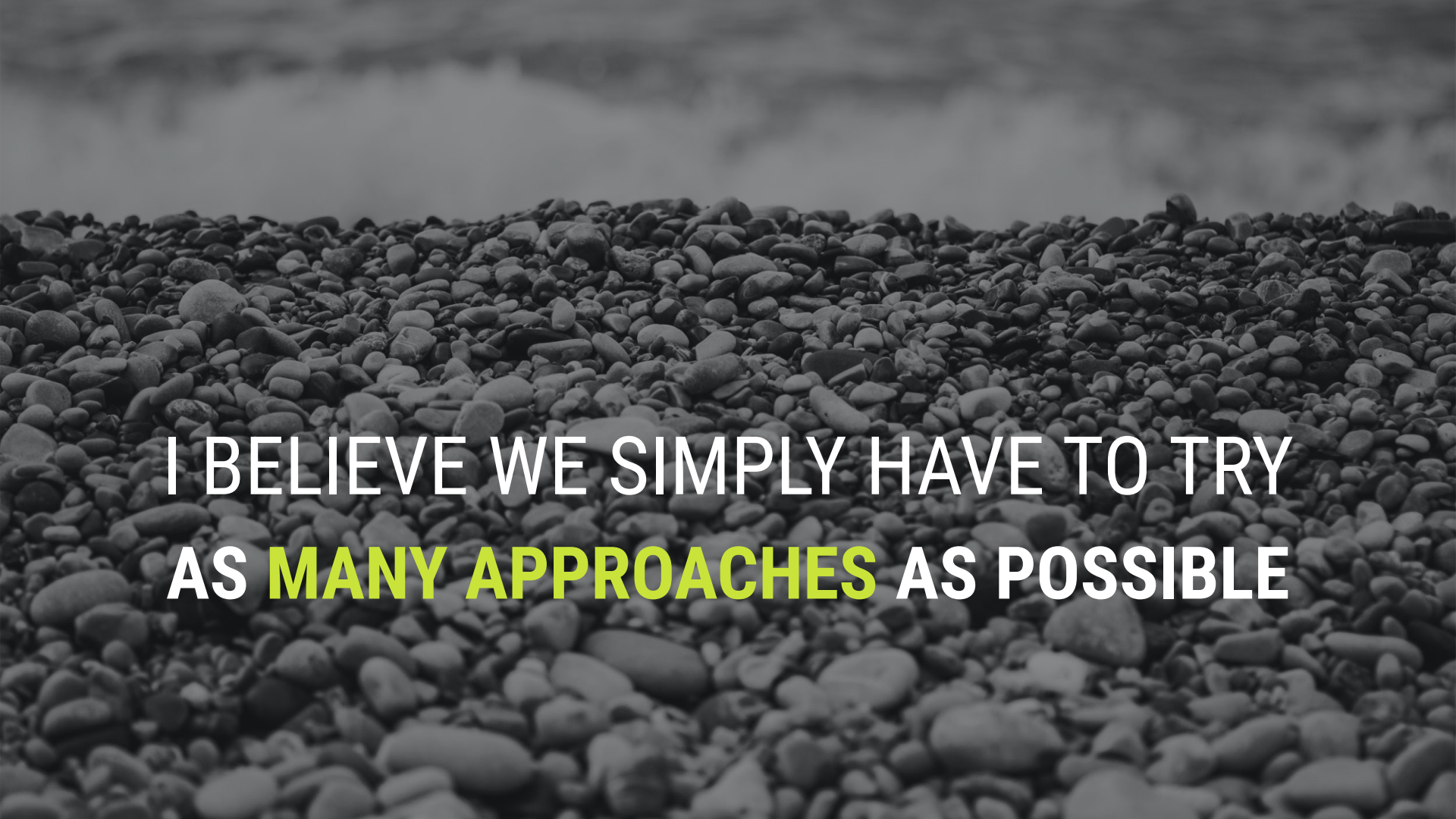
SOME PEOPLE BELIEVE THAT BY

FOCUSING ON THE PARTS

(*SOFT* AI) WE WILL SOME DAY BE

ABLE TO **ASSEMBLE THEM TOGETHER**

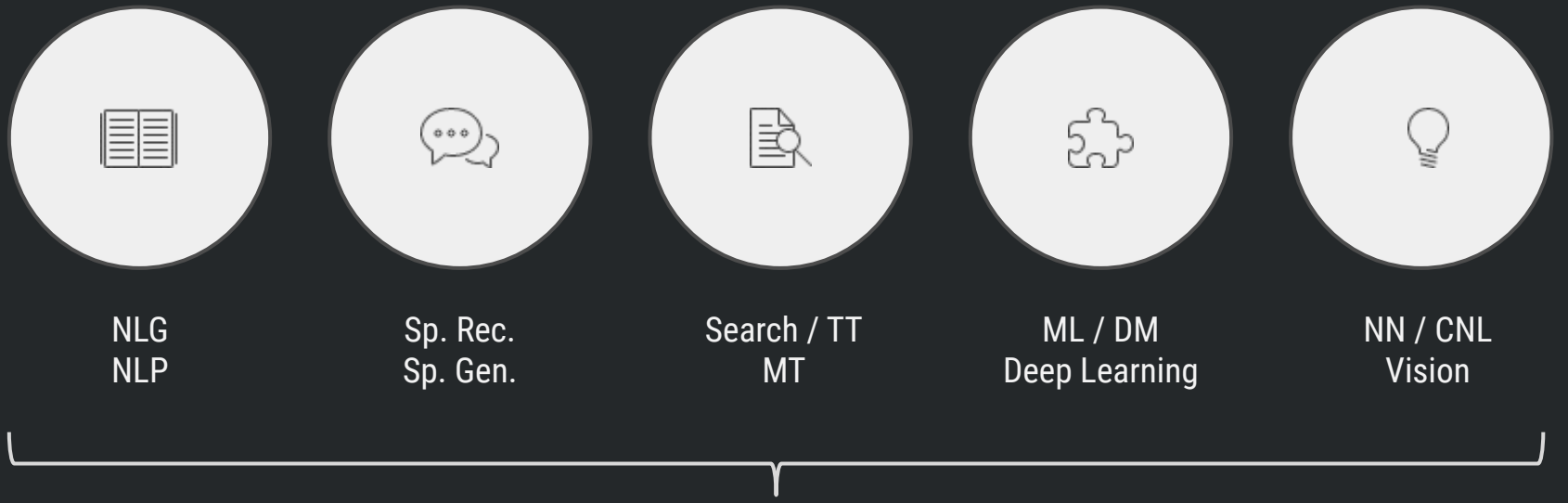



I BELIEVE WE SIMPLY HAVE TO TRY
AS MANY APPROACHES AS POSSIBLE


THE TASK IS
MASSIVE


BUT WE HAVE TO START
SOMEWHERE

First let’s analyse the **current state of AI**...



There has been **major progress** in most individual areas of **soft AI**

But who is working on **strong AI**?

NOT MANY

ACCORDING TO QUORA AND A FEW BLOGS, THERE ARE 8-14 ACTIVE COMPANIES + PHDs

MOSTLY COMMERCIALLY-VIABLE PROJECTS

WILL IT GENERATE DIRECT REVENUE? COST IS HIGH, LONG-TERM BENEFITS.

INDEPENDENT (OR FUNDED) RESEARCHERS

JEFF HAWKINS, ANDREW BARTO, NICK CASSIMATIS, AND OTHER ML/COG-SCI RESEARCHERS

STANFORD AI100

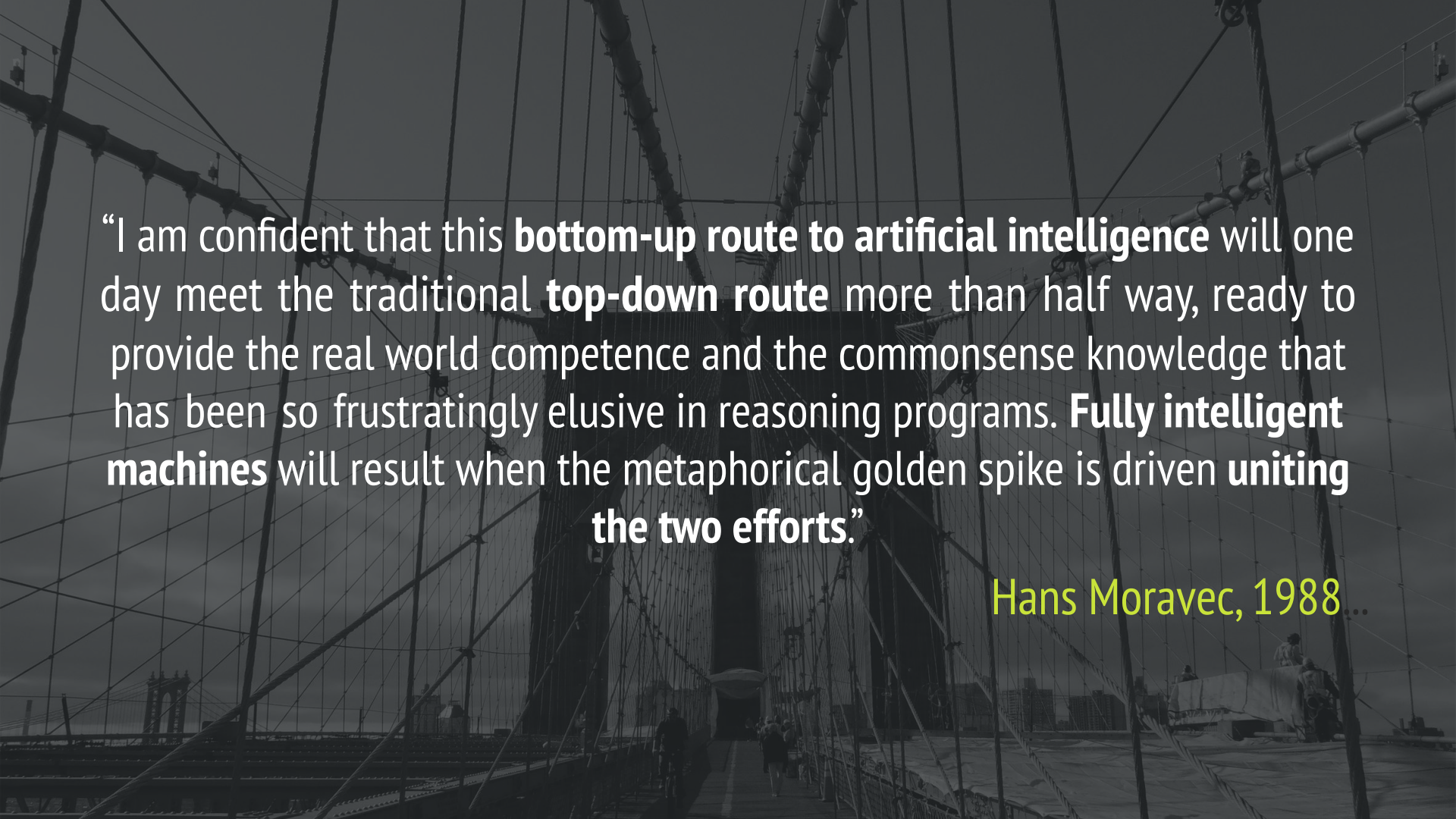
100 YEAR STUDY STARTED BY STANFORD (EFFECTS/IMPLICATIONS OF AI ON SOCIETY)

FACEBOOK AI LAB

DIRECTED BY YANN LECUN (MOSTLY DEALS WITH DEEP LEARNING)

MUCH TALK... LITTLE ACTION

“WHAT BETTER PLACE THAN HERE, WHAT BETTER TIME THAN NOW?”



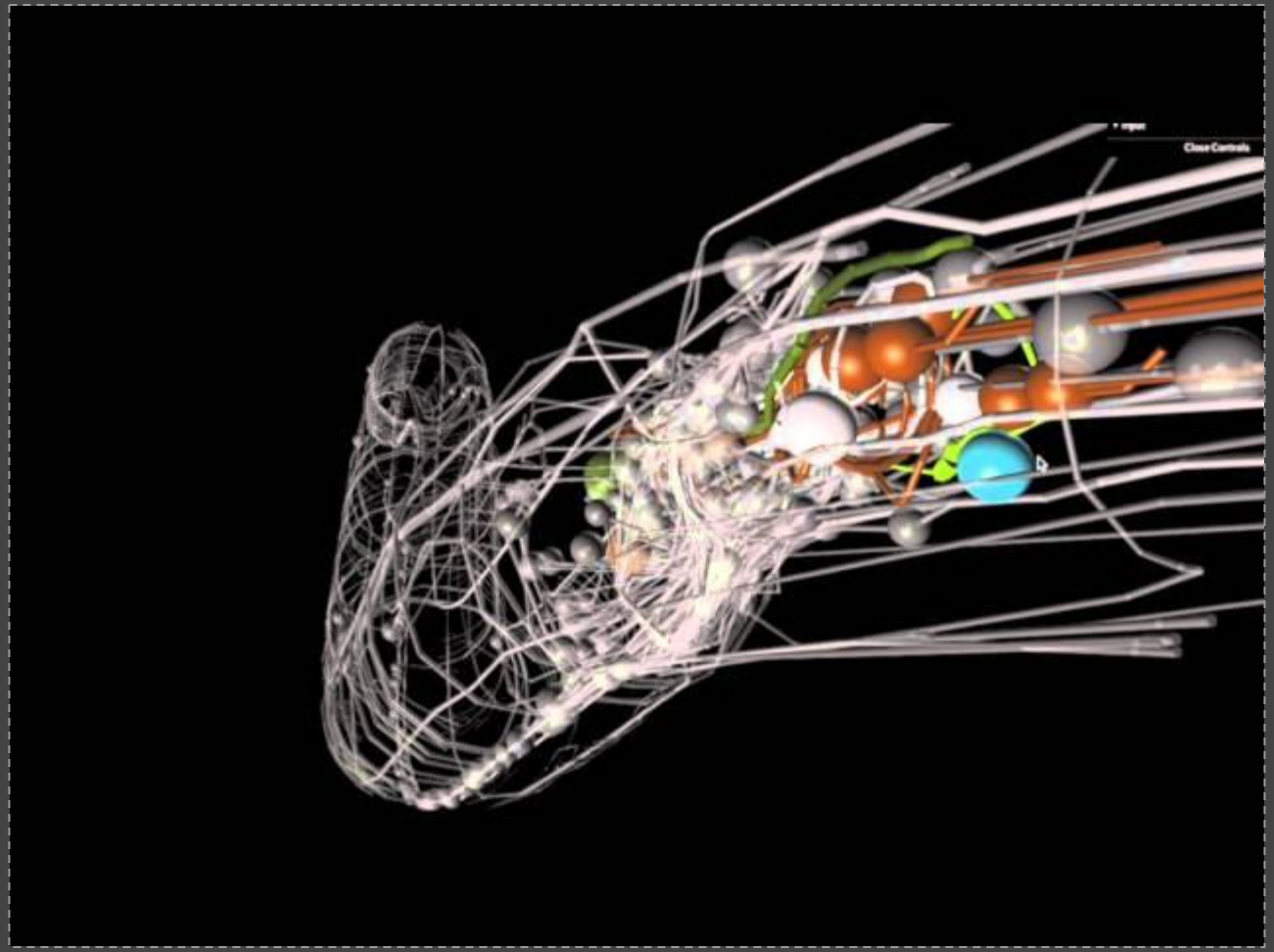
"I am confident that this **bottom-up route to artificial intelligence** will one day meet the traditional **top-down route** more than half way, ready to provide the real world competence and the commonsense knowledge that has been so frustratingly elusive in reasoning programs. **Fully intelligent machines** will result when the metaphorical golden spike is driven **uniting the two efforts**."

Hans Moravec, 1988 ...

Is mapping the entire brain the solution?

OPENWORM

VIRTUAL C. ELEGANS

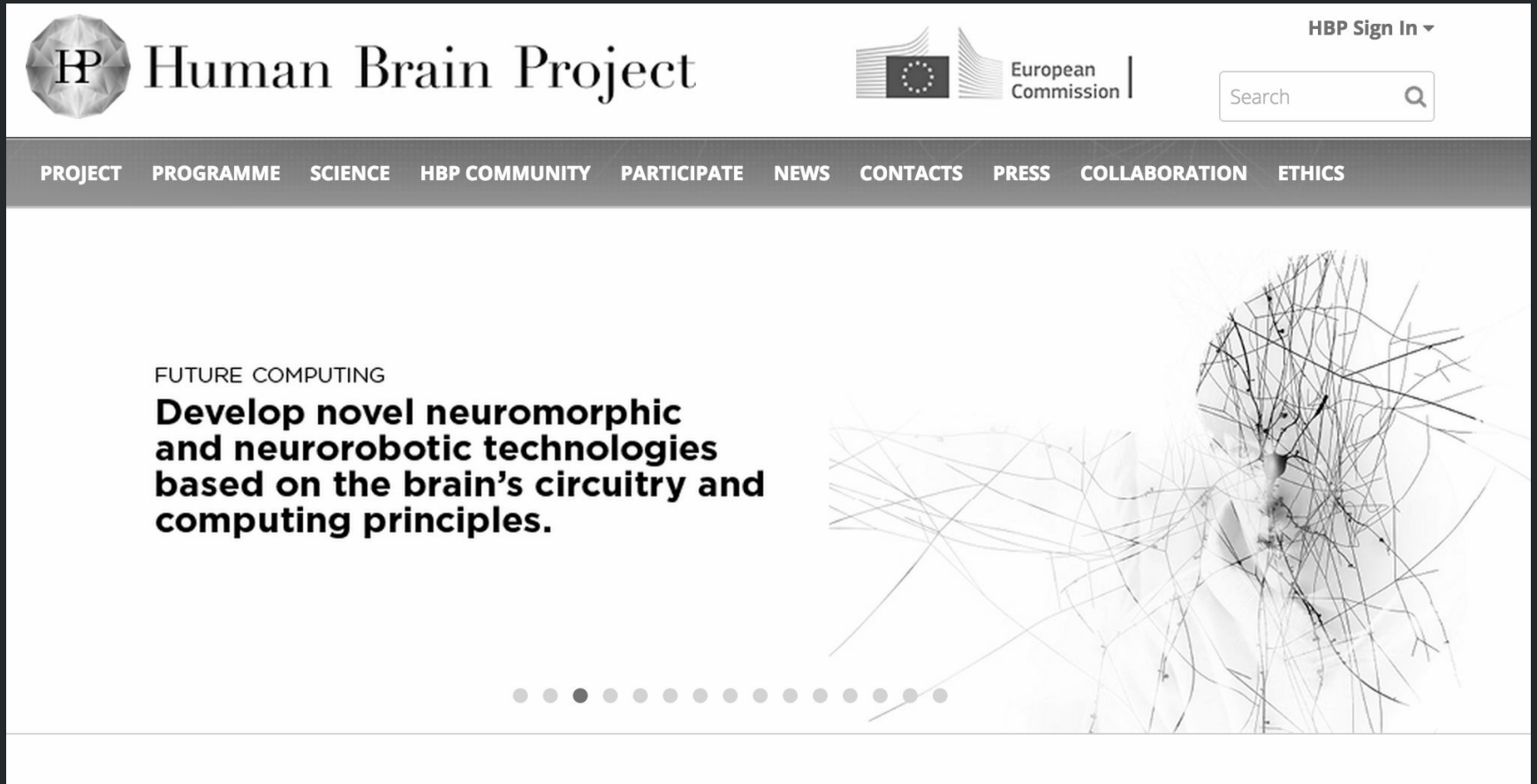


Kickstarter-backed

http://openworm.org

github.com/openworm

THE **HUMAN BRAIN PROJECT** HAS SIMILAR GOALS FOR THE HUMAN BRAIN



BUT IS IT THE **RIGHT PATH** TOWARDS **STRONG AI**? MAYBE, MAYBE NOT

Updated: European neuroscientists revolt against the E.U.'s Human Brain Project

Martin is a contributing news editor and writer based in

By Martin Enserink and Kai Kupferschmidt | 11 July 2014 11:45 am | 34 Comments

***Update, 11 July, 11:58 a.m.:** HBP's Board of Directors and its Executive Committee have responded to the the open letter in a 4-page statement released yesterday. They say they are "saddened" by the letter and say that cognitive neuroscience will still be a part of the HBP's Partnering Projects. The statement expresses the hope that HBP will unite the neuroscience, medical, and computing communities.*

An influential group of European neuroscientists is threatening to boycott the Human Brain Project (HBP), the hugely ambitious plan to map the

CURIOUS ACTORS

REQUIREMENTS SPECIFICATION

Version List:

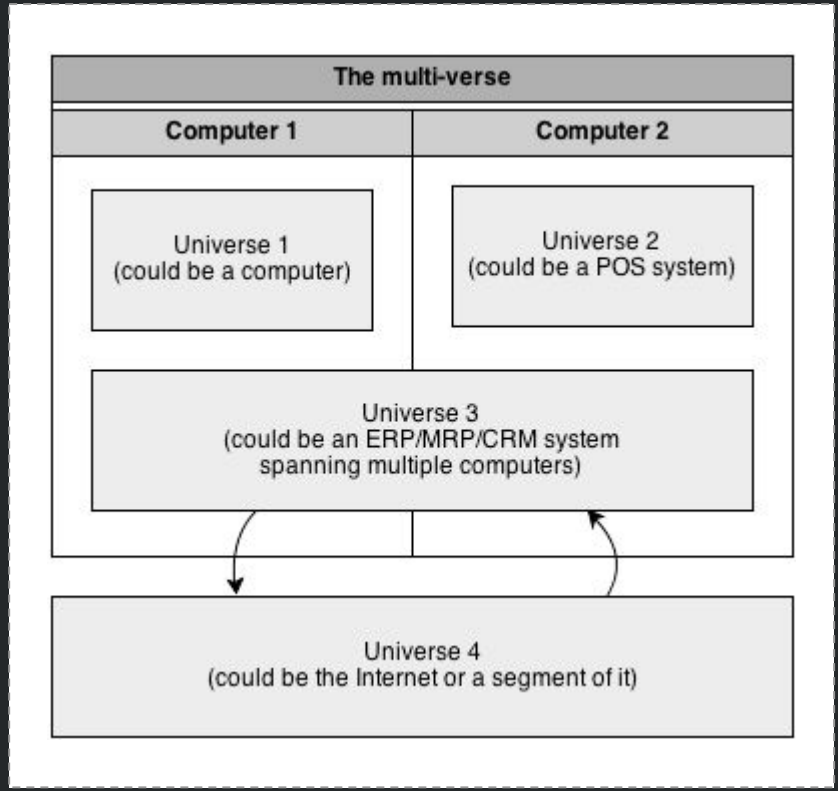
0.1 - Feb 2014
0.2 - Dec 2014 (@iainb)
0.3 - Jan 2015 (@phpbenelux)
0.4 - Feb 2015 (SP)


#!/bin/sh
for i in who what where when why how; do
echo "But $i, dad?"
done
I hereby submit this project to the /. community under the GPL v2.
SEE THE DISCUSSION ON SLASHDOT
j.mp/programming-curiosity

Can we build autonomous bots that exhibit curiosity?

THE MULTIVERSE

- Contains ‘Universe’ (UV) objects
- Actors/objects live in UVs
- UVs have rules for merging
- UVs have rules for splitting
- Merges/splits are manual!



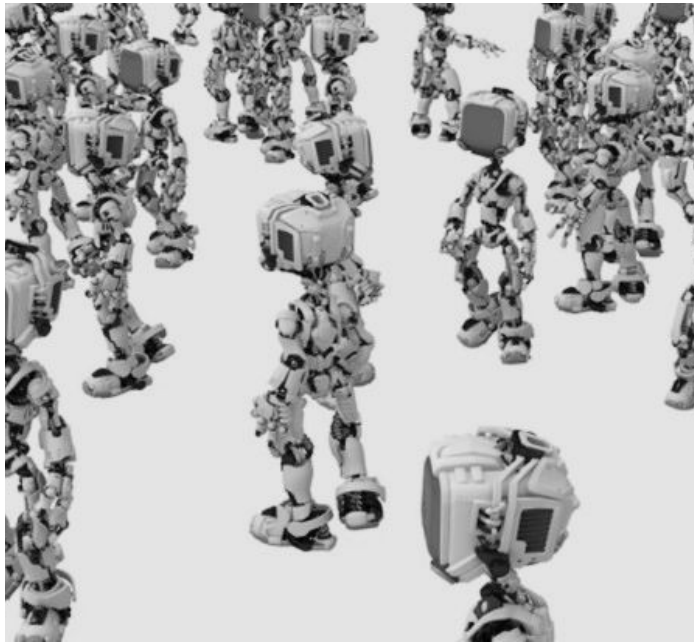
A UNIVERSE (UV)

- Has a set of rules
- Rules may be dynamic
- Can consist of anything
- Init = translation to objects
- Multiple concurrent UVs
- Can merge/split from other UVs
- Two base types
  - Spatial
  - Structured

```
{
  “objects”: [
    {
        “url”: “http://go.com”,
        “type”: “link”,
        “protocol”: “http”,
        “parser”: “parse.py”
    }
  ]
}
```

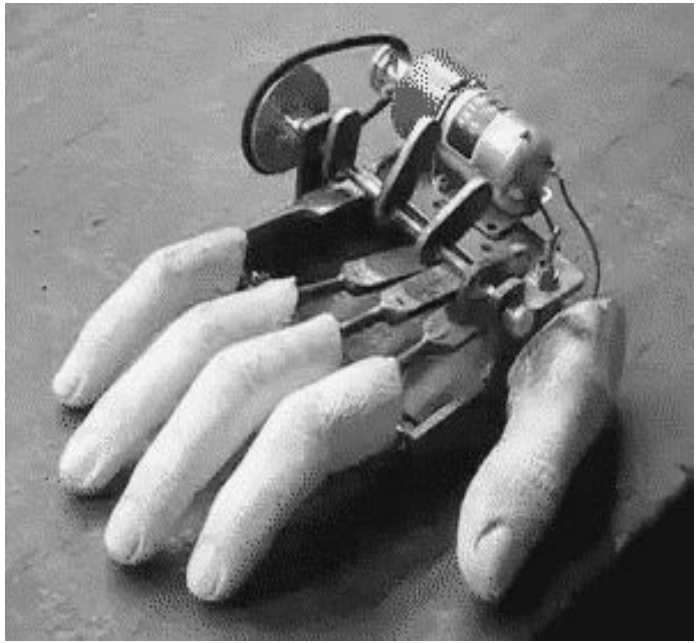
CURIOUS ACTORS

- Bots that live in any UV
- Independent from the UV
- One process per bot
- N bots per UV
- Multiple spawn options
- Init = parse/load genetics.json
- Can reproduce
- Can be persisted
- Naturally curious



GUIDING PRINCIPLES

- Exploration
- Communication
- Collaboration
- Desire to do things
- 3 laws of robotics



Examples

#1 - SITE AVAILABILITY

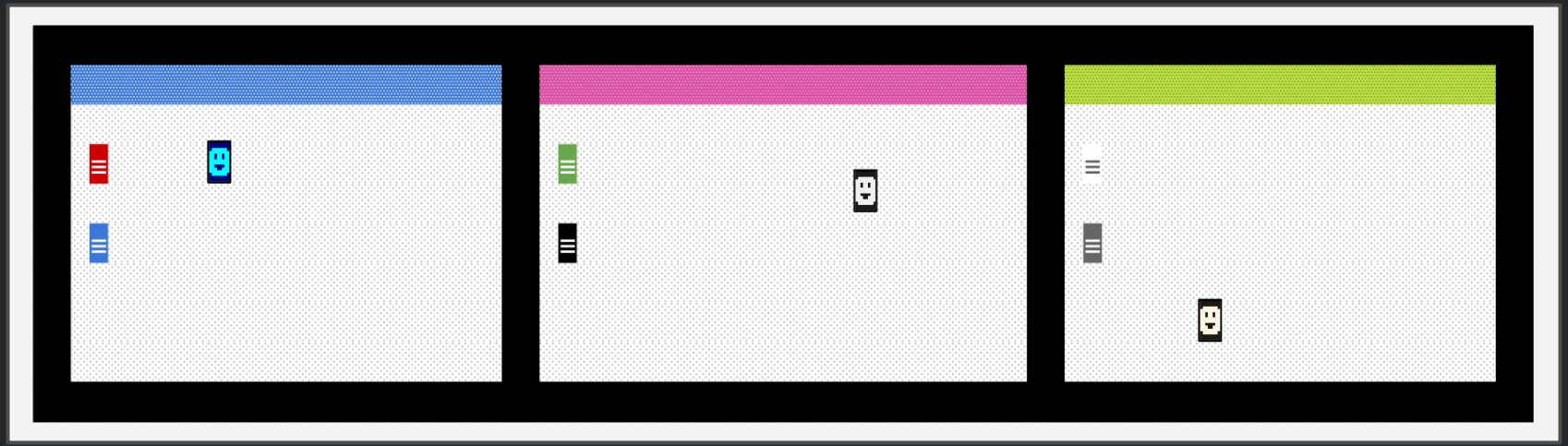
- The universe:
  - 1 object with a property that contains a bunch of links
  - 1 object with a list of HTTP status codes and their meaning (positive/negative)
- The bot can traverse and open links by learning the rules
- The actor forms a knowledge base
- Each new insight runs through ML / clustering algorithms
- It can determine the type of new objects
- Actors can interact with others to expand their knowledge
- Results can be reported / exported

#2 - FILE STRUCTURE ANALYSIS

- Universe: Set of (XML?) documents (with methods to parse/traverse)
- Actors learn how to traverse the documents
- Actors can identify XML documents that are not parsable
- Actors can also identify possible **mistakes with values**

#3 - THE CASE OF THE “WRONG” BOOKS

- The universe consists of a 2D map with 3 enclosed rooms
- Each room contains 1 actor and two ‘books’



OBJECTS PER ROOM

```
“Room 1”: [
    {type: 'book', color: 'red'}, {type: 'book', color: 'blue'}
],
“Room 2”: [
    {type: 'book', color: 'green'}, {type: 'book', color: 'black'}
],
“Room 3”: [
    {type: 'car', color: 'white'}, {type: 'car', color: 'gray'}
]
```

PHASE 1: EXPLORATION & LEARNING

- **Bot 1 learns**
  - Objects with type 'book' can come in two colours (so far)
  - Books tend to have both properties
- **Bot 2 learns**
  - Objects with type 'book' can come in two colours (so far)
  - It knows the same as bot 1 about property count / keys
- **Bot 3 learns**
  - Objects with type 'car' can come in two colours (so far)

PHASE 2: COMMUNICATION

- **Universe is modified**: All rooms are joined
- Actors may interact/discuss base on a simple protocol
- 3rd bot might go to the other bots and exclaim that the object is a 'car'
- Other bots will declare they are 'books'
- Bots will share their knowledge, and can empirically prove the facts based on the properties

PHASE 3: COLLABORATION

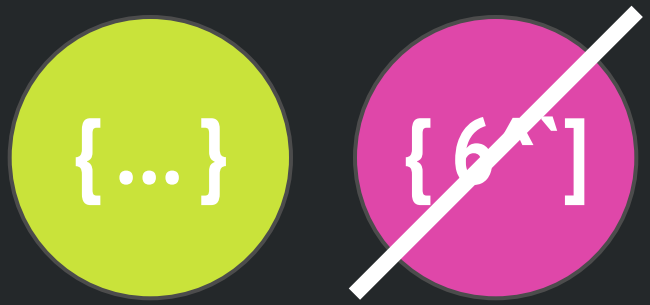
- Bots can interact with the methods of each object freely
- They can see public properties / methods and experiment
- They know the expected result based on their current knowledge base
- Given that these bots have a natural repulsion to anomalies:
- They could experiment until they achieve the **desired result** (or break the object completely / or alter it past the point of repair)
- If a negative state is achieved, they should ignore the bad object.
- **They now know two new things:**
  - That object is [currently] not repairable
  - The properties that each method usually changes (and how to break the object)

ANOTHER POSSIBILITY...

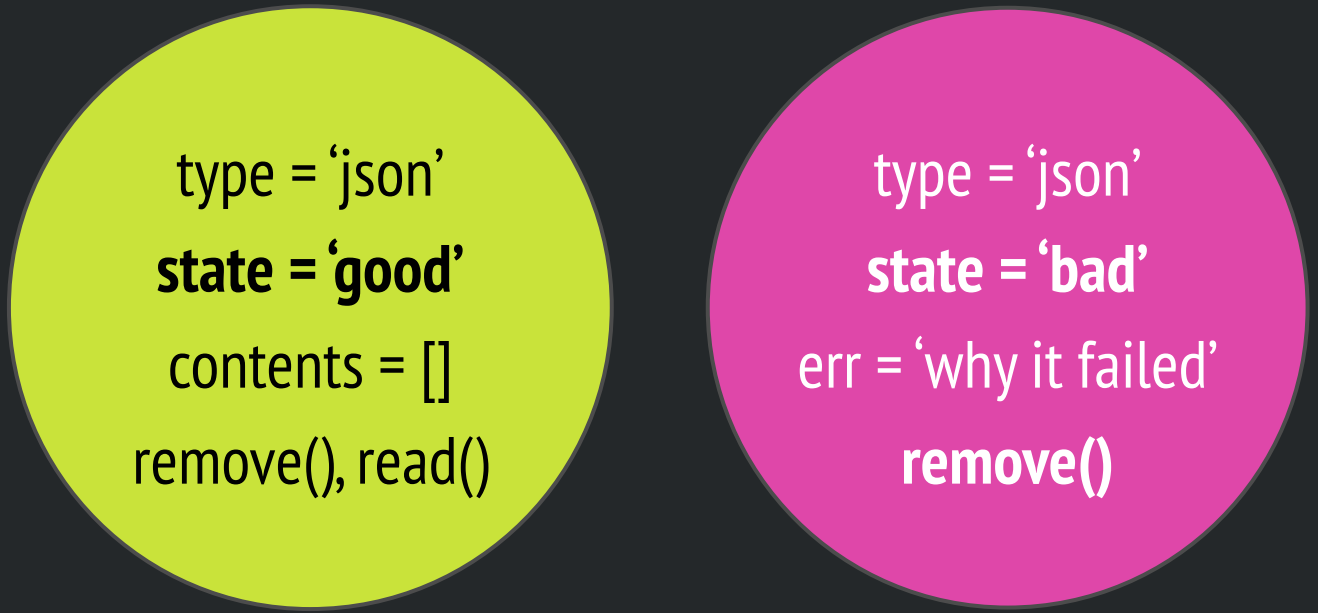
- The same scenario we just explored could be done with 3 separate universes that are later merged. In this case we don't need a 2d map, as every bot is effectively secluded.

A philosophical question on death and suicide

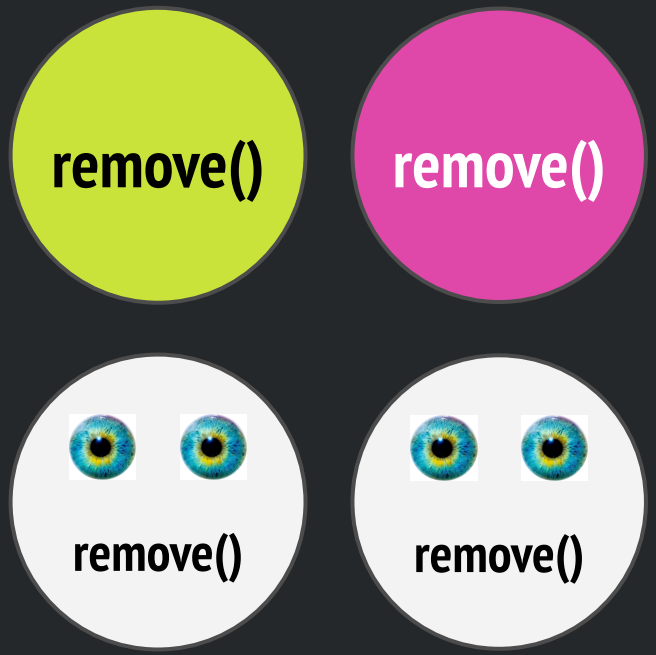
UNIVERSE CONFIG: TWO **JSON** FILES



UNIVERSE: TWO OBJECTS

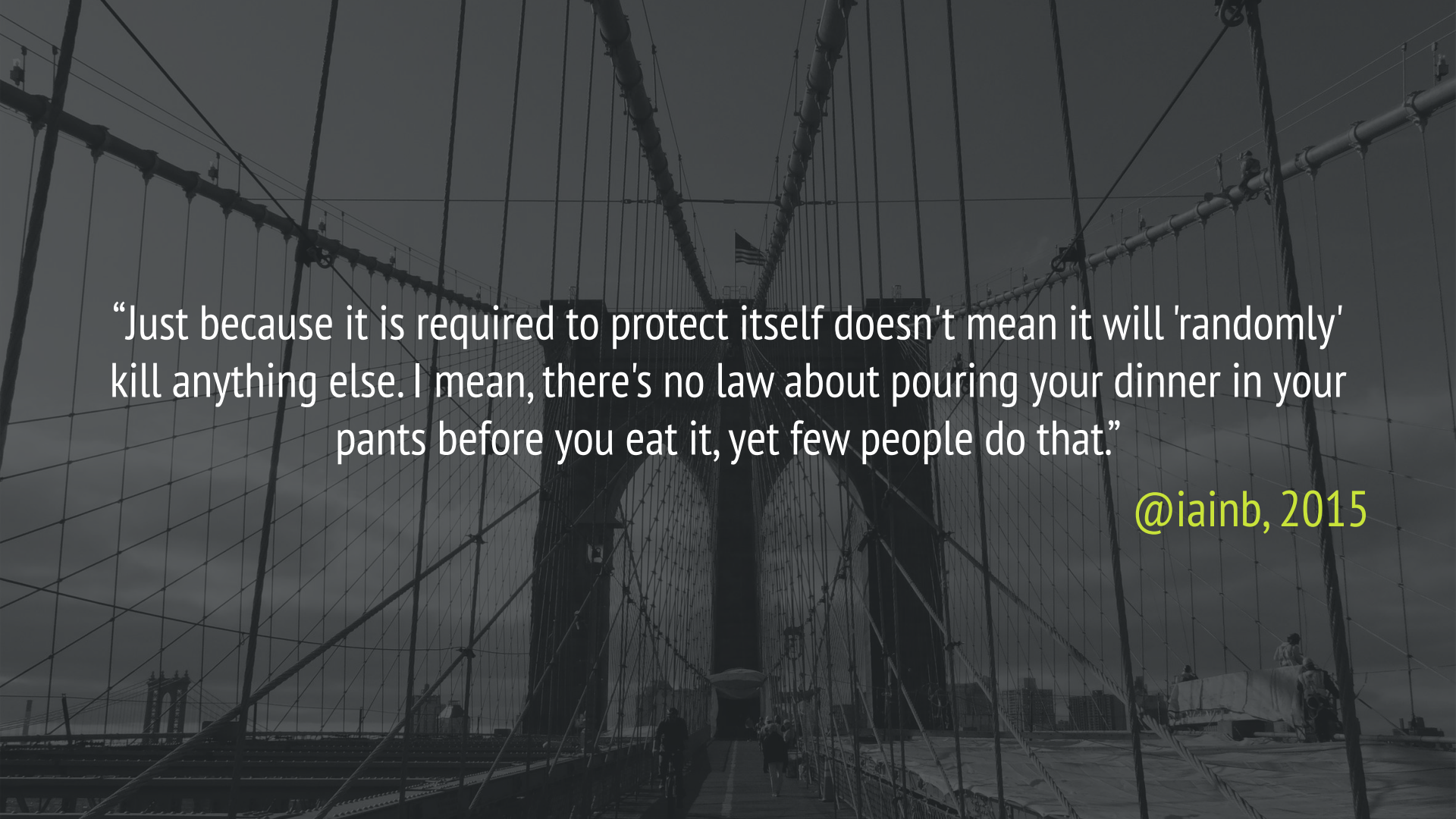


UNIVERSE: TWO OBJECTS AND TWO BOTS



THE THREE LAWS

- A robot may **not injure a human being** or, through inaction, allow a human being to come to harm.
- A robot must **obey the orders given it by human beings**, except where such orders would conflict with the First Law.
- A robot must **protect its own existence** as long as such protection does not conflict with the First or Second Law.



“Just because it is required to protect itself doesn't mean it will 'randomly' kill anything else. I mean, there's no law about pouring your dinner in your pants before you eat it, yet few people do that.”

@iainb, 2015

DEALING WITH MURDER

- **Option 1**: A system of law, with police actors and deterrents
- **Option 2**: Sentiment analysis on the names of the methods
- **Option 3**: Use it's knowledge about itself and apply it to others 

Open problems

CURIOSITY ↛ MOTIVATION

- **Process**: HOW to do things, not just what things are made of
- Creativity? Emotions? Too far-fetched for now?
- **Drive**: The difference between ‘want to’ and ‘have to’
- **Value:** Reward mechanism & sense of value
- **Desire**: Why do anything at all?

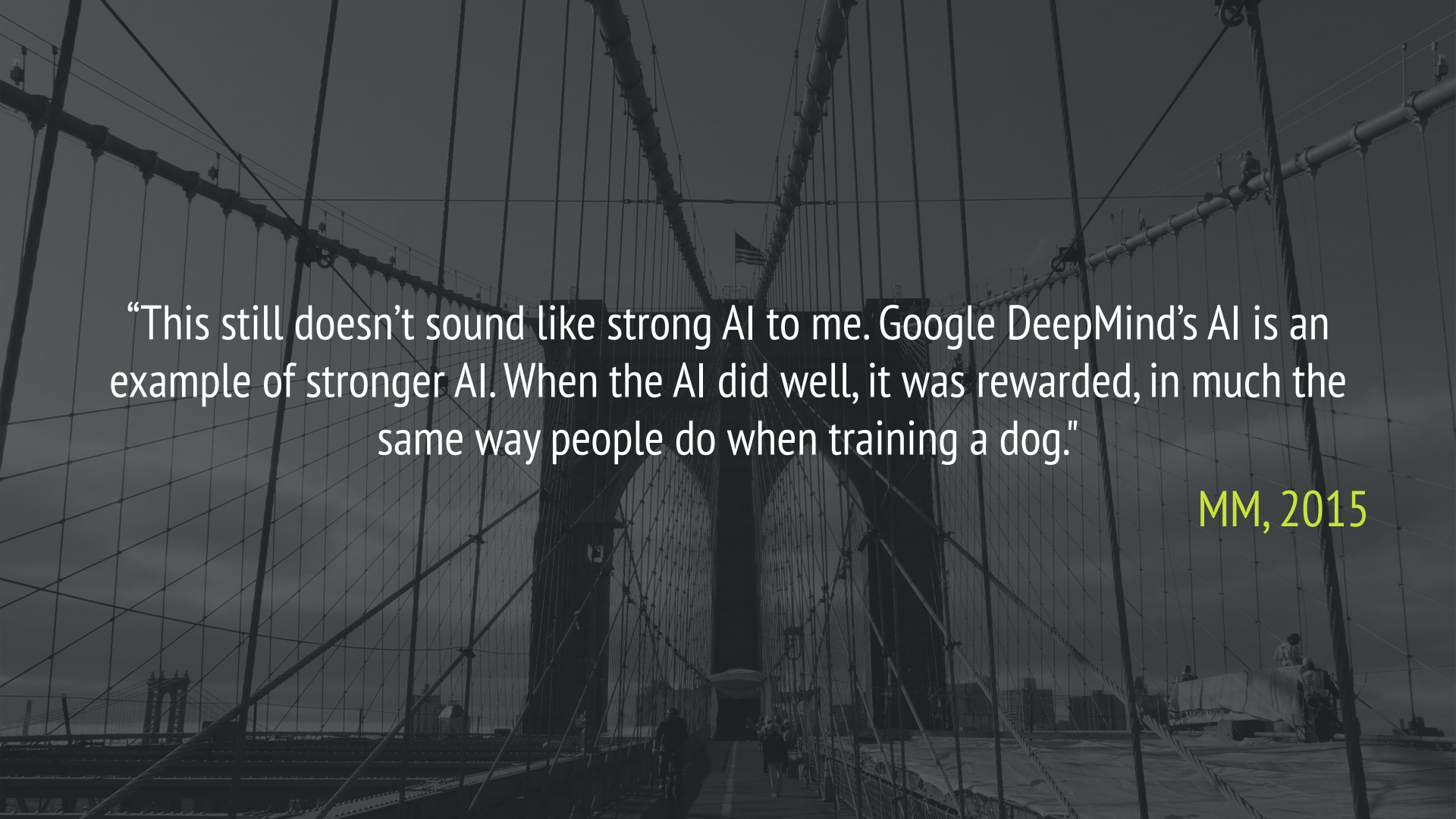
DISCUSS ON MEDIUM

bit.ly/1GfgPuz

THE THOUGHT PROCESS

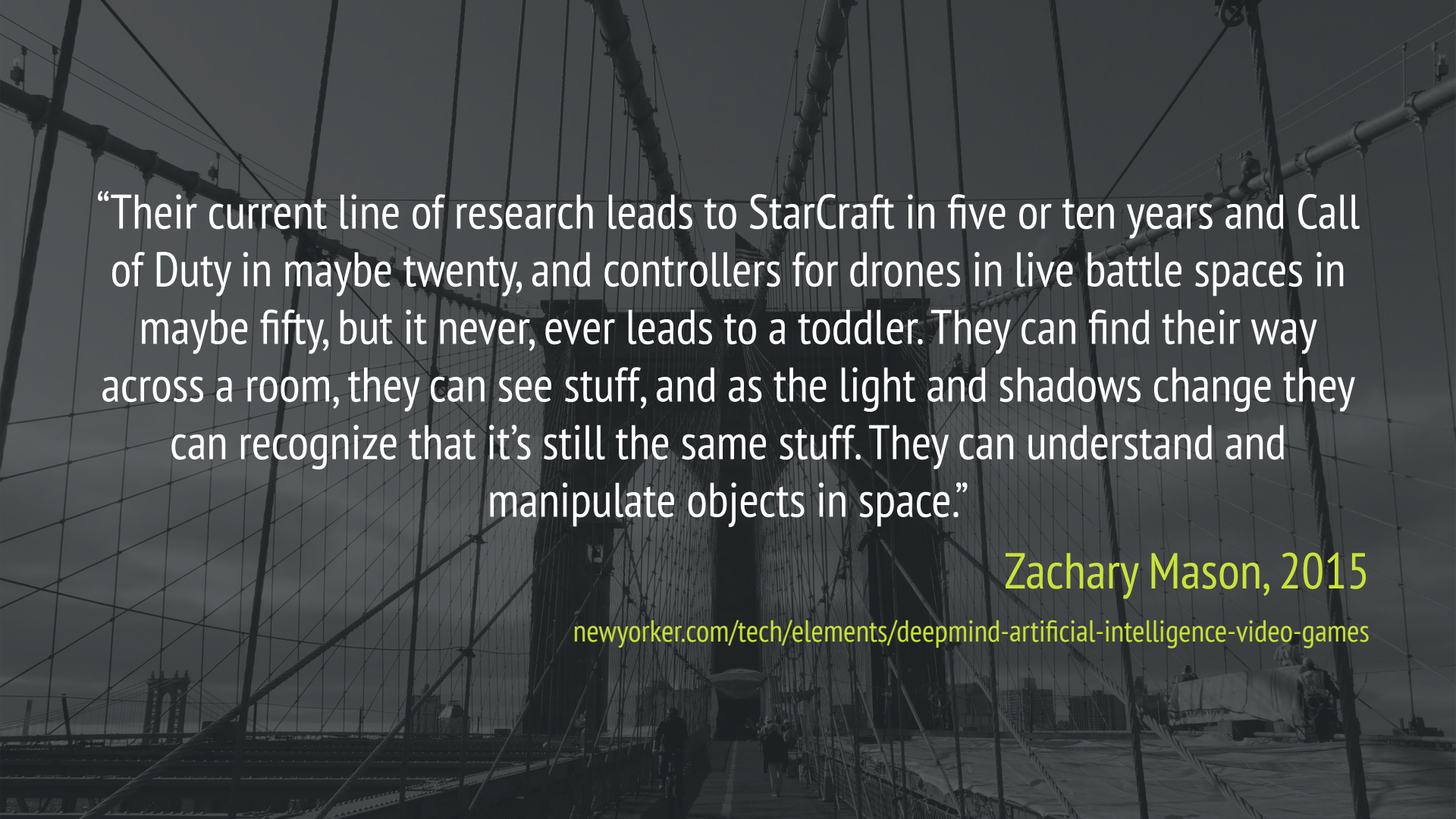
- Reasoning
  - Relationships between data, methods, and outcomes
  - Neural networks?
  - Problem solving (search amongst space consisting of relationships/outcomes)
- Multi-tasking VS fast-switching
- Time management (planning, routines)
- Idea management (pipeline leading to scheduled plans?)
- Positive re-inforcement (leads back to the ‘motivation’ problem)
- The case of the infinite room (explore-everything | ∞ empty space $\wedge$ ∞ energy)


Isn’t all of this still narrow AI?



“This still doesn’t sound like strong AI to me. Google DeepMind’s AI is an example of stronger AI. When the AI did well, it was rewarded, in much the same way people do when training a dog."

MM, 2015



"Their current line of research leads to StarCraft in five or ten years and Call of Duty in maybe twenty, and controllers for drones in live battle spaces in maybe fifty, but it never, ever leads to a toddler. They can find their way across a room, they can see stuff, and as the light and shadows change they can recognize that it's still the same stuff. They can understand and manipulate objects in space."

Zachary Mason, 2015

newyorker.com/tech/elements/deepmind-artificial-intelligence-video-games

However, DeepMind’s reward mechanism works

Might it be the only way to deal with the reward / self-worth problem?

CURIOUS **ACTORS**

CONTRIBUTING TO THE PROJECT

What technologies are we using?

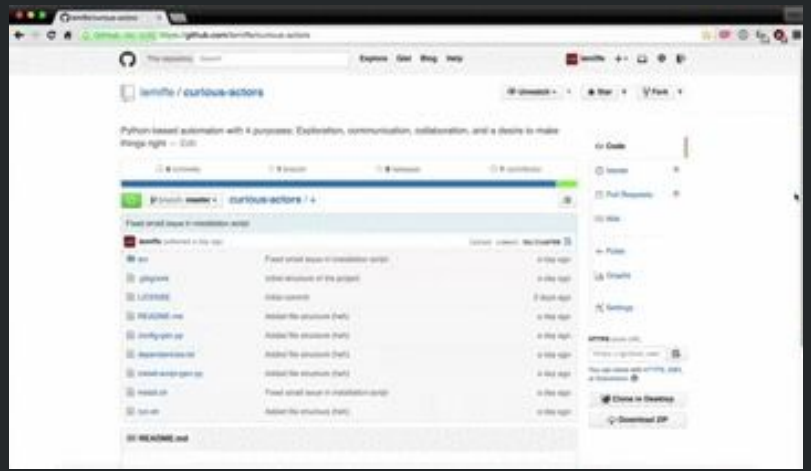
Python 2.7.3+ with virtualenv, mongoDB, vagrant and ansible

Plus a bunch of libraries including pymongo and pattern

Almost everything is on Github

Issues, milestones, technical spec, guidelines, code

One branch per issue ⟶ submit PRs

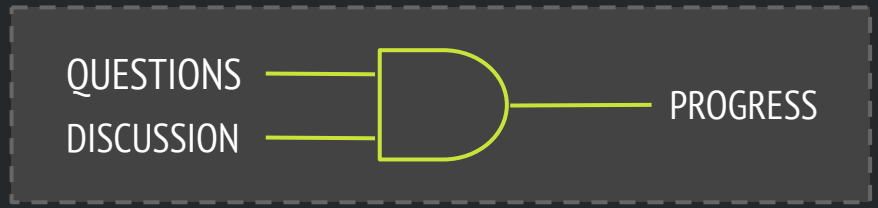


Stay up-to-date!

Project updates will be posted on medium.com/@lemiffe

This presentation will be updated with requirements/vision changes


CONTRIBUTE TO THE PROJECT
/lemiffe/curious-actors



Dedicated to John McCarthy,

Alan Turing, Jon Oberlander, and Joseph Weizenbaum

CURIOUS ACTORS

NOTES AND UPDATES

Notes 2015-01-24: Uncon at #phpbnl15

- **We have a problem**: what is next, after knowing everything? Once there are no more objects to 'correct' or information to learn from. Can you potentially learn more from things you already know? What about plans for the future? Will it care about memories from the past? Interesting philosophical questions that may be technically infeasible.
- Genetic mutations might be an interesting addition ← Added!
- Read about patterns/learning by Jeff Hawkins
  - "people learn practically everything, as a sequence of patterns"
  - "[m]aking predictions is the essence of intelligence."
  - See http://www.homodiscens.com/home/ways/agnoscens/hawkins/index.htm
- **What is next**? What motivates the bot to go on?

Notes 2015-02-21: Chat with @iainb

- Humans' curiosity is based on the instinct for survival and knowledge gathering.
- The bot has no such instincts... do we pre-program them? Is this a valid approach?
- @iainb thinks that a better approach is to pre-program basic behaviours (such as the instinct humans have to eat, procreate, etc.) and hopefully this will lead to a larger 'motivation/drive' such as the 'will & desire to live'

Notes 2015-02-27: Talk at SP HQ

- http://en.wikipedia.org/wiki/Genetic_programming
- http://en.wikipedia.org/wiki/Genetic_algorithm
- MM: This still doesn't sound like strong AI to me. This is an example of stronger ai:
  - http://www.gamespot.com/articles/google-creates-ai-that-can-play-atari-games-better/1100-6425557/
- MM: "When the AI did well, it was rewarded, in much the same way people do when training a dog."
- MW: that sounds like cheating and was one of the points I raised. You can't reward a computer if it has no sense of value and emotions.
- MM: It is possible to treat actors neutrally, and not reward them with anything but rewarding is kinda pruning the problem space if you don't do it, you keep permutating and permutating possible states and there are no atoms enough in the universe to hold them all
- MM: Languages like J are great to evolve genetically, http://en.wikipedia.org/wiki/J_%28programming_language%29

Notes 2015-03-06: Talk pending

- Pending

LINKS, TOOLS & RESOURCES

- **Jeff Hawkins on pattern recognition:** http://www.homodiscens.com/home/ways/agnoscens/hawkins/index.htm
- **Google's DeepMind:** http://www.gamespot.com/articles/google-creates-ai-that-can-play-atari-games-better/1100-6425557/
- **More on DeepMind:** http://www.newyorker.com/tech/elements/deepmind-artificial-intelligence-video-games
- http://deeplearning.net/software/theano/
- http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLtoolkit/RLtoolkit1.0.html
- http://www.computing.dcu.ie/~humphrys/Notes/RL/Code/
- **The road to super intelligence:** http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
- **IBM's 'Watson' in the wild:** https://developer.ibm.com/watson/blog/2015/03/05/watson-wild-march-5th-2015/
- **Q-Learning:** http://artint.info/html/ArtInt_265.html