Diversity-Awareness – The Key to Human-Like Computing?

Michael Rovatsos
mrovatso@inf.ed.ac.uk

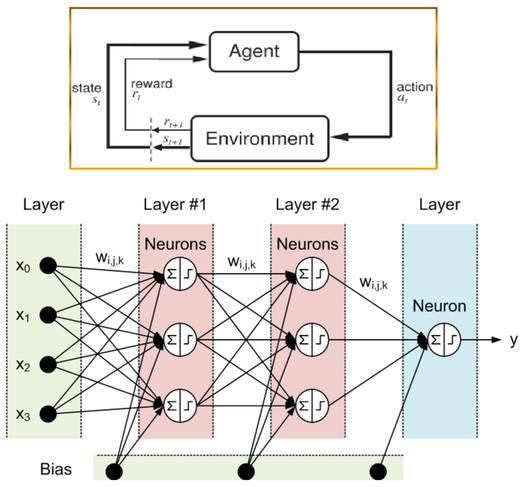
Abstract: While AI has recently produced impressive systems that achieve human-like performance at challenging tasks, these systems tell us very little about how human intelligence works. In particular, they do not address the problem of composing knowledge and behaviour incrementally – a phenomenon that is pervasive in individual and collective human intelligence. We argue that achieving more human-like AI requires focusing on diversity of reasoning and behaviour among humans and artificial agents, and that developing systems capable of dealing with such diversity is key to achieving more human-like AI. In these systems intelligence will not measured in terms of how a system performs at a certain task, but in terms of the properties of the process by which each component combines its knowledge and behaviour with that of others.

Why is AI not human-like (yet)?

"Standard model" or 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 can often be guaranteed **in the limit**.

A vanilla-flavour example of this is the combination of neural networks + reinforcement learning, used in data-driven task optimisation algorithms like DeepQ:



This model and its assumptions have little in common with human intelligence:

- Humans pursue **different, vaguely defined, even conflicting goals** in parallel
- Humans **satisfice** much more often than they optimise
- Many human reasoning and learning processes do not exhibit **steady progress**
- Performance improvement often occurs with **very little additional experience**
- **Heterogeneous reasoning processes** control overall behaviour
- These processes may **complement or compete** with each other

Why diversity?

- To become more human-like, AI needs to overcome the "static model" view of adaptation and aim at accommodating **"model change"**.
- **Diversity** is the fundamental requirement underlying such model change:
  The choices that need to be made lie ***beyond* the boundaries** of one's current view of the problem domain.
- 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 radically alter its view of reality?
- We need to look for **meta-level criteria** that allow us to implement methods of assessing what model changes to perform

Integrating existing approaches

To gain a deeper understanding of **how to deal with and exploit diversity**, we should utilise approaches that have previously studied it in different contexts:

- Hierarchical and hybrid inference systems
- Semantic web and ontologies
- Non-monotonic and defeasible reasoning
- Mechanism design and social choice
- Language evolution and emergent semantics
- Cross-lingual approaches to natural language understanding
- Teamwork and collaborative multiagent systems
- Human-AI and human-robot collaboration methods
- Crowdsourcing and human computation

Toward Diversity-Aware AI

The cornerstones of diversity-aware AI are **meaningful interaction among semantically autonomous agents** and **value-based, context-aware, incremental sensemaking and interpretation** beyond task-rational behaviour:

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 does not violate the values held by the agents involved
5. **Values** are internal 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. The **incremental update of internal semantic structures** with new information is crucial to this process
8. Agents must be capable of **representing others' input distinctly from their own internal structures**

Example 1: Recommender Systems

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.
- Initial approach: aggregate users into **types** (fewer preference profiles) introduce **coarse** preferences (fewer relevant task details)
- This does not solve **diversity problem**: the system might is not aware of features relevant to different users, and how their importance varies among users
- We are now looking into allocation mechanisms that are stable under *limited reporting*, acknowledging that we can only partially understand users' preferences
- The next step will be to allow users to **change their profile instantly** and effect **model change** in the system

We anticipate that improved diversity-awareness will help address **"long tail"** and **"cold start"** problems of recommender systems.

Example 2: 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 allignments** between knowledge structures – we want agents to learn which interpretations of others' symbols are **useful** to them
- We formulate the alignment adoption problem as a **multi-armed bandit** problem, but huge number of possible mappings
- How to **evaluate usefulness** of an adopted alignment, and how to encode prior knowledge about likely alignments?

We are currently conducting experiments which utilise different families of **kernels** to bias learning, and apply **information-theoretic measures** to determine the usefulness of candidate hypotheses.