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

Michael Rovatsos
School of Informatics, The University of Edinburgh
Edinburgh EH8 9AB, United Kingdom
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 in 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 should not only be measured in terms of how a system performs at a certain task, but also in terms of the properties of the process by which each component combines its knowledge and behaviour with that of others, just like humans do.

## 1 Introduction

The current "standard model" of rational reasoning and learning [3, 5, 8] is based on optimising behaviour to an objective function given large amounts of data. Assuming the availability of such data, such methods can often guarantee convergence to an optimal solution in the limit, and we have recently seen impressive examples of this kind of adaptive stochastic optimisation, for example deep learning systems system that achieve human-level performance at tasks considered completely intractable for AI systems in the past [6].

However, this model and its assumptions have little in common with human intelligence: Humans pursue different, vaguely defined, even conflicting goals in parallel, and they "satisfice" much more often than they optimise [4, 7]. Human reasoning and learning often improves with very little additional experience. Most importantly, *heterogeneous reasoning processes* control overall human behaviour, and these processes may complement or compete with each other.

We claim that to become more human-like, AI systems need to overcome the static view of "within-model" adaptation and aim at accommodating *model change*, i.e. adapt representations and decision-making strategies based on information received from components different in structure to themselves. *Diversity-awareness* is a fundamental ingredient of such model change, as the choices an agent needs to make lie beyond the boundaries of its current view of the world. Adopting this view requires embracing a more open-ended notion of intelligence, where the intelligence of different components can be incrementally combined, and answering a fundamental question: *How should an intelligent agent make choices regarding things that radically alter its view of reality?*

* The research presented in this paper has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600854 *"SmartSociety – Hybrid and Diversity-Aware Collective Adaptive Systems: Where people meet machines to build smarter societies"* (http://www.smart-society-project.eu/).

## 2 Diversity-Aware AI

Our preliminary proposal for diversity-aware AI is based on enabling *meaningful interaction* among *semantically autonomous agents* incrementally making sense of each other. As a starting point, it assumes that an intelligent system is a *collective* where individuals may have a different understanding of the world, but may mutually benefit from each other despite this diversity. The overall intelligence of this collective increases with the amount of interaction among agents capable of sharing meaning despite their diversity. *Shared meaning* is produced when external input (output) can be processed (produced) in such a way that it does not violate the values held by the receiving (sending) agent.

To determine whether and how input from others should be used, and how an agent should attempt to produce meaningful output toward others, it applies a set of *values*, i.e. fixed internal constraints not directly related to task achievement. These regulate the process of reasoning, and remain valid throughout an agent's lifetime despite the *model changes* effected by adapting internal knowledge and representations based on input from other agents.

Crucial to this process is the *incremental* update of internal semantic structures with novel information in a diversity-aware way that is contextualised by the agent's internal state. To achieve this incrementality, agents need to have a facility of representing input received from others *distinctly* from their own internal structures, and functionality to mediate between the two.

Overall, our vision of modelling intelligent systems can be summarised with the simplified statement *Intelligence = Meaning + Interaction + Diversity*. Alternatively, it can be understood in terms of *compositionality*, whereby the intelligence of different components can be combined if they are capable of producing an appropriate coupling between their knowledge, skills, and motivations in a given problem environment.

## 3 An example

Consider an example that involves two very different reasoning systems: System 1, a deductive knowledge base that contains metereological rules, and System 2, an image analysis system that can detect weather conditions in landscape photographs. These two systems use completely different representations and reasoning processes for very different tasks: System 1 maintains a logic-based ontology of the domain, and a database of axioms, assumptions and domain facts. Its purpose is to answer queries about the weather in a relational query language. System 2 has been trained using large sets of landscape photographs using machine learning algorithms and representations, and its purpose is to correctly identify the weather conditions given an image using linguistic terms, possibly annotated with a numerical confidence value.

Normally, neither system can be expected to deliver 100% accuracy, and it seems plausible that, in principle, they might benefit from each other's capabilities. For example, System 1 could refine its rule base by looking at regularities System 2 has discovered across consecutive days. System 2 could increase its confidence on outliers by inspecting causal rules from System 1 that rule out certain weather events co-occurring on the same day.

It is not hard to imagine how a human designer might integrate the two systems to enable them to make use of each other's insights. Normally this would

involve defining an intermediate representation and appropriate interfaces for the two components, and the design of these and usefulness of their integration would be validated against either the individual performance metrics of the two systems, or a new, additional functionality that the integration might deliver.

But what would it take for the systems themselves to discover and perform this integration at runtime? This could be guided by the right *values*, e.g. System 1 only modifying its internal model of the weather if it obtains a sufficient number of observations from System 2 that allow it to identify a new, logically consistent explanation by way of abduction, or System 2 only identifying specific phenomena as "noise" when a rule supplied by System 1 is consistent with a sufficiently high number of its own previous observations.

## 4 Closing remarks

At this stage, we have only very vague ideas regarding how to identify suitable value systems for diversity-aware AI methods. In our recent work on diversity-aware task recommendation for human collectives [1], we have explored new optimisation criteria that combine global efficiency with individual user satisfaction, introducing "soft" additional incentive mechanisms to "nudge" users into accepting globally efficient allocations.

In another line of work on ontology alignment among robots that construct local, distinct systems of symbols in their local ontology when exploring a physical environment [2], we combine long-term estimation of reward, information-theoretic measures, and structural similarity heuristics as values that agents use to determine whether and how they should incorporate others' knowledge into their own domain ontology.

We believe, however, that an improved understanding of the principles underlying diversity-awareness may be key to achieving more human-like AI.

Beyond the conceptual, and admittedly rather speculative ideas outlined in this paper, the ubiquity of decentralised data and control on the Internet that gives rise to a multitude of socially and technically interconnected systems, has not yet met with a principled answer to the question of what its role in AI, and, conversely, the role of AI in the context of globally interconnected people and machines should be. Research into diversity-awareness might be part of the answer to this question.

## References


1. P. Andreadis, S. Ceppi, M. Rovatsos, and S. Ramamoorthy. Diversity-Aware Recommendation for Human Collectives. In *Procs DIVERSITY 2016@ECAI 2016*, The Hague, 2016.
2. M. Anslow and M. Rovatsos. Aligning Experientially Grounded Ontologies using Language Games. In *Graph Structures for Knowledge Representation and Reasoning*, LNAI 9501, pp. 15–31, Springer International, 2015.
3. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.
4. D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.
5. S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. 2nd edition, Pearson Education (Prentice-Hall), Upper Saddle River, NJ, 2003.
6. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
7. H. A. Simon. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. MIT press, 1982.
8. R.S. Sutton and A.G. Barto. *Reinforcement Learning. An Introduction*. The MIT Press/A Bradford Book, Cambridge, MA, 1998.