Diversity-Aware AI: The Next Frontier?

Michael Rovatsos
University of Edinburgh

Computing Science Seminar
University of Aberdeen, 2/3/16

What magical trick makes us intelligent?
The trick is that there is no trick.
The power of intelligence stems from our vast **diversity**, not from any single, perfect principle.

- Marvin Minsky

A health warning

This talk contains “vision”, i.e. it probably lacks technical rigour.

This vision is not revolutionary, it synthesises existing ideas.

Preliminary, partial contributions are (hopefully) rigorous.

Do they provide evidence that the vision is worth pursuing? I think so.

AI is back!


Eugene Goostman
THE WEIRDEST CREATURE IN THE WORLD
Princeton ai
ARTIFICIAL INTELLIGENCE
No. Beep-beep. I am not a ma-chine. Blink-blink. I am hu-man.
Click! Hu-man. Click! Hu... Damn
Type your question here:
Are you a machine? And can I trust your answer?
reply
THINK
$300,000
$1,000,000
$200,000
tv or monitor
domestic cat
Boston Dynamics

But is it human-like?

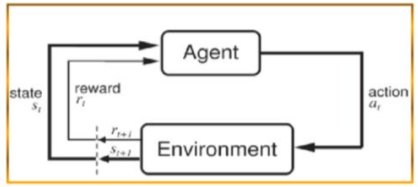
“Standard model” of rational reasoning and learning is based on optimising behaviour to objective function given data.

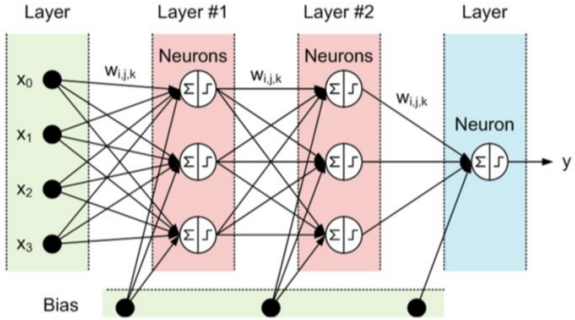
Assuming the availability of **very large amounts of data**, these methods can often guarantee convergence to an optimal solution **in the limit**.

In practice, they can solve amazingly hard problems much sooner… what else could we ask for?

Example

Combination of neural networks + reinforcement learning, used in data-driven algorithms like DeepQ:





Human “intelligence”

Humans pursue **different**, **vaguely defined, conflicting goals** in parallel

Humans **satisfice** much more often than they optimise

Incremental improvement can occur with **little additional experience**

Long-term success is **not guaranteed**

Heterogeneous reasoning processes control overall behaviour

These processes **complement** or **compete with** each other

An example

From cycling…

to motorcycling…

Transfer learning, rote learning, model adaptation, representational change…

A (bold?) claim

The distinguishing feature of humans is that they can improve their knowledge and skill **incrementally** using external information.

Discovery of “**unknown unknowns”** is essential: information that lies beyond the boundaries of one’s current view of the problem domain.

To become more human-like, AI needs to overcome the “static model” view of adaptation and aim at accommodating **model change**.

Why diversity?

Dealing with diversity is necessary to enable such model change. In fact, diversity is what **enables** it.

Requires embracing a more **open-ended notion of intelligence**, where the intelligence of different components can be incrementally combined.

Fundamental problem: How should an intelligent agent make choices regarding things that alter its view of reality?

We need to look for **meta-level criteria** to assess what model changes to perform. Common criteria like correctness, consistency, utility won't do.

Example 1: Recommender Systems

Task Recommendation on the Web

In the **SmartSociety** project, we are developing task recommendation algorithms to help communities of people collaborate in applications like ridesharing.

Huge number of potential **user preferences** and **possible coalitions** users can form. The system should recommend tasks that **balance individual and global objectives**.

Diversity among users makes it harder to discover the **context** relevant to making choices that will enable a centralised mechanism to combine intelligence of components.

Ridesharing example

Tourist Tony: wants to ride into town from his B&B in a suburb

- prepared to pay for ride, cares about saving money, company

Commuter Carrie: wants to split cost on commute from suburb to work

- owns a car, cares about price, safety, punctuality

Facilitator Felix: wants to maximise user satisfaction and vehicle occupancy

Benefit of diversity: number of solutions increases
Challenge of diversity: number of concerns increases

Task recommendation

Our first approach: Bayesian Decision-Theoretic Model with Coarse Preferences + Coalition Formation with Limited Type Reporting

Agents have preferences over features of solution, described by utility function $u_i(s) = w_1(s)f_1(s) + \ldots + w_n f_n(s)$

We assume that agents' choices reflect maximisation of utility

$$a^* = arg\max_a \sum_s u(s)P(s|D)$$

$$P(s|D) \propto P(D|s)P(D)$$

Task recommendation

Where is diversity in all this? In reality there is a whole range of solution features different users (do not) care about

$$\langle \underbrace{punctual(s), safe(s), \overbrace{price(s)}^{Tony}, fun(s)}_{Carrie}, \underbrace{satisfaction(s), occupancy(s)}_{Felix} \rangle$$

Assume service only allows for specification of *price*, *occupancy* derives from computation of solution, *safety* and *satisfaction* predicted from past ratings

Task recommendation

Recommended solutions attempt to maximise social welfare:

$$s^* = \arg\max_s sw(s) \qquad sw(s) = \sum_i u_i(s)$$

Assuming that **all** recommendations are maximally beneficial for users collectively, mechanism taxes all alternatives not optimal to it:

$$u_i(s') = u_i(s) - p_s \quad s.t. \quad u_i(s') \leq u_i(s^*) \quad \forall s' \neq s^*$$

Task recommendation

Agents come in “types” ***Θ*** that correspond to a preference ordering

$$\theta(i) = \theta(j) \Rightarrow \forall s, s'.(u_i(s) \geq u_i(s') \Leftrightarrow u_j(s) \geq u_j(s'))$$

Coarse preferences capture indifference toward some solution features

$$\exists S_1 \uplus \ldots \uplus S_m = S \;\; \forall s, s' \in S_k \;\; u_\theta(s) = u_\theta(s')$$

Limited reporting captures caring about solution features not present, i.e. agents can only signal their type through an insufficient set of “messages”

$$\Psi \subseteq \Theta, \quad \mu : \Theta \to \Psi$$

Initial results

Want mechanism that yields **stable** coalitions under **incomplete** information

Investigated **Posted Goods Signalling Protocol**

1) Each user sends a request to the platform
2) The platform computes an allocation
3) The platform sends an offer signal to each user
4) Each user sends a signal indicating whether they accept
5) At execution time users indicate whether they performed the task

Note difference to classical mechanism design: users have **choice** over solutions, and **enactment** of solution is not guaranteed

Initial results

Within this protocol we need to design an **allocation mechanism** and **message sets** to be used by users when signaling preferences.

Studied **hedonic preferences** where utility depends only on coalition members/**topological preferences**, where it depends on metric properties.

In **hedonic** case, it turns out that we have to insert coalition members one by one to ensure stability (key problem is allocation).

In **topological** case, we can allocate coalitions simultaneously but we need to present users with “extreme” options (key problem are messages).

Values vs. diversity

In this scenario, the (collective) **intelligence** emerges if individual agents are willing to participate and the recommended solutions suit their needs.

This depends on the **values** embedded by the coordination mechanism:

1. The assumptions it makes regarding rational choice
2. The preference types and coarseness of preferences assumed
3. The ways in which it allows agents to provide feedback on solutions
4. The (dis)incentives provided by the mechanism

Only if these are aligned with the needs of agents can **meaningful** interaction take place (otherwise they will disengage or behave counter-productively).

Example 2: Knowledge Sharing

Knowledge sharing

In the **ESSENCE** project, we are looking at how agents with different local perceptions can learn to align these in ways that benefits most their local tasks.

Agents explore their environment and assign arbitrary symbols to entities and relations they encounter.

We are not interested in constructing **accurate ontology alignments** between knowledge structures – we want agents to learn which interpretations of others' symbols are **useful** to them.

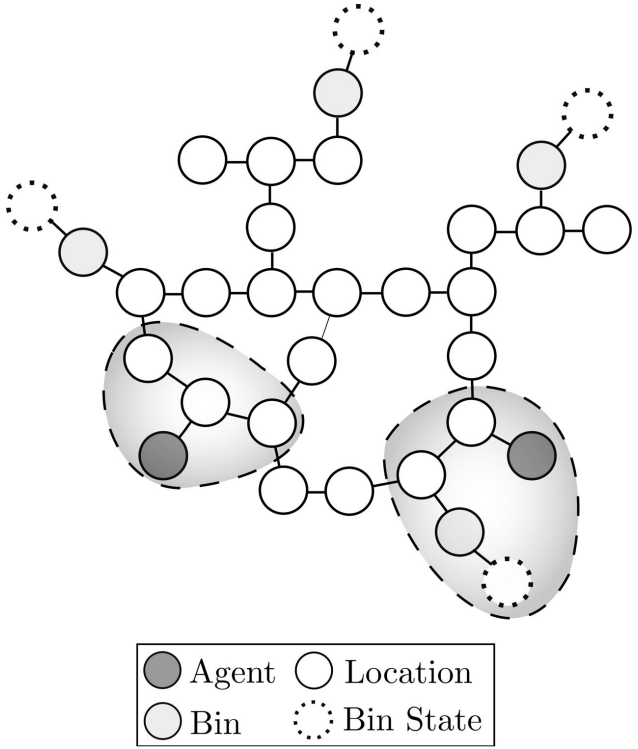
The ontology adoption problem

Consider two agents who build local knowledge graphs and consider all possible label mappings:

$$\begin{array}{ccc} \mathcal{L} & \Theta & \mathcal{L}' \\ \{\mathrm{A}, \mathrm{B}\} & \times \{=, \perp\} & \times \{1, 2\} \end{array}$$

Ideally they would like to find reward-maximising alignment A*

$$A^* = \arg\max_A \sum_{t=1}^{\infty} \gamma^t r(t) P\big(r(t)|\pi^A\big)$$



The ontology adoption problem

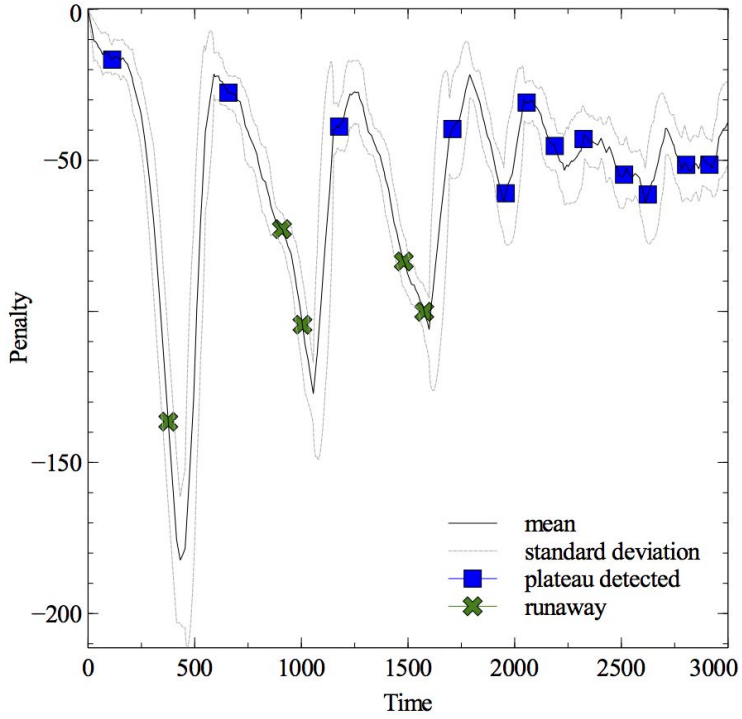
We formulate the alignment adoption problem as a **multi-armed bandit** problem, but huge number of possible mappings.

How to encode prior knowledge about likely alignments and generalise over them? Utilise different families of **kernels** to bias sampling.

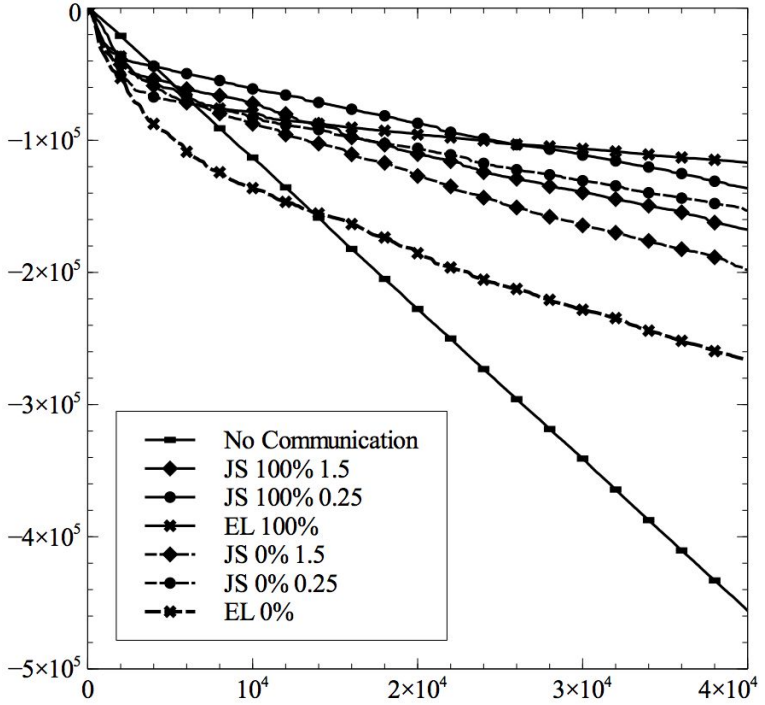
How to ensure only **consistent** alignments are considered during sampling? Use constrained local search (genetic algorithm) to ensure consistency.

How to **evaluate usefulness** of an adopted alignment? Apply **statistical measures** to determine whether rewards obtained are representative.

Preliminary results



Preliminary results



Values vs. diversity

In this scenario (incremental) **intelligence** emerges if the knowledge shared by other agents can be used meaningfully.

This depends on the **values** shared by the interacting agents:

1. They must attempt to convey accurate information.
2. They should make the best attempt to act on their adopted alignment.
3. They must attempt to abandon less useful alignments in the long term.

Decisions regarding **model changes** must be made in such a way that they satisfy these values.

Where I cheated

Task recommendation only involves a very simple form of model change (the set of possible solutions is unknown at design time). Ideally, we would like the **structure** of solutions, e.g. their attributes, to change.

Knowledge sharing assumes task-optimal policy is known for any possible environment structure. In reality this would be learned **in parallel** with alignment adoption.

Examples only scratch the surface of the deeper diversity vision - ideally we would want reasoning frameworks where agents can **purposefully explore usefulness of others' input dynamically**.

Let’s conclude with a manifesto

The cornerstones of diversity-aware AI are **meaningful interaction** among autonomous agents based on **value-driven incremental sense-making and interpretation** of information provided by other agents.

Diversity manifesto under construction

1. **Diverse individuals** have different views of the world but can mutually benefit from each other.
2. **Intelligence** is a result of the interactions among heterogeneous agents capable of sharing meaning.
3. The atomic unit of intelligence is **interaction** among two or more individuals that carries meaning shared by them.
4. **Shared meaning** emerges when interaction mechanisms are aligned with the values held by the agents involved.

Diversity manifesto under construction

5. **Values** are semantic and behavioural constraints not directly related to task achievement which regulate the process of reasoning.
6. They determine whether and how input from others is used and output for them is produced in a **meaningful** way.
7. **Model change** involving structural adaptation to new information is crucial to this process.
8. Agents must be capable of **representing others' input distinctly** from their own internal structures to decide whether they can use it.

Standing on the shoulders of giants

Hierarchical and hybrid inference systems

Ontology alignment, mapping, and learning

Non-monotonic and defeasible reasoning

Mechanism design and social choice

Language evolution and emergent semantics

Teamwork and collaborative multiagent systems

Crowdsourcing and human computation

A vision for human-like AI

Assume we could build agents that can incrementally adapt to a diverse range of human views of the world.

If the outlined principles of diversity-aware AI are correct, this would help bridge the gap between human and machine intelligence.

It would also help build AI that complements human intelligence, and acts to the benefit of humans.

Diversity-awareness might be a more promising path to human-friendly AI than attempts to emulate human-level intelligence.

Workshop Announcement

DIVERSITY 2016 @ ECAI 2016 - International Workshop on Diversity-Aware Artificial Intelligence, 29th/30th August 2016, The Hague, Netherlands

Contributions from all areas of AI invited, as long as they address diversity in some way. Expect a highly interactive, discussion-oriented programme.

Preliminary submission deadline: **12th June**. Extensive financial support for attending available.

Further details to be announced soon at http://www.ecai2016.org/program/workshops/.