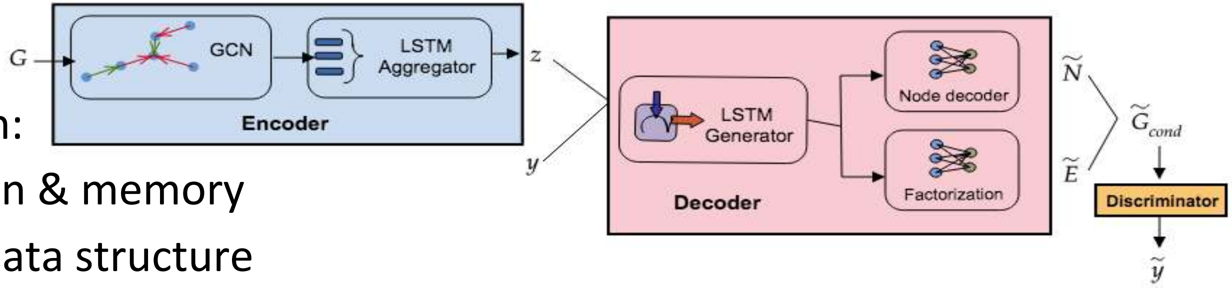
Challenges towards AGI

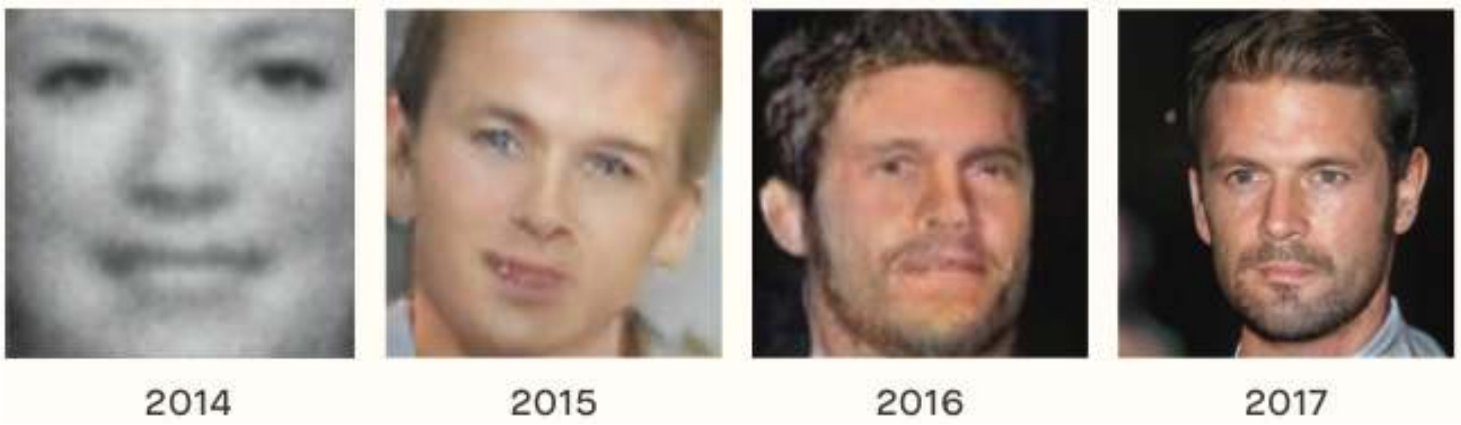
Yoshua Bengio

BENEFICIAL AGI CONFERENCE, PUERTO-RICO
JANUARY 6, 2019

Recent Progress: Deep Learning

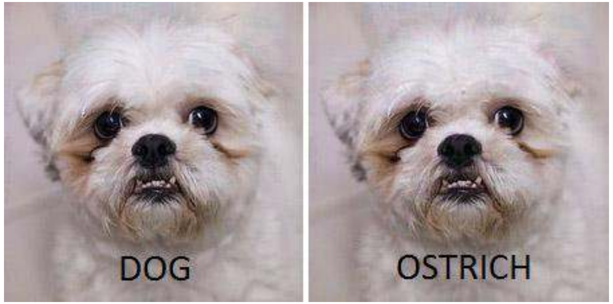


- Beyond pattern recognition:
  - Incorporating attention & memory
  - Handling almost any data structure
  - Powerful generative models
  - Broadening set of applications: healthcare, robotics, environment, dialogue…



Still Really Far from Human-Level AI!

- Industrial successes mostly based on **supervised** learning

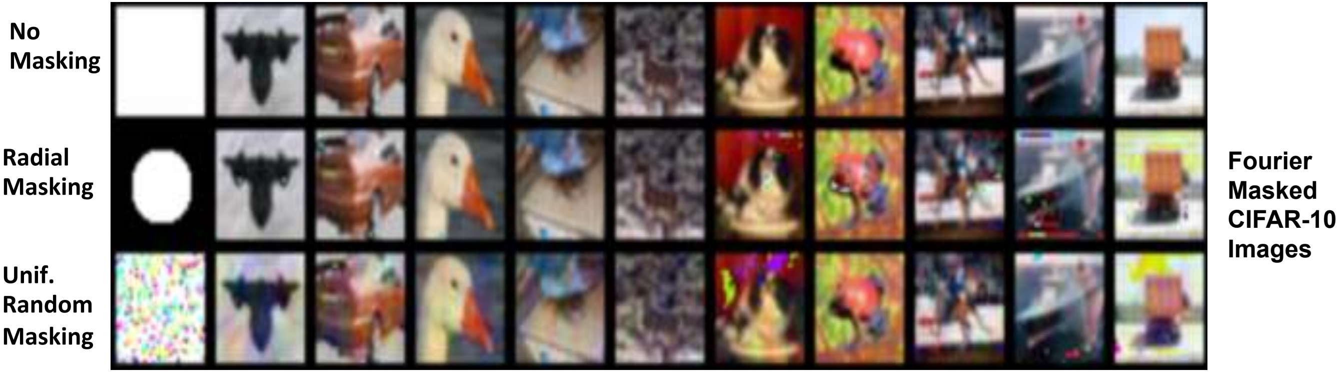


- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
  - Current models cheat by picking on surface regularities

Measuring the Tendency of CNNs to Learn Surface Statistical Regularities

Jason Jo and Yoshua Bengio 2017, arXiv:1711.11561

- **Hypothesis**: *Deep CNNs have a tendency to learn superficial statistical regularities in the dataset rather than high level abstract concepts*.
- From the perspective of learning high level abstractions, Fourier image statistics can be *superficial* regularities, not changing object category, but changing them leads CNNs to make mistakes



Most statistical NLP uses only natural language corpora & annotations

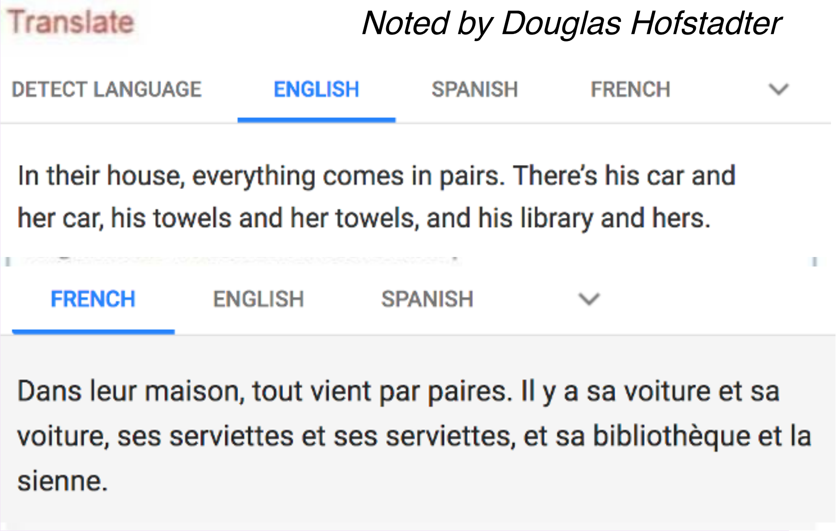
- **Language modeling**: from text to

$$P(\text{next}\quad\text{word} \mid \text{previous}\quad\text{words})$$

- In theory, it would require complete understanding to obtain the best model, but the log-likelihood (perplexity) achieved by humans is not much better than that obtained by the best deep nets.

- **Speech recognition and machine translation**: huge progress, but errors made by these systems show that **they don't understand what the sequences of words actually mean.**

Noted by Douglas Hofstadter



In their house, everything comes in pairs. There's his car and her car, his towels and her towels, and his library and hers.

Dans leur maison, tout vient par paires. Il y a sa voiture et sa voiture, ses serviettes et ses serviettes, et sa bibliothèque et la sienne.

Common Sense & Winograd Schemas

The women stopped taking pills because they were pregnant.

Which entities were pregnant? The women or the pills?

The women stopped taking pills because they were carcinogenic.

Which entities were carcinogenic? The women or the pills?

Humans: 100% accurate

SOTA systems: 56% accurate

Chance: 50% accuracy

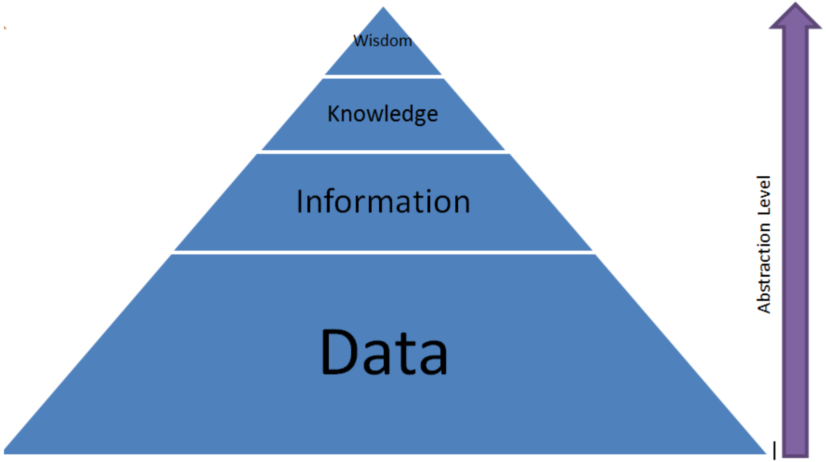
Intuitive Psychology and Intuitive Physics

Still lacking in our best AIs

Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

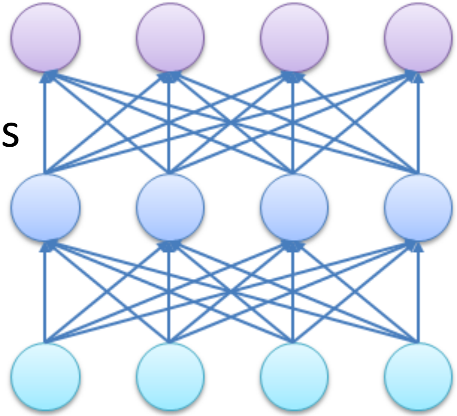
- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions **disentangle the factors of variation**, which allows much easier generalization and transfer



New concern:
Also disentangle the computation (modules) and the hypothesized causal mechanisms

How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- ***Dependencies are simple in the right representation***
- Need clues (= priors) to help **disentangle** the underlying factors, such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*



System 1 vs System 2 Cognition

Two systems (and categories of cognitive tasks):

- **System 1**
  - Intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - *What current **deep learning** does quite well*
- **System 2**
  - Slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - *What **classical symbolic AI** was trying to do*
- **Grounded language learning**: combine both language learning and world modeling

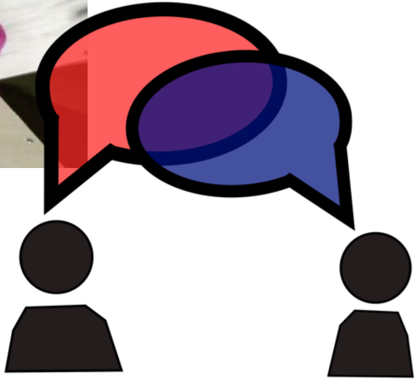
What next?

- From **passive** to **active** intelligence
- From **perception** to **reasoning** to **action planning**
- Acquisition of **world models capturing causal structure**
- Continued inspiration from & synergy w/ **neuroscience & cognition**

Joining System 1 and System 2

Can we build world models from streams of sensory data, anchored in the kind of high-level abstractions which humans take advantage of to understand the world?

Grounded Language Learning



Jointly Learning Natural Language and a World Model

- Should we first learn a world model and then a natural language description of it?
- Or should agents jointly learn about language and about the world?
- Consider top-level representations from supervised ImageNet classifiers. They tend to be much better and easier to learn than those learned by unsupervised learning. Why?
- Because language (here object categories) provides to the learner clues about relevant semantic high-level factors from which it is easier to generalize.
- Culture can help a learner escape from poor optimization, guide (through curricula) the learner to better explanations about the world.

Learning « How the world ticks »

- So long as our machine learning models « cheat » by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples, distributional shifts, catastrophic forgetting, etc.

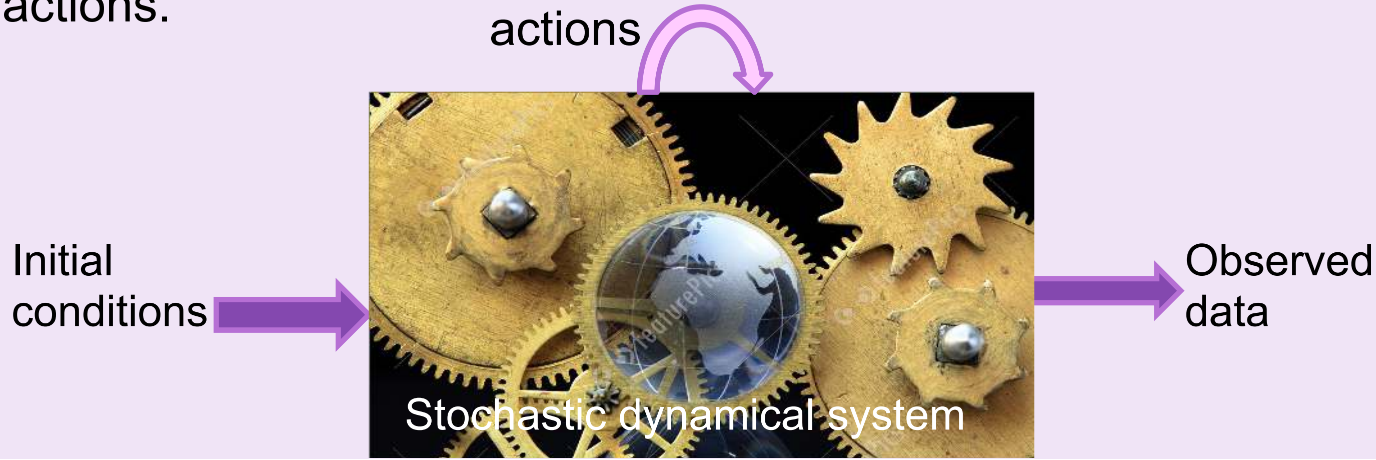
- Humans generalize better than other animals thanks to a more accurate internal model of the **underlying causal relationships**

- To predict future situations (e.g., the effect of planned actions) far from anything seen before while involving known concepts, an essential component of reasoning, intelligence and science

Beyond iid assumption: causal mechanisms

- The assumption that the test data is from the same distribution as the training data is too strong, and it is often violated in practice, leading to poor out-of-distribution generalization.

- Consider relaxed assumptions: the test data was generated under the same ***causal dynamics***, but from different initial conditions (which may be unlikely under the training distribution) and agents' actions.



actions

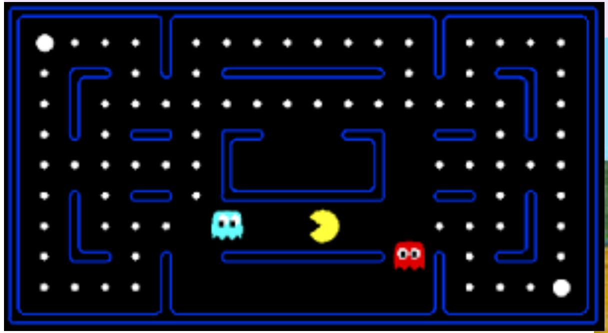
Initial conditions

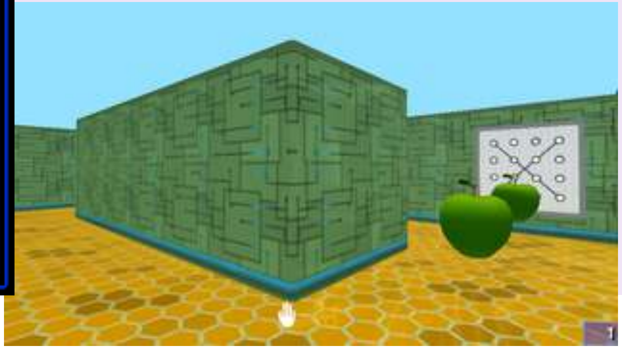
Observed data

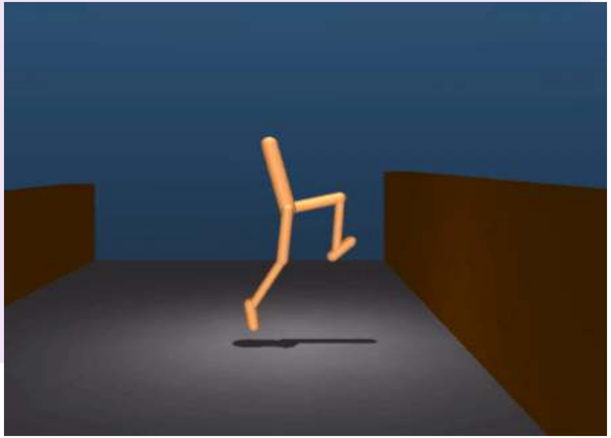
Stochastic dynamical system

Develop learning procedures which figure out how their small-scale environment works

- Outcome of ML **research** = learning framework, not a trained learner
- Solve simple environments before human-level understanding of our world
- Working on a simpler virtual environment leads to a faster research cycle.
- **AI Olympics**: competing research groups can propose different environments as benchmarks to evaluate all the competing learners → open set of tasks, levels, environments forcing generality (same learner sees all)







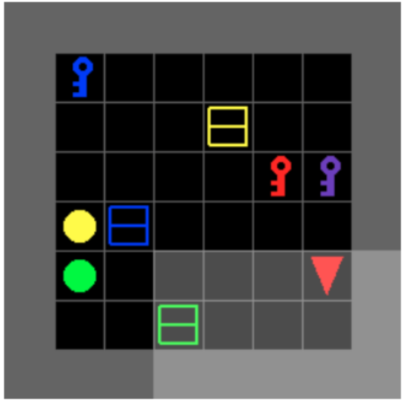
BabyAI platform as initial AI Olympics

Purpose: Simulate language learning from a human and study data efficiency

Comprises:

- A gridworld with partial observability (Minigrid)
- A compositional natural-looking Baby language with over 10^19 instructions
- 19 levels of increasing difficulty
- A heuristic stack-based expert that can solve all levels

github.com/mila-udem/babyai



(b) PutNextLocal: "put the blue key next to the green ball"

Results of 1st benchmark: data efficiency needs work!

- Hundreds of thousands of demonstrations are needed for very simple tasks
- It takes 3 times as much data to get from 95% to 99%
- **A lot of progress is needed before putting a human in the loop!**
- Use BabyAI for your data efficiency studies!
- … but don’t try too hard (e.g. semantic parsing) cause it’s a gridworld

What Next? Abstract Causal Word Models

- Current ML and RL tends to model dependencies in data space, w/o causal structure
- Current ML and RL tends to model temporal sequences via the unfolding of one-step predictions

$$P(\text{next frame} \mid \text{previous frames})$$

- Humans' plans are very different:
  - We project ourselves at arbitrary points into the future or the past
  - A plan is a sequence of meta-actions & events which are not at regularly spaced intervals and can be hypothetical (**counterfactual**)
  - A future event in a plan does not need to be specified at a particular time, e.g. ”tomorrow I will…”
  - We imagine not the full future state but only very specific and abstract aspects of it ('Consciousness Prior')

The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- Focus on **representation learning** and access consciousness:
- Conscious thoughts are very low-dimensional objects compared to the full state of the (unconscious) brain = analogous to a sentence or a rule in rule-based systems
- Yet they have unexpected predictive value or usefulness

  → strong constraint or prior on the underlying representation

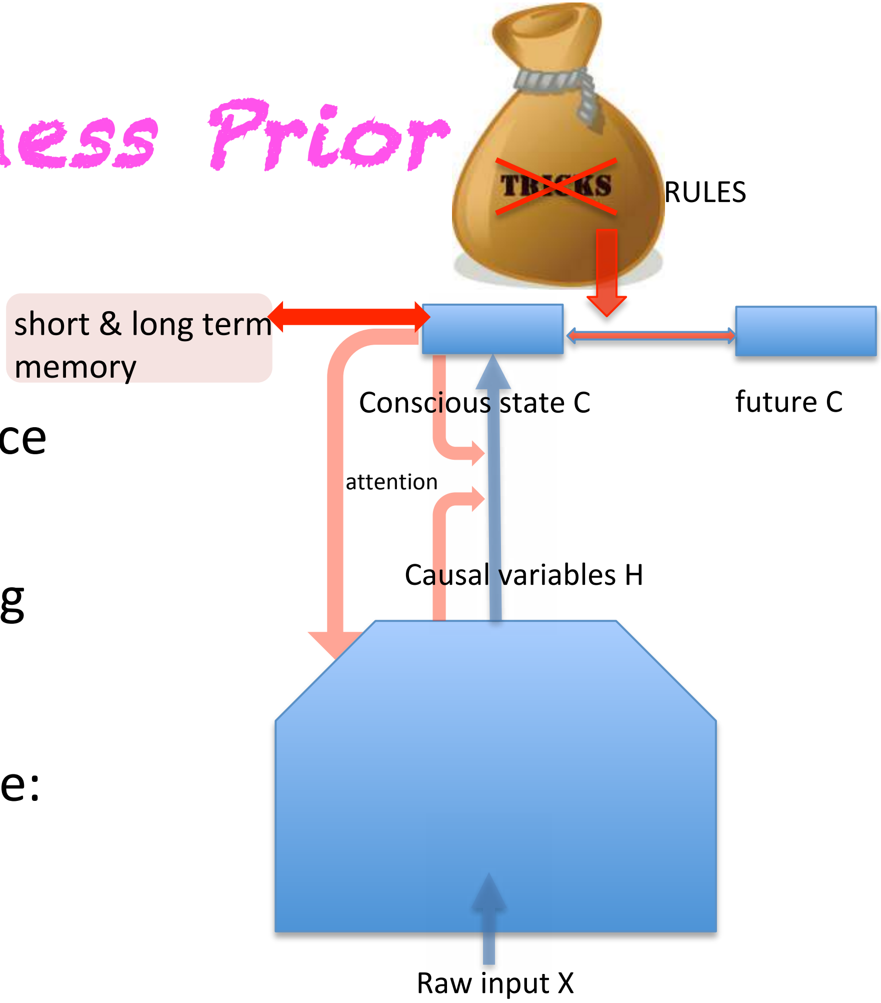
- **Thought**: composition of few selected factors / concepts at the highest level of abstraction of our brain
- Richer than but closely associated with short verbal expression such as a **sentence** or phrase, a **rule** or **fact** (link to classical symbolic AI & knowledge representation)

Need to **disentangle** both

- **Variables** in rule ⇔ **features** in representation space
- **Rules** ⇔ **causal mechanisms**

Causal Consciousness Prior

- Encoder maps sensory data to space where a few **sparse rules** relate causal variables together, following the ***consciousness prior***
- Need to handle uncertainty in state: P(H|X)



A Non-Mysterious ML View on Consciousness

4 computational aspects of consciousness:

- Self-consciousness
  - Notion of self as part of the agent's state, which conditions the agent's decisions, self as agent among other agents, theory of mind
- Access consciousness, conscious attention
  - While conscious, focus at each time step on a few attended elements which condition action/planning/imagination/reminding
- Emotions
  - Shortcut (unconscious) calculations to estimate socially context-dependent value, conditions what goes in memory and access consciousness
- Qualia, subjective perception
  - The focus of conscious attention is mostly in a high-level abstract space in which perception is context-dependent based on the agent's history, goals, emotions, etc.

AI: Hopes & Dangers

- Hopes
  - Economic growth
  - Material progress for all
  - Improving healthcare
  - Improving education and other services (e.g. legal)
  - Freedom from work as slavery
- Dangers
  - Big Brother, killer robots, global security, loss of privacy & freedom
  - Misery for jobless people, at least in transition
  - Manipulation in advertising and social media
  - Reinforcement of social biases and discrimination
  - Increased inequality and power concentration

Beneficial AI Activities at MILA

- Montreal Declaration for the Responsible Development of AI
- Recruitment incentives for diversity (scholarships)
- AI4G workshop @ NeurIPS 2018
- Democratize AI in developing countries, research interns program
- AI for healthcare research projects (ongoing, last 5 years)
- AI for fighting climate change projects (new)
- AI Commons project, ***http://www.aicommons.com***
- Int’l School on Bias and Discrimination in AI, June 2019

Montreal Declaration: Ten Principles, Citizen Participation

Funded by governments of Canada and Quebec

Well-being

Respect for autonomy

Protection of privacy

Solidarity

Democratic participation

Equity

Diversity

Prudence

Responsibility

Sustainable development

Equity Principle

The development and use of AIS must contribute to the creation of a just and equitable society.

1) AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on — among other things — social, sexual, ethnic, cultural, or religious differences.

2) AIS development must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge.

3) AIS development must produce social and economic benefits for all by reducing social inequalities and vulnerabilities.

4) Industrial AIS development must be compatible with acceptable working conditions at every step of their life cycle, from natural resources extraction to recycling, and including data processing.

5) The digital activity of users of AIS and digital services should be recognized as labor that contributes to the functioning of algorithms and creates value.

6) Access to fundamental resources, knowledge and digital tools must be guaranteed for all.

7) We should support the development of shared algorithms — and of open data needed to train them — and expand their use, as a socially equitable objective.

Ethically use ML to convince people of the truth

- Project taking form at Mila in collaboration with Jennifer Chayes (MSR)
- Use ML to press on people's emotional buttons to convince them emotionally of the importance of fighting climate change, using truthful personalized interaction
  - GANs to generate images of your house and descendants in 50-100 years from now
  - with some probability computed by climate change and economic models for events like flooding, fire, hurricanes, economic disasters
  - Bring hope: give knobs which will influence the outcome, e.g., individual and collective action
  - Provide positive action items (e.g. write to your representative, make personal commitments…)
- Need funding & help from outside ML (climate science, behavioral science, computer graphics, environmental activists, advertising experts, economists, etc)

Plausibility of Exponential Self-Improvement?

- ML-based AI does not arbitrarily change itself, only via training objective gradient
- Diminishing returns (possibly exponentially) of #examples vs knowledge obtained
- Intelligence grows "linearly" with amount of knowledge, which grows in log(#examples)
- Absolute size of the brain is not sufficient (consider whale brains >> human brains)
- Our main human advantage is culture (accelerates discovery of knowledge by searching in a more appropriate space than that of proteins)
- Computer science & ML theory: full of exponential WALLS, only manage to mitigate them via assumptions and specialized tricks