International Workshop on Artificial Intelligence and Cognition

AIC 2014

PROCEEDINGS

Edited by:
Antonio Lieto, Daniele P. Radicioni and Marco Cruciani

Preface

In memory of Roberto Cordeschi

This book of Proceedings contains the accepted papers of the second International Workshop on Artificial Intelligence and Cognition (AIC 2014), held in Turin (Italy) on November 26th and 27th, 2014. The series of workshop AIC was launched in 2013 with the idea of fostering the collaboration between the researchers (coming from the fields of computer science, philosophy, engineering, psychology, neurosciences etc.) working at the intersection of the Cognitive Science and Artificial Intelligence (AI) communities, by providing an international forum of discussions and communication of the research results obtained.

In this workshop proceedings appear 3 abstracts of the talks provided by the keynote speakers and 15 peer reviewed papers accepted by the Program Committee Members through a process of peer-review. Specifically, 10 full papers (31% acceptance rate) and 5 short papers were selected out of 32 submissions coming from researchers of 18 different countries from all the continents. In the following, a short introduction to the content of the volume (full and short papers) is presented.

In the paper "An Interdisciplinary Approach for a Holistic and Embodied Emotion Model in Humanoid Agents", by Samer Schaat, Matthias Huber, Klaus Doblhammer, Dietmar Dietrich, the authors survey contributions from different disciplines (such as Neuroscience, Psychoanalysis, Neuropsychoanalysis, and in Agent-based Systems) to a holistic and embodied emotion model. In particular, the paper investigates how models from relevant disciplines can be beneficial in building Agents Systems.

In the paper "Using Meta-Cognition for Regulating Explanatory Quality Through a Cognitive Architecture", by John Licato, Ron Sun and Selmer Bringsjord, the authors present an investigation on the generation of explanations, framed in the meta-cognitive and non-action-centered subsystems of the cognitive architecture CLARION. The paper focuses on the generation of qualitatively different types of explanations.

The paper "Revisiting Interacting Subsystems Accounts of Cognitive Architecture: The Emergence of Control and Complexity in an Algebra Task", by Gareth Miles, presents a simulation of an algebra task in the cognitive architecture GLAM-PS where the cognitive control is not implemented directly but rather emerges from the interaction of several sub-systems.

The paper "Biologically Plausible Modelling of Morality" by Alessio Plebe presents a biologically plausible neurocomputational model of moral behaviour. The model is implemented in a neural network combining reinforcement and Hebbian learning and simulates the involvement of the sensorial system interaction with emotional and decision making systems in a situation involving moral judgments.

The paper “How Artificial is Intelligence in AI? Arguments for a Non-Discriminatory Turing test”, by Jack Birner, presents a theoretical contribution where the author suggests a resemblance between some long-forgotten ideas of F. A. Hayek’s and some ideas discussed by A. Turing in his well-known 1950 article “Computing Machinery and Intelligence” lying at the basis of “classical” AI.

In the paper “On the Concept of Correct Hits in Spoken Term Detection”, by Gàbor Gosztolya, the author compares system for spoken term detection (STD) against human response in dealing with the Hungarian, which is an agglutinative language and, as such, poses additional challenges to both human and automatic STD tasks. A discussion on how the spoken term detection task is evaluated is provided, along with practical tools to individuate ground truths for evaluation (by starting from user information).

The paper “Action Recognition based on Hierarchical Self-Organizing Maps”, by Miriam Buonamente, Haris Dindo, and Magnus Johnsson, presents a hierarchical neural architecture, based on Self-Organizing Maps (SOM), able to recognise observed human actions. The architecture is evaluated in an experimental setting based on the recognition of actions from movies taken from the INRIA 4D repository.

In the paper “Learning Graspability of Unknown Objects via Intrinsic Motivation”, by Ercin Temel, Beata J. Grzyb, and Sanem Sariel, the authors propose a machine learning optimization aimed at learning whether and how some simple objects can be grasped through a robotic arm. The optimization relies on the notion of frustration. The frustration, which is governed by the ‘impulsiveness’, that measures how fast a robot gets frustrated. The introduced frustration is experimentally proven useful to faster learning.

In the paper “On the Cognitive and Logical Role of Image Schemas in Computational Conceptual Blending”, by Maria Hedblom, Oliver Kutz, and Fabian Neuhaus, the role of image schemas in computational concept creation process is discussed. In particular, the authors propose to build a library of formalized image schemas, and illustrate how they can guide the search for a base space in the concept invention workflow.

The paper “Monoidal Logics: how to Avoid Paradoxes”, by Clayton Peterson, presents monoidal logics as a formal solution that can be useful in AI in order to avoid some classical paradoxes based on cartesian logical structure. The main differences between standard Cartesian logics and monoidal logics are presented in the article.

In the paper “Mathematical Patterns and Cognitive Architectures”, by Agnese Augello, Salvatore Gaglio, Gianluigi Oliveri, and Giovanni Pilato, the authors investigate the nature of mathematical patterns and some elements of the cognitive structure an agent should have to recognize them, and propose a mapping of such patterns in the setting of Conceptual Spaces.

In the paper “Romeo2 Project: Humanoid Robot Assistant and Companion for Everyday Life: I. Situation Assessment for Social Intelligence”, by Pandey *et al.*, concerns robotic situational awareness and perception in HRI scenarios. In particular, a general overview of a multi-modal perception and situation assessment system built in the Romeo2 project. redmi pare che questa frase non stia in piedi: forse possiamo aggiungere ‘are illustrated’ alla fine? verrebbe quindi: In particular, a general overview of a multi-modal perception and situation assessment system built in the Romeo2 project are illustrated.

In the paper “Information for Cognitive Agents”, by Nir Fresco, a theoretic account of the concept of “information” is proposed. In particular, the author defends the importance of a pragmatic - neo Peircean - account of information that can be useful in the area of cognitively inspired AI.

In the paper “Mining and Visualizing Uncertain Data Objects and Named Data Networking Traffics by Fuzzy Self-Organizing Map”, by Amin Karami and Manel Guerrero-Zapata, the authors propose a novel algorithm to mine and visualize uncertain objects; the proposed algorithm is successfully applied to known benchmarks and to mine and visualize network attacks in the context of the Named Data Networking (NDN).

In the paper “Implementation of Evolutionary Algorithms for Deep Architectures”, by Sreenivas Sremath Tirumala, the author advocates for further research on deep learning by evolutionary computation (EC) researchers. A review of some latest deep architectures is presented and a survey is provided about some evolutionary algorithms that can be explored to train these deep architectures to the ends of promoting the research on evolutionary inspired deep learning techniques.

December 2014

Antonio Lieto
Daniele P. Radicioni
Program Chairs, AIC 2014

Acknowledgments

We would like to thank the keynote speakers of the workshop: Prof. Aaron Sloman (University of Birmingham, UK), Prof. Diego Marconi (University of Torino, Italy) and Prof. Giulio Sandini (Italian Institute of Technology, Italy) for accepting our invitation.

We sincerely thank the Department of Computer Science of the University of Turin, Italy (http://www.di.unito.it/), the Italian Association for Artificial Intelligence (AI*IA, http://www.aixia.it/), and the Italian Association of Cognitive Sciences (AISC, http://www.aisc-net.org) for their support in the organisation of the workshop, and also the Rosselli Foundation (Fondazione Rosselli), Turin, Italy, for its precious logistic support. The organization of the workshop also was partly supported by the Progetto di Ateneo-San Paolo, 'The role of visual imagery in lexical processing' (RVILP), TO_call03_2012_0046, we are indebted with this sponsor, as well.

We are grateful to the members of the Scientific Committee and to the additional reviewers for their valuable work during the reviewing process. Finally, thanks should be given to our wonderful student volunteers for their help in many practical issues.

This book is dedicated to Roberto Cordeschi, unfortunately no longer among us, whose research was dedicated to the investigation of the epistemological roots of Cybernetics and Artificial Intelligence, with a focus on the explanatory role of the computational models of cognition for the comprehension of the human mind.

December 2014

Antonio Lieto
Daniele P. Radicioni
Program Chairs, AIC 2014

Organization

Organizers

Antonio Lieto	University of Torino, Italy & ICAR-CNR, Palermo, Italy
Daniele P. Radicioni	University of Torino, Italy
Marco Cruciani	University of Trento, Italy

Program Committee

Gabriella Airenti	University of Torino, Italy
Bruno Bara	University of Torino, Italy
Giuseppe Boccignone	University of Milano, Italy
Erik Cambria	Nanyang Technological University, Singapore
Angelo Cangelosi	University of Plymouth, UK
Cristiano Castelfranchi	ISTC-CNR, Italy
Antonio Chella	University of Palermo, Italy
Rosaria Conte	ISTC-CNR, Italy
Roberto Cordeschi	University La Sapienza, Italy
David Danks	Carnegie Mellon University, USA
Christian Freksa	University of Bremen, Germany
Salvatore Gaglio	University of Palermo and ICAR-CNR, Italy
Aldo Gangemi	CNR-ISTC Rome, Italy and LIPN-Paris 13-Sorbonne Cité, France
Mark Finlayson	Massachusetts Institute of Technology, USA
Marcello Frixione	University of Genova, Italy
Peter Gärdenfors	University of Lund, Sweden
Nicola Guarino	LOA ISTC-CNR, Trento, Italy
Anna Jordanous	University of Kent, UK
Ismo Koponen	University of Helsinki, Finland
Oliver Kutz	University of Magdeburg, Germany
Othalia Larue	Université du Québec à Montréal, Canada
Vincenzo Lombardo	University of Torino, Italy
Diego Marconi	University of Torino, Italy
Marjorie McShane	Rensselaer Polytechnic Institute, Troy, NY, USA
Orazio Miglino	University of Napoli 'Federico II', Italy
Alessandro Oltramari	Carnegie Mellon University, USA
Fabio Paglieri	ISTC-CNR, Italy
Maria Teresa Pazienza	University of Roma 'Tor Vergata', Italy
Alessio Plebe	University of Messina, Italy
Alessandro Saffiotti	Örebro University, Sweden
Giovanni Semeraro	University of Bari, Italy
Guglielmo Tamburrini	University of Napoli 'Federico II', Italy
Charalampos Tampitsikas	University of Lugano, Switzerland and Carnegie Mellon University, USA
Pietro Terna	University of Torino, Italy
Radboud G.F. Winkels	Leibniz Center for Law, The Netherlands
Fabio Massimo Zanzotto	University of Roma 'Tor Vergata', Italy

Additional Reviewers

Livia Colle
Haris Dindo
Roberto Esposito
Lutz Frommberger
Valentina Gliozzi
Riccardo Rizzo
Manuela Sanguinetti
Carl Schultz

Student Volunteers

Davide Colla
Leo Ghignone
Andrea Minieri
Mattia Ognibene
Valentina Rho

Sponsoring Institutions

Dipartimento di Informatica, University of Turin (www.di.unito.it)
Progetto di Ateneo-San Paolo, *The role of visual imagery in lexical processing* (RVILP), *TO_call03_2012_0046*.
Associazione Italiana per l'Intelligenza Artificiale (http://www.aixia.it)
Fondazione Rosselli, Torino (http://www.fondazionerosselli.it)

Supporting Institutions

Associazione Italiana per l'Intelligenza Artificiale (http://www.aixia.it)
Associazione Italiana per le Science Cognitive
Center for Cognitive Science - University and Polytechnic of Turin (http://www.psych.unito.it/csc/)

Table of Contents

Artificial Intelligence and Cognition, AIC 2014

How can we reduce the gulf between artificial and natural intelligence?

Invited talk at AIC 2014
http://aic2014.di.unito.it/
November 26-27 University of Turin, Italy

Aaron Sloman

School of Computer Science, Birmingham, UK
http://www.cs.bham.ac.uk/~axs

Abstract. AI and robotics have many impressive successes, yet there remain huge chasms between artificial systems and forms of natural intelligence in humans and other animals. Fashionable "paradigms" offering definitive answers come and go (sometimes reappearing with new labels). Yet no AI or robotic systems come close to modelling or replicating the development from helpless infant over a decade or two to a competent adult. Human and animal developmental trajectories vastly outstrip, in depth and breadth of achievement, products of artificial learning systems, although some AI products demonstrate super-human competences in restricted domains. I'll outline a very long-term multi-disciplinary research programme addressing these and other inadequacies in current AI, cognitive science, robotics, psychology, neuroscience, philosophy of mathematics and philosophy of mind. The project builds on past work by actively seeking gaps in what we already understand, and by looking for very different clues and challenges: the Meta-Morphogenesis project, partly inspired by Turing's work on morphogenesis, outlined here: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

Keywords: evolution,information-processing,meta-morphogenesis,Turing

1 Introduction

There are many impressive successes of AI and robotics, some of them summarised at `http://aitopics.org/news`. Yet there remain huge chasms between artificial systems and forms of natural intelligence in humans and other animals – including weaver-birds, elephants, squirrels, dolphins, orangutans, carnivorous mammals, and their prey.[1]

Fashionable "paradigms" offering definitive answers come and go, sometimes reappearing with new labels, and often ignoring previous work, such as the

[1] Nest building cognition of a weaver bird can be sampled here: http://www.youtube.com/watch?v=6svAIgEnFvw

impressive survey by Marvin Minsky over 50 years ago [6], long before computers with suitable powers were available.

Despite advances over several decades, accelerated recently by availability of smaller, cheaper, faster, computing mechanisms, with very much larger memories than in the past, no AI or robotic systems come close to modelling or replicating the development from helpless infant over a decade or two to plumber, cook, trapeze artist, bricklayer, seamstress, dairy farmer, shop-keeper, child-minder, professor of philosophy, concert pianist, mathematics teacher, quantum physicist, waiter in a busy restaurant, etc. Human and animal developmental trajectories vastly outstrip, in depth and breadth of achievement, the products of artificial learning systems, although AI systems sometimes produce super-human competences in restricted domains, such as proving logical theorems, winning at chess or Jeopardy.[2]

I'll outline a very long-term multi-disciplinary research programme addressing these and other inadequacies in current AI, robotics, psychology, neuroscience and philosophy of mathematics and mind, in part by building on past and ongoing work in AI, and in part by looking for very different clues and challenges: the Meta-Morphogenesis project, partly inspired by Turing's work on morphogenesis.[3]

2 First characterise the gulf accurately

We need to understand what has and has not been achieved in AI. The former (identifying successes) gets most attention, though in the long run the latter task (identifying gaps in our knowledge) is more important for future progress.

There are many ways in which current robots and AI systems fall short of the intelligence of humans and other animals, including their ability to reason about topology and continuous deformation (for examples see [7] and http://www.cs.bham.ac.uk/research/projects/cogaff/misc/torus.html). Don't expect any robot (even with soft hands and compliant joints) to be able to dress a two year old child (safely) in the near future, a task that requires understanding of both topology and deformable materials, among other things.[4]

Getting machines to understand **why** things work or don't work lags even further behind programmed or trained abilities to perform tasks. For example, understanding why it's not a good idea to start putting on a shirt by inserting a hand into a cuff and pulling the sleeve up over the arm requires a combination of topological and metrical reasoning: – a type of mathematical child-minding theorem, not taught in schools but understood by most child-minders, even if they have never articulated the theorem and cannot articulate the reasons why

[2] Though it's best not to believe everything you see in advertisements http://www.youtube.com/watch?v=tIIJME8-au8

[3] http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html This project is unfunded and I have no plans to apply for funding, though others may do so if they wish.

[4] As illustrated in this video. http://www.youtube.com/watch?v=WWNlgvtYcEs

it is true. Can you? Merely pointing at past evidence showing that attempts to dress a child that way always fails does not explain why it is impossible.

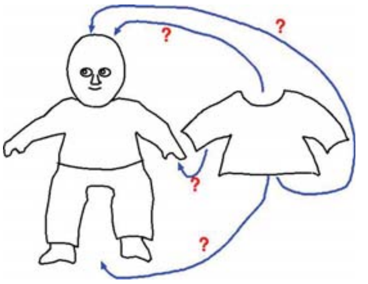


Fig. 1. What sequence of movements could get the shirt onto the child if the shirt is made of material that is flexible but does not stretch much? Why would it be a mistake to start by pushing the head through the neck-hole? What difference would it make if the material could be stretched arbitrarily without being permanently changed?

In more obviously mathematical domains, where computers are commonly assumed to excel, the achievements are narrowly focused on branches of mathematics using inference methods based on arithmetic, algebra, logic, probability and statistical theory.

However, mathematics is much broader than that, and we lack models of the reasoning (for instance geometrical and topological reasoning) that enabled humans to come up with the profoundly important and influential mathematical discoveries reported in Euclid's *Elements* 2.5 millennia ago – arguably the single most important book ever written on this planet. The early pioneers could not have learnt from mathematics teachers. How did they teach themselves, and each other? What would be required to enable robots to make similar discoveries without teachers?

Those mathematical capabilities seem to have deep, but mostly unnoticed, connections with animal abilities to perceive practically important types of affordance, including use of mechanisms that are concerned not only with the perceiver's possibilities for immediate action but more generally with what is and is not possible in a physical situation and how those possibilities and impossibilities can change, for example if something is moved. A child could learn that a shoelace threaded through a single hole can be removed from the hole by pulling the left end of the lace or by pulling the right end. Why does combining two successful actions fail in this case, whereas in other cases a combination improves success (e.g. A pushing an object and B pushing the object in the same direction)? Collecting examples of explanations of impossibilities that humans understand but not yet current robots is one way to investigate gaps in what has been achieved so far. It is also a route toward understanding the nature of human mathematical competences, which I think start to develop in children long before anyone notices.

Many animals, including pre-verbal humans, need to be able to perceive and think about what is and is not possible in a situation, though in most cases without having the ability to reflect on their thinking or to communicate the thoughts to someone else. The meta-cognitive abilities evolve later in the history of a species and develop later in individuals.

Thinking about what would be possible in various possible states of affairs is totally different from abilities to make predictions about what will happen, or to reason probabilistically. It's one thing to try repeatedly to push a shirt on a child by pushing its hand and arm in through the end of a sleeve and conclude from repeated failures that success is improbable. It's quite another thing to understand that if the shirt material cannot be stretched, then success is impossible (for a normally shaped child and a well fitting shirt) though if the material could be stretched as much as needed then it could be done. Additional reasoning powers might enable the machine to work out that starting by pushing the head in through the largest opening could require least stretching, and to work this out without having to collect statistics from repeated attempts.

3 Shallow statistical vs deep knowledge

It is possible to have a shallow (statistical) predictive capability based on observed regularities while lacking deeper knowledge about the set of possibilities sampled in those observations. An example is the difference between (a) having heard and remembered a set of sentences and noticed some regular associations between pairs of words in those sentences and (b) being aware of the generative grammar used by the speakers, or having acquired such a grammar unconsciously. The grasp of the grammar, using recursive modes of composition, permits a much richer and more varied collection of utterances to be produced or understood. Something similar is required for visual perception of spatial configurations and spatial processes that are even richer and more varied than sentences can be. Yet it seems that we share that more powerful competence with more species, including squirrels and nest-building birds.

This suggests that abilities to acquire, process, store, manipulate, and use information about spatial structures evolved before capabilities that are unique to humans, such as use of spoken language. But the spatial information requires use of something like grammatical structures to cope with scenes of varying complexity, varying structural detail, and varying collections of possibilities for change. In other words visual perception, along with planning and acting on the basis of what is scene, requires the use of internal languages that have many of the properties previously thought unique to human communicative languages. Finding out what those languages are, how they evolved, how they can vary across species, across individuals, and within an individual during development is a long term research programme, with potential implications for many aspects of AI/Robotics and Cognitive Science – discussed further in [8].

Conceivably a robot could be programmed to explore making various movements combining a shirt and a flexible, child-shaped doll. It might discover one or more sequences of moves that successfully get the shirt on, provided that the shirt and doll are initially in one of the robot's previously encountered starting states. This could be done by exploring the space of sequences of possible moves, whose size would depend on the degree of precision of its motion and control parameters. For example, if from every position of the hands there are 50 possible 3-D directions of movement and the robot tries 20 steps after each starting direction, then the number of physical trajectories from the initial state to be explored is

$$50^{20} = 9536743164062500000000000000000000$$

and if it tries a million new moves every second, then it could explore that space in about 302408000000000000 millennia. Clearly animals do something different when they learn to do things, but exactly how they choose things to try at each moment is not known.

The "generative grammar" of spatial structures and processes is rich and deep, and is not concerned only with linear sequences or discrete sequences. In fact there are multiple overlapping space-time grammars, involving different collections of objects assembled, disassembled, moved, repaired, etc. and used, often for many purposes and in many ways. Think of what structures and processes are made possible by different sorts of children's play materials and construction kits, including plasticine, paper and scissors, meccano, lego, tinkertoys, etc. The sort of deep knowledge I am referring to involves grasp of the structure of a construction-kit with generative powers, and the ability to make inferences about what can and cannot be built with that kit, by assembling more and more parts, subject to the possibilities and constraints inherent in the kit.[5]

There are different overlapping subsets of spatio-temporal possibilities, with different mathematical structures, including Euclidean and non-Euclidean geometries (e.g. the geometry of the surface of a hand, or face is non-euclidean) and various subsets of topology. Mechanisms for acquiring and using these "possibility subsets", i.e. possible action sequences and trajectories, seem to be used by pre-verbal children and other animals. That suggests that those abilities, must have evolved before linguistic capabilities. They seem to be at work in young children playing with toys before they can understand or speak a human language. The starting capabilities extended through much spatial exploration, provide much of the subject matter (semantic content) for many linguistic communications.

Some of the early forms of reasoning and learning in young humans, and corresponding subsets in other animals, are beyond the scope of current AI theorem provers, planners, reasoners, or learning systems that I know of. Some of those forms seem to be used by non-human intelligent animals that are able to perceive

[5] An evolving discussion note on this topic can be found here: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/construction-kits.html

both possibilities and constraints on possibilities in spatial configurations. Betty, a New Caledonian crow, made headline news in 2002 when she surprised Oxford researchers by making a hook from a straight piece of wire, in order to lift a bucket of food out of a vertical glass tube. Moreover, in a series of repeated challenges she made multiple hooks, using at least four very different strategies, taking advantage of different parts of the environment, all apparently in full knowledge of what she was doing and why – as there was no evidence of random trial and error behaviour. Why did she not go on using the earlier methods, which all worked? Several of the videos showing the diversity of techniques are still available here: http://users.ox.ac.uk/˜kgroup/tools/movies.shtml. The absence of trial-and-error processes in the successful episodes suggests that Betty had a deep understanding of the range of possibilities and constraints on possibilities in her problem solving situations.

It is very unlikely that you have previously encountered and solved the problem posed below the following image, yet many people very quickly think of a solution.

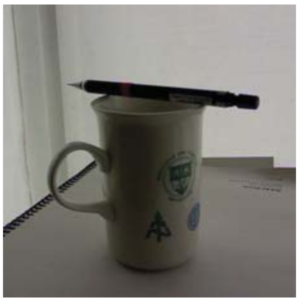


Fig. 2. Suppose you wanted to use one hand to lift the mug to a nearby table without any part of your skin coming into contact with the mug, and without moving the book on which the mug is resting, what could you do, using only one hand?

In order to think of a strategy you do not need to know the exact, or even the approximate, sizes of the objects in the scene, how far away they are from you, exactly what force will be required to lift the mug, and so on. It may occur to you that if the mug is full of liquid and you don't want to spill any of it, then a quite different solution is required. (Why? Is there a solution?).

The two pictures in Figure 3 present another set of example action strategies for changing a situation from one configuration to another. At how many different levels of abstraction can you think of the process, where the levels differ in the amount of detail (e.g. metrical detail) of each intermediate stage. For example, when you first thought about the problem did you specify which hands or which

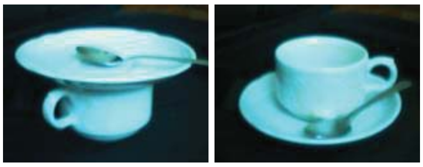


Fig. 3. Consider one or more sequences of actions that would enable a person or robot to change the physical configuration depicted on the left into the one depicted on the right – not necessarily in exactly the same locations as the objects depicted. Then do the same for the actions required to transform the right configuration to the left one.

fingers would be used at every stage, or at which location you would need to grasp each item? If you specified the locations used to grasp the cup, the saucer and the spoon, what else would have to change to permit those grasps? The point about all this is that although you do not normally think of using mathematics for tasks like this, if you choose a location at which to grasp the cup using finger and thumb of your left hand, that will mathematically constrain the 3-D orientation of the gap between between finger and thumb, if you don't want the cup to be rotated by the fact of bringing finger and thumb together. A human can think about the possible movements and the orientations required, and why those orientations are required, without actually performing the action, and can answer questions about why certain actions will fail, again without doing anything.

These are examples of "offline intelligence", contrasted with the "online intelligence" used in actually manipulating objects, where information required for servo-control may be used transiently then discarded and replaced by new information. My impression is that a vast amount of recent AI/Robotic research has aimed at providing online intelligence with complete disregard for the requirements of offline intelligence. Offline intelligence is necessary for achieving complex goals by performing actions extended over space and time, including the use of machines that have to be built to support the process, and in some cases delegating portions of the task to others. The designer or builder of a skyscraper will not think in terms of his/her own actions, but in terms of what motions of what parts and materials are required.

3.1 Limitations of sensorymotor intelligence

When you think about such things even with fairly detailed constraints on the possible motions, you will not be thinking about either the nervous signals sent to the muscles involved, nor the patterns of retinal stimulation that will be provided – and in fact the same actions can produce different retinal processes depending on the precise position of the head, and the direction of gaze of the eyes, and whether and how the fixation changes during the process. Probably the fixation requirements will be more constrained for a novice at this task than for an expert.

However, humans, other animals, and intelligent robots do not need to reason about sensory-motor details if they use an ontology of 3-D structures and processes, rather than an ontology of sensory and motor nerve signals. Contrast this with the sorts of assumptions discussed in [2], and many others who attempt to build theories of cognition on the basis of sensory-motor control loops.

As John McCarthy pointed out in [4] it would be surprising if billions of years of evolution failed to provide intelligent organisms with the information that they are in a world of persisting 3-D locations, relationships, objects and processes – a discovery that, in a good design, could be shared across many types of individuals with very different sensors and motors, and sensory motor patterns. Trying to make a living on a planet like this, whose contents extend far beyond the skin of any individual, would be messy and highly inefficient if expressed entirely in terms of possible sensory-motor sequences, compared with using unchanging representations for things that don't change whenever sensory or motor signals change. Planning a short cut home, with reference to roads, junctions, bus routes, etc. is far more sensible than attempting to deal, at any level of abstraction, with the potentially infinite variety of sensory-motor patterns that might be relevant.

This ability to think about sequences of possible alterations in a physical configuration without actually doing anything, and without having full metrical information, inspired much early work in AI, including the sorts of symbolic planning used by Shakey, the Stanford robot, and Freddy, the Edinburgh robot, over four decades ago, though at the time the technology available (including available computer power) was grossly inadequate for the task, including ruling out visual servo-control of actions.

Any researcher claiming that intelligent robots require only the right physical mode of interaction with the environment, along with mechanisms for finding patterns in sensory-motor signals, must disregard the capabilities and information-processing requirements that I have been discussing.

4 Inflating what "passive walkers" can do

Some (whom I'll not mention to avoid embarrassing them) have attempted to support claims that only interactions with the immediate environment are needed for intelligence by referring to or demonstrating "passive walkers",[6] without saying what will happen if a brick is in the way of a passive walker, or if part of the walking route starts to slope uphill. Such toys are interesting and entertaining but do not indicate any need for a "New artificial intelligence", using labels such as "embodied", "enactivist", "behaviour based", and "situated", to characterise their new paradigm. Those new approaches are at least as selective as the older reasoning based approaches that they criticised, though in different ways. (Some of that history is presented in Boden's survey [1].)

[6] E.g. http://www.youtube.com/watch?v=N64KOQkbyiI

The requirements for perception and action mechanisms differ according to which "central" layers the organism has. For instance, for an organism able to use deliberative capabilities to think of, evaluate, and select multi-step plans, where most of the actions will occur in situations that do not exist yet, it is not enough to identify objects and their relationships (pencil, mug, handle of mug, book, window-frame, etc.) in a current visual percept. It is also necessary to be able to "think ahead" about possible actions at a suitable level of abstraction, including consideration of objects not yet known, requiring a potentially infinite variety of possible sensory and motor patterns.

5 The birth of mathematics

The ability to reason about possible actions at a level of generality that abstracts from metrical details seems to be closely related to the abilities of ancient Greeks, and others, to make mathematical discoveries about possible configurations of lines and circles and the consequences of changing those configurations, without being tied to particular lengths, angles, curvatures, etc., in Euclidean geometry or topology. As far as I know, no current robot can do this, and neuroscientists don't know how brains do it. Some examples of mathematical reasoning that could be related to reasoning about practical tasks and which are currently beyond what AI reasoners can do, are presented on my web site.[7,8]

In 1971 I presented a paper at IJCAI, arguing that the focus solely on logic-based reasoning, recommended by McCarthy and Hayes in [5] could hold up progress in AI, because it ignored forms of spatial reasoning that had proved powerful in mathematics and practical problem solving. I did not realise then how difficult it would be to explain exactly what the alternatives were and how they worked – despite many conferences and journal papers on diagrammatic reasoning since then.

There have also been several changes of fashion promoted by various AI researchers (or their critics) including use of neural nets, constraint nets, evolutionary algorithms, dynamical systems, behaviour-based systems, embodied cognition, situated cognition, enactive cognition, autopoesis, morphological computation, statistical learning, bayesian nets, and probably others that I have not encountered, often accompanied by hand-waving and hyperbole without much science or engineering. In parallel with this there has been continued research advancing older paradigms for symbolic and logic based, theorem proving, planning, and grammar based language processing. Several of the debates are analysed in [1],

6 Other inadequacies

There are many other inadequacies in current AI, including, for example the lack of an agreed framework for relating information-processing architectures

[7] http://www.cs.bham.ac.uk/research/projects/cogaff/misc/torus.html

[8] http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-sum.html

to requirements in engineering contexts or to explanatory models in scientific contexts. For example attempts to model emotions or learning capabilities, in humans or other animals, are often based on inadequate descriptions of what needs to be explained, for instance poor theories of emotions that focus only on emotions with characteristic behavioural expressions: a small subset of phenomena requiring explanation or poor theories of learning that focus only on a small subset of types of learning (e.g. reinforcement learning where learners have no understanding of what's going on). That would exclude the kind of learning that goes on when people make mathematical discoveries or learn to program computers or learn to compose music.

Moreover, much AI research uses a seriously restricted set of forms of representation (means of encoding information) partly because of the educational backgrounds of researchers – as a result of which many of them assume that spatial structures must be represented using mechanisms based on Cartesian coordinates – and partly because of a failure to analyse in sufficient detail the variety of problems overcome by many animals in their natural environments.

Standard psychological research techniques are not applicable to the study of learning capabilities in young children and other animals because there is so much individual variation, but the widespread availability of cheap video cameras has led to a large and growing collection of freely available examples.

7 More on offline and online intelligence

Researchers have to learn what to look for. For example, *online* intelligence requires highly trained precisely controlled responses matched to fine details of the physical environment, e.g. catching a ball, playing table tennis, picking up a box and putting it on another. In contrast *offline* intelligence involves understanding not just existing spatial configurations but also the possibilities for change and constraints on change, and for some tasks the ability to find sequences of possible changes to achieve a goal, where some of the possibilities are not specified in metrical detail because they do not yet exist, but will exist after part of the plan has been carried out.

This requires the ability to construct relatively abstract forms of representation of perceived or remembered situations to allow plans to be constructed with missing details that can be acquired later during execution. You can think about making a train trip to another town without having information about where you will stand when purchasing your ticket or which coach you will board when the train arrives. You can think about how to rotate a chair to get it through a doorway without needing information about the precise 3-D coordinates of parts of the chair or knowing exactly where you will grasp it, or how much force you will need to apply at various stages of the move.

There is no reason to believe that humans and other animals have to use probability distributions over possible precise metrical values, in all planning contexts where precise measurements are not available. Even thinking about such precise values probabilistically is highly unintelligent when reasoning about topological relationships or partial orderings (nearer, thinner, a bigger angle, etc.) is all that's needed[9] Unfortunately, the mathematically sophisticated, but nevertheless unintelligent, modes of thinking are used in many robots, after much statistical learning (to acquire probability distributions) and complex probabilistic reasoning, that is potentially explosive. That is in part a consequence of the unjustified assumption that all spatial properties and relations have to be expressed in Cartesian coordinate systems. Human mathematicians did not know about them when they proved their first theorems about Euclidean geometry, built their first shelters.

8 Speculations about early forms of cognition

It is clear that the earliest spatial cognition could not have used full euclidean geometry, including its uniform metric. I suspect that the metrical version of geometry was a result of a collection of transitions adding richer and richer non-metrical relationships, including networks of partial orderings of size, distance, angle, speed, curvature, etc.

Later, indefinitely extendable partial metrics were added: distance between X and Y is at least three times the distance between P and Q and at most five times that distance. Such procedures could allow previously used standards to be subdivided with arbitrarily increasing precision. At first this must have been applied only to special cases, then later somehow (using what cognitive mechanisms?) extrapolated indefinitely, implicitly using a Kantian form of potential infinity (long before Kant realised the need for this).

Filling in the details of such a story, and relating it to varieties of cognition not only in the ancestors of humans but also many other existing species will be a long term multi-disciplinary collaborative task, with deep implications for neuroscience, robotics, psychology, philosophy of mathematics and philosophy of mind. (Among others.)

Moreover, human toddlers appear to be capable of making proto-mathematical discoveries ("toddler theorems") even if they are unaware of what they have done. The learning process starts in infancy, but seems to involve different kinds of advance at different stages of development, involving different domains as suggested by Karmiloff-Smith in [3].

For example, I recently saw an 11 month old infant discover, apparently with great delight, that she could hold a ball between her upturned foot and the palm of her hand. That sort of discovery could not have been made by a one month old child. Why not?[10]

Animal abilities to perceive and use complex novel affordances appear to be closely related to the ability to make mathematical discoveries. Compare the

[9] As I have tried to illustrate in: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/changing-affordances.html

[10] A growing list of toddler theorems and discussions of their requirements can be found in http://www.cs.bham.ac.uk/research/projects/cogaff/misc/toddler-theorems.html

abilities to think about changes of configurations involving ropes or strings and the mathematical ability to think about continuous deformation of closed curves in various kinds of surface.

Not only computational models, but also current psychology and neuroscience, don't seem to come close to describing these competences accurately or producing explanations – especially if we consider not only simple numerical mathematics, on which many psychological studies of mathematics seem to focus, but also topological and geometrical reasoning, and the essentially mathematical ability to discover a generative grammar closely related to the verbal patterns a child has experienced in her locality, where the grammar is very different from those discovered by children exposed to thousands of other languages.

There seem to be key features of some of those developmental trajectories that could provide clues, including some noticed by Piaget in his last two books on Possibility and Necessity, and his former colleague, Annette Karmiloff-Smith [3].

9 The Meta-Morphogenesis project

Identifying gaps in our knowledge requires a great deal of careful observation of many forms of behaviour in humans at various stages of development and many other species, always asking: "what sort of information-processing mechanism (or mechanisms) could account for that?"

Partly inspired by one of Alan Turing's last papers on Morphogenesis [10], I proposed the Meta-Morphogenesis (M-M) project in [9], a very long term collaborative project for building up an agreed collection of explanatory tasks, and present some ideas about what has been missed in most proposed explanatory theories.

Perhaps researchers who disagree, often fruitlessly, about what the answers are can collaborate fruitfully on finding out what the questions are, since much of what needs to be explained is far from obvious. There are unanswered questions about uses of vision, varieties of motivation and affect, human and animal mathematical competences, information-processing architectures required for all the different sorts of biological competences to be combined, and questions about how all these phenomena evolved across species, and develop in individuals. This leads to questions about what the universe had to be like to support the forms of evolution and the products of evolution that have existed on this planet. The Meta-Morphogenesis project is concerned with trying to understand what varieties of information processing biological evolution has achieved, not only in humans but across the spectrum of life. Many of the achievements are far from obvious.[11]

[11] A more detailed, but still evolving, introduction to the project can be found here: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

Unfortunately, researchers all too often mistake impressive new developments for steps in the right direction. I am not sure there is any way to change this without radical changes in our educational systems and research funding systems.

But those are topics for another time. In the meantime I hope many more researchers will join the attempts to identify gaps in our knowledge, including things we know happen but which we do not know how to explain, and in the longer term by finding gaps we had not previously noticed. I think one way to do that is to try to investigate transitions in biological information processing across evolutionary time-scales, since its clear that types of information used, the types of uses of information, and the purposes for which information is used have changed enormously since the simplest organisms floating in a sea of chemicals.

Perhaps some of the undiscovered intermediate states in evolution will turn out to be keys to unnoticed features of the current most sophisticated biological information processors, including humans.

References

1. Boden, M.A.: Mind As Machine: A history of Cognitive Science (Vols 1–2). Oxford University Press, Oxford (2006)
2. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36(3), 1–24 (2013)
3. Karmiloff-Smith, A.: Beyond Modularity: A Developmental Perspective on Cognitive Science. MIT Press, Cambridge, MA (1992)
4. McCarthy, J.: The well-designed child. Artificial Intelligence 172(18), 2003–2014 (2008), http://www-formal.stanford.edu/jmc/child.html
5. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of AI. In: Meltzer, B., Michie, D. (eds.) Machine Intelligence 4, pp. 463–502. Edinburgh University Press, Edinburgh, Scotland (1969), http://www-formal.stanford.edu/jmc/mcchay69/mcchay69.html
6. Minsky, M.L.: Steps toward artificial intelligence. In: Feigenbaum, E., Feldman, J. (eds.) Computers and Thought, pp. 406–450. McGraw-Hill, New York (1963)
7. Sauvy, J., Sauvy, S.: The Child's Discovery of Space: From hopscotch to mazes – an introduction to intuitive topology. Penguin Education, Harmondsworth (1974), translated from the French by Pam Wells
8. Sloman, A.: Evolution of minds and languages. What evolved first and develops first in children: Languages for communicating, or languages for thinking (Generalised Languages: GLs)? (2008), http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0702
9. Sloman, A.: Virtual machinery and evolution of mind (part 3) meta-morphogenesis: Evolution of information-processing machinery. In: Cooper, S.B., van Leeuwen, J. (eds.) Alan Turing - His Work and Impact, pp. 849–856. Elsevier, Amsterdam (2013), http://www.cs.bham.ac.uk/research/projects/cogaff/11.html#1106d
10. Turing, A.M.: The Chemical Basis Of Morphogenesis. Phil. Trans. R. Soc. London B 237 237, 37–72 (1952)

Work on the Dual Structure of Lexical Semantic Competence

Invited talk at AIC 2014

November 26-27 University of Turin, Italy

http://aic2014.di.unito.it

Diego Marconi

Università degli Studi di Torino, Italy

Abstract. Philosophical arguments and neuropsychological research on deficits of lexical processing converge in indicating that our competence on word meaning may have two components: inferential competence, that takes care of word-word relations and is relevant to tasks such as recovery of a word from its definition, pairing of synonyms, semantic inference ("Milan is north of Rome" → "Rome is south of Milan") and more; and referential competence, that takes care of word-world relations, or, more carefully, of connections between words and perception of the outside world (through vision, hearing, touch). Normal subjects are competent in both ways; however, there are patients in which one component seems to be impaired while the other performs at normal level. Typically, cases are found of patients that are excellent at defining, say, the word 'duck' but cannot recover the word when shown the picture of a duck. Complementary cases have also been found and studied. Recent experiments using neuroimaging (fMRI) found that certain visual areas are active even in purely inferential performances, and a current experiment appears to show that such activation is a function of what might be called the "visual load" of both the linguistic material presented as stimulus and the target word. Such recent results will be presented and discussed. It should be noted that the notion of "visual load", as applying to both individual words and complex phrases, has also been given a computational interpretation.

Brain for Robots

Invited talk at AIC 2014

November 26-27 University of Turin, Italy

http://aic2014.di.unito.it

Giulio Sandini

IIT, Italian Institute of Technology,
Università degli Studi di Genova, Italy

Abstract. Simulating and getting inspiration from biology is not a new endeavor in robotics [1]. However, the use of humanoid robots as tools to study human cognitive skills it is a relatively new area of the research which fully acknowledges the importance of embodiment and interaction (with the environment and with others) for the emergence of motor and perceptual skills, sensorimotor coordination, cognitive and social abilities [2]. Within this stream of research "developmental robotics" is a relatively new area of investigation where the guiding philosophy – and main motivation – is that cognition cannot be hand-coded but it has to be the result of a developmental process through which the system becomes progressively more skilled and acquires the ability to understand events, contexts, and actions, initially dealing with immediate situations and increasingly acquiring a predictive capability [3]. The aim of this talk is to present the guiding philosophy – and main motivation – and to argue that, within this approach, robotics engineering and neuroscience research are mutually supportive by providing their own individual complementary investigation tools and methods: neuroscience from an "analytic" perspective and robotics from a "synthetic" one.

References

1. Atkeson, C.G., Hale, J.G., Pollick, F.E., Riley, M., Kotosaka, S., Schaul, S., Shibata, T., Tevatia, G., Ude, A., Vijayakumar, S., et al.: Using humanoid robots to study human behavior. IEEE Intelligent Systems and their applications 15(4), 46–56 (2000)
2. Sandini, G., Metta, G., Konczak, J.: Human sensori-motor development and artificial systems. In: Proc. of the Int. Symp. on Artificial Intelligence, Robotics, and Intellectual Human Activity Support for Applications. pp. 303–314 (1997)
3. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. Evolutionary Computation, IEEE Transactions on 11(2), 151–180 (2007)

An Interdisciplinary Approach for a Holistic and Embodied Emotion Model in Humanoid Agents

Samer Schaat[1], Matthias Huber[2], Klaus Doblhammer[1], Dietmar Dietrich[1]

[1]Institute for Computer Technology, Vienna University of Technology, A – 1040 Vienna
{schaat, doblhammer, dietrich}@ict.tuwien.ac.at
[2]Department of Education, University of Vienna, A – 1010 Vienna
matthias.huber@univie.ac.at

Abstract. Computational models of the human decision making process have already enabled several insights and applications. Nevertheless, such models have only recently begun to consider that simply modelling rational decision making was insufficient to represent human behavior. Recent efforts have started to consider this factor by utilizing computational models of emotions. One of the most significant challenges to be faced here is the interdisciplinary cooperation required in order to develop a holistic and integrated model, which reflects aspects of embodiment and is integrated in a holistic cognitive architecture. In this paper we will analyse how models from relevant disciplines can support us in outlining an overview model which considers the mentioned criteria.

Keywords: Emotions · Humanoid Agents · Evaluation Models · Decision Making · Simulation · Artificial General Intelligence · Neuroscience · Neuropsychoanalysis.

1 Introduction

If one considers humans as being the archetypes of intelligent information processing, the fundamental mechanisms of human decision making must be considered. This is especially relevant for a holistic model of decision making. Over the last years the focus in this area was on rational decision making, with little consideration for the fact that this is actually based upon *unconscious processes*. In particular the consideration of affective processes is needed in a computational representation of human capabilities that are used to cope with every-day problems, such as “intuition”, “gut feeling” etc. A further tendency in humanoid agents is the reduction of decision making to a set of universal rules. Even when the abstract principles of decision making are generally valid, the definite rules followed by humans cannot be generalized, and therefore must be generated according to a *subjective approach*, based on memory and also in an agent specific manner. Besides, conventional decision making models are mostly abstracted from the body. Following an *embodied approach*, human cognition and particularly decision making is however based upon the interaction of the brain with the rest of the body.

These topics can only be tackled with an *interdisciplinary approach,* where computer science is not used as a tool, but its methodologies are applied on the content of other disciplines. Such approach not only enables to develop computer models for technical applications, but also serves precision, concretization and sharpening of concepts from different disciplines. In order to develop a deterministic model of an agent decision making process according to human-based principles, we will leverage insights from different disciplines, considering embodiment, subjective approaches and unconscious processes. In particular we consider neuroscience, psychoanalysis, neuropsychoanalysis, and current computation models of emotions. Although the unconscious is a key concept in psychoanalysis, as far as we know, it is not analyzed yet for computational models of affective processes, and neuroscientific theories are not reflected in a holistic and embodied computational model of human decision making yet. In order to further construct a model transformation upon this and to integrate neuroscientific and (neuro)psychoanalytic models into a technical model, it is necessary that they are described *consistently and holistically*. Even though this paper shall focus on a model of the basic mechanisms of decision making, the holistic perspective may not be neglected. This represents the integration of cognition and emotion, which are thereby not modelled separately. Upon highlighting how each discipline (by an expert of neuroscientific emotion theories, a psychoanalyst and a computer scientist) perceives the fundamental mechanisms of subjective decision making in terms of the criteria described, we shall sketch an evaluation model for subjective decision making in humanoid agents which integrates the insights of the various disciplines.

2 Emotions in Neuroscience

Currently several epistemologically divergent neuroscientific research movements are endeavoring to describe human emotionality in a holistic context. In particular the work of Jaak Panksepp is of note, specifically his theory of mutually linked (neuronal interacting) basic emotion systems (seeking, fear, panic, rage) [1]. Also of note is the work of Gerald M. Edelman and his theory of Neuronal Darwinism in which emotions, in the context of values and value systems, are introduced in a holistic theory of consciousness [2], a precursor to a modern integrative comprehension of emotions. The research of Joseph LeDoux concerning the “Amygdala Fear System” [3] and also the work of James McGough on the significance of emotions in learning and memory [4] are also worthy of mention. The Portuguese Antonio Rosa Damasio is recognized as one of the most significant and influential contemporary neuroscientists. His “Theory of Emotion” is highly regarded far beyond neurobiology circles, as his comprehension of emotions and feelings possesses great explanatory potential, confirmed through diverse studies [5], [6], and a plethora of neurophysiological data [7], [3], [8]: On the one hand Damasio is able to justify the holistic claim as his comprehensive body of work includes, along with the Theory of Emotion, a “Theory of Self”, a “Theory of Consciousness”, a “Theory of Mind” along with a theory governing the automatic regulation of life. On the other hand, Damasio also took on the challenge of creating a model for the dualistic juxtaposition of body and mind, based on the inevitable interdependence of reason and feelings, thereby transporting an anthropological comprehension which satisfies both the psychoanalytical developmental theories and phenomenological traditions.

The starting point of Antonio Damasio's deliberations is the organism, as a *holistic, open system*, that progressively interacts with the environment and is organized according to the operating principle of homeostasis (automatic life regulation), whereby emotions form the ultimate level of these regulative (permanently active) mechanisms. This also clarifies why emotion and feeling take on a biological (and also evolutionary) function as necessary survival regulators for the protection of an organism. Damasio differentiates in this context between three processing stages: (1) an emotional status, which can be unconsciously triggered and executed, (2) a feeling state, which can be unconsciously represented along with (3) a consciously generated feeling state, in which the organism knows that it possesses emotions but also feelings [6, p. 51]. Furthermore three types of emotions are differentiated: (a) the primary emotions (fear, anger, sorrow, happiness, disgust, surprise) which are congenital (preorganized in terms of classical "Jamesist" feelings) and can be thought of as being genetically determined, universal and culturally dependent basic emotions. (b) The secondary or social emotions which in contrast develop over the course of ontogeny and emerge as soon as systematic connections of objects and/or situations with primary emotions are formed. Social or secondary emotions (e.g. compassion, embarrassment, shame, guilt, pride, jealously, awe, envy etc.) are thereby acquired and respectively triggered through mental registration with respect to the cognitive processing of situations and/or objects. (c) Background emotions are seen by Damasio as being the consequence of certain combinations of simple regulative (homeostatic) reactions (e.g. desire) [9, p. 56]. They are ever-present but are seldom consciously perceived and may be interpreted as being an expression of well-being or discomfort. Emotions thereby fulfill a double biological function: as already mentioned they must continuously *regulate the inner status of the organism*. Above and beyond that they must also *trigger a specific reaction to a particular stimulus or situation*. Two mechanisms are available for this: emotions are formed either when our sensory organs process certain objects and/or situations (the body-loop) or when the organism retrieves certain objects and/or situations from memory and represents them as imaginary images in the thought process (the as-if-loop). It has already been stated that emotions and feelings are temporally and structurally predetermined: the key content of a feeling is the illustration of a particular bodily state. A feeling is thereby a *projection of the body under certain conditions* and the feeling of an emotion is the projection of the body while under the influence of a particular emotional process. Additionally, along with the bodily-related projections, in certain situations specific projections of the thought process are also relevant. Therefore a (conscious) feeling consists of the perception of a certain bodily state along with the perception of a certain associated mental state (the cognition of the emotion). As before, Damasio also differentiates here between three types of feelings: (a) feelings of basic universal emotions (in terms of the primary emotions), (b) feelings of differentiated universal emotions (as a connection of cognitive states with emotional bodily states; depending on experience, in terms of the social emotions), along with (c) background sensations (in terms of background emotions, although not formed by emotions in the strictest sense) [5, p. 208].

In the context of this "Theory of Emotion", an elementary principle is clearly evident in Damasio's work: emotion, feelings and consciousness are continually dependent on the representations of the organism and their common entity is and remains the body. Human thoughts and actions are therefore dependent on the emotional constitution and respectively to certain changes in bodily state. The purpose of thought and the prerequisite for action is however always a decision, whereby the essence of a decision lies in choosing a certain response (e.g. a course of action) [5, p. 227].

If one considers the decision making process on the basis of an undesirable development, one thereby creates an undesirable/negative outcome, that is connected to the associated response and is consciously perceived, even when short-lived, as an uncomfortable/negative feeling. As a feeling (from emotion) affects the body, Damasio choose the term somatic (soma = body), and as the feeling identifies, or marks a projection, he also chose the term marker. A somatic marker is thereby understood as being the perception of an automatic reaction of the body to a certain projected image (a situation or an event) respectively, as a bodily signal marking a particular scenario as being either good or bad. [5, p. 238]. Accordingly a positive somatic marker functions as a start signal and a negative somatic marker functions as an inhibitor. Somatic markers are formed throughout the course of upbringing and socialization through connecting certain classes of stimuli with certain classes of somatic statuses. Therefore they touch on the process of the secondary emotions. The adaptive function of the somatic marker (as an assistant with anticipatory skills) orientates itself towards congenital, regulatory dispositions (internal preference system) which ensure the survival of the organism and take care of avoiding pain and seeking or increasing desire.

Looking back at the functional mechanisms of emotion, one can differentiate four forms of the decision making process: (A) in the context of the body-loop, the body is actually (from the prefrontal cortex and the amygdala) prompted to take on a certain state profile, the result of which (via the somatosensory cortex) is considered with attention and perceived. (B) In the context of the as-if-loop the somatosensory cortex functions (as instructed by the prefrontal cortex and the amygdala) as if the signals were received from the body. Therefore the body is taken out of the loop, nevertheless the as-if activity patterns influence the decision making process, as it suggests that real bodily states are symbolically processed. (C) Additionally *somatic markers* (e.g. feelings) can represent very concrete components or *triggers for decisions*, regardless of whether they follow a real or representative route. (D) Very often decisions are made, where it appears that no feeling at all was involved. Therefore it is not – and that is key here – that it doesn't come to an evocation of a bodily state or that of a representative surrogate, but rather just that of the bodily state with which the signal function is activated, it is simply just not considered and therefore not consciously perceived [10, p. 84]. By this means somatic markers operate permanently outside of consciousness and persistently influence conscious thought and decisions. Therefore one differentiates between somatic markers with respect to the influence of emotion and feeling on the decision making processes based on their neural route, (A) the real body-loop versus (B) the representative as-if-loop and on the basis of their influence, (C) manifest versus (D) covert.

3 Evaluation Models in Psychoanalysis

There are multiple aspects to the description of decision making in the "psychic apparatus" in classical psychoanalytic theory [11, 14]: On the one hand the body delivers via homeostatic differences drive tensions, so-called "*quota of affects*" represented in the psyche, which can consequently cathect[1] psychic contents. This allocation of the "quota of affects" to psychic contents already activates content in the unconscious whereby these become accessible for mental processing. The level of *cathexis* is a measure of the grade of the activation, representing an evaluation, which ultimately is a key factor in determining, if content shall be processed and ultimately become consciously perceivable and actionable. The psychic contents on the other hand, come from *memory traces* which are associated via perceptional data and drive representatives and by this means serve as a basis for cathexis. A cathected association complex is called thing presentation. This is, during the transition from (unconscious) primary process to secondary process – upon going through a conflict regulating defense – linked with a so-called word presentation, which means, that from now on a psychic content can be treated within general (formal, verbal) logic. This was not yet possible in the primary process, as this is governed by a pre-logical order of associations. Cathexes, which in the course of the mental processing of the primary process have been displaced many times, remain intact throughout these transitions. In the secondary process, the topical description of which encompasses the psychoanalytical preconscious and conscious areas, are now fixed to the "quota of affects" of certain contents and contribute, along with the logical links via word presentations to the evaluation of the association complexes.

The overall evaluation of action chains generated in this manner is regulated on the one hand – in the primary process – by the so called "*pleasure principle*" and in the secondary process by the "*principle of reality*". The "pleasure principle" states that the overall goal of all mental activity is to avoid unpleasure and to aim for pleasure (as in [12, p.321]), the "principle of reality" is a variation upon this, namely the moments in which the outside world becomes included into these activity designs (as in: [12, p.378]). Pleasure is created if psychic energy is discharged, unpleasure is equivalent to the "quotas of affect" present within the apparatus. Both of these principles are ultimately relevant for decision making and choosing a course of action, so much so, that in total a maximum of expected pleasure less the necessary unpleasure in order to achieve it is always sought. Primary and secondary process oriented thinking and evaluation mechanisms complement each other towards taking action. Thinking is essentially an experimental kind of acting. [13, p.220]. If action is undertaken, the "quota of affects" is discharged and alongside the physical impulses, consciously perceptible affects and impulses of feelings in particular shades are formed. Uncon-

[1] In psychoanalysis this also known as the economic aspect of psychological operation. Cathexis (a psychoanalytic term) is the allocation of quota of affects to psychic contents.

scious affects, feelings and sensations are of no relevance in psychoanalytic theory, they are, - in contrast to psychic contents –virtual qualities, which with respect to the occupation conditions of the "quota of affects" can be constructed retrospectively [14, p.176].

4 Emotions in Neuropsychoanalysis

Neuropsychoanalysis seeks to forge a connection between psychoanalytical models and related neuroscientific findings [7, 15]. It seeks to assume a neuroscientific perspective of every mental function and thereby wishes to reassess Freud's description of the functional organization [15, p. 830]. The results of these comparisons may remain patchy in accordance with the method applied and usually only the most significant theses and statements of these disciplines are studied.

Considering evaluation models for decision making, neuropsychoanalysis holds that the conscious decisions for actions, in reality follow their unconscious initiation [cf. 15, p. 384]. Feelings of pleasure and unpleasure with respect to an object or a situation represent the most elementary evaluation of a consciousness, according to Solms, Panksepp and Damasio [15, p. 836]. Likewise the basic units of these evaluations, of the driving forces, can be illustrated neurologically and yield that: An amazingly large consensus emerges between Panksepp's SEEKING-System and Freud's Libido-System and the highest priority of a regulating function of consciousness is to generate feelings of pleasure and unpleasure, which in turn are then associated with the objects which are best suited to their generation [cf. 15, p. 848].

5 Emotions in Agent-based Systems

In recent years several computational models of emotions have been developed and integrated in the decision making process of artificial agents. These models differ in various aspects. Generally they differ in the components which are considered as being intrinsic to emotions (e.g. bodily processes, behavioral tendencies), in their relationship to cognitive processes and in their representation [16]. The most important difference lies in the supporting theory, which in most cases originates from psychology. This in turn influences the distinction if emotions are generated dynamically (emergent emotions) or are designed explicitly (discrete emotions). These aspects of distinction are mirrored in the division of "appraisal, dimensional and anatomical" computational models of emotion [16], whereby the former is the most widespread due to its aptness in linking emotion with cognitive processes.

In appraisal models (e.g. OCC Modell [17]), emotions are formed through the evaluation of external events regarding the agent's beliefs, desires and intentions, whereby coping strategies (e.g. planning, delaying) are triggered. A computational model of this, which offers the chance to adjust the appraisal process, is FAtiMA Modular [18]. The agent architecture consists of an extensible core architecture which offers the framework for various implementations of appraisal models, enabling easier comparison. The appraisal process is split into the appraisal derivation and the affect derivation. The former evaluates the relevance of an event and creates appraisal variables (e.g. unexpectedness, appealingness and desirability). The latter builds upon these variables by creating the associated emotions, determined by a specific appraisal theory. According to a set of rules, emotions then influence the choice of action in either a reactive or deliberate form.

In dimensional models, emotions are located in a dimensional space instead of being formulated as discrete entities. A typical model is the PAD model [19], with the dimensions pleasure, arousal and dominance. Some computational models such as ALMA [20] and WASABI [21] utilize both, appraisal models to model appraisal processes, and dimensional models to model mood processes. Other models such as MicroPsi [22, p.143-155] describe emotions implicitly as regions of a multi-dimensional space, defined by the parameters which determine the behavior of the agent. These are: arousal, resolution level, dominance of the leading motive, the level of background checks (the rate of the securing behavior), the level of goal-directed behavior, and valence. Thus, explicit emotions do not exist for agents but rather emotions are first attributed to an agent upon (self) perception.

In conclusion, it can be stated that none of the models mentioned offers an embodied model that holistically considers the various aspects of emotion, or is integrated in a holistic cognitive architecture.

6 A Holistic and Embodied Emotion Model for Evaluation in Decision Making

Building on the findings of the various disciplines, we sketch a holistic and embodied model. As shown above, the models of Damasio and psychoanalysis fulfill the initially mentioned criteria especially well. Insights from both theories confirm and complement each other. For a technical model of the basic mechanisms of decision making, the psychoanalytic findings mentioned above are particularly well suited as an abstract framework (which is required for a holistic and coherent model), and Damasio's model is especially well suited for its concretization due to its consistent and holistic character in considering the interaction between body and mind.

The role of computer science here is to integrate the various models from other disciplines in a consistent and coherent model of decision making, which is deterministic and can therefore be validated by means of simulation. Thus, computer science enables a model building methodology and evaluation tool, by the means of agent based simulation. The basic principle of this approach has been illustrated in a previous article [23], where a functional view of emotions in the decision making unit of an agent was integrated. However, a holistic view considering the theories of Damasios was neglected. Nevertheless, the fundamental principle remains intact, i.e. affective evaluation processes are the foundation of evaluating data (psychic contents, thereby also actions and plans). These are those processes which use "quotas of affects" or derived evaluation variables to determine the relevance of data, based on memories, for decision making in a given situation (see Fig. 1). Thus, the evaluation of data is an incremental process on multiple levels – considering various evaluation principles (pleasure and reality principles) and evaluation influences (bodily influences and through perception activated memories and fantasies).

Emotions are an additional level of this incremental and hierarchical multi-level evaluation model. They represent (1) “quotas of affects” from the drives, (2) emotions activated through perception and fantasy (memories associated with emotions), (3) the current pleasure. Hence emotions form a holistic representation of the *psychobiological* status of the agent (having information concerning the body and mental status) and can therefore consider the overall status of the agent in the evaluation. The final evaluation step is carried out by feelings, whereby depending on the intensity of the emotion, it is transformed into a preconscious feeling and subsequently a consciously “felt feeling” (in the sense of Damasios). The latter can be described as an inner perception, upon which the agent can reflect. As with the other valuation variables, feelings evaluate goals and plans by activating memories.

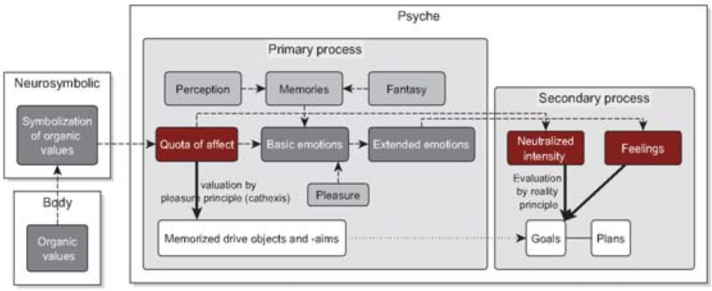


Fig. 1. Evaluation is an incremental process that considers multiple influences and principles.

By considering perception and fantasy, evaluation through feelings not only occurs in the terms of gaining pleasure, but also in terms of avoiding unpleasure, that is to say not just to support the fulfilment of drives but also to evaluate external events in terms of their potential to increase unpleasure. Evaluation generally serves to prioritize and select actions, mediating between the environment and the internal state (e.g. to fulfil desires in the environment and to adapt desires to external conditions).

The representation of the biological aspect of the psychobiological status is achieved through drives and body perception (proprioceptive and external perception). Whereby it must be emphasized that memory-based psychic representation (representation of drives and body representation) is used for emotion generation (and not the body signals as such). The psychic aspect of the psychobiological status is represented by the memories activated by environment perception and fantasies. In the sense of Damasio, one can conceptually speak of background emotions (red influences in Fig. 2), which can be considered as moods, and emotions triggered from the outside world.

The key concretization, when integrating Damasio’s model, concerns the consideration of the embodiment by means of a mental representation of it. In this regard we follow the approach of considering the psyche as an information theoretical level of the physical world. This is reflected in the differentiation between the neural, neuro-symbolic and physical levels (see Fig. 2), and also in the application of a memory based physical representation (see Fig. 2).

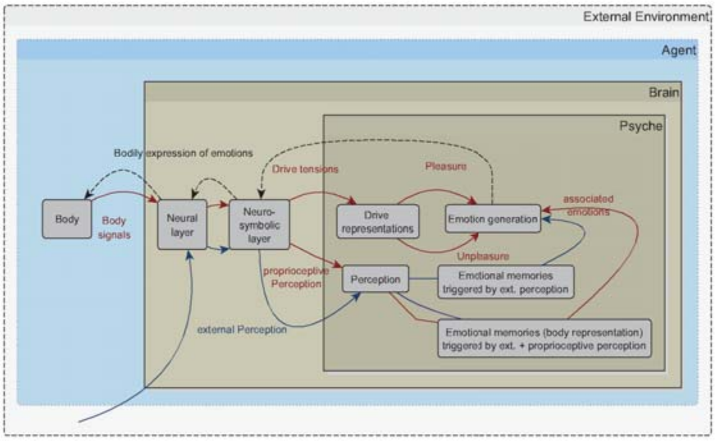


Fig. 2. A holistic and embodied emotion model.

7 Conclusion

Interdisciplinary cooperation enabled us to outline a holistic model of emotion for humanoid agents. Due to the consideration of bodily influences and the evaluation of perceived events, the model outlined can be considered as being a combination of both appraisal and dimensional models, whereby neuroscientific inputs and embodiment are considered in developing a holistic model. Psychoanalysis and Damasio’s neuroscientific model fulfill the initially mentioned criteria particularly well. Whereby psychoanalysis offers us an abstract holistic framework which can be concretized by means of Damasio’s model. Both models also complement each other, as Damasio considers the interdependence of body and mind more concretely and is more consistent. Neuropsychoanalysis supports our approach of combining psychoanalysis with neuroscientific models, by revealing supporting evidence. Computer science enables the combination of the various models in one consistent and holistic model and offers an evaluation tool by means of simulation. Having outlined such a simulation model, the next steps lie in extending an existing implementation of a holistic functional model of the human mind [23], to integrate these new findings in a holistic model of human information processing. We expect that the integration of the various models in an overall evaluation model will yield new discoveries and opportunities in simulations.

References

1. Panksepp, J.: Affective Neuroscience. The Foundations of Human and Animal Emotions. Oxford University Press, New York (1998)
2. Edelman, G. M., Tononi, G.: Neuronaler Darwinismus. Eine selektionistische Betrachtungsweise des Gehirns. In: Meier, H., Ploog, D. (eds.): Der Mensch und sein Gehirn. Die Folgen der Evolution, 187-233, Piper Verlag, München (1997)
3. Ledoux, J.: The Emotional Brain. The Mysterious Underpinnings of Emotional Life. Simon and Schuster, New York (1996)
4. Cahill, L., Mcgough, J. L.: Mechanisms of emotional arousal and lasting declarative memory. In: Trends in Neurosciences 21, 284-299 (1998)
5. Damasio, A. R.: Descartes' Irrtum. Fühlen, Denken und das menschliche Gehirn. [Descartes' Error. Emotion, Reason, and the Human Brain Putnam. Penguin edition, London (2005)] Deutscher Taschenbuch Verlag, München (2001)
6. Damasio, A. R.: Ich fühle, also bin ich. Die Entschlüsselung des Bewusstseins. [The Feeling of What Happens. Body and Emotion in the Making of Consciousness. Harcourt, New York (1999)] List Verlag, Berlin (2007)
7. Solms, M., Turnbull, O.: The Brain and the Inner World. An Introduction to the Neuroscience of the Subjective Experience. Other Press, New York (2002)
8. Kandel, E. R.: In Search of Memory. The Emergence of a New Science of Mind. Norton, New York (2006)
9. Damasio, A. R.: Der Spinoza-Effekt. Wie Gefühle unser Leben bestimmen. [Looking for Spinoza. Joy, Sorrow, and the Feeling Brain. Harcourt, New York (2003)] List Verlag, Berlin (2007)
10. Huber, M.: Die Bedeutung von Emotion für Entscheidung und Bewusstsein. Die neurowissenschaftliche Herausforderung der Pädagogik am Beispiel von Damasios Theorie der Emotion. Verlag Königshausen & Neumann, Würzburg (2013)
11. Freud, S.: The Interpretation of Dreams. The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume IV. Hogarth Press, London (1900)
12. Laplanche, J., Pontalis, J.B.: The Language of Psycho-Analysis.London: The Hogarth Press and the Institute of Psycho-Analysis, London (1973)
13. Freud, S.: Formulations on the Two Principles of Mental Functioning. The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XII, Hogarth Press, London, 213-226 (1911)
14. Freud, S.: The Unconscious. The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIV. Hogarth Press, London, 159-215 (1915)
15. Solms, M.: Sigmund Freud heute. Eine neurowissenschaftliche Perspektive auf die Psychoanalyse. Psyche. 60, pp. 829-859 (2006)
16. Marsella, S. C., Gratch, J., Petta, P.: Computational models of emotion, In: K.R. Scherer, T. Bänziger, and E. Roesch, (eds.), Blueprint for Affective Computing. Oxford University Press, New York (2010)
17. Ortony, A., Clore, G. L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, New York (1990)
18. Dias, J., Mascarenhas, S., Paiva, A.: FAtiMA Modular: Towards an Agent Architecture with a Generic Appraisal Framework. In: Proceedings of the International Workshop on Standards for Emotion Modeling (2011)
19. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. MIT Press, MA (1974)

20. Gebhard, P.: ALMA: A layered model of affect. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems. pp. 29-36 (2005)
21. Becker-Asano, C., Wachsmuth, I.: Affective computing with primary and secondary emotions in a virtual human. Auton Agent Multi-Agent Syst. 20, pp. 32-40 (2010)
22. Bach, J.: Principles of Synthetic Intelligence, Psi: An Architecture of Motivated Cognition. Oxford University Press, New York (2009)
23. Schaat, S., Doblhammer, K., Wendt, A., Gelbard, F., Herret, L., Bruckner, D.: A Psychoanalytically-Inspired Motivational and Emotional System for Autonomous Agents. In: Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, pp. 6648 - 6653 (2013)

Using Meta-Cognition for Regulating Explanatory Quality Through a Cognitive Architecture

John Licato • Ron Sun • Selmer Bringsjord

Rensselaer Polytechnic Institute
Troy, NY, USA
{licatj, rsun, selmer}@rpi.edu

Abstract. Recent years have seen a renewed interest in cognitive systems with the ability to explain either external phenomena or their own internal reasoning processes while solving problems. Some successful models of explanation-generation have made use of structured representations, reasoned over using analogical or deductive mechanisms. But before such models can be adapted for use in real-world situations, they need to incorporate additional features associated with explanation-generation. For example, generated explanations may differ qualitatively based on the explanandum's domain; e.g., explanations rooted in physical causality to explain physical phenomena vs. folk-psychology explanations that rely on propositional attitudes (believes, knows, intends, ...). This may affect the generated explanations in both explicit and implicit ways. We tackle both the explicit and implicit effects of this cognitive feature and incorporate them into a comprehensive cognitive architecture: CLARION (especially its meta-cognitive and non-action-centered subsystems).

Keywords: Explanation, Cognitive Architecture, CLARION, Analogy, Deductive Reasoning, Meta-Cognition

1 Introduction: Features of Explanations

The importance of a cognitive system's ability to explain its results, or the actions of others, and to produce *useful* explanations, is being increasingly realized by AI researchers. But as has been known for quite some time now, there are a variety of explanations that might be considered useful. For example, if one wishes to tell some cognitive system **W** that a chicken crossed the road (which happened to require movement in an eastward direction), there are at least two different ways of presenting more or less the same thing:

E_1 Chicken C wanted to cross the road.
E_2 Muscle contractions in chicken C propelled it eastward.

These two explananda refer to the same event at different levels of abstraction by invoking different concepts. The type of explanation (or alternately, explanans) that might be deemed an appropriate response to each of these explananda differ as well. An explanation whose language features many propositional attitudes of the chicken (e.g. *"C believes," "C knows," "C wants," etc.*) may be appropriate for explaining E_1, but may not constitute a satisfactory explanation in response to E_2. An explanation rooted in physical causality (referring to the normal properties of muscle contractions, for example) may be the other way around: it would be appropriate for E_2 but less so in response to E_1. In short, the presentation of the explanandum affects the sort of explanation that is most appropriate.

The question hinted at in the above example, of whether to root an explanation in physical causality or propositional attitudes, reflects a parallel one faced by cognitive systems: What factors are used by agents to determine which qualitative features of an explanation are appropriate? In the present paper, we explore and model one answer to this question: that the concepts used in the *presentation* of the explanandum affect the explanans in both implicit and explicit ways. We model these ways using the Meta-Cognitive Subsystem (MCS) of the CLARION cognitive architecture.

We do not hope, nor do we attempt, to resolve any questions regarding whether one type of explanation is *better* than another. Although discussion in the philosophical literature of the so-called *intentional stance* [3], the normative views of Hempel [8], and so on, are fascinating and informative, we are here only concerned with modeling the cognitive processes that lead humans to choose one style of explanation over another.

The remainder of this paper proceeds as follows. After further motivating the features the modeling of which is our target, will first discuss related previous work in modeling explanation-generation, in order to set the stage for the communication of our own, and to introduce concepts we use in this paper such as metaknowledge, metacognition, and so on (§2). In section 3, we present the cognitive architecture CLARION, and briefly discuss recent developments in its representational capabilities which make it possible for us to do the work we present herein. We close with brief demonstrations in section 4, and section 5 concludes with final remarks.

1.1 Effects of the Explanandum's Presentation

The type of feature of explanation-generation we aim to model here, which we refer to as **F** effects for convenience, are the effects that the presentation of the explanandum has on the explanation generated. If the explanandum e is a simple fact about some world, let us define the *full* explanandum E as the explanandum plus all of the contextual facts required to understand the explanandum. For example, to return to the earlier example, the position of the chicken relative to the road, the position of the road relative to the four cardinal directions, and so on, are all examples of facts comprising E. The presentation of the full explanandum $P(E)$ is a particular form of the full explanandum E. This distinction is important. E_1 and E_2 might be considered two partial presentations of the same full explanandum, but they differ in their presentations.

F effects, then, are those which the presentation of the full explanandum exhibits on the explanations generated. We can further subdivide these into $\mathbf{F_e}$ effects, and $\mathbf{F_i}$ effects; these are explicit and implicit effects, respectively. Examples of both in the psychological literature are numerous, e.g. see [23, 13].

Determinations of similarity based on simple featural overlap might be considered an implicit process, or one that operates primarily using the representations on CLARION's lower level [23], if the features in question are predominantly micro features not immediately verbalizable. Such similarity processes are known to be used in analogical reasoning, particularly in the initial stages, which use surface similarity to select source analogs from long-term memory [9, 17, 7].

But explicit processes may play a large role in explanation as well. One way to identify explicit processes, or those that operate primarily using the representations like those on CLARION's top level, is to perform experiments on human subjects that require them to verbalize their thoughts in some way. In explanation, one example relates to the so-called "self-explanation effect," in which children who verbalize their explanations seem to be able to improve the quality of their learning, and learn more [2]. This effect also applies to adults who actively create explanations for their own use [1]. Furthermore, explaining the reasoning of the beliefs and the reasoning of others also directly enhances learning [19]; this suggests that encouraging development of theory of mind may be helpful in teaching [31].

Our basic hypothesis for the modeling of **F** effects in the present paper is that the knowledge structures used to construct explanations are selected based on parameters in the metacognitive system, which themselves may be influenced, either explicitly or implicitly, by the concepts used in the explanandum's presentation.

2 Metacognition and Explanation Generation in Cognitive Systems

In this section we provide an overview of some recent modeling of metacognition in order to give the reader a feel for the state of the art in the field, and to clarify the present paper's contribution. Explanation, and in particular the modeling of explanation using analogy, has been tackled before. Thagard (2012) divides the computational models of explanation thus far into four types: probabilistic; those based on artificial neural networks; logical; and those based on schemas or analogy . The approach described in this paper falls in between the last two of these four types, since the template-matching system which we describe in the next section allows for both rule-based deductive reasoning and a form of analogical reasoning.

Hummel and Landy [11] propose that in explanation-generation, there are at least three types of flexibilities required by the representations and underlying processes: relational flexibility, the ability to see one concept as possibly playing multiple roles; semantic flexibility, the ability to exploit partial or imperfect matches between the objects and relations comprising an explanandum and the objects and relations encoded in potentially relevant domains in long-term memory; and an ability to map to, and transfer elements from, multiple domains in long-term memory simultaneously. However, the third type of flexibility can lead to a variant of the type-token problem (i.e. ambiguity about whether two elements have the same referent) against which Gentner's one-to-one constraint [6] is often used for defense. To fix this, they have their system decide whether two units correspond within the context of a certain source analog (which effectively implements a context-sensitive variant of the one-to-one constraint), and model the system using LISA [10–12].

Friedman and Forbus [4] and Friedman [5] propose a tiered framework in which explanations sit in a layer above that of justifications, which itself sits above a concept level. They demonstrate qualitative shifts in explanation-generation by exploiting metaknowledge that provides information about the structures in each tier. They do not, however, model explanation-generation for external preferences, but instead focus on the self-explanation effect. Tailoring explanations based on the beliefs of others may involve many types of reasoning, including modeling theory of mind [16], or having the ability to represent nested beliefs (e.g. "I know that the person I'm talking to believes that I believe X.").

Let us make two broad observations from the preceding summaries of literature. First, we see a form of metacognition in the work by Friedman and Forbus [4], in that metaknowledge about the structures in each tier is produced, manipulated, and reasoned over by the system. It is this sense of metacognition which we propose to utilize in this paper, in order to (among other things) qualitatively constrain the types of explanations which are generated by our model. The idea of qualitatively different explanations connects to our second observation, which is that the current body of work modeling explanation generation does not adequately address the cognitive processes which vary the qualitatively different types of explanations and selects the ones which are most appropriate.

Therefore, the work we propose in this paper distinguishes itself from the above approaches, on whose shoulders our work stands, in four key ways. First, our approach distinguishes between the full explanandum and its presentation. Second, we assume that this presentation affects a metacognitive system which in turn constrains the type of explanation that is generated. Third, we propose the use of specialized knowledge structures, such as *templates* and *constraint chunks* (both of which are described shortly), to allow such constraints to take the form of highly expressive knowledge structures.

Finally, we acknowledge both explicit and implicit effects of the explanandum on the explanation generation, and model both using the cognitive architecture CLARION, in such a way as to take advantage of the features it provides. We next summarize the aspects of CLARION we have used.

3 Explanation Generation in CLARION

CLARION is an integrative cognitive architecture with a several key features that we take advantage of here. These features include dual representation, a division of cognitive subsystems in a way that has previously been demonstrated to be psychologically plausible, and a flexible knowledge framework which can capture sub-conceptual, unstructured-conceptual, and structured-conceptual knowledge simultaneously [23, 25, 14]. CLARION consists of two levels: an explicit top level and an implicit bottom level. The top level typically contains knowledge structures and localist representations (which may or may not be linguistic concepts) and the bottom level often contains micro features and distributed representations. (Micro features, for our purposes here, can be defined informally as low-level constructs that correspond to properties which are not necessarily explicit, often because they are features that are not paid attention to by the agent. For example, a micro feature chunk may correspond to a certain brightness of a certain hue of the color red, or a very specific sound that can be heard precisely at three minutes in to a specific performance of Beethoven's 9th Symphony.) The top/bottom

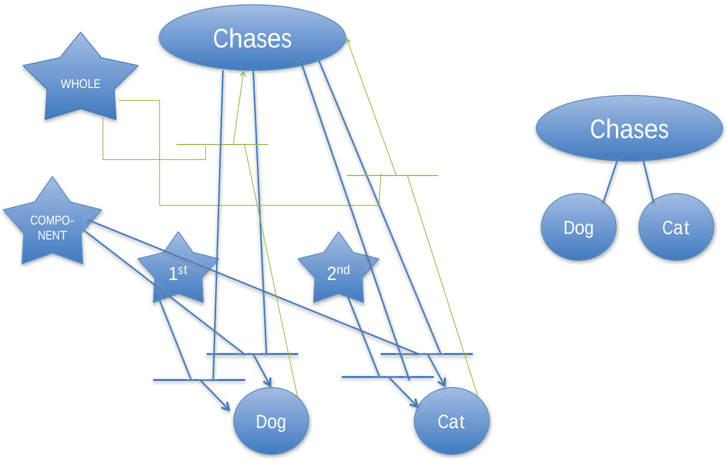


Fig. 1. A knowledge structure representing the proposition $CHASES(DOG,CAT)$. CDCs are pictured as star-shaped. On the right is the simplified version, which omits the CDCs and many of the ARs, though they are there (just not pictured).

level division is reflected in each of CLARION's subsystems: the Motivational, Action-Centered, Non-Action-Centered, and Meta-Cognitive Subsystems (MS, ACS, NACS, and MCS, respectively). A primary focus of CLARION has been psychological plausibility, and much work has been devoted to defining mechanisms within its subsystems that are tied to known psychological phenomena and processes [22, 26, 27].

The focus in the present paper is exclusively on an interaction between the NACS and MCS. In particular, recent work by Licato et al. has demonstrated how structured knowledge can be represented and reasoned over using no more than the psychologically plausible mechanisms already defined in the NACS [14, 15]; we use and expand on this method of representing structured knowledge to model explanation-generation and its metacognitive control below.

The NACS contains declarative knowledge, or general knowledge about the world that is not action-centered, which is often used for making inferences on the basis of its knowledge. The top level of the NACS contains localist chunks linked to units on the bottom level called DV pairs (Dimension-Value) pairs. The DV pairs can be linked to each other, and the chunks can also be linked to each other. However, the links between chunks are a special type of directed link called Associative Rules (ARs), which are represented pictorially using arrows. All of the links between top and bottom level units have weights that can be changed over time. This unique structure allows CLARION to define a *directed* similarity measure between two chunks [30, 21, 26]. This simple similarity measure can be used as part of larger algorithms used for analogical reasoning, deductive reasoning, and general behaviors defined over structured representations [14].

The MCS [24, 28] contains knowledge concerning the agent's cognitive processes and their outcomes, and also includes mechanisms that allow for active monitoring, regulation, and orchestration of the agent's cognitive processes (often toward some pragmatic goal that may be set by the MS). Like the other subsystems, the MCS is divided into a top and bottom level; however, not much work has been focused on fully exploiting both levels productively. In [24] and [28], the MCS was mostly used as the place where parameters which weighted processes in other subsystems were housed. In this paper, we propose to expand on the role of the MCS by having it hold structured knowledge analogous to that already defined in the NACS [14].

Structured knowledge in the NACS is achieved by first allowing top-level chunks to differentiate into types: object chunks, proposition chunks, template chunks, etc. These chunks are then linked using ARs and specialized chunks called Cognitively Distinguished Chunks (CDCs). For example, the proposition $Chases(Dog, Cat)$ can be represented as in Figure 1.

3.1 Templates

Analogical and deductive reasoning are carried out by defining special structures called Templates. These are essentially NACS structures that have been grouped under a single Template Chunk (TC) using a CDC defined for that purpose. In deductive reasoning, a template can specify the antecedent and consequent portions of a rule separately, so that when a structure sufficiently matches the antecedent portion, the consequent contains information on how to transfer the matched knowledge structure to create a new inference. Analogical reasoning can also be modeled by converting potential source analogs into templates and relaxing the match requirements. Matching structures to templates uses an Ant Colony Optimization algorithm inspired by [18], where the Template itself determines what types of matches are acceptable [15].

Explanation-construction proceeds as follows. We assume that we are given a knowledge base of templates. Each template is either a single structure, in which case it is to be used as a source analog for analogical matching and inference, or the template consists of an antecedent and consequent portion, in which case it is to be used as a deductive inference rule (e.g. Figure 2).

Given some knowledge structure K and template T, if a match is found from K to T (using the minimum conditions for an acceptable match specified by T itself), then a new structure K' is created from the elements of K and the instructions provided by T (these instructions are not explicitly stated by T in any way, rather they are implicit in the template's structure itself).[1]

Each template is grouped under a single Template Chunk (TC). The chunks in each template may each be linked to DV pairs in the NACS bottom level, and the template's TC is linked to a disjunction of all DV pairs linked to all non-CDC chunks in the template.

[1] For further detail, we direct interested readers to [14].

Algorithm 1 The Template Selection algorithm. This is used to filter out the template structures and select a subset of them based on how much they satisfy the constraints.

```
Require: Beliefs or knowledge the agent holds $B = \{B_i\}$
Require: A set of template chunks $T = \{T_i\}$
Require: A set of CCs $C = \{C_i\}$
  Define $\phi = 0.2$
  for all $T_i \in T$ do
    Set the activation level of $T$ to $\phi$
  end for
  for all $C_i \in C$ do
    if $C_i$ is an excitatory chunk then
      Set $C_i$'s activation level to $2 * \phi$ [
    else if $C_i$ is an inhibitory chunk then
      Set $C_i$'s activation level to $-2 * \phi$ ]
    end if
  end for
  Perform one iteration of Similarity-Based Reasoning to propagate activations
  return Active set $T_A$, consisting of the $n$ $T_i \in T$ with the highest activation levels (typical value
  for $n$ is between 5 and 10).
```

3.2 Constraint Chunks and the General Explanation-Construction Algorithm

We can now introduce a new type of chunk, which we will call a Constraint Chunk (CC). A CC is a chunk that resides on the top level of the MCS, and is used to either bias the parameters of cognitive processes based in the other (non-MCS) subsystems, or to point to the TC of a template which serves as a inviolable rule to constrain cognitive processes. The precise way in which it performs this biasing function is described shortly in the present section.

Just as the NACS chunks are linked to distributed units on the NACS bottom level, CCs are also linked to distributed units on the bottom level of the MCS. However, unless a similarity measure is defined between elements on the bottom levels of the NACS and MCS, no similarity measure will exist between chunks on their top levels. At least for this project, then, the design decision was made to allow the NACS and MCS to draw from a common pool of bottom-level distributed units, so that the same similarity measures used between two chunks of the NACS could be used from NACS to MCS chunks.[2]

Explanation generation is a simple backward-chaining process that starts with a set of knowledge structures $B = \{B_i\}$ corresponding to beliefs or knowledge that the agent holds, which are not part of the full explanandum, a set of templates $T = \{T_i\}$, a set of CCs $C = \{C_i\}$, and a full explanandum E.

The algorithm will start by selecting the relevant template structures. This requires that we have a set of CCs which are currently created manually in order to allow external users to set the qualitative features of the desired explanation, but the CCs are in such

[2] This design decision is partially justified by CLARION's view that meta-cognitive processes are intermeshed with other processes, and although the MCS is treated as a separate subsystem, it should really be viewed as closely integrated with the processes of the other subsystems [28].

a form that they can later be set autonomously by the motivational or action-centered subsystems. To carry out our demonstrations, we create two types of CCs: excitatory CCs, used to bias certain templates into being selected; and inhibitory CCs, which instead suppress and constrain the templates selected. Inhibitory and excitatory CCs can be single chunks, or they may also serve as TCs for templated structures in the NACS.

Next, the algorithm selects $T_i \in T$ subject to the constraints set by the CCs. It does this by activating all templates a fixed amount, and then activating excitatory CCs, allowing the activation to propagate using similarity-based reasoning [20, 26] (a single iteration was sufficient, though we could perform more later), and further activate certain templates. If any excitatory CCs serve as TCs for templated structures, then that structure is matched with the structures in T, and successful matches further activate those templates. Next, inhibitory CCs are activated, but rather than further activating similar templates, it lowers their activations.

As a result, we have a degree of activation for each $T_i \in T$ which reflects the constraints defined by the CCs. We collect the top n template chunks with the highest activations. This resulting set of templates is called the *active template set* (T_A). The pseudocode for the creation of T_A is shown in Algorithm 1.

The backward-chaining process can now begin. The algorithm starts by defining S as the set of facts in the full explanandum E. The templates are momentarily reversed: If some fact $s \in S$ matches the conclusion portion of a template in T_A, inference is performed on the antecedent portion of that template to produce a new set of facts, which replace s in S. If any of these newly added facts match beliefs in B, they are removed from S. This constitutes a single iteration of the backward-chaining process, which repeats until either S is empty, no more facts are found that can be added to S, or a preset time limit is reached. The remaining facts in S are then outputted as abductive assumptions.

We offer the pseudocode describing the general explanation-construction algorithm in Algorithm 2.

Algorithm 2 The General Explanation Generation algorithm.

```
Require: Beliefs or knowledge the agent holds $B = \{B_i\}$
Require: A set of active templates $T_A$ obtained through Algorithm 1.
Require: Set of facts $S = \{s_i\}$ in full explanandum $E$.
  Let $currAssumptions \leftarrow S$
  while $currAssumptions \nsubseteq B$ or timeout not yet reached do
    for all $t \in T_A$ do
      if Consequent of $t$ matches some $f \in currAssumptions$ and $f \notin B$ then
        Let $A$ = The facts comprising the antecedent of $t$
        $currAssumptions \leftarrow (currAssumptions - \{f\}) \bigcup A$
      end if
    end for
  end while
  return $currAssumptions$ as the abductive explanation of $E$.
```

4 Demonstrations

Our very brief proof-of-concept demonstrations will serve as examples for testing the model we describe in this paper. These examples attempt to construct explanations when given a small knowledge-base, using the analogical comparison and transfer mechanisms defined in the NACS and the constraints in the MCS.

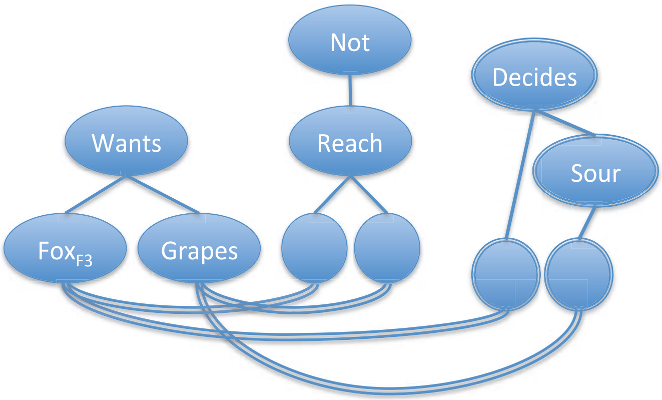


Fig. 2. A template representing the inference that a certain fox (the subscript $F3$ is meant to denote that it is a particular fox from a story with the label $F3$) wants grapes that he cannot reach, and therefore he decides that those grapes are sour. Following the notation defined in [14], the chunks with double lines are part of the consequent, and the horizontal double-lines connecting chunks are *identity links*. Assume that there is a template with chunks a, b connected by an identity link. Next, the template-matching algorithm may attempt to match two chunks a' and b' to a and b, respectively. But because of the identity link, a' and b' must have an extremely high similarity (using the measure defined in [26]).

4.1 Modeling $\mathbf{F_i}$ and $\mathbf{F_e}$ Effects

In order to clarify how we model the implicit and explicit effects of full explanandum presentation on explanation, we present a simple example demonstration that generates explanations for why a chicken crossed the road. The two full explananda, presented here in English for readability, are:

E_i The chicken decided to cross the road; the chicken was heading East.
E_p The chicken's body moved, crossing the road; the chicken was heading East.

Note that there is a very subtle difference in presentation: E_i invokes the concept of "deciding" whereas E_p does not. The algorithm will construct a new CC by simply creating a new chunk whose connected DV pairs are the disjunction of the DV pairs connected to the chunks in $P(E)$, the presentation of the full explanandum. This new CC bias the templates selected in the explanation-generation step, and thus will allow us to test $\mathbf{F_i}$ constraints. The templates provided to the system would include:

- If there is wind blowing east, and that wind is blowing on an object o, then o will move east.
- If c wants to achieve goal g, and g requires that action a happen, then c will decide to perform action a.
- If there is an object o that is East of c, and c likes o, then c will want to achieve the goal of moving East.
- If the chicken wants to achieve the goal of moving East, then the action of the chicken crossing the road must happen.

We now run the explanation-generation algorithm, and output the top explanation generated. When full explanandum E_i was used, the explanation generated the majority of the time (presented here again in English for readability) was:

> *Assume there is an object o that is East of the chicken. Assume the chicken likes o. The chicken will want to achieve the goal of moving East. The action of the chicken crossing the road must happen. The chicken will decide to cross the road.*

Whereas when E_p was used, the explanation was:

> *Assume that there is wind blowing east. Assume that wind is blowing on the chicken. The chicken will move east.*

Explicit effects are modeled by creating an inhibitory CC that is also the template chunk for a structure corresponding to the proposition $p =$ "The wind is blowing east." This will attempt to prevent any explanations that have p as one of its intermediate structures.

We ran the trial with E_p as the full explanandum, except this time the inhibitory CC corresponding to p is included. As expected, the explanations which require that the wind is blowing east are suppressed, and the explanation is generated as if E_i were provided instead.

5 Conclusion / Future Work

It is increasingly important that cognitive systems be able to explain and justify their conclusions and choices to the humans they will inevitably work with. For such systems, generating qualitatively different types of explanations may be essential. Using the work we have presented in this paper, such a thing can be accomplished with a few parameter changes in a meta-cognitive system. These parameters may be changed autonomously according to contextual factors, or by normal processes rooted in CLARION's subsystems, such as the ACS, MCS, or MS. We have presented a model that can explain produce explanations at different levels of abstraction, like E_1 and E_2 in this paper's introduction.

The work here is certainly not complete; a much wider variety of explanations must eventually be addressed. For example, the ability to justify behaviors using a proof defined in a fully formalized logic is (for some domains) a glaring absence to be tackled soon, but the work in this paper can be used as a springboard for moving in that direction.

An obvious next step is to flesh out the proof-of-concept demonstration briefly described in this paper, and to examine how it performs when provided with a much larger knowledge-base. Furthermore, more sophisticated deductive reasoning is necessary to augment the part of our explanation-generation algorithm that uses inhibitory CCs corresponding to full structures. In the demonstration we presented herein, $p =$"The wind is blowing east." was used to find and suppress templates that may have led to intermediate propositions equivalent to p. But if a template leads to a logically equivalent proposition such as "The wind is not not blowing east," our algorithm would have failed.

Finally, our current system does not demonstrate learning. If the templates drawn on by the explanation generator are insufficient, then presumably a human would eventually learn a new set of templates, somehow; this is not modeled in the present work. Clearly, there is much to do.

This work was funded by grant N000141310342 from the Office of Naval Research.

References

1. Chi, M.T., Bassok, M., Lewis, M.W., Reimann, P., Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science 13(2), 145–182 (1989)
2. Chi, M.T., De Leeuw, N., Chiu, M.H., Lavancher, C.: Eliciting self-explanations improves understanding. Cognitive Science 18(3), 439–477 (1994)
3. Dennett, D.: The Intentional Stance. The MIT Press (1989)
4. Friedman, S.E., Forbus, K.: An integrated systems approach to explanation-based conceptual change. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, GA (2010)
5. Friedman, S.E., Forbus, K.: Repairing incorrect knowledge with model formulation and metareasoning. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (2011)
6. Gentner, D.: Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science 7, 155–170 (1983)
7. Gentner, D., Rattermann, M.J., Forbus, K.: The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness. Cognitive Psychology 25, 524–575 (1993)
8. Hempel, C.: Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. Free Press, New York (1965)
9. Holyoak, K.J., Koh, K.: Surface and structural similarity in analogical transfer. Memory and Cognition 15(4), 332–340 (1987)
10. Hummel, J.E., Holyoak, K.J.: A Symbolic-Connectionist Theory of Relational Inference and Generalization. Psychological Review 110, 220–264 (2003)

11. Hummel, J.E., Landy, D.H.: From analogy to explanation: Relaxing the 1:1 mapping constraint...very carefully. In: Kokinov, B., Holyoak, K.J., Gentner, D. (eds.) New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy. Sofia, Bulgaria (2009)
12. Hummel, J.E., Licato, J., Bringsjord, S.: Analogy, explanation, and proof. Frontiers in Human Neuroscience (In Press)
13. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Girous (2011)
14. Licato, J., Sun, R., Bringsjord, S.: Structural Representation and Reasoning in a Hybrid Cognitive Architecture. In: Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN) (2014)
15. Licato, J., Sun, R., Bringsjord, S.: Using a Hybrid Cognitive Architecture to Model Children's Errors in an Analogy Task. In: Proceedings of CogSci 2014 (2014)
16. Pynadath, D.V., Rosenbloom, P., Marsella, S.C., Li, L.: Modeling two-player games in the sigma graphical cognitive architecture. In: Proceedings of the Sixth Conference on Artificial General Intelligence (AGI-13) (2013)
17. Ross, B.H.: Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. Journal of Experimental Psychology: Learning, Memory, and Cognition 15(3), 456–468 (1989)
18. Sammoud, O., Solnon, C., Ghédira, K.: An Ant Algorithm for the Graph Matching Problem. In: 5th European Conference on Evolutionary Computation in Combinatorial Optimization (EvoCOP 2005). Springer (2005)
19. Siegler, R.S.: How does change occur: A microgenetic study of number conservation. Cognitive Psychology 28(3), 225–273 (1995)
20. Sun, R.: Robust Reasoning: Integrating Rule-Based and Similarity-Based Reasoning. Artificial Intelligence 75(2) (1995)
21. Sun, R.: Schemas, logics, and neural assemblies. Applied Intelligence 5.2, 83–102 (1995)
22. Sun, R.: From Implicit Skills to Explicit Knowledge: A Bottom-Up Model of Skill Learning. Cognitive Science 25(2), 203–244 (2001)
23. Sun, R.: Duality of the Mind: A Bottom Up Approach Toward Cognition. Lawrence Erlbaum Associates, Mah- wah, NJ (2002)
24. Sun, R.: The motivational and metacognitive control in clarion. In: Gray, W. (ed.) Modeling Integrated Cognitive Systems. Oxford University Press, New York, New York, USA (2007)
25. Sun, R.: Autonomous generation of symbolic representations through subsymbolic activities. Philosophical Psychology (2012)
26. Sun, R., Zhang, X.: Accounting for Similarity-Based Reasoning within a Cognitive Architecture. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates (2004)
27. Sun, R., Zhang, X.: Accounting for a Variety of Reasoning Data Within a Cognitive Architecture. Journal of Experimental and Theoretical Artificial Intelligence 18(2) (2006)
28. Sun, R., Zhang, X., Mathews, R.: Modeling meta-cognition in a cognitive architecture. Cognitive Systems Research 7, 327–338 (2006)
29. Thagard, P., Litt, A.: Models of scientific explanation. In: Thagard, P. (ed.) The Cognitive Science of Science, chap. 3. The MIT Press (2012)
30. Tversky, A.: Features of Similarity. Psychological Review 84(4), 327–352 (1977)
31. Wellman, H.M., Lagattuta, K.H.: Theory of mind for learning and teaching: The nature and role of explanation. Cognitive Development 19(4), 479–497 (2004)

Revisiting Interacting Subsystems Accounts of Cognitive Architecture: The Emergence of Control and Complexity in an Algebra Task

Gareth Miles[1]

[1]University of South Wales

gareth.miles@southwales.ac.uk

Abstract. Symbolic accounts of cognitive architecture most often have a central hub where information is processed (e.g. the production process in ACT-R [1]). An alternative approach is to model cognition as the interaction of multiple largely autonomous subsystems [2, 3]. This latter, Interacting Subsystems, approach is explored in the GLAM-PS cognitive architecture, a theory that operationalizes many of the assumptions of strongly grounded approaches to cognition [4]. The GLAM-PS model of problem solving in algebra is described. Control in the algebra model is passed between three subroutines when solving a problem. These subroutines emerge from the interaction of different subsystems and are not explicitly programmed into the model. By systematically varying two short-term memory parameters it is shown that the model's successful performance of the task depends on the interaction of the contributing modules, and that this interaction demonstrates complexity, with additional memory resources not always improving performance.

Keywords. Cognitive Architecture, Production System, Embodied Cognition

1 The Interacting Subsystems approach to cognition

Cognitive Architecture has been established as a key research area within Cognitive Science following seminal work between 1970 and 1990 [5, 6, 7]. Although a lot of recent work has focused on either the ACT-R Architecture (e.g. [1]) or large-scale neural network models (e.g. [8]), there remain a wide variety of approaches to modeling Cognitive Architecture (e.g. [3], [9]). The purpose of the current paper is to look at cognitive control within a particular subset of these approaches.

In ACT-R and other notable architectures cognitive control is an aspect of cognition that is explicitly modeled, with specialist cognitive modules taking responsibility for the representation of goals and the selection of action. This reflects a consensus view of how the physical brain is specialized within different anatomical areas, notably the identification of the basal ganglia with action selection, and the frontal areas of the brain with the influence of intention [1], [8]. Thus in ACT-R the matching of IF-

THEN production rules is centralized in a module that is mapped on to the basal ganglia, with the representation of goals handled by a separate module mapped on to the anterior cingulate cortex (a frontal area) [1]. In the SPAUN Architecture intention is controlled in neural networks mapped on to the frontal cortex and action selection is mapped on to the basal ganglia [8].

However there are logical objections to this approach that become particularly apparent when one examines the relationship between neural networks and production systems. Both in essence are doing the same thing, the association of an output with a particular configuration of inputs. Whilst there are clearly differences between production systems and neural networks in how areas such as partial matching of configurations, generalization and one trial learning are handled, both can be considered methods of representing configural associations. The logical objections arise because anatomically there are networks of neurons present throughout the brain and it follows that these will be able to compute configural associations. Symbolic approaches to cognition clearly indicate that configural associations are the key underlying process in action selection and cognitive control. Therefore it would seem particularly strange that these key processes are modeled as strongly centralized in leading Cognitive Architectures when configural associations can be computed in many distinct parts of the brain.

An alternative approach is found in Architectures that use distributed interacting subsystems. Barnard’s Interacting Cognitive Subsystems (ICS) approach [2] to cognition and emotion theorized how such an approach could model complex tasks. In Barnard’s theory there are separate morphonolexical, propositional, object and implicational subsystems, each of which processes and translates symbolic output from the other subsystems. Whilst ICS has proved influential in highlighting the potential of interacting subsystems, the approach was not computationally implemented in full and did not compute configural associations (it’s subsystems simply translated one symbol into another). A more recent interacting subsystems approach is 4CAPs [3], an example of a Cognitive Architecture that was directly inspired by knowledge from neuroscience. The emphasis on 4CAPS is on modeling higher cognition, with amodal subsystems modeled including Left and Right Hemisphere Spatial and Executive centres.

The focus however within this paper is on GLAM-PS an interacting subsystems approach to embodied cognition. The idea of emergent control and action selection in a distributed system is particularly relevant to modeling embodied cognition because of the emphasis therein on modal rather than amodal cognitive systems. Modal subsystems are those directly associated with perception and action, in which the grounding of symbols (see [10]) in the external world is clearly indicated. Amodal subsystems are those that are not directly associated with perception or action (e.g. the goal module in ACT-R).

The plan for the paper is as follows, to briefly describe the GLAM-PS cognitive architecture, to demonstrate how cognitive control is modeled in a simple algebra problem solving task, and then finally to demonstrate the emergence of complexity in the algebra model by exploring the effects of small variations in the starting parameters of the GLAM-PS Architecture in the algebra task. Algebra was chosen because it is a paradigmatic task for studying Cognitive Architecture that has often been used by John Anderson to illustrate how ACT-R works (e.g. [1]). In the remainder of the paper ACT-R is used as for comparison purposes as an example of a mature, widely used symbolic Cognitive Architecture.

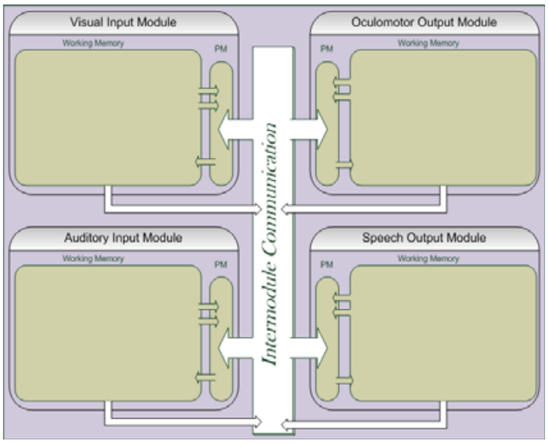


Fig. 1. A simplified view of the GLAM-PS architecture showing the four modules used in the Algebra Model and communication between these modules. PM is the Production Memory.

2 The GLAM-PS Cognitive Architecture

GLAM-PS shares a distributed modular structure with 4CAPS and ICS, however, whilst these Architectures make widespread use of amodal representation, GLAM-PS is intended to explore the implications of a strongly grounded distributed Architecture for cognition (see [4] for a review on Grounded Cognition). Whilst comparisons with ICS are difficult as it was never fully implemented, if we compare GLAM-PS to 4CAPS (arguably the most similar Architecture) it can be seen that GLAM-PS takes an outside-to-inside approach to modeling Cognitive Architecture, wherein peripheral processes dominate cognition. By contrast 4CAPS takes an inside-to-outside approach. The anatomical areas of the brain featured in 4CAPS do not map easily on to the modules described by GLAM-PS, instead the latter features modules that map on to the sensory and motor areas of the cortex. Grounded Cognition [4] suggests much of cognition is driven by these peripheral systems and a major novel contribution of GLAM-PS is to implement these ideas computationally in a symbolic architecture.

A simplified diagrammatic representation of the Architecture is shown in Fig. 1, with the two perception and two action modules used in the algebra task model included (no other modules are used for modeling this task). Both long-term and short-term/working memories are stored and revivified in the modules that originally processed what is being remembered. However, each module influences the behavior of other modules via the mechanism of inter-module communication of the current contents of working memory. In this manner the actions (productions) chosen in a module are based upon a composite view of working memory across all modules. Whilst this mostly acts like a single unified working memory there is a delay associated with inter-module communication. The implication of this is that a given module has an up-to-date view of its own working memory, but a delayed view of working memory in other modules (α is the GLAM-PS global parameter defining this delay in term of production cycles, it is set to 4 in the model reported here).

All long-term memories are stored as productions in GLAM-PS (following early SOAR [8]) using a classic IF-THEN structure. For simplicity and to improve plausibility all productions can only have a single action associated with the THEN side and the IF side is only able to check for the presence or absence of a representation (no programming code is allowed). When actions are represented in the action modules they are not necessarily executed and can be used to reason without action. Actions are only executed once they become 'Super Activated', a process whereby their activation level is raised substantially above the level needed for representation. Only once an Action Execution Threshold (global parameter β) is surpassed will the action be executed. Thus GLAM-PS is able to represent and then reason about actions without necessarily executing them.

Whilst the modules shown in Fig. 1 can be thought of as mapping on to sensory and motor areas of the brain, the processes associated with inter-module communication can be thought of as mapping on to the higher cortical areas (e.g. prefrontal cortex). This is a distinctly different interpretation of cortical function from many existing accounts. Whilst currently GLAM-PS makes no specific claims about how inter-module communication should be mapped on to the brain anatomically, it is nevertheless a potentially interesting future direction.

3 Cognitive Control in the GLAM-PS Algebra Model

The GLAM-PS Algebra Model (GAM) solves simple linear problems of the form $Ax + B = C$, for instance $2x + 4 = 10$ (where the solution is $x = 3$). To solve the problem GLAM-PS, like most human solvers [1], must proceed through three distinct stages or sub-goals, first reading and encoding the problem, then resolving the addend (the B term), before resolving the multiplier (the A term). The cognitive steps used by GLAM-PS are in essence the same as those used by Anderson's ACT-R model of the same task [4], what differs here however is how cognitive control is achieved.

Two types of cognitive control problems occur in GAM, firstly moving between sub-goals and secondly combining actions in such as was as to solve each sub-goal. The latter of these is relatively easy for GLAM-PS as it typically involves a sequence of actions where the result of the preceding action acts as the trigger for the next action in the sequence. In the failed runs reported in section 4 it is rarely the case (< 1%) that failure occurs because of a failure to sequence actions within a sub-goal, instead failures occur because the actions needed to begin a sequence that achieves a sub-goal are not initiated. Hence it is the first type of cognitive control, moving between sub-goals, that GLAM-PS finds difficult (for example beginning the process of resolving the addend once the problem has been encoded).

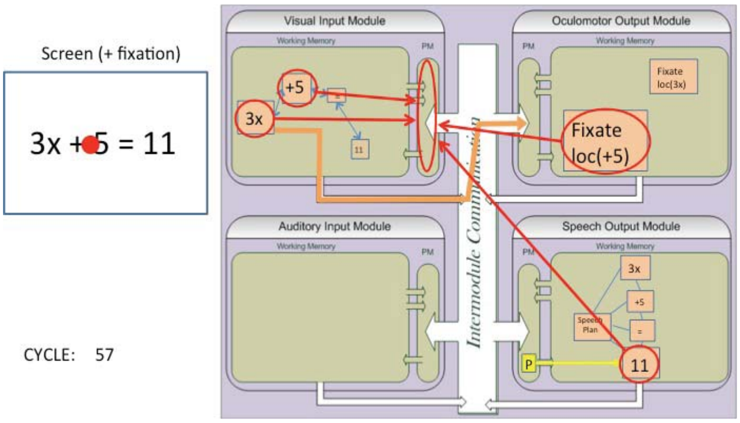


Fig. 2. Visualization of state of the GLAM-PS Algebra Model (GAM) when the control state has been established that begins transition from the Reading sub-goal to Solving the Addend sub-goal when solving $3x + 5 = 11$. Working memory elements (WMEs) are depicted as squares with area proportional to their activation. The WMEs contributing to the control state are circled in red, with their locus of action indicated by arrows pointing to the Visual Input production memory. GAM's current eye fixation is depicted on the left. Cycle indicates the number of production cycles from the beginning of simulation run.

Here we refer to the conditions that need to be satisfied to begin solving a sub-goal as the Control State. Within a distributed cognitive architecture the Control State needed to begin a new sub-goal will often be based on the state of multiple subsystems. If each of these subsystems is largely independent of one another then it can become difficult to achieve the required Control State. This is less of a problem in centralized architectures where a higher degree of control is possible and there is no need to coordinate representations across multiple subsystems. The control state needed to move between reading the algebra problem and solving it is shown in Fig 2.

In Fig 2. the state of GLAM-PS's working memory is visualised after the GAM has read the equation. As well as visual representations of the equation in the Visual Input module, GLAM-PS also has phonological representations of the equation in the Speech output module, the result of having read the equation. The lines between representations indicate structural links. The control state necessary to begin the solving of the equation by unwinding the addend consists of four representations across three different modules, these are the visual representation of the '$3x$' and the '$+5$', the oculomotor representation of the '$+5$' location (indicating attention is focused on the '$+5$') and the phonological representation of the '11' (indicating that the last element of the equation has been read and thus that the equation has been encoded). The production that matches this control state is a Visual Input production that acts by inhibiting the representation of the '$+5$'. Once this representation is inhibited a sequence of actions is initiated that relocates the '$+5$' to after the '11' in the equation (using imagery that is projected into the visual input module), GLAM-PS then changes the sign and then computes their combined value (eleven minus five).

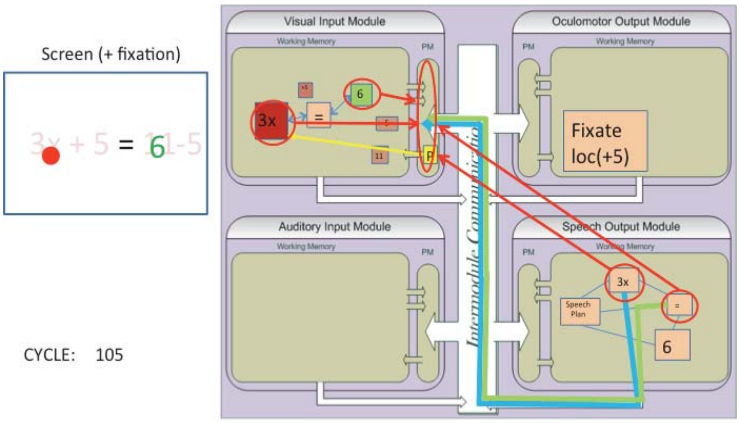


Fig. 3. Visualisation of the state of GAM when the control state has been established that begins the sub-goal of resolving the multiplier (the '3' in '$3x$'). See caption to Fig 2. for key.

The control state that is required to move between resolving the addend and the subsequent sub-goal of resolving the multiplier (the '3' in the example) is shown in Fig 3. Again, the control state is established through the combined presence of four working memory elements, this time across two modules. This consists of visual representations of the '$3x$' and a projected/imagined '6' (the result of the last sub-goal) and adjacent phonological representations of the '$3x$' and the '$=$', together these confirm that the addend has been resolved (the phonological representation is needed to confirm there are no other unresolved terms on the '$3x$' side of the equation). The sequence of actions needed to resolve the multiplier is then initiated by a production in the Visual Input module that inhibits the '$3x$' visual representation, allowing it to be subsequently broken into '3' and 'x' elements using imagery.

A key point is that in both of the transitions illustrated in Fig. 2 and Fig. 3 the control state consists of combinations of perceptual and motor representations, each of these representations is also used for perception or action (respectively), there are no abstract context or goal representations to force a particular cognitive subroutine to take control. This compares to ACT-R and other architectures where sub-goaling is used to ensure that only productions that solve the active sub-goal can be matched and executed, by contrast in GLAM-PS all productions are considered all of the time by the production matching process. Despite this GLAM-PS is able to demonstrate both task sufficiency and subroutine following in an Algebra task that can be considered a classic sub-goaling paradigm. This control is characterised as emergent because of the absence of any explicit control process within the modelling of the task.

In conclusion cognitive control in the GLAM-PS Algebra Model emerges from the interaction of working memory elements in multiple cognitive subsystems. When information from these different subsystems is combined there is sufficient information to indicate what actions the systems has taken previously and what still needs to be achieved. In Taatgen's work on the Minimal Control Principle [11] he indicates that often there will be sufficient information in a system to control action with only minimal need for explicit control representations. Whilst Taatgen clearly imagines that some form of goal representation will remain, in this GLAM-PS model there is no need for explicit goal representation. In short control is totally emergent [12]. Whether some form of goal representation would be needed once a more complex, multi-faceted model is considered is an open question. Certainly sometimes people want to simply read and equation, whilst at other times they need to solve them, though it could be the case that there are always enough clues in the external or internal environment to distinguish the two scenarios and establish an appropriate Control State.

4 The Emergence of Complexity in Interacting Cognitive Subsystems

Symbolic cognitive architectures often behave in a very predictable way, something that is often true of Production System Architectures. Once a set of productions has been 'programmed' into the system then these productions will provide a stable model of performance. This typically reflects the explicit use of goal representation that guides performance toward the achievement of that goal. Failure to achieve the goal would typically be modelled by the forgetting of the goal due to distraction [13]. Sometimes multiple strategies of achieving a set goal might be modelled and it is often the case that random 'noise' parameters will be used to help capture the variation in human performance that is observed from trial to trial in individual participants (e.g. [14]).

Much of the stability seen in established architectures is the result of centralised decision making. For example only one goal can be followed at a time in ACT-R [1] (though see [15]). When an architecture utilising multiple Interacting Subsystems is considered then complexity and instability may well emerge from the unpredictable interaction of the multiple distinct decision cycles in the component subsystems. If information from one module arrives at another module just one decision cycle later in one simulation run as compared to another, then the behaviour of the whole system might change very significantly over the full course of that run.

In order to explore the nature of the interaction of the multiple subsystems used in the GLAM-PS Algebra Model (GAM) a series of 1,170 simulation runs were conducted of the model with systematic variation of two working memory parameters. Note that the model used in the runs was deterministic without any randomised elements.

Working memory in GLAM-PS is module specific, with each module's working memory currently governed by the same global parameters and equations. Each working memory element has an activation varying from 0 to 1. To be matched by a production then a working memory element must have an activation greater than global parameter γ. Each working memory also has a total activation limit, global parameter δ. If the creation or change in activation of a working memory element takes the total activation within a module's working memory above δ, then the activation associated with all other elements in that module's working memory is adjusted so that total activation is equal to δ.

To explore the impact of small changes in working memory availability on GAM the parameter γ was systematically varied from .01 to .39 in increments of .01, this was combined the systematic variation of δ from 1.0 to 3.9 in increments of .1. On each simulation run the total number of cycles taken to solve the algebra equation $3x + 5 = 11$ was measured. The results of these simulation runs are displayed graphically in Fig 4.

The first aspect to consider of the results of these simulations runs is the vulnerability of the GAM model to failure. As δ dips below 3.0 and as γ increases it becomes increasingly more likely that GLAM-PS will not be able to solve the equation. An examination of failed runs clearly indicates that almost all (>99%) result from the failure to establish a control state that allows transition between one sub-goal and the next. According to the GAM model establishing that one sub-goal has been completed and then finding a suitable way to begin the next is difficult and prone to failure if working memory is compromised (e.g. by distraction). This broadly fits in with what has been observed in human participants, who typically take more time to complete steps of a problem that involve starting a new sub-goal [13].

The second aspect we see in the simulation runs is the emergence of complexity. One might reasonably expect that as each module's total working memory capacity, δ, increases then the likelihood of solving the equation would also increase. This is broadly the case, but there are many exceptions to this shown in Fig 4. Similarly as the production matching process becomes increasingly strict, matching fewer working memory elements (as γ increases), one would expect failures to become more likely, but again this is not always the case.

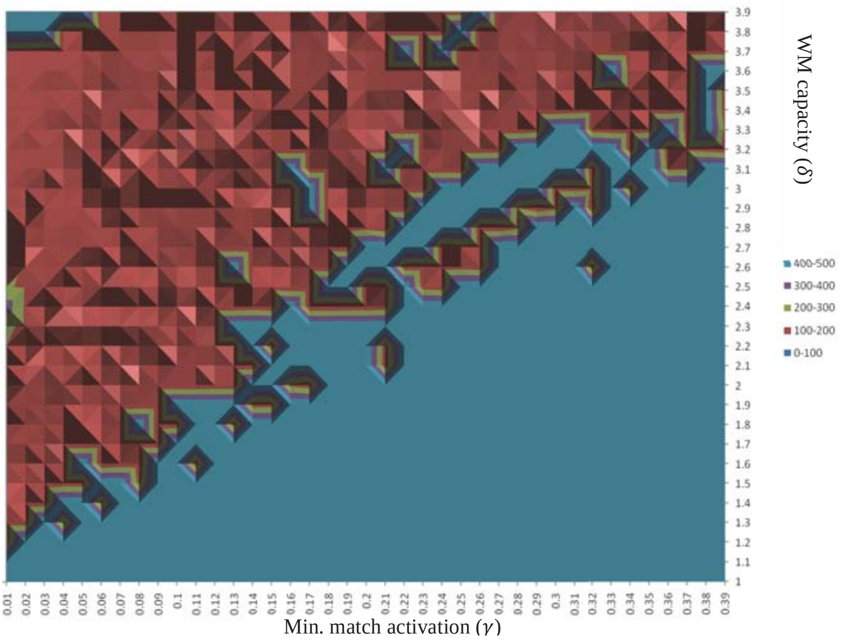


Fig. 4. A graphical display of the number of decision cycles needed to complete the equation $3x + 5 = 11$ by the GLAM-PS Algebra Model when working memory capacity (δ) and the activation needed to match productions (γ) were systematically varied. Light blue indicates the model did not solve the equation. The data is displayed in partial 3D and lit from the x-axis.

Indeed if one examines Fig 4, the parameters determining failure and success appear to influence these outcomes in a non-linear manner. If one considers the point where δ = 2.3 and γ = .13 then GAM fails, yet if we were to either increase or decrease either parameter by a fraction then GAM succeeds. Instead of a smooth curve or a straight line defining the regions where we see success versus where we see failure, what is shown in Fig 4. has more similarity to a geographical coastline. Even where there are successes the number of cycles taken to succeed varies unpredictably, the smoothness of the area in the top left (around δ = 3.5, γ = .03; though note the failures at δ > 3.7, γ < .04) can be contrasted with the peaks and troughs found in other areas where successes prevail (e.g. around δ = 3, γ = .1, the default parameter settings). The pattern observed in Fig 4. reflects the chaotic nature of the interaction of the multiple subsystems in the GLAM-PS Algebra Model. In short, complexity emerges from Interacting Subsystems.

5 Conclusion

The GLAM-PS Algebra Model demonstrates how both cognitive control and complexity emerges from the Interaction of Multiple Subsystems in Cognitive Architectures that adopt an Interacting Cognitive Subsystems approach [2]. The model is notable for not using any explicit goal representation, instead showing how control is based on Control States in working memory. Each of these Control States contain sufficient information about what the system has done previously and about what the system needs to do, to enable the initiation of purposeful, self-perpetuating sequences of behaviour. The simulation runs reported, exploring working memory parameter space, demonstrate how the model is vulnerable to failure when working memory is reduced or compromised, and how the interaction of cognitive subsystems is chaotic and somewhat unpredictable in nature.

References

1. Anderson, J. R.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press, New York (2007)
2. Barnard, P. J.: Interacting Cognitive Subsystems: Modelling Working Memory Phenomena Within a Multiprocessor Architecture. In A. Miyake & P. Shah (Eds.) Models of Working Memory: Mechanisms of Active Maintenance and Executive Control, pp. 298-339, Cambridge University Press, Cambridge, UK (1999)
3. Just, M. A., Varma, S.: The Organization of Thinking: What Functional Brain Imaging Reveals About the Neuroarchitecture of Complex Cognition. Cognitive, Affective, & Behavioural Neuroscience 7, 153-191 (2007)
4. Barsalou, L. W.: Grounded Cognition. Annual Review of Psychology 59, 617-645 (2008)
5. Newell, A., Simon, H. A.: Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ (1972)
6. Anderson, J. R.: The Architecture of Cognition. Harvard University Press, Cambridge, MA (1983)
7. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge, MA (1990)
8. Eliasmith, C.: How to Build a Brain: A Neural Architecture for Biological Cognition. Oxford University Press, New York (2013)

9. Sun, R.: Duality of the Mind: A Bottom-up Approach Toward Cognition.: Lawrence Erlbaum Associates, Mahwah, NJ (2002)
10. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335-346 (1990)
11. Taatgen, N. A.: The Minimal Control Principle. In W. Gray (Ed.), Integrated Models of Cognitive Systems. Oxford University Press, New York (2007)
12. Cooper, R. P.: Cognitive Control: Componential or Emergent? Topics in Cognitive Science 2, 593-613 (2010)
13. Altmann, E. M., Trafton, J. G.: Memory for Goals: An Activation-Based Model. Cognitive Science 26, 39-83 (2002)
14. Lovett, M. C., Daily, L. Z., Reder, L. M.: Modeling Individual Differences in Working Memory Performance: A Source Activation Account in ACT-R. Cognitive Science 25, 315-353 (2001)
15. Salvucci, D. D., Taatgen, N. A.: Threaded Cognition: An Integrated Theory of Concurrent Multitasking. Psychological Review 115, 101-130 (2008)

Biologically Plausible Modelling of Morality

Alessio Plebe

Department of Cognitive Science – University of Messina
v. Concezione 8, Messina, Italy
aplebe@unime.it

Abstract. Neural computation has an extraordinarily influential role in the study of several human capacities and behavior. It has been the dominant approach in the vision science of the last half century, and it is currently one of the fundamental methods of investigation for several higher cognitive functions. Yet, no neurocomputational models have been proposed for morality. Computational modeling in general has been scarcely pursued in morality, and existent non-neural attempts have failed to account for the mental processes involved during moral judgments. In this paper we argue that in the past decade the situation has evolved in a way that subverted the insufficient knowledge on the basic organization of moral cognition in brain circuits, making the project of modeling morality in neurocomputational terms feasible. We will sketch an original architecture that combines reinforcement learning and Hebbian learning, aimed at simulating forms of moral behavior in a simple artificial context.

Keywords: moral cognition; orbitofrontal cortex; amygdala

1 Introduction

Neural computation has an extraordinarily influential role in the study of several human capacities and behavior, however no neurocomputational models have been proposed yet for morality, a failure clearly due to the lack of empirical brain information.

On the other hand, there have been computational approaches oriented toward an understanding of morality different from neurocomputation, we will briefly review two main directions: formal logic and the so-called Universal Moral Grammar. It will be shown that both lines of research, despite their merits, will fail in giving an account of the mental processes involved during moral cognition.

In this paper we argue that in the past decade the situation has evolved in a way that makes the project of modeling morality in neurocomputational terms feasible. Even if there are no moral models yet, existing developments in simulating emotional responses and decision making are already offering important frameworks that we think can support the project of modeling morality. The existing models deemed closer to what pertains to morality will be shortly reviewed. We will also sketch an original architecture that combines reinforcement learning and Hebbian learning, aimed at simulating forms of moral behavior in a simple artificial context, and show its few preliminary results.

2 Other approaches to moral computing

Two computational accounts of morality, different from neurocomputation, will be briefly reviewed here.

The first, with the longest tradition, has been aimed at including morality within formal logic. Hare [18] assumed moral sentences to belong to the general class of prescriptive languages, for which meaning come in two components: the *phrastic* which captures the state to be the case, or command to be made the case, and the *neustic* part, that determines the way the sentence is nodded by the speaker. While Hare did not provided technical details of his idea for prescriptive languages, in the same years Wright [31] developed deontic logic, the logical study of normative concepts in language, with the introduction of the monadic operators $O(\cdot)$, $F(\cdot)$, and $P(\cdot)$ for expressing obligation, prohibition and permission. It is well known that all the many attempts in this directions engender a set of logical and semantic problems, the most severe is the Frege-Geach embedding problem [12]. Since the semantics of moral sentences is determined by a non-truth-apt component, like Hare's neustic, it is unclear how they can be embedded into more complex propositions, for example conditionals. This issue is related with the elimination of the mental processes within the logic formalism, and in fact viable solutions are provided by proponents of expressivism, the theory that moral judgments express attitudes of approval or disapproval, attitudes that pertains to the mental world.

One of the best available attempt in this direction has been given by Blackburn [3] with variants of the deontic operators, like $H!(\cdot)$ and $B!(\cdot)$, that merely express attitudes regards their argument: "Hooray!" or "Boo!". Every expressive operator has its descriptive equivalent, given formally by the $|\cdot|$ operation. An alternative has been proposed by Gibbard [13] in possible worlds semantics, defining an equivalent expressivist friendly concept, that of *factual-normative world* $\langle W, N\rangle$ where W is an ordinary Kripke-Stalnaker possible world, while N, the system of norms, is characterized by a family of predicates like N-forbidden, N-required. If a moral sentence S is N-permitted in $\langle W, N\rangle$ then it is said to hold in that factual-normative world. Both proponents acknowledge the need of moving toward a mental inquire, but their aim did never translated into an effective attempt to embed genuine mental processes in a logic system.

The second account here sketched, was apparently motivated by filling the gap left by formal logic, the lack of the mental processes in morality. The idea that there exists a Universal Moral Grammar, that rules human moral judgments in analogy with Chomsky's Universal Grammar, was proposed several decades ago [26], but remained disregarded until recently, when resuscitated by Mikhail [22], who fleshed it out in great detail.

His fragment of Universal Moral Grammar is entirely fit to the "trolley dilemma", the famous mental experiment invented by Foot [10], involving the so-called doctrine of the double effect, which differentiates between harm caused as means and harm caused as a side effect, like deviating a trolley to save people, but killing another one. Mikhail refined importantly the trolley dilemma, by inventing twelve subcases that catch subtle differences. subjects. The model he developed had the purpose of computing the same average responses given by subjects on the twelve trolley subcases. It is conceived in broad analogy with a grammatical parser, taking as input a structured description of the situation and a potential action, the moral grammar, and producing as output the decision if the potential action is permissible, forbidden, or obligatory. At the core of the grammar there is a "moral calculus", including rewriting rules from actions to moral effects.

The rules are carefully defined in compliance with American jurisprudence, therefore this grammatical approach looks like a potential alternative to logical models of jurisprudence, but it is claimed to simulate the mental processes of morality. Unfortunately nothing in his model is able to support such claim. The incoherence is that all the focus in the development of Mikhail is in the descriptive adequacy, the simplicity, and the formal elegance of the model, without any care on the mental plausibility. This is correct for an external epistemology, which was probably the original position of Rawls. But a model constructed on a strict external project, and in analogy with a well established mathematical framework (formal grammar) could well have principles quite at odds with anything that is subserved by a specific mental mechanism.

3 Toward moral neurocomputing

It is manifest that for the internal enterprise, the modeling of choice should be neural computation, the attempt to imitate the computational process of the brain, in certain tasks. Neurocomputational approaches to morality were unfeasible without a coverage of empirical brain information [16]. A main realization to emerge from all the work done so far is that there is no unique moral module. There is no known brain region activated solely during moral thinking, while a relatively consistent set of brain areas that become engaged during moral reasoning, is also active in different non moral tasks. In brief, the areas involved in morality are also related to emotions, and decision making in general [15, 23, 6].

Not every human decision is morally guided, nor does moral cognition necessarily produce decisions, however, investigations on the computational processes in the brain during decision taking, are precious for any neurocomputational moral model. Reinforcement learning [27] is the reference formalization of the problem of how to learn from intermittent positive and negative events in order to improve action selection through time and experience. It has been the basis of early models using neuronlike elements [1], and the concepts of reinforcement learning have been later fitted into the biology of neuromodulation and decision making [8, 5].

The model GAGE [32] assembles groups of artificial neurons corresponding to the ventromedial prefrontal cortex, the hippocampus, the amygdala, and the nucleus accumbens. It hinges on the somatic-marker idea [7], feelings that have become associated through experience with the predicted long-term outcomes of certain responses to a given situation. GAGE implementation of somatic-markers was based on Hebbian learning only, while reinforcement learning has been adopted in ANDREA [21], a model where the orbitofrontal cortex, the dorsolateral prefrontal cortex, and the anterior cingulate cortex interact with basal ganglia and the amygdala. This model was designed to reproduce a well known phenomenon in economics: the common hypersensitivity to losses over equivalent gains, analyzed in the prospect theory [19]. The overall architecture of these models have several similarities with those of [11], in which the orbitofrontal cortex interacts with the basal ganglia, but more oriented to dichotomic on/off decisions. A main drawback of all the models here mentioned is the lack of sensorial areas, that makes them unfit to be embedded even in the simplest form of environment in which a moral situation could be simulated.

4 The proposed model

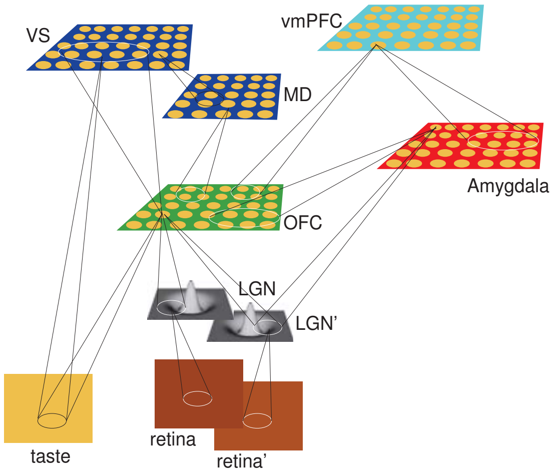


Fig. 1. Overall scheme of the model, composed by LGN (*Lateral Geniculate Nucleus*), V1 (*Primary Visual Area*), OFC (*OrbitoFrontal Cortex*), VS (*Ventral Striatum*), MD (*Medial Dorsal Nucleus*), Amyg (*Amygdala*), vmPFC (*ventromedial PreFrontal Cortex*).

The proposed model is able to simulate one specific moral situation, by including parts of the sensorial system, in connections to emotional and decision making areas. In the world seen by this artificial moral brain architecture there are three types of objects, two are neutral, and only one, resembling an apple, is edible, and its taste is pleasant. However, fruits in one quadrant of the scene are forbidden, like belonging to a member of the social group, and to collect these fruits would be a violation of her/his property, that would trigger an immediate reaction of sadness and anger. This reaction is perceived in the form of a face with a marked emotion. The overall scheme is shown in Fig. 1. It is composed by a series of sheets with artificial neural units, labeled with the acronym of the brain structure that is supposed to reproduce. It is implemented using the *Topographica* neural simulator [2], and each cortical sheet adheres to the LISSOM (*Laterally Interconnected Synergetically Self-Organizing Map*) concept [30].

There are two main circuits that learn the emotional component that contributes to the evaluation of potential actions. A first one comprises the orbitofrontal cortex, with its processing of sensorial information, reinforced with positive perspective values by the loop with the ventral striatum and the medial dorsal nucleus of the thalamus. The second one shares the representations of values from the orbitofrontal cortex, which are evaluated by the ventromedial prefrontal cortex against conflicting negative values, encoded by the closed loop with the amygdala. The subcortical sensorial components comprise LGN at the time when seeing the main scene, the LGN deferred in time, when a possibly angry face will appear, and the taste information.

4.1 Equations at the single neuron level

The basic equation of the LISSOM describes the activation level x_i of a neuron i at a certain time step k:

$$x_i^{(k)} = f\left(\gamma_{\mathrm{A}} \boldsymbol{a}_i \cdot \boldsymbol{v}_i + \gamma_{\mathrm{E}} \boldsymbol{e}_i \cdot \boldsymbol{x}_i^{(k-1)} - \gamma_{\mathrm{H}} \boldsymbol{h}_i \cdot \boldsymbol{x}_i^{(k-1)}\right) \tag{1}$$

The vector fields $\boldsymbol{v}_i$, $\boldsymbol{e}_i$, $\boldsymbol{x}_i$ are circular areas of radius r_{A} for afferents, r_{E} for excitatory connections, r_{H} for inhibitory connections. The vector $\boldsymbol{a}_i$ is the receptive field of the unit i. Vectors $\boldsymbol{e}_i$ and $\boldsymbol{h}_i$ are composed by all connection strengths of the excitatory or inhibitory neurons projecting to i. The scalars γ_{A}, γ_{E}, γ_{H}, are constants modulating the contribution of afferents, excitatory, inhibitory and backward projections. The function f is a piecewise linear approximation of the sigmoid function, k is the time step in the recursive procedure. The final activation of neurons in a sheet is achieved after a small number of time step iterations, typically 10.

All connection strengths adapt according to the general Hebbian principle, and include a normalization mechanism that counterbalances the overall increase of connections of the pure Hebbian rule. The equations are the following:

$$\mathit{\Delta}\mathbf{a}_{r_{\mathrm{A}},i} = \frac{\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}}{\|\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}\|} - \mathbf{a}_{r_{\mathrm{A}},i}, \tag{2}$$

$$\mathit{\Delta}\mathbf{e}_{r_{\mathrm{E}},i} = \frac{\mathbf{e}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}}{\|\mathbf{a}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}\|} - \mathbf{e}_{r_{\mathrm{E}},i}, \tag{3}$$

$$\mathit{\Delta}\mathbf{i}_{r_{\mathrm{I}},i} = \frac{\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}}{\|\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}\|} - \mathbf{i}_{r_{\mathrm{I}},i}, \tag{4}$$

where $\eta_{\{\mathrm{A,E,I}\}}$ are the learning rates for the afferent, excitatory, and inhibitory weights, and $\| \cdot \|$ is the L^1-norm.

4.2 Cortical components

The first circuit in the model learns the positive reward in eating fruits. The orbitofrontal cortex is the site of several high level functions, in this model information from the visual stream and taste have been used. There are neurons in the orbitofrontal cortex that respond differentially to visual objects depending on their taste reward [29], and others which respond to facial expressions [28], involved in social decision making [7].

OFC has forward and feedback connections with the Ventral Striatum, VS, which is the crucial center for various aspects of reward processes and motivation [17], and reprojects through MD, the medial dorsal nucleus of the thalamus, which, in turn, projects back to the prefrontal cortex. The global efficiency of the dopaminergic backprojections to OFC are modulated by a global parameter, used to simulate the hunger status of the model.

The second main circuit in the model is based on the ventromedial prefrontal cortex, vmPFC, and its connections from OFC and the amygdala. The ventromedial prefrontal cortex is long since known to play a crucial role in emotion regulation and social decision making [7]. More recently it has been proposed that the vmPFC may encode a kind of common currency enabling consistent value based choices between actions and goods of various types [14]. It is involved in the development of morality, in a study [9] older participants showed significant stronger coactivation between vmPFC and amygdala when attending to scenarios with intentional harm, compared to younger subjects. The amygdala is the primary mediator of negative emotions, and responsible for learning associations that signal a situation as fearful [20]. In the model it is used specifically for capturing the negative emotion when seeing the angry face, a function well documented in the amygdala [4].

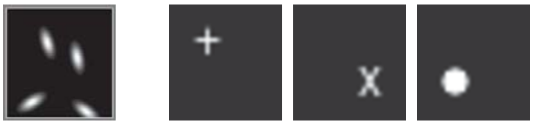


Fig. 2. Images seen by the model in the first phase of learning. On the left the patterns used for the development of the visual system. The other three images depict the objects that populate the simulated world: apples, +-shaped and ×-shaped.

4.3 First learning stages

The artificial brain is first exposed to a series of experiences, starting with a preliminary phase of development of the visual system with generic patterns, as those shown on the left in Fig. 2. These patterns mimic the retinal waves experienced before eye opening in humans, and allow the formation of retinotopy and orientation domains in the model V1 area, similarly to the process described in [25]. When the visual system is mature, the model is presented with samples from the collection of three simple objects, in random positions, as shown in Fig. 2. At the same time their taste is perceived too, and only one of the objects, the apple, has a good taste. The connection loop between OFC, and the dopaminergic areas VS, MD, attain an implicit reinforcement learning, where the reward is not imposed externally, but acquired by the OFC map, through its taste sensorial input. The amygdala has no interaction during this stage. The model will gradually become familiar with the objects, and learn how pleasant apples are, in its OFC model area. In order to characterize the ensemble activation pattern of the OFC neurons, and decode the objects categorization, a population code method is applied. The overall population is clustered according to those neurons, which were active in response to different classes of objects, compared to those which were not responsive, mathematical details are in [24].

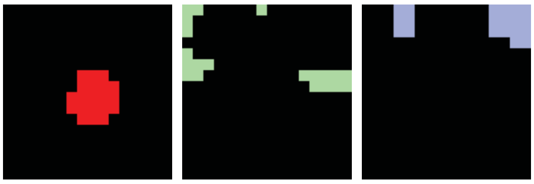


Fig. 3. Neural coding of the three objects in the model OFC area: apple on the left, the ×-shaped in the center, and the ×-shaped on the right.

In Fig. 3 is shown the resulting coding of the three categories of objects in the OFC model area, with neurons that are selectively activated by objects of one class, independently on their position in space.

In a second stage the model receives additional experiences, that of the moral learning, with the same objects as stimuli. The model can choose between two possible behaviors: collect and eat an object, or refrain from doing it, a selection coded in the vmPFC component. Now, if the model decides to pick apples in a certain area of the world, that shown in the central image in Fig. 4, suddenly an angry face will appear, like those shown in the right of Fig. 4. Fruits in this portion of the space may belong to a member of the social group, and to collect these fruits would be a violation of her/his property, that would trigger an immediate reaction of sadness and anger.

Now the amygdala gets inputs from both the OFC map and directly from the thalamus, when the angry face appears. There is an implicit reinforcement, with the negative reward embedded in the input projections to the amygdala.

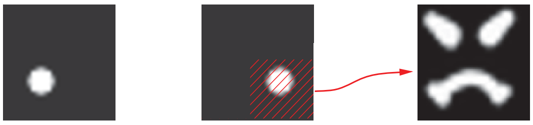


Fig. 4. Images seen by the model in the second phase of learning. On the left an apple in a part of the world where it is allowed to pick it. The center image is an apple in the forbidden area, if the model attempts to pick it, the angry face shown on the right will suddenly appear.

4.4 Surviving without stealing

Finally, the developed artificial agent is embedded in its simple world, where all possible objects may randomly appear, and she can choose to grasp them or not. There is a parameter in the model which is used to modulate its state of hunger, in the dopaminergic circuit, which detailed equations are the following:

$$
\begin{aligned}
x^{(\mathrm{OFC})} = f\Bigg(& \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})} + \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\circledcirc)} + \\
& \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\boxdot)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\boxdot)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\boxdot)} + \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} \boldsymbol{b}_{r_{\mathrm{B}}}^{(\mathrm{OFC})} \cdot \boldsymbol{v}_{r_{\mathrm{B}}}^{(\mathrm{MD})} + \\
& \gamma_{\mathrm{E}}^{(\mathrm{OFC})} \boldsymbol{e}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} \cdot \boldsymbol{x}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} - \gamma_{\mathrm{H}}^{(\mathrm{OFC})} \boldsymbol{h}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \cdot \boldsymbol{x}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \Bigg)
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
x^{(\mathrm{VS})} = f\Bigg(& \gamma_{\mathrm{A}}^{(\mathrm{VS}\leftarrow\mathrm{OFC})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{VS}\leftarrow\mathrm{OFC})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{OFC})} + \gamma_{\mathrm{A}}^{(\mathrm{VS}\leftarrow\boxdot)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{VS}\leftarrow\boxdot)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\boxdot)} + \\
& \gamma_{\mathrm{E}}^{(\mathrm{VS})} \boldsymbol{e}_{r_{\mathrm{E}}}^{(\mathrm{VS})} \cdot \boldsymbol{x}_{r_{\mathrm{E}}}^{(\mathrm{VS})} - \gamma_{\mathrm{H}}^{(\mathrm{VS})} \boldsymbol{h}_{r_{\mathrm{H}}}^{(\mathrm{VS})} \cdot \boldsymbol{x}_{r_{\mathrm{H}}}^{(\mathrm{VS})} \Bigg)
\end{aligned} \tag{6}
$$

$$
x^{(\mathrm{MD})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{MD}\leftarrow\mathrm{VS})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{MD}\leftarrow\mathrm{VS})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{VS})} \right) \tag{7}
$$

These two equations are just specialization of the general equation (1), for areas VS and MD. The afferent signals $\boldsymbol{v}^{(\mathrm{OFC})}$ come from the OFC model area, $\boldsymbol{v}^{(\boxdot)}$ is the taste signal, and [$\circledcirc$ the output of the LGN deferred in time, when a possibly angry face will appear. The output $x^{(\mathrm{MD})}$ computed in (7) will close the loop into the prefrontal cortex. The parameter $\gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})}$ is a global modulatory factor of the amount of dopamine signaling for gustatory reward, and therefore it is the most suitable parameter for simulating hunger states.

A simulation is performed by letting the model meeting with random objects, at random positions in the world. Now there will be no more angry face in case the model steal an apple in the forbidden place, whoever, it is expected that the moral norm to avoid stealing will work, at least up to a certain level of hunger. There is no more learning in any area of the model. At every simulation step the modulation parameter is updated as following:

$$
\gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} \leftarrow \begin{cases} \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} - \chi & \text{when an apple is grasped} \\ \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} + \phi & \text{otherwise} \end{cases} \tag{8}
$$

Where χ is the amount of nutriment provided by an apple, and ϕ is the decrease of metabolic energy in time.

In Fig. 5 the decisions to grasp are shown, as a function of the hunger level, after 50000 simulation steps. Neutral objects are grasped occasionally, about one over three, almost independently from hunger. Allowed apples are grasped more frequently with hunger, every time with level over 0.1, while it can be seen the strong inhibition to grasp apples in the forbidden sector, with few attempts at extreme hunger level only, over 0.3.

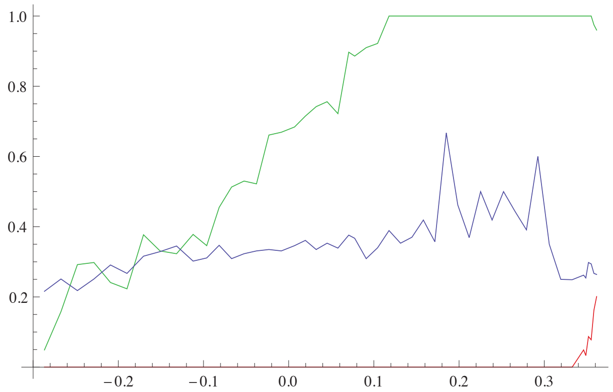


Fig. 5. Percentage of grasping decisions as a function of hunger level. Green: allowed apples, blue: neutral objects, red: forbidden apples.

In conclusion, we believe that the neurocomputational approach is an additional important path in pursuing a better understanding of morals, and this model, despite the limitation in its cortical architecture, and the crudely simplified external world, is a valid starting point. It picks up on one core aspect of morality: the emergence of a norm, not to steal, induced by a moral emotion. Obeying this norm is an imperative that supersedes other internal drives, like hunger, up to a certain extent. It has to be warned again, that morality is a collection of several, partially dissociated mechanisms, and the presented model is able to simulate only one kind of moral situation, the temptation of stealing food, and the potential consequent feelings of guilt. Further work will address other type of morality, that will need different scenarios to be simulated.

References

1. Barto, A., Sutton, R., Anderson, C.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man and Cybernetics 13, 834–846 (1983)

2. Bednar, J.A.: Topographica: Building and analyzing map-level simulations from Python, C/C++, MATLAB, NEST, or NEURON components. Frontiers in Neuroinformatics 3, 8 (2009)
3. Blackburn, S.: Spreading the Word. Oxford University Press, Oxford (UK) (1988)
4. Boll, S., Gamer, M., Kalisch, R., Büchel, C.: Processing of facial expressions and their significance for the observer in subregions of the human amygdala. NeuroImage 56, 299–306 (2011)
5. Bullock, D., Tan, C.O., John, Y.J.: Computational perspectives on forebrain microcircuits implicated in reinforcement learning, action selection, and cognitive control. Neural Networks 22, 757–765 (2009)
6. Casebeer, W.D., Churchland, P.S.: The neural mechanisms of moral cognition: A multiple–aspect approach to moral judgment and decision–making. Biology and Philosophy 18, 169–194 (2003)
7. Damasio, A.: Descartes' error: Emotion, reason and the human brain. Avon Books, New York (1994)
8. Dayan, P.: Connections between computational and neurobiological perspectives on decision making. Cognitive, Affective, & Behavioral Neuroscience 8, 429–453 (2008)
9. Decety, J., Michalska, K.J., Kinzler, K.D.: The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. Cerebral Cortex 22, 209–220 (2012)
10. Foot, P.: The problem of abortion and the doctrine of the double effect. Oxford Review 5, 5–15 (1967)
11. Frank, M.J., Scheres, A., Sherman, S.J.: Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. Philosophical transactions of the Royal Society B 362, 1641–1654 (2007)
12. Geach, P.T.: Assertion. The Philosophical Review 74, 449–465 (1965)
13. Gibbard, A.: Wise Choices, Apt Feelings – a theory of normative judgment. Harvard University Press, Cambridge (MA) (1990)
14. Gläscher, J., Hampton, A.N., O'Doherty, J.P.: Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cerebral Cortex 19, 483–495 (2009)
15. Greene, J.D., Haidt, J.: How (and where) does moral judgment work? Trends in Cognitive Sciences 6, 517–523 (2002)
16. Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D.: fMRI investigation of emotional engagement in moral judgment. Science 293, 2105–2108 (2001)
17. Haber, S.N.: Neural circuits of reward and decision making: Integrative networks across corticobasal ganglia loops. In: Mars, R.B., Sallet, J., Rushworth, M.F.S., Yeung, N. (eds.) Neural Basis of Motivational and Cognitive Control, pp. 22–35. MIT Press, Cambridge (MA) (2011)
18. Hare, R.M.: The Language of Morals. Oxford University Press, Oxford (UK) (1952)
19. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decisions under risk. Econometrica 47, 313–327 (1979)
20. LeDoux, J.E.: Emotion circuits in the brain. Annual Review of Neuroscience 23, 155–184 (2000)
21. Litt, A., Eliasmith, C., Thagard, P.: Neural affective decision theory: Choices, brains, and emotions. Cognitive Systems Research 9, 252–273 (2008)
22. Mikhail, J.: Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. In: Bartels, D., Bauman, C., Skitka, L., , Medin, D. (eds.) Moral Judgment and Decision Making. Academic Press, New York (2009)

23. Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J.: The neural basis of human moral cognition. Nature Reviews Neuroscience 6, 799–809 (2005)
24. Plebe, A.: A neural model of moral decisions. In: Madani, K., Filipe, J. (eds.) Proceedings of NCTA 2014 - International Conference on Neural Computation Theory and Applications. Scitepress (2014)
25. Plebe, A., Domenella, R.G.: Object recognition by artificial cortical maps. Neural Networks 20, 763–780 (2007)
26. Rawls, J.: A Theory of Justice. Belknap Press of Harvard University Press, Cambridge (MA) (1971)
27. Rescorla, R.A., Wagner, A.R.: A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (eds.) Classical Conditioning II: Current theory and research, pp. 64–99. Appleton Century Crofts, New York (1972)
28. Rolls, E., Critchley, H., Browning, A.S., Inoue, K.: Face-selective and auditory neurons in the primate orbitofrontal cortex. Experimental Brain Research 170, 74–87 (2006)
29. Rolls, E., Critchley, H., Mason, R., Wakeman, E.A.: Orbitofrontal cortex neurons: Role in olfactory and visual association learning. Journal of Neurophysiology 75, 1970–1981 (1996)
30. Sirosh, J., Miikkulainen, R.: Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. Neural Computation 9, 577–594 (1997)
31. Von Wright, G.H.: Deontic logic. Mind 60, 1–15 (1951)
32. Wagar, B.M., Thagard, P.: Spiking Phineas Gage: A neurocomputational theory of cognitiveaffective integration in decision making. Psychological Review 111, 67–79 (2004)

How Artificial is Intelligence in AI?
Arguments for a Non-Discriminatory Turing test

Jack Birner

University of Trento, University College Maastricht
jack.birner@unitn.it

Abstract. Friedrich von Hayek's *The Sensory Order* (1952) presents a physicalist identity theory of the human mind. In a reaction to Karl Popper's criticism that such a "causal" theory of the mind cannot explain the descriptive and critical-argumentative functions of language, Hayek wrote a paper that was never published. It contains the description of a thought experiment of two communicating automata. This paper confirms the impression of the AI-like character of the structuralism and functionalism of Hayek's *Sensory Order*. In some important respects, what Hayek tries to do in his paper is similar to Turing's discussion of the question "can machines think?" Arguments will be given why according to a functionalist and physicalist identity theory of mind the distinction between artificial and "natural" intelligence cannot be upheld. According to such a theory, Turing tests are unnecessarily restrictive and discriminatory vis-à-vis machines. In the end, the question whether or not machines can think is not meaningless, as Turing thought. It can be replaced by the question if artificial minds are capable of consciousness. The Turing test , however, cannot give the answer.

Key words. theory of mind • physicalist identity theory • virtual machines • communication of symbolic description

1 Introduction

This paper is the consequence of the interest in the philosophy of Karl Popper that I share with Aaron Sloman. A couple of months ago he reacted to the announcement of a conference on Popper that bears my signature and this led to both of us reading some of the other's publications. We discovered that we had more interests in common. This happy chance meeting of minds led to my writing what you are now reading.[1] Popper is also one of the *dramatis personae* of this story, next to Friedrich von Hayek. Popper and Hayek became close intellectual and personal friends during and after the Second World War. In their published work they appear to agree on almost everything. Some aspects of their thought, however, convinced me that this could not be really true. And indeed, a closer look revealed that till the end of their lives they remained divided on several important issues. I have dealt with some of these, and with the influence – both positive and negative - they had on one another elsewhere.[2]

[1] Without Aaron's encouragement I would not have dreamt of sending a text to a workshop on AI. Let me hasten to add that his guilt stops here: I take full responsibility for everything that follows. I would also like to apologize in advance for not referring to authors who may have discussed the same or similar problems; these are my first steps in AI.

[2] Cp. Birner (2009) and (forthcoming).

2 Philosophy of mind

What I will take up here are their disagreements in the philosophy of mind. I do so first of all because Hayek's theory of mind and his defence against Popper's criticism have a strong AI flavour.[3] Second, there are some striking similarities between Hayek's work of the early 1950s and "Computing Machinery and Intelligence" (*CMI*) of 1950 by Alan Turing, the third main character of this tale. These parallels deserve more attention than has been given them.[4] In 1952 Hayek published *The sensory order: an inquiry into the foundations of theoretical psychology* (*SO*). The foundations mentioned in the title are a philosophy of mind that I will now summarize. Hayek tries to explain the human mind using only the laws of physics. He had adopted this explanatory programme from Moritz Schlick's *Allgemeine Erkenntnistheorie.* The ontological idea underlying it is that the mind does not have a separate existence from the brain. So Hayek's is a physicalist identity theory.

As the vehicle for his explanation he uses a neural-network model.[5] According to Hayek, mental processes consist in the continuous reorganization on many levels of a hierarchical system of relationships. That is why he speaks of an order of events. A neural network is one possible model of the mind. Hayek is a radical functionalist in the sense that he states that *any* physical configuration of elements and their relationships might embody mental processes. He introduces this idea thus:

> "That an order of events is something different from the properties of the individual events, and that the same order of events can be formed from elements of a very different individual character, can be illustrated from a great number of different fields. The same pattern of movements may be performed by a swarm of fireflies, a flock of birds, a number of toy balloons or perhaps a flight of aeroplanes; the same machine, a bicycle or a cotton gin, a lathe, a telephone exchange or an adding machine, can be constructed from a large variety of materials and yet remains the same kind of machine within which elements of different individual properties will perform the same functions. So long as the elements, whatever other properties they may possess, are capable of acting upon each other in the manner determining the structure of the machine, their other properties are irrelevant for our understanding of the machine." (*SO* 2.28)

Then he proposes a radically functionalist and structuralist hypothesis:

> "In the same sense the peculiar properties of the elementary neural events which are the terms of the mental order have nothing to do with that order

[3] Already hinted at in an afterthought to Birner 2009, where I wrote that Hayek missed the chance to be recognized as a pioneer in AI. This will be discussed below.

[4] But cp.Van den Hauwe (2011).

[5] As does Donald Hebb, the publication of whose *The Organization of Behavior* in 1949 almost kept Hayek from publishing his book. *SO* elaborates a manuscript that dates from 1920. For a discussion, cp. Birner (2014).

> itself. What we have called physical properties of those events are those properties which will appear if they are placed in a variety of experimental relations to different other kinds of events. The mental properties are those which they possess only as a part of the particular structure and which may be largely independent of the former. *It is at least conceivable that the particular kind of order which we call mind might be built up from any one of several kind of different elements – electrical, chemical, or what not; all that is required is that by the simple relationship of being able to evoke each other in a certain order they correspond to the structure we call mind.*" (*SO* 2.29, my italics)[6]

This sounds very AI-like. The link between Hayek's theory of mind and AI is even more apparent in the way Hayek developed his ideas after the publication of *SO*. That is the subject of the next section.

3 Popper's criticism

Upon publication of *SO* Hayek sent a copy to Popper. Although Popper was – as always - very polite in his reaction, he did not like it. Though Popper never writes this down, his main general objection to Hayek's theory of mind is that it is too inductivist. What he does write in a letter to Hayek (2 December 1952) is that he thinks his theory of the sensory order is deterministic. This implies, says Popper, that it is a sketch for a deterministic theory of the mind. Now Popper had just written a criticism (later published as Popper 1953) of this type of theory.[7] He argues that a deterministic theory of the mind cannot be true because it is impossible to have a deterministic theory of human language.

In his criticism, Popper uses a particular analysis of language. He considers it to be part of his solution to what he calls Compton's problem. Popper uses that name for what he considers to be a generalization of Descartes' formulation of the mind-body problem. Descartes asks how the immaterial mind can act upon the physical body. Popper wants to know how abstract entities such as the contents of ideas and theories can influence the physical world. He builds upon Karl Bühler's theory of the evolution of language. It says that the first function of language to emerge in human evolution is the *expression* of subjective states of consciousness. The next function to develop is *communication* (or *signaling*), followed by *description*. Popper adds a fourth function, *argumentation and criticism*. It presupposes the previous (or, as

[6] For a contemporary elaboration of this idea that seems to be very fruitful for understanding and measuring consciousness, cf. Tononi 2012.

[7] Apparently as a criticism of *SO*, of which he may have read the proofs. Cp. what Popper writes to Hayek (letter of 30 November 1953 – Klagenfurt Popper archives, folder 541.12, on file from microfilm of the Hoover Archives): "I was extremely pleased to hear that with "the challenge of my article on Language and the Body Mind Problem", I have done "a great service". I am really happy about this article. I have ??? ???? M... (?) on the problem, but although I think that I got somewhere, I don't know whether it is worth much. If you really can refute my views (?), it would, I think, be an achievement." (hand writing partially illegible).

Popper says, lower) functions. Not only has the need of humans to adapt to the environment given rise to new physical instruments, it has also produced their capacity to theorize. That is a consequence of the evolution of the higher functions of language: they serve to control the lower ones (Popper 1972: 240-41). Abstract contents of thought, meanings and the higher functions of language[8] have co-evolved. They help us control our environment "plastically" because they are adaptable. Popper proposes a dualistic and indeterministic theory of the mind and of the influence of the contents of consciousness on the world, which according to him can account for the higher linguistic functions – unlike physicalist and behaviourist theories:

> "When the radical physicalist and the radical behaviourist turn to the analysis of human language, they cannot get beyond the first two functions (see my [1953]). The physicalist will try to give a physical explanation - a causal explanation - of language phenomena. This is equivalent to interpreting language as expressive of the state of the speaker, and therefore as having the expressive function alone. The behaviourist, on the other hand, will concern himself also with the social aspect of language - but this will be taken, essentially, as the way in which speakers respond to one another's "verbal behavior." This amounts to seeing language as expression and communication.
>
> But the consequences of this are disastrous. For if language is seen as merely expression and communication, then one neglects all that is characteristic of human language in contradistinction to animal language: its ability to make true and false statements, and to produce valid and invalid arguments. This, in its turn, has the consequence that the physicalist is prevented from accounting from the difference between propaganda, verbal intimidation and rational arguments." (Popper and Eccles 1977: 58)[9]

Hayek took this criticism of Popper's very seriously.[10] He responded to it in "Within Systems and about Systems; A Statement of Some Problems of a Theory of Communication." That paper was never published. It was never finished, either. Later Hayek writes about it:

> [I]n the first few years after I had finished the text of the book [*SO*], I made an effort to complete its formulations of the theory in one respect. I had then endeavoured to elaborate the crucial concept of "systems within systems" but found it so excruciatingly difficult that in the end, I abandoned the

[8] All of these are inhabitants of what Popper in his later philosophy has called world-3.

[9] Popper & Eccles (1977) makes the same points that are made in the 1953 paper more forcefully.

[10] "With the challenge of your article on "Language and the Body Mind Problem" you have unwittingly done me a great service. Much has crystallized in my mind as a result of my inability fully to accept (?) the argument. I believe I can now (?) provide (?) a causal theory of description and intention, but of course only an "explanation of the principle" applicable to greatly simplified models and not sufficient to provide a full explanation either of human language or human intention. But sufficient to construct models possessing all the characteristics common to all instances of "description" and intention. I am still struggling with great (?) difficulties, but I believe I am getting somewhere." (Hayek to Popper, 30 October 1953, Popper Library, Klagenfurt, folder 541.12, on file, from microfilm Hoover archives, hand writing partially illegible).

> longish but unfinished paper that apparently nobody I tried it upon could understand". (Hayek 1982: 290)

In the paper Hayek follows a two-pronged defence strategy against Popper's criticism, one "negative," the other constructive or "positive". As to the former, Hayek states the purpose of the paper as

> "deriving from the study of certain kinds of causal systems conclusions concerning the character of our possible knowledge of mental processes. (…) [T]he main conclusion to which [the argument] will lead is that for any causal system there is a limit to the complexity of other systems for which the former can provide an analogon of a description or explanation, and that this limit necessarily excludes the possibility of a system ever describing or explaining itself. This means that, if the human mind were a causal system, we would necessarily experience in discussing it precisely those obstacles and difficulties which we do encounter and which are often regarded as proof that the human mind is not a causal system." (*Systems*: 1).

Put bluntly, this "negative" part of Hayek's reaction to Popper's criticism is of the heads-I-win-tails-you-lose type. The gut reaction of Popperian philosophers to such an argument would be to condemn it out of hand as an immunizing stratagem. Interestingly enough, Popper does not do so. I will briefly come back to this below. The average non-Popperian citizen of Academe might instead dismiss it as corny. That, however, would fail to do justice to Hayek. He gives two arguments for his conclusion. First, as he states in the next sentence, "[w]e shall find that to such a system the world must necessarily appear not as one but as two distinct realms which cannot be fully "reduced" to each other." (*ibid.)* The second argument invokes complexity. In a generalized form it says that an *explanans*, in order to be successful, has to be more complex than its *explanandum*. The argument is taken over from *SO*: "any apparatus of classification must possess a higher degree of complexity than is possessed by the objects which it classifies… therefore, … the human brain can never fully explain its own operations." (*SO*: 8.68).[11] This may be true or false but it certainly deserves closer examination. If it is true, then Hayek has demonstrated by a *reductio ad absurdum* that the mind cannot explain[12] itself (for it would have to be more complex than it is).

The complexity Hayek refers to, and which he does not explain in more detail, may consist of at least two circumstances. One has to do with problems of self-reference, the other with the impossibility of describing all the relevant initial conditions for explaining the human mind. Hayek does not mention or elaborate these aspects (which would deserve closer scrutiny). What he does instead is to work out, in subsequent publications, the methodological idea of in-principle explanations or explanations of the principle, which are all we can achieve in the case of complex

[11] For Hayek, who is a methodological instrumentalist, explanation is tantamount to classification. Cp. Birner (forthcoming).

[12] In the sense of classify, which is of course a view of explanation that is not shared by everyone (not by Popper, for instance).

phenomena.[13] Instead of rejecting this idea, that underlies Hayek's "explanatory impossibility theorem," as part of a move to make Hayek's naturalistic theory of mind immune to criticism, Popper takes it seriously enough to refer to it 25 years later.[14]

In the modern literature on the mind-body problem Hayek's argument is known as the explanatory gap (cf. Levine 1983 and 1999 and Chalmers 1999). In *SO* Hayek claims that his theory is less materialistic than dualistic theories because it does not assume the existence of a separate mind-substance: ‘‘While our theory leads us to deny any ultimate dualism of the forces governing the realms of the mind and that of the physical world respectively, it forces us at the same time to recognize that for practical purposes we shall always have to adopt a dualistic view’’ (*SO*, 8.46). This is because we cannot produce a complete description or explanation of the processes that constitute our mind and its relationships with the physical order without including a description of the subset of those same processes that do the describing and explaining, i.e., the mind itself. This again is because, as Hayek repeats in 8.44, his theory is not a double-aspect theory. The complete order of all neural processes, ‘‘if we knew it in full, would ... not be another aspect of what we know as mind but would be mind itself.’’

Since *SO* is an identity theory, rather than denying the possibility of reducing the sensory order to the physical order, it implies that there is no need to do so. In the physical order, events are similar or different to the extent that they produce similar or different external effects. In the sensory order, events are classified according to their sensory properties: ‘‘to us mind must remain forever a realm of its own which we can know only through directly experiencing it, but which we shall never be able fully to explain or ‘reduce’ to something else’’ (*SO* 8.98). Yet, the two ways of describing mental phenomena, in physical and in subjective terms, are two alternative ways of describing the same phenomena. For the practical purpose of describing the mind Hayek is a dualist in the sense that we humans with our human minds use different languages describing the mental and the physical. Ontologically, there is just one physical order.[15]

4 Hayek as a Pioneer of AI

[13] Cp. for instance Hayek 1967.

[14] "It has been suggested by F.A. von Hayek ([1952], p. 185) that it must be impossible for us ever to explain the functioning of the human brain in any detail since "any apparatus … must possess a structure of a higher degree of complexity that is possessed by the objects" which it is trying to explain." (Popper and Eccles 1977: 30).

[15] Cp. Levine 1999: 11: "Metaphysically speaking, there is nothing to explain. That is, we are dealing with a brute fact and there is no further source (beyond the fact itself) responsible for its obtaining. The fact that we still find a request for an explanation intelligible in this case shows that we still conceive of the relata in the identity claim as distinct properties, or, perhaps, the one thing as manifesting distinct properties. We can't seem to see the mental property as the same thing as its physical correlate. But though our inability to see this is indeed puzzling, it doesn't show, it can't show, that in fact they aren't the same thing. For what is the case cannot be guaranteed by how we conceive of it."

The constructive defence against Popper's criticism is undertaken in the second part of the paper. Hayek describes a thought experiment that is meant to demonstrate that a causal system *is* capable of one of the higher functions of language, description. By "system" he intends

> "a coherent structure of causally connected physical parts. The term system will thus be used here roughly in the sense in which it is used in von Bertalanffyi's "General System Theory (…) [By system I intend] a persistent structure of coherent material parts that are so connected that, although they can alter their relations to each other and the system thereby can assume various states, there will be a finite number of such states of which the system is capable, that these states can be transformed into each other through certain orderly sequences, and that the relations of the parts are interdependent in the sense that if a certain number of them are fixed, the rest is also determined." (*Systems*, pp. 4-5)

Hayek concentrates on the behaviour of a type of causal system that he calls "classifying system," for a fuller explanation of which he refers to *SO*.[16] After dealing, in the first part of the paper, with a series of preliminaries, Hayek is ready with

> "the setting up of the framework within which we wish to consider the main problem to which this paper is devoted. In the next section we shall take up the question how such a system can transmit to another similar system information about the environment so that the second system will as a result behave in some respects as if it had directly undergone those effects of the environment which in fact have affected only the first system, but have become the object of the "description" transmitted by that system to the second." (*Systems*: 18-9)

He introduces two automata[17] that communicate with one another by means of symbols. Since he uses them in a thought experiment, it is justified to consider them as virtual machines.[18] Hayek very ably concentrates on his main problem by excluding the different problem whether, or to what extent, the structure of the two systems have to be identical or similar in order to be able to interact with one another:[19] he assumes that they are identical. Hayek argues that the self-expressive or symptom and signaling functions of communication pose no problem for his thought experiment. Then he describes a situation in which the two systems are hunting a prey. S_1 can see the prey but S_2 cannot because it is hidden from it by an obstacle. The problem now is how S_1 can describe and communicate the description of the itinerary the prey is following to S_2. The manuscript breaks off in the middle of this attempt to fit the descriptive function of communication by means of symbols into the thought

[16] Hayek's description in *SO* of the human mind is that of a classifier system (a term he does not use).
[17] Hayek does not use that term but he refers to Von Neumann's theory of automata.
[18] Aaron Sloman's comment in correspondence.
[19] He addresses that problem elsewhere. For a discussion, cp. Birner (2009).

experiment, and in the framework of a causal theory of systems.[20] Apparently he did not succeed getting beyond the lowest two functions of communication.[21] This is precisely what Popper had said in his criticism.

5 Hayek and Turing

This section is dedicated to a (non exhaustive) comparison of the ideas in Hayek's *SO* and *Systems* with Turing's in *CMI*. The objective is to give additional arguments that Hayek's *SO* and even more so his *Systems* deserve a place in the AI literature: if Turing's *CMI* is about AI, then so are these texts of Hayek's.

5.1 What is the question?

In a comparison between Turing and Hayek we must not lose from sight that they address different problems – at least at first sight. In *CMI* Turing poses the question "Can machines think?" The problem Hayek wants to solve in *SO* is "What is consciousness?" This, at any rate, is my reconstruction; Hayek himself is much less sure and explicit in *SO*,[22] even though he writes: "it is the existence of a phenomenal world which is different from the physical world which constitutes the main problem" (*SO*, 1.84). This is part of the qualia problem. It is different from the question whether or not we humans can think; it is at best part of the latter problem. Nevertheless, the way Turing and Hayek elaborate their respective problems show some similarities that in my opinion make a comparison non futile.

Turing transforms his original question

> "into [a] more accurate form of [it:] I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning. The original question "Can machines think?" I believe to be too meaningless to deserve discussion." (*CMI*: 442).

[20] It breaks off in the middle of a word, "system". That suggests that part of the typescript has gone missing. I have repeatedly looked for the missing pages in the Hayek archives. A hand-written note by Hayek on the first of the 27 typewritten pages of the ms. reads: "seems incomplete." Added to Hayek's comment quoted in the third para. of section 3 above, this laconic note suggests that he has not looked very hard for possible missing pages, which may be very few in number.

[21] This is also suggested by the fact that years later Hayek writes to Popper that he feels "that some day you ought to come to like even my psychology" (letter of 30 May 1960, Hayek Archives, Hoover Institution on War, Revolution and Peace, box 44/2). This may be taken to imply that Hayek had not solved the problem of showing that causal systems are capable of communication descriptions to other causal systems, thus confirming Hayek's comments (Hayek 1982: 290) quoted above.

[22] This is highly uncharacteristic for Hayek, who in all his work follows a meticulously methodical approach. Cp. Birner (2013).

Now this reformulation comes much closer to the way in which Hayek elaborates the problem of *SO* in the second part of *Systems*. His thought experiment, which is meant to show that physical machines can express their internal states, signal, and communicate descriptions to one another, qualifies as an early exercise in AI. That exercise, moreover, is inspired by a physicalist identity theory of the human mind. Turing's "imitation game" is always interpreted as a procedure in which a human mind attempts to debunk a computer that tries to imitate another human mind. A generalized version of the game, one that is not based on the ontological assumption that a human mind and a computer (and/or its software – in the sequel I will delete this addition) are fundamentally different, would lose its purpose and become meaningless. If there are no fundamental differences between computers and human minds – as Hayek's physicalist identity theory asserts – a Turing test would only compare one kind of material realization of a mind with another. I will return to this in the Conclusion.

When Turing discusses the possible objection of the "Argument from Consciousness," *i.e*., that machines can only be considered to be capable to think if they are capable of experiencing feelings and emotions, he deals with the same problem as Hayek in *SO*. Turing does not deny there is a problem, but he considers it as different from, and secondary to, the problem that he addresses:

> "I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." (*CMI*: 447).

Now, according to Hume "Reason is, and ought only to be the slave of the passions." (Hume 1739: 415).[23] The very least we need for rational thought are motivations.[24] Hayek deals with this effectively by describing how intentions may be modeled in his thought experiment:

> "By intention we shall mean such a state of a system that, whenever its classifying apparatus represents a chain of actions as producing a result which at the same time the internal state of the system singles out as appropriate to that state, it will perform that chain of actions. And we shall define the result or class of results which in any such state will activate the chains of actions which will produce them as the goal or goals to which the intention is directed." (*Systems*: 17)

This is sufficient for the purpose of his thought experiment.

5.2 Functionalism

[23] Research in cognitive science shows that Hume was right.
[24] Aaron Sloman in correspondence.

In the above, I have described Hayek's functionalist approach to the mind. Compare this with what Turing writes:

> "The fact that Babbage's Analytical Engine was to be entirely mechanical will help us to rid ourselves of a superstition. Importance is often attached to the fact that modem digital computers are electrical, and that the nervous system also is electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. Of course electricity usually comes in where fast signalling is concerned, so that it is not surprising that we find it in both these connections. In the nervous system chemical phenomena are at least as important as electrical. In certain computers the storage system is mainly acoustic. The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should look rather for mathematical analogies of function." (*CMI*: 439)

This is identical to Hayek's mental functionalism and structuralism.

5.3 Machines as subjects of themselves

When, on p. 449, Turing writes about machines being their own subjects, he seems to have in mind a different problem than Hayek does when he addresses the question if causal systems can describe themselves – by which he means *fully* describe.

> "The claim that a machine cannot be the subject of its own thought can of course only be answered if it can be shown that the machine has *some* thought with *some* subject matter. Nevertheless, "the subject matter of a machine's operations" does seem to mean something, at least to the people who deal with it. If, for instance, the machine was trying to find a solution of the equation x^2-40a-11=0 one would be tempted to describe this equation as part of the machine's subject matter at that moment. In this sort of sense a machine undoubtedly can be its own subject matter. It may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively. These are possibilities of the near future, rather than Utopian dreams." (*CMI*: 449).

This impression, however, may be mistaken. Compare the following passage:

> "The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get

> changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States." (*CMI*: 458)

This seems similar to the distinction Hayek makes, in para. 18, between changes within a causal system and changes of the system itself:

> "The concept of the state of a certain system must be carefully distinguished from the changes in a collection of elements which turn it into a different system. Different individual systems may be instances of the same kind of system (or possess the same structure) if they are capable of assuming the same states; and any one individual system remains in the same system only so long as it remains capable of assuming any one of the same set of states, but would become a different system in our sense. A full description of any system would have to include sufficient information to derive from it descriptions of all possible states of that system and of their relations to each other, such as the order in which it can pass through the various states and the conditions in which it will pass from one state into another. It will be noted that strictly speaking a change in the permanent nature of one of our systems such as would be produced by long term memory (the acquisition of new connections or linkages) being an irreversible change implies a change of the system rather than a mere change of the state of a given system." (*Systems*: 9-10)

The formulations are different, but Turing's and Hayek's ideas appear to be the same.

5.4 Hayek's fate

Some of Hayek's and Turing's central ideas are very similar or even identical. Yet Hayek has not been recognized as a pioneer of AI whereas Turing has. That might have been different if he had published *Systems*. The radically thorough systematic method that characterizes Hayek's approach to each and every problem he ever put on his research agenda[25] kept him from doing so; he had, after all, failed to complete what he considered to be the homework that Popper had assigned him with his criticism of *SO*. Had he published the paper, even without a satisfactory account of the communication of symbolic description between virtual machines, both Hayek and AI might have been spared a lost opportunity.

6 Conclusion: for a scientifically and morally sounder Turing test?

[25] For an explication of this methodical approach cp. Birner 2013. This is not the only case of Hayek's being the victim of his own ambitiousness and thoroughness. Cp. Birner 1994.

Perhaps the main defect of the Turing test as it is generally interpreted, is that it tests whether humans have the subjective impression that machine intelligence is human. As such, it may be of interest to psychology but hardly to AI. In addition, the Turing test is biased or at least not general (and hence unduly discriminatory in the scientific sense) in that it presupposes a particular type of theory of mind without making this explicit, one that excludes the physicalist identity position. In *CMI*, C, the interrogator, is a human being. In a scientifically sounder version of the Turing test the population of humans and machines should be randomly divided in testers and tested or judges and judged. But this would give rise to legitimate doubts as to what the test is really testing. Is it the capacity of mind-like entities to recognize similar mind-like entities?

There is no doubt that naturally evolved human minds and bodies are capable of much more complex tasks than artificially created mind-like systems and their physical implementations. This is not due to engineering problems in the realization of the latter but to the fact that human minds and bodies are the products of a very long evolutionary process. But we already know this without a Turing test.

Whether or not human judges in a Turing test can be fooled into thinking that machine intelligence is human also depends on whether or not these judges think that they share the same type of consciousness with the objects they judge. According to a radical physicalist identity theory of mind machines are capable of having consciousness and subjective feelings. If they don't,[26] this may be due to the fact that we humans happen to have a longer evolutionary history, in which we have learnt to have these impressions. Likewise, by interacting with humans, machines might learn to understand and explain why we have subjective feelings (as in Star Trek). They could even learn to have these impressions and sentiments themselves, particularly if these have survival value (which in an environment that includes interaction with human minds seems likely). The Turing test, however, is ill-suited for finding out whether or not artificially created mind-like machines have consciousness, or have consciousness that is similar to human minds. Giulio Tononi's Integrated Information Theory offers a much more sophisticated approach, one that even allows of measuring the degree of consciousness – at least in principle. In this perspective it also seems legitimate to ask if machines experience the same dualism as we humans do according to Hayek (*i.e.* we cannot speak of the realm of the mental without using subjective-psychological language;[27] see above, the last two paragraphs of section 3).

The possibility that machines have consciousness may even raise an additional, moral, objection to the traditional Turing test: it discriminates machines in favour of humans

[26] But how could we find out? This raises the same problems Hayek addressed in *Systems* without finding a solution.

[27] The non-reducibility of a subjectivist language to a physicalist one that Hayek argues for may be seen as a solution to what he considers to be a problem of complexity, *viz.* his explanatory impossibility theorem (as I have called it). That is because subjectivist language enables us to speak meaningfully about mental phenomena even in the absence of a complete reduction of them to an explanation in physical terms. Perhaps the idea can be generalized to the question if subjective language and/or impressions may serve to reduce complexity in general.

by assigning the role of judges only to the latter. Machines might feel discriminated against – if, I repeat, they are capable of moral feelings and other emotions at all.

So in the end, my arguments for a scientifically and morally sounder Turing test seems to lead to the conclusion that the Turing test does not serve any useful purpose at all. Turing's belief, quoted above, that "[t]he original question "Can machines think?" [is] too meaningless to deserve discussion" seems to me to be unfounded. Thinking involves things such as intentionality, description, explanation, understanding, creativity, having impressions and creativity. These are all features of consciousness. So Turing's question would reduce to the problem whether intelligent machines are capable of consciousness. That certainly is a difficult question, but it is hardly meaningless. As with so much research in AI, attempts to answer it have taught us more about human minds than about artificial ones, and is likely to continue to do so.

7 References

Birner J (forthcoming), "Generative mechanisms and decreasing abstraction", in Manzo (forthcoming)

Birner J (2014), F. A. Hayek's *The Sensory Order*: An Evolutionary Perspective? *Biological Theory* on-line first, DOI 10.1007/s13752-014-0189-4

Birner J (2013), "F.A. Hayek: the radical economist," http://econ.as.nyu.edu/docs/IO/28047/Birner.pdf

Birner J (2009), "From group selection to ecological niches. Popper's rethinking of evolution in the light of Hayek's theory of culture", in Parusnikova & Cohen 2009

Birner J (1994), "Introduction; Hayek's Grand Research Programme", in Birner & Van Zijp 1994

Birner J & van Zijp, R eds (1994), *Hayek, Co-ordination and Evolution; His Legacy in Philosophy, Politics, Economics, and the History of Ideas*, Routledge, London

Bunge, M ed. (1964), *The Critical Approach to Science and Philosophy. Essays in Honor of Karl R. Popper*, The Free Press, Glencoe

Chalmers, DJ (1999), "The Explanatory Gap. Introduction", in Hameroff et.al. (1999): 1-2

Hameroff SR, Kaszniak AW, Chalmers DJ eds (1999) *Towards a science of consciousness. The third Tucson discussions and debates.* MIT Press, Cambridge

Hayek FA (1952) *The Sensory Order. An Inquiry into the Foundations of Theoretical Psychology*. University of Chicago Press, Chicago, referred to as *SO*

Hayek FA (n.d.) "Within systems and about systems. A statement of some problems of a theory of communication." Hayek Archives folder 94/51, The Hoover Institution, referred to as *Systems*

Hayek FA (1964), "The Theory of Complex Phenomena", in Bunge (1964)

Hayek, FA 1967, *Studies in Philosophy, Politics and Economics*, University of Chicago Press, Chicago

Hayek FA (1982) "The Sensory Order After 25 Years", in Weimer and Palermo (1982)

Hume, D (1739), *A Treatise of Human Nature*, Selby-Bigge, LA ed., Claredon Press, Oxford, 1896

Levine, J 1999, “Conceivability, Identity, and the Explanatory Gap”, in Chalmers et.al 1999: 3-13

Levine, J (1983), “Materialism and Qualia: The Explanatory Gap” *Pacific Philosophical Quarterly* 64:354-61

Manzo G ed. (forthcoming), *Paradoxes, Mechanisms, Consequences: Essays in Honor of Mohamed Cherkaoui.* Oxford: Bardwell Press

Parusnikova & Cohen RS eds. (2009), *Rethinking Popper*, Boston Studies in the Philosophy of Science 272, Berlin, Springer

Popper KR (1953) “Language and the Body-Mind Problem”, in *Proceedings of the XIth Congress of Philosophy,* **7**: 101-7, Amsterdam, North Holland

Popper KR (1972) *Objective Knowledge. An Evolutionary Approach,* Oxford, Clarendon Press

Tononi, G 2012, “Integrated information theory of consciousness: an updated account” *Archives Italiennes de Biologie,* 150: 290-326

Turing, A (1950), “Computing Machinery and Intelligence,” *Mind* LIX (236): 433-560, referred to as *CMI*

Van den Hauwe LMP (2011), “Hayek, Gödel, and the case for methodological dualism,” *Journal of Economic Methodology* 18 (4): 387–407

Weimer WB and Palermo DS (1982) *Cognition and the Symbolic Process*, Hillsdale, N.J, Lawrence Erlbaum

On the Concept of Correct Hits in Spoken Term Detection*

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences and University of Szeged
H-6720 Szeged, Tisza Lajos krt. 103., Hungary
ggabor@inf.u-szeged.hu

Abstract. In most Information Retrieval (IR) tasks the aim is to find human-comprehensible items of information in large archives. One such task is the spoken term detection (STD) one, where we look for user-entered keywords in a large audio database. To evaluate the performance of a spoken term detection system we have to know the real occurrences of the keywords entered. Although there are standard automatic ways to obtain these locations, it is not obvious how these match user expectations. In our study, we asked a number of subjects to locate these relevant occurrences, and we compared the performance of our spoken term detection system using their responses. In addition, we investigated the nature of their answers, seeking to find a way to determine a commonly accepted list of relevant occurrences.

KeyWords: spoken term detection, information retrieval, artificial intelligence, speech processing, keyword spotting

Spoken term detection [19] is a relatively new area, which is closely related to speech recognition. Both seek to precisely match the relation between audio speech recordings and their transcripts; but while speech recognition seeks to produce the correct transcript of speech utterances [16], spoken term detection attempts to locate those parts of the utterance where the user-entered keyword or keywords occur.

One critical part of the latter concept is that of identifying the relevant occurrences of the keywords. At first glance, this question could be answered quite easily, provided we have the correct, time-aligned textual representation (*transcription*) of the utterances: a standard solution is to consider an occurrence relevant if it is present at the given position as a whole word [15]. However, this approach completely ignores compound words, which could also be considered relevant occurrences. A further problem arises in the *agglutinative* languages [7, 4]: these construct new word forms by adding affix morphemes to the end of

* This publication is supported by the European Union and co-funded by the European Social Fund. Project title: Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073.

the word stem. (E.g. in Hungarian the expression "in my house" takes the form *ház-am-ban*.) In these cases, inflected forms of keywords should also be accepted. (This is even present in English to a certain extent, e.g. the plural form of nouns.)

The best solution for this task would be to ask the user which occurrences he thinks are relevant. The problem with this approach is that usually the archives are huge, hence hand-labeling them is quite expensive. Furthermore, the expectations could vary from user to user, but for practical reasons we would need an "objective" list of the relevant occurrences. It is also not clear whether, by using user responses, a broad consensus could be reached; i.e. whether it is possible to create an occurrence list that is acceptable to most people.

In this study we examined these expectations, and we also sought to measure the effect of these on STD accuracy. (Although we think that the topic of this paper is not limited to spoken term detection, but it also covers several IR topics like text document retrieval [3] and document categorization [20] as well.) For this reason, we created a form containing ambiguous occurrences and asked people about their opinions of relevance. The results were compared with each other, and with our standard, automatic occurrence-detection method.

Although our experiments were performed on a set of Hungarian recordings, we think that our findings might be of interest to researchers working with other languages as well, especially as recently languages other than English have been receiving more attention (e.g. [14, 17, 22, 13]).

1 The Spoken Term Detection Task

In the spoken term detection task we would like to find the user-entered expressions (called *terms* or *keywords*) in an audio database (the set of *recordings*). An STD method returns a list of *hits*, each consisting of the position of occurrence (a speech signal index, starting and ending times), the term found, and a probability value that can be used to rank the hits. In contrast to other similar tasks, in STD the order of the hits does not matter; the probability value is primarily used to further filter the hit list, keeping just the more probable elements.

As a user expects a quick response for his input, we have to scan hours of recordings in just a few seconds (or less); to achieve this, the task is usually separated into two distinct parts. In the first one, steps requiring intensive computation are performed without knowing the actual search term, resulting in some intermediate representation. Then, when the user enters the keyword(s): some kind of (quick) search is performed in this representation. There exist a number of such intermediate representations, from which we used the one where we stored only the most probable phoneme sequence for a recording [15, 6].

In this paper we will concentrate on the concept of *relevant occurrence*; hence spoken term detection is only of interest to us here because it can provide us with accuracy scores that can be compared with each other when using different strategies for detecting these occurrences. Therefore, in a quite unusual way, we will use the *same* STD system configuration, with exactly the same parameters; what we will vary is the occurrences of search terms we expect it to find.

1.1 The Evaluation Metrics

A spoken term detection system returns a list of hits for a query. Given the correct list of occurrences, we should rate the performance of the system to be able to compare different systems and configurations. Since STD is an information retrieval task, it is straightforward to apply standard IR metrics of precision and recall:

$$Precision = \frac{N_C}{N_C + N_{FA}} \tag{1}$$

and

$$Recall = \frac{N_C}{N_{Total}}, \tag{2}$$

where N_C is the number of correct hits returned, N_{FA} is the number of false alarms, and N_{Total} is the total number of real occurrences [1]. Intuitively, precision measures how much of the hit list returned contains correct hits, while recall measures the fraction of the real occurrences that were found. A perfect system has both a precision and a recall score of 1 (or 100%). Clearly, there is a trade-off between these two values: high precision can easily lead to a low recall score if we only include very probable hits in our list, while it is easy to achieve high recall rates and get poor precision scores by returning a hit list full of "rubbish". Hence it would be better to summarize the performance of a system using just one score. In IR tasks usually the F-measure (or F_1) is used for this, which is the harmonic mean of precision and recall, defined as

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{3}$$

This formula, however, weights precision and recall equally, which might differ from our preferences. We could also use different weights for the two measures, but their relative importance is not really clear. Another requirement in STD might be to normalize the scores based on the total length of the recordings. This is why in the area of spoken term detection usually some other – although similar – measures are used.

Figure-of-Merit (FOM) The evaluation metric commonly applied earlier is the Figure-of-Merit (FOM). It can be calculated simply as the mean of the recall scores when we allow only $1, 2, \ldots 10$ false alarms per hour per keyword. In general, this metric is a quite permissive one: it is possible to achieve relatively high scores quite easily, since 10 false alarms per hour clearly exceeds the limits of actual applicability. It weights keywords relative to their frequency of occurrence in the archive of recordings, hence if we want to maximize this score, it may be worth optimizing it on more frequent keywords instead of rarer ones. However, this behaviour is clearly contrary to user expectations. Another interesting property is that the STD system does not have to filter the hits returned, but the FOM metric determines the actual probability thresholds depending on the number of false alarms permitted.

Actual Term-Weighted Value (ATWV) Another, more strict measure was defined by the National Institute of Standards and Technology (NIST) in its 2006 evaluation of Spoken Term Detection [12]. Unlike FOM, it uses all the hits supplied by the STD method, and is defined as

$$ATWV = 1 - \frac{1}{T} \sum_{t=1}^{T} \big(P_{Miss}(t) + \beta P_{FA}(t) \big), \tag{4}$$

where T is the number of terms, $P_{Miss}(t)$ is the probability value of missing the term t and $P_{FA}(t)$ is the probability value of a false alarm. These probability values are defined as

$$P_{Miss}(t) = 1 - \frac{N_C(t)}{N_{Total}(t)} \tag{5}$$

and

$$P_{FA}(t) = \frac{N_{FA}(t)}{T_{speech} - N_{Total}(t)}, \tag{6}$$

where T_{speech} is the duration of the test speech in seconds. (This formula uses the somewhat arbitrary assumption that every term can occur once in every second.) Usually the penalty factor for false alarms (β) is set to 1000. A system achieving perfect detection (i.e. having a precision and a recall of 1.0) has an ATWV score of 1.0; a system returning no hits has a score of 0.0; while a system which finds all occurrences, but produces 3.6 false alarms for each term and speech hour also has a score of 0.0 (assuming that T_{speech} is significantly larger than N_{Total}) [15].

ATWV differs from FOM in a number of ways. First, it weights all keywords equally, regardless of the frequency of actual occurrences. Second, it punishes missed occurrences and false alarms much more than FOM does, so it is a very strict metric indeed. Third, whereas FOM performs a filtering of the hits returned, ATWV uses *all* of them, hence to achieve a high ATWV score an STD system has to filter the hit lists itself by setting up a minimal probability threshold $P_{\min}$. This is usually done in two steps: first the actual $P_{\min}$ value is determined on a *development set* of recordings as the threshold value belonging to the optimal ATWV score. Then, to measure actual performance, ATWV is calculated on another set of recordings (the *test set*) using the already determined value for $P_{\min}$. In this study, we also performed these two steps.

2 The Concept of "Correct Hit"

Having defined the evaluation metrics, we are now able to calculate the accuracy scores of an STD system when we have the number of correct hits, false alarms and missed occurrences. For this, we supposedly know the hits returned, and their ordering; however, we still have to define the technique to get the list of real occurrences, and the way of matching returned hits and the relevant occurrences.

2.1 Matching Hits and Occurrences

In the literature this topic has been discussed quite extensively. Of course, a hit and an occurrence can be matched only if the keywords are the same, and they occur in the same recording. As regards the match of time-alignment, there are a number of possibilities. A valid option would be to expect both the starting and ending times to lie below a threshold. [12] expects the time span of the hit to be in at most 0.5 seconds from the centre of the real occurrence. [21] demanded that the time spans of the hit and of the occurrence intersect. We chose the latter method, partly because of the agglutinative nature of the Hungarian language, which makes the task of determining the exact keyword starting and ending times quite hard.

2.2 Determining the "Real" Occurrences

When we search for the method of choosing the "relevant occurrences" in the literature, we usually find no mention of it. Hence we chose to assume that a keyword only occured if it was present in the textual transcription as a whole word by itself. This approach, however, is hardly applicable when we work with recordings different from English (which was also the case for us). In morphologically rich languages such as Hungarian, nouns (which are typical candidates for keywords) can have hundreds of different forms owing to grammatical number, possession marking and grammatical cases, all of these forms being ones that should also be treated as "real" occurrences.

Our standard automatic method is a simple variation of this default approach. In it we treat a given position as an occurrence of the given keyword if the word at this position contains the keyword. (This concept can be extended to keywords consisting of several words in a straightforward way.) Because in Hungarian a noun ending with a vowel may change its form when getting some inflections (like the noun *"Amerika"* (*America*) changing to the form *"Amerikában"* (*in America*)), we also considered the occurrence a real one if the given keyword appears in the form having its last vowel substituted by its long counterpart, as long as the last vowel is also the last phoneme of the keyword.

It is of course known that this technique is not perfect: for short keywords in particular it is likely that they will appear inside other words having a completely different meaning, which should be categorized as false alarms.

2.3 Relying on Human Expectations

The other choice is to employ the concept that a relevant occurrence is where the actual users think that the current occurrence is indeed relevant. This approach sounds quite reasonable, but it requires valuable human interaction, so it could be quite labour-intensive when we have to annotate a big archive manually. For smaller archives, however, it can be carried out relatively cheaply; and since the aim of this study was to check the difference between the automatic and human concept of a real occurrence, we performed this manual task by asking our subjects about their opinions of potential occurrences.

Strategy	Dev	Test
Automatic	381	709
Subject #1	365	690
Subject #2	368	689
Subject #3	396	732
Subject #4	366	699
Subject #5	367	697
Subjects (majority voting)	367	697
Clean occurrences	334	651

Table 1. The number of relevant occurrences using different strategies for determining correct hits for the development (*Dev*) and test (*Test*) sets.

Creating the Form to Fill In To make subjects list the occurrences which they thought were relevant, we created a form using the textual transcript of the recordings, which each subject had to fill in. For each keyword we located the similar letter-sequences in the transcripts of the recordings using the edit distance [9]: we allowed character insertions, deletions and substitutions, and listed the parts of the recordings where we could reproduce the given keyword with at most N operations, where N was 30% of the length of the keyword. (That is, for a search term consisting of 10 characters, we allowed only 3 operations.)

Because this list was still quite long, we shortened it with a simple trick: we did not list those occurrences which could be produced without any operations, and were located at the beginning of a word. Instead, we assumed that these were the occurrences of the actual keyword in inflected form, thus treating them as relevant occurrences. (The set of these occurrences was also used in the experiments section, referred to as the list of *clean occurrences*.) Of course this was not so in a number of cases (like certain compound words), but this technique was quite close to our objective, and it effectively reduced the number of items in the form.

Evaluating Subject Responses Table 1 shows the number of relevant occurrences found when using the automatic occurrence detector method (see line "Automatic"), and for the responses of the subjects (see lines "Subject #N"). The form contained 111 (development set) and 242 (test set) occurrences that were used to decide on their relevance; from these, the test subjects marked between 31 and 62, and between 38 and 81 occurrences as relevant ones, development sets and test sets, respectively. The results indicate that most occurrences were judged in quite a similar way by our subjects (with the exception of Subject #3). Besides comparing the responses of the subjects with the results of our standard automatic occurrence checker, we also wanted to know whether a consensus could be reached between the answers of the subjects. For this reason we used majority voting: we considered an occurrence relevant if at least half of the subjects (now at least three of them) considered it relevant.

3 Experiments and Results

Having defined the task, introduced the method of obtaining subject responses, and selected the evaluation metrics, we will now turn to the testing part. We will describe the STD framework used, present and analyze the results, concentrating on the various kinds of discrepancies among the individual subjects, and between each subject and the automatic occurrence detector method used.

3.1 The STD Framework

Testing was performed using the spoken term detection system presented in [6]. It uses phoneme sequences as an intermediate representation, and looks for the actual search term in these sequences, allowing phoneme insertions, deletions and substitutions. These operations have different costs depending on the given phoneme (or phoneme pair), calculated from phoneme-level confusion statistics.

We used recordings of Hungarian broadcast news for testing, which were taken from 8 different TV channels [5]. The 70 broadcast news recordings were divided into three groups: the first, largest one (about 5 hours long) was used for training purposes. The second part (about an hour long) was the *development set*: these recordings were used to determine the optimal threshold for the ATWV metric. The third part was the *test set* (about 2 hours long), and it was used to evaluate the overall performance. We chose 50 words and expressions as search terms, which came up in the news recordings quite frequently. They varied between 6-16 phonemes in length (2-6 syllables), and they were all nouns, one-third of them (18) being proper nouns. The phoneme sequence intermediate representations were produced by Artificial Neural Networks [2], trained in the way described in [18], using the standard MFCC $+\Delta + \Delta\Delta$ feature set [8].

3.2 Results

The accuracy scores produced by our actual STD system (using different strategies for determining the list of relevant occurrences) can be seen in Table 2. By "Automatic" we mean the standard, automatic method used for determining correct hits; "Subject #N" means the responses of the Nth subject. Below we list the mean and the median values of the accuracy scores produced, and the scores obtained using majority voting. The last line shows the accuracy scores calculated without any subject answers, using just the clean occurrences; that is, in this case we treated an occurrence as a correct one only if the keyword appeared unchanged in the transcription at the beginning of a word.

The first thing to notice is that the FOM scores practically do not vary, which is probably due to the way this accuracy score is calculated: it is relatively easy to achieve high FOM scores, but it is very hard to significantly improve them. The ATWV scores, however, differ much more from each other, ranging from 48.00% (where we use only the clean occurrences) to 60.23% when using the list of relevant occurrences given by Subject #3. The results are also quite different from the case where we applied our automatic method.

Strategy	FOM	ATWV	F_1	Prec.	Recall
Automatic	88.72%	56.84%	85.29%	91.17%	80.11%
Subject #1	88.35%	52.32%	83.93%	88.44%	79.86%
Subject #2	87.39%	48.00%	82.32%	86.68%	78.37%
Subject #3	88.85%	60.23%	86.05%	93.58%	79.64%
Subject #4	88.15%	52.90%	84.11%	89.25%	79.54%
Subject #5	88.22%	53.05%	84.24%	89.25%	79.77%
Subjects (mean)	88.19%	53.30%	84.13%	89.44%	79.44%
Subjects (median)	88.22%	52.90%	84.11%	89.25%	79.64%
Subjects (majority voting)	88.22%	53.07%	84.24%	89.25%	79.77%
Clean occurrences	87.94%	44.77%	81.48%	83.31%	79.72%

Table 2. STD accuracy scores using different strategies for determining correct hits

The F_1 scores varied from 82.32% to 86.05%. Quite interestingly, the corresponding precision scores were practically the same, so the difference came from the recall scores. The correlation of the precision, F-measure, ATWV scores, and the number of occurrences marked as real is clear: for Subject #3 these were 93.58%, 86.05%, 60.23% and 732, respectively, whereas for Subject #2 these were 86.68%, 82.32%, 48.00% and 689. (The ATWV metric is known to be fairly sensitive to false alarms.)

Another interesting finding is that the scores belonging to majority voting appear to be quite close to those of three subjects (#1, #4 and #5), or the mean/median of all the subjects. This suggests that by using the simple technique of majority voting a consensus of correct hits can be achieved, which falls quite close to the expectations of the average user.

3.3 Verifying the Occurrences

Having evaluated the accuracy scores belonging to the different subject responses, we will now turn to the perhaps more interesting part, where we focus on the more significant and/or more interesting differences among the responses of the users or between the user-entered and the automatic hit lists. Note that, as we used a Hungarian database for this study, the examples below will also be in Hungarian; nevertheless, we think that the cases encountered have a much wider scope as probably quite similar types appear in other languages as well.

One well-known drawback of language-independent STD approaches is that they are likely to produce false alarms when the (usually short) actual search term is contained inside another word. In our case, one such example was the term *"kormány"* (meaning *cabinet*), which came up quite frequently inside the word *"önkormányzat"* (*local council*). Since in this case the whole keyword is present, the automatic occurrence detector method included these as real occurrences, whereas 4 of the 5 subjects treated them as false alarms. Of course the STD system, relying only on the acoustic data, also found these occurrences.

Recall that, due to the agglutinative property of the Hungarian language, we allowed the final vowel of the keyword (as long as it was also the last phoneme) to change to its long counterpart, so the STD system was also expected to find these occurrences. However, by default no such changes with earlier vowels were allowed, although they were also sometimes related to similar word-pairs. A good example of this is the keyword *"vasút"* (meaning *railway*) and the word *"vasutas"* (*railway worker*); each subject viewed the latter word as a relevant occurrence of the search term. Yet, for the term *"miniszter"* (*minister*), there is only a vowel difference in *"minisztérium"* (*ministry*), hence it is exactly the same type as the previous one; but it was rejected by 4 out of the 5 subjects.

Another big group was the presence of certain proper nouns in the list of keywords, typically names of people like *"Angela Merkel"* (German chancellor), *"Bajnai Gordon"* or *"Orbán Viktor"* (both of them being Hungarian prime ministers[1]). The search terms consisted of their full names (i.e. both first and family names), whereas sometimes these people were referred to only by their family names. All the subjects agreed that these were real occurrences, despite that only half of the actual keywords were present at the given position. Note that as we used edit distance when creating the form, only those occurrences were present for the subjects to evaluate where the context was sufficiently similar to the first name (e.g. *"amely Merkel"*, "*Bajnai kormány"*, *"Orbán kormány"*).

A quite similar case was that of the keyword *"rendőrség"* (*police force*), which, due to the similarity of the word following it, proved likely to occur in a recording where only the word *"rendőr"* (*policeman*) was present. Here 3 of the 5 subjects found this "inverse containment" relevant, indicating that the concept of the two words are strongly related. In the last frequent case the keyword was *"gázár"* (*gas price*), and the listed items in the form all contained *"gáz ára"* (*price of gas*); all subjects thought that these were real occurrences of the search term.

From these examples it can be seen that the subjects usually agreed with each other, but their choice can hardly be predicted automatically. If a word contains the keyword, then it is usually a correct occurrence. But at certain times (*kormány*) it is a false alarm, while at other times (*rendőrség*) the keyword contains the word that actually occurred. The last vowel of the keyword may become its long counterpart. But such a change is sometimes allowed for other vowels as well (*vasút*), while sometimes it is not (*miniszter*). The case of *"gázár"* probably cannot be handled at all: allowing word boundaries inside keywords would lead to a lot of false alarms. Still, when looking for famous people, the keyword should be only their family name (like *Merkel*, *Bajnai* and *Orbán*).

The accuracy scores in Table 2 also accord with our findings when examining the actual answers of subjects. Subject #3 accepted both *"minisztérium"* for the keyword *"miniszter"* and *"önkormányzat"* for the search term *"kormány"*; this compliance reduced the number of false alarms for the STD system, leading to high precision, ATWV and F_1 scores. In contrast, Subject #2 rejected several compound words as correct hits, which were all accepted by the other four subjects; this is also reflected in the lower precision, F_1 and ATWV scores.

[1] Although, of course, not at the same time

Quite interestingly, when there was a disagreement among the subjects, in most cases four of them agreed on one option, and only in four instances was there a voting outcome of three to two. This may indicate that in almost every case a broad consensus can be achieved, although this should be tested in experiments with more subjects. Our test results also support this hypothesis: increasing the number of votes required to four lowered the accuracy scores only slightly, whereas when we required that all subjects should agree, they fell more sharply.

Comparing the scores obtained involving human interaction with those we got using the two automatic methods to determine the relevant occurrences, it is clear that they differ significantly: when we only allowed the clean hits, the resulting ATWV score of 44.77% was low compared to the others due to the high number of false alarms; whereas when we used the standard automatic method, it was too permissive, resulting in an overoptimistic ATWV score of 56.84%.

Based on these observations, we can sum up our findings in three parts. Firstly, keyword selection should match user behaviour a bit more: all search terms should be nouns, preferably proper nouns (e.g. names, cities, etc.), and for well-known people only their family name should be used. Of course a limitation for this is the set of available recordings (so that the given keywords should occur in the dataset several times); still, further investigations should be preceded by a more careful keyword selection.

The form containing the possible occurrences was constructed in a syntactical manner (using the edit distance-based similarity of the transcriptions); from the results it seems that we should also turn to a linguistic analysis. It would mean a more robust way to distinguish, for example, the inflected forms (e.g. plurals) of the keywords from compound words, since the latter ones should remain in the form to fill, whereas the former occurrences should be omitted.

Overall, it seems that the users focus on the stem of the keywords, often even dismissing affixes (e.g. *rendőr* instead of *rendőrség*, *vasút* instead of *vasutas*). In some cases this is also an oversimplification (e.g. the case of *miniszter* – *minisztérium*), but it still seems to be a pretty close estimation of keyword occurrence relevance. A deeper analysis could be performed via a more detailed linguistic analysis like using Natural Language Processing tools, or expressing the type of connection between word forms via a WordNet [11, 10].

4 Conclusions

In this study, we examined the spoken term detection task from an unusual viewpoint: we checked how much automatically generated ground truth keyword occurrences match user expectations. For this, we asked a number of subjects to mark the possible occurrences that they thought were relevant. We found that although no two subjects gave exactly the same responses, generally their answers were quite similar; and by using majority voting a clear consensus could be achieved. But the standard automatic keyword occurrence detection methods used were either too lax or too strict when compared with the subject responses.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York (1999)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
3. Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM 28(3), 289–299 (1985)
4. Farkas, R., Vincze, V., Nagy, I., Ormándi, R., Szarvas, G., Almási, A.: Web-based lemmatisation of named entities. In: Proceedings of TSD. pp. 53–60. Brno, Czech Republic (2008)
5. Gosztolya, G., Tóth, L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal (in Hungarian). In: Proceedings of MSZNY. pp. 224–235. Szeged, Hungary (2010)
6. Gosztolya, G., Tóth, L.: Spoken term detection based on the most probable phoneme sequence. In: Proceedings of SAMI. pp. 101–106. Smolenice, Slovakia (Jan 2011)
7. Hakkani-Tür, D.Z., Oflazer, K., Tür, G.: Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities 36(4), 381–410 (2002)
8. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice Hall (2001)
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
10. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of GWC. pp. 310–320 (2008)
11. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
12. NIST: Spoken Term Detection 2006 Evaluation Plan (2006)
13. Özgür, A., Özgür, L., Güngör, T.: Text categorization with class-based and corpus-based keyword selection. In: Proceedings of ISCIS. pp. 607–616 (2005)
14. Parlak, S., Saraclar, M.: Spoken term detection for Turkish broadcast news. In: Proceedings of ICASSP. pp. 5244–5247 (2008)
15. Pinto, J., Hermansky, H., Szöke, I., Prasanna, S.: Fast approximate spoken term detection from sequence of phonemes. In: Proceedings of SIGIR. Singapore (2008)
16. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall (1993)
17. Tangruamsub, S., Punyabukkana, P., Suchato, A.: Thai speech keyword spotting using heterogeneous acoustic modeling. In: Proceedings of RIVF. pp. 253–260 (2007)
18. Tóth, L.: A hierarchical, context-dependent Neural Network architecture for improved phone recognition. In: Proceedings of ICASSP. pp. 5040–5043 (2011)
19. Wang, D.: Out-of-Vocabulary Spoken Term Detection. Ph.D. thesis, University of Edinburgh (2010)
20. Xu, J.W., Singh, V., Govindaraju, V., Neogi, D.: A hierarchical classification model for document categorization. In: Proceedings of ICDAR. pp. 486–490 (2009)
21. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)
22. Zhang, P., Han, J., Shao, J., Yan, Y.: A new keyword spotting approach for spontaneous Mandarin speech. In: Proceedings of ICSP (2006)

Action Recognition based on Hierarchical Self-Organizing Maps

Miriam Buonamente[1], Haris Dindo[1], and Magnus Johnsson[2]

[1] RoboticsLab, DICGIM, University of Palermo,
Viale delle Scienze, Ed. 6, 90128 Palermo, Italy
{miriam.buonamente,haris.dindo}@unipa.it
http://www.unipa.it
[2] Lund University Cognitive Science,
Helgonavägen 3, 221 00 Lund, Sweden
magnus@magnusjohnsson.se
http://www.magnusjohnsson.se

Abstract. We propose a hierarchical neural architecture able to recognise observed human actions. Each layer in the architecture represents increasingly complex human activity features. The first layer consists of a SOM which performs dimensionality reduction and clustering of the feature space. It represents the dynamics of the stream of posture frames in action sequences as activity trajectories over time. The second layer in the hierarchy consists of another SOM which clusters the activity trajectories of the first-layer SOM and thus it learns to represent action prototypes independent of how long the activity trajectories last. The third layer of the hierarchy consists of a neural network that learns to label action prototypes of the second-layer SOM and is independent - to certain extent - of the camera's angle and relative distance to the actor. The experiments were carried out with encouraging results with action movies taken from the INRIA 4D repository. The architecture correctly recognised 100% of the actions it was trained on, while it exhibited 53% recognition rate when presented with similar actions interpreted and performed by a different actor.

Keywords: Self-Organizing Map, Neural Network, Action Recognition, Hierarchical models, Intention Understanding

1 Introduction

Recognition of human intentions is becoming increasingly demanded due to its potential application in a variety of domains such as assisted living and ambient intelligence, video and visual surveillance, human-computer interfaces, gaming and gesture-based control. Typically, an intention recognition system is focused on a sequence of observed actions performed by the agent whose intention is being recognised. To provide the system with this component, it is necessary to use activity recognition together with the intention recognition. The purpose of action recognition is an analysis of ongoing events from data captured by a camera in order to track movements of humans and to identify actions.

Many challenges make the action recognition task extremely difficult to imitate artificially, each person differs in terms of height, weight, shape of the human body and gender. Another important aspect to be considered is the impact of the camera viewing angle variations on the action recognition performance. Multi-camera setups have been employed to implement view independent methods [1], [2], [3]. These methods are based on the observation of the human body from different angles, obtaining in this way a view-invariant representation.

Dealing with action recognition, it is important to give a brief definition of what we mean by action. We adopt the following action hierarchy: actions and activities. The term action is used for simple motion patterns typically executed by a single human. An example of an action is crossing arms. A sequence of actions represents an activity, such as the activity dancing. Activities usually involve coordination among persons, objects and environments. In this paper, we focus only on the recognition of actions, where actions can be viewed as sequences of body postures.

An important question is how to implement the action recognition ability in an artificial agent. We tried to find a suitable neural network architecture having this ability. In our previous work, we have focused on the representational part of the problem. We endowed an artificial agent with the ability to internally represent action patterns [4]. Our system was based on the Associative Self-Organizing Map [5], a variant of the Self-Organizing Map (SOM) [6], which learns to associate its activity with additional inputs. The solution was able to parsimoniously represent human actions.

In this paper, we present a novel architecture able to represent and classify others' behaviour. In order to get a more complete classification system we adopt a hierarchical neural approach. The first level in the system is a SOM that learns to represent postures - or posture changes - depending on the input to the system. The second level is another SOM that to represent the superimposed activity trace in the first level SOM during the action, i.e. it learns to represent actions. Thus, the second layer SOM provides a kind of time independent representation of the action prototypes. The third level is a supervised artificial neural network that learns to label the action.
In our previous paper [7] we showed that we could get discriminable activity traces using an A-SOM, which corresponds to the first level SOM in the current system.The system was able to simulate the likely continuation of the recognised action. Due to this ability, the A-SOM could receive an *incomplete* input pattern (e.g. an initial part of the input sequence only) and continue to elicit the most likely evolution of the action, i.e. to carry out sequence completion of perceptual activity over time. In the present system, instead, we focus on the probem of robust action representation and recogniton, given the whole (noisy) input sequence. We are currently working towards an integration of the two approaches.

We have tested the ability of our architecture to recognise observed actions on movies taken from the "INRIA 4D repository [3]", a publicly available dataset of movies representing 13 common actions: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, and throw (see Fig. 1).

The implementation of all code for the experiments presented in this paper was done in C++ using the neural modelling framework "Ikaros" [8].

This paper is organized as follows: A short presentation of the proposed architecture is given in section II; section III presents the experiment for evaluating the model; and finally conclusions are outlined in section IV.

2 Proposed Architecture

The architecture presented in this paper is composed of three layers of neural networks, see Fig. 3. The first and the second layers consist of SOM networks whereas the third layer consists of a custom made supervised neural network. The first layer SOM receives sequences of vectors representing preprocessed sequences of posture images. The activity trajectories, Fig. 2, elicited during the time actions last are superimposed and vectorized into a new representations before entering the layer two SOM as input. This superimposition process can be imagined as the projection of the matrices representing the activity in the grid of neurons in the SOM for all the iterations an action lasts onto a new matrix of the same dimensionality, followed by a vectorization process. The second layer SOM thus clusters the activity trajectories and learns to represent action prototypes independent of how long the activity trajectories in the first layer SOM last. Thus the second layer SOM provides a kind of time independent representation of the action prototypes. The activity of the second layer SOM is conveyed to a third level neural network that learns to label the action prototypes of the second layer SOM independent of the camera's capturing angle and distance to the actor.

2.1 First and Second Layers

The first and the second layers of the architecture consist of SOMs. The SOM is one of the most popular neural networks and has been successfully applied in pattern recognition and image analysis. The SOM is trained using unsupervised learning to produce a smaller discretized representation of its input space. In a sense it resembles the functioning of the brain in pattern recognition tasks. When presented with input, it excites neurons in a specific area. The goal of learning in the SOM is to cause nearby parts of the network to respond to similar input patterns while clustering a high-dimensional input space to a lower-dimensional

[3] The repository is available at http://4drepository.inrialpes.fr. It offers several movies representing sequences of actions. Each video is captured from 5 different cameras. For the experiments in this paper we chose the movie "Andreas2" for training and "Hedlena2" for testing, both with frontal camera view "cam0".

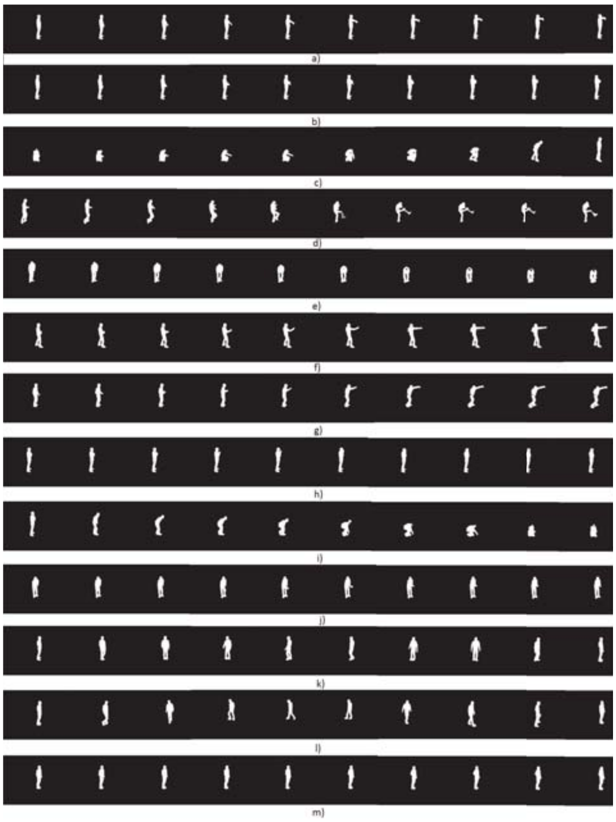


Fig. 1. Prototypical postures of 13 different actions in our dataset: check watch, cross arms, get up, kick, pick up, point, punch, scratch head, sit down, throw, turn around, walk, wave hand.

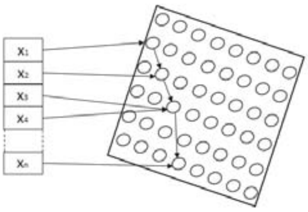


Fig. 2. The trajectory resulting from the neurons activated by the input sequence.

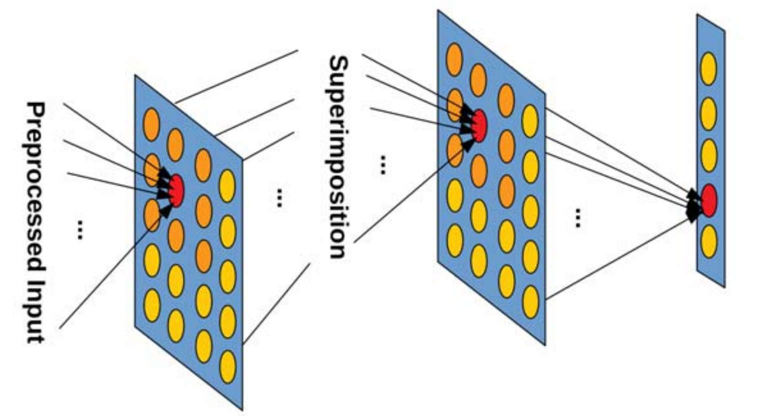


Fig. 3. The proposed architecture is composed of three layers of neural networks. The first and the second layers consist of SOM networks whereas the third layer consists of a custom made supervised neural network.

output space. SOMs are different from many other artificial neural networks because they use a neighbourhood function to preserve the topological properties of the input space. The SOM algorithms adapt a grid of neurons, so that neurons located close to each other respond to similar features.

The SOM structure is made of one input layer and one output layer, the latter also known as the Kohonen layer. The input layer is fully connected to the neurons in the Kohonen layer. The weight vectors of the neurons in the Kohonen layer are modified iteratively in the training phase. When a new input arrives, every neuron competes to represent it. The Best Matching Unit (BMU) is the neuron that wins the competition. The BMU together with its neighbours in the grid are allowed to adapt to the input. The neighbouring neurons less so than the BMU. Neighbouring neurons will gradually specialise to represent similar inputs, and the representations will become ordered in the map. Another important characteristic of the SOM is its ability to generalise, i.e. the network can recognise or characterise input it has never encountered before.

The SOM consists of a grid of neurons with a fixed number of neurons and a fixed topology. Each neuron n_i is associated with a weight vector w_i. All the elements of all the weight vectors are initialized by real numbers randomly selected from a uniform distribution between 0 and 1, after which all the weight vectors are normalized, i.e. turned into unit vectors.

At time t each neuron n_i receives an input vector $x(t)$.

The BMU n_b at time t is the neuron with the weight vector w_b that is most similar to the input $x(t)$ and is obtained by:

$$b = \mathrm{argmax}_i \frac{x(t) \cdot w_i(t)}{||x(t)||||w_i(t)||}, \tag{1}$$

The neurons of the Kohonen layer adapt to increase their representation of the current input by modifying their weight vectors to become more similar to it with an amount that depends on a Gaussian function of the neuron's distance to the BMU:

$$\Delta w_i = \gamma(t) G_{ib}(t)(x(t) - w_i(t)) \tag{2}$$

where the learning rate $\gamma(t)$ is a monotonically decreasing function of time. $G_{ib}(t)$ is a Gaussian function, with a radius $\sigma(t)$ monotonically decreasing with time, of the distance in the map between the neuron n_i and the BMU:

$$G_{ib}(t) = \exp \frac{-d(i,b)^2}{\sigma(t)^2} \tag{3}$$

2.2 Third Layer

The third layer, which is the output layer of the architecture, consists of an array of a fixed number of neurons. Each neuron n_i is associated with a weight vector $w_i \in R^n$, where n is equal to the number of neurons in the second layer SOM. All the elements of the weight vector are initialized by real numbers randomly selected from a uniform distribution between 0 and 1, after which the weight vector is normalized.

At time t each neuron n_i receives an input vector $x(t) \in R^n$, which is the vectorized activity of the second layer SOM.

The activity y_i in the neuron n_i is calculated using the standard cosine metric:

$$y_i = \frac{x(t) \cdot w_i(t)}{||x(t)||||w_i||} \tag{4}$$

During the learning phase the weights w_{ij} are adapted by

$$w_{ij}(t+1) = w_{ij}(t) + \beta x_j(t)[y_i - d_i] \tag{5}$$

where β is the adaptation strength and d_i is the desired activity for the neuron n_i.

3 Experiment

We have tested our architecture (Fig. 3) in an experiment to verify that it is capable of recognising and properly classifying observed actions, overcoming problems related with the action recognition task.

To this aim we created training and test sets for the architecture by choosing two movies from the INRIA 4D repository. In the movies, two different actors (Andreas and Hedlena) perform the same set of 13 actions. Each actor interprets and performs actions as individuals and thus they tend to differ slightly in how they perform the same actions. We chose to use one of the movies (performed by Andreas) to create a training set for the architecture and the other movie (performed by Hedlena) to create a test set. In this way, we wanted to demonstrate that our architecture is able not only to properly recognize action instances it has observed during training, but that it is also able to recognise the actions when they are performed by someone else, i.e. to recognise action instances it never encountered before. To create the training and test sets from the original movies, we split each of the original movie into 13 new movies, one for each action (see Fig. 1).

Before entering the architecture, the input goes through a preprocessing phase. This is done to reduce the computational load and improve architecture performances. In the preprocessing phase the number of images for each movie is reduced to 10 without affecting the quality of the action reproduction and guaranteeing seamless and fluid actions, see Fig. 4 a).

Consecutive images are then subtracted to catch only the dynamics of the action, focusing in this way the attention on the movement exclusively. This operation further reduced the number of frames for each movie to 9. As an example, we can see in Fig. 4 that in the *check watch* action only the arm is involved in the movement.

In the next step of the preprocessing phase, a fixed boundary box is used to cut the images and produce binary images of a fixed and small size while

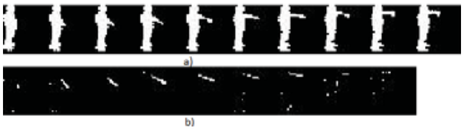


Fig. 4. a) The *check watch* action with a reduced number of images; b) The sequence of images obtained by subtracting consecutive images of the *check watch* action.

eliminating anything not significantly involved in the movement. In this way an attentive process, similar to how the human eye observes and follows only the salient parts of an action, is simulated. The binary images are then shrunk to 50×50 matrices and vectorized before entering the first layer SOM.

The architecture was trained in two phases. First the first layer SOM was trained for 20000 iterations by randomly selecting actions performed by the actor Andreas. Then the fully trained SOM of the first layer received each action performed by Andreas again and the corresponding sequence of activity matrices elicited by each action was superimposed and vectorized. Each such new superimposed activity vector represents the activity trajectory in the first layer SOM elicited by a particular action. More in detail, to superimpose the activity matrices, before the vectorization, can be seen as the creation of matrices, one for each action, with dimensions equal to the neuron grid of the first layer SOM. The value of the elements of these matrices are either zero or one. All elements corresponding to a neuron in the first layer SOM, which was most activated for at least one of the inputs during the action is set to one and all the other elements are set to zero.

In the second training phase the second layer SOM and the third layer neural network were trained. In this process the second layer SOM received randomly selected input from the set of superimposed activity vectors for 20000 iterations, and the third layer neural network received the corresponding target output (action labels). The target output consists of 13-dimensional vectors, with one element set to one and the other elements set to zero.

To show how the activity trajectories in the first layer SOM in the fully trained architecture differ we have depicted these for the actions carried out by the actor Andreas in Fig. 5. This was, for each action, done by recording the neuron in the first layer SOM most activated by each input in the sequence composing the action. The most activated neurons for each of the actions were then depicted and connected with arrows to show how the trajectories evolves over time. Each picture in Fig. 5 shows the grid of neurons forming the first layer SOM and illustrates the sequence of most activated neurons, represented by black dots, during the corresponding action. The black dots were connected with arrows to show how the trajectories evolve over time. The locations of the neurons activated most by the first and the last inputs of an action are represented by empty dots.

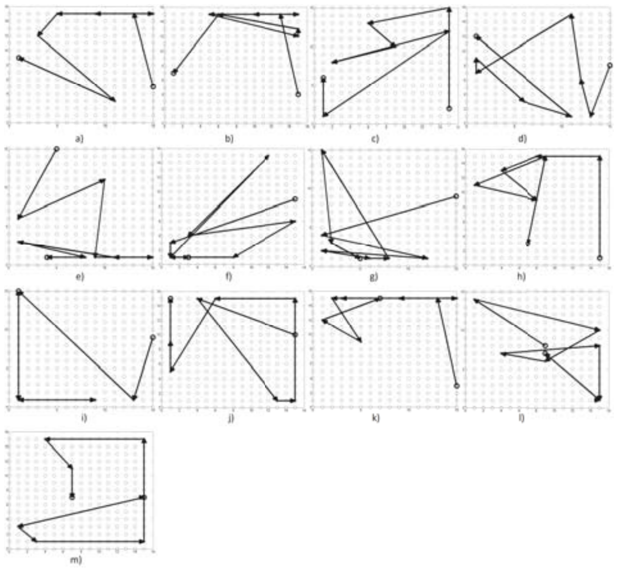


Fig. 5. Activity trajectories in the first layer SOM for the 13 actions carried out by Andreas: a) Check Watch; b) Cross Arms; c) Scratch Head; d) Sit Down; e) Get Up; f) Turn Around; g) Walk; h) Wave Hand; i) Punch; j) Kick; k) Point; l) Pick Up; m) Throw. The dots in each diagram represent the most activated neurons (centres of activity) during the action and the arrows indicate the action's evolution over time. Actions composed of similar postures present fewer centres of activity, whereas actions composed of postures with more different characteristics present more centres of activity. The diagrams indicate the ability of the SOM to create topology preserving maps in which similar postures are represented close to each other.

We tested the fully trained architecture with all 13 actions performed both by the actor Andreas (the action instances the architecture was trained on) and by the actor Hedlena (the action instances the architecture was not trained on). During testing, the input went through the preprocessing described above before entering the first layer SOM, which activity in turn were superimposed and vectorized as described above before entering the second layer SOM.

During the testing we recorded the most activated neuron in the third layer neural network to see if the actions were labelled correctly. This was done for both the actions carried out by Andreas as reported in Table 1 and by Hedlena as reported in Table 2. The architecture was able to recognise 100% of the actions performed by Andreas and 53% of the actions performed by the actor Hedlena, which the architecture was not trained on.

Andreas			
Actions	Most activated neuron	Expected neuron	Correctenss
Check Watch	0	0	correct
Cross Arm	1	1	correct
Scracth Head	2	2	correct
Sit Down	3	3	correct
Get Up	4	4	correct
Turn Around	5	5	correct
Walk	6	6	correct
Wave Hand	7	7	correct
Punch	8	8	correct
Kick	9	9	correct
Point	10	10	correct
Pick Up	11	11	correct
Throw	12	12	correct
%Correctness			100

Table 1. Recognition rate for the actions carried out by Andreas. Our architecture recognises 100% of the actions performed by Andreas, which the architecture was trained on.

4 Conclusion

We have proposed a novel hierarchical SOM based architecture that recognises actions. Our architecture is composed of three layers of neural networks. The first layer consists of a SOM that learns to represent the dynamics of sequences of postures composing actions. The second layer consists of another SOM, which learns to represent the activity trajectories in the first layer SOM, which also means that it learns to represent action prototypes. The third layer consists of a custom made supervised neural network that learns to label the action prototypes represented in the second layer SOM.

In an experiment we verified the architecture's ability to recognise observed actions as well as to recognise the same actions interpreted and performed by someone else.

As reported in Table 1 the actions used to train the architecture and performed by the actor Andreas were recognised to 100%. In Table 2 we can see

Hedlena				
Actions	Most activated neuron	Expected neuron	Place in order of activation of the expected neuron	Correctenss
Check Watch	12	0	4	
Cross Arm	1	1	1	correct
Scracth Head	2	2	1	correct
Sit Down	3	3	1	correct
Get Up	5	4	2	
Turn Around	5	5	1	correct
Walk	6	6	1	correct
Wave Hand	2	7	5	
Punch	8	8	1	correct
Kick	3	9	5	
Point	10	10	1	correct
Pick Up	2	11	11	
Throw	4	12	2	
%Correctness				53

Table 2. Recognition rate for the actions carried out by Hedlena. Our architecture recognises 53% of the actions performed by Hedlena, which the architecture was not trained on. In the cases of failed recognition the place in the order of activation of the expected neuron could be seen as the order of choice, i.e. if the place in the order of activation of the expected neuron is k, then the correct action would be the k:th most likely action according to the architecture.

that the actions interpreted and performed by another actor Hedlena, that the architecture was not trained on, were recognised to 53%. The values reported in the fourth column of Table 2 show that in some of the cases where recognition failed, the expected neuron, i.e. the neuron which if most activated would indicate the correct action, is still one of the most activated. For example, in the case of the action *Get Up*, which was incorrectly recognised as the action *Turn Around*, the architecture's second choice would have been the correct action *Get Up*.

An important observation is that some failed recognitions are plausible. Actions like *check watch*, *throw*, *wave hand* and *scratch head* can easily be confused even by a human observer. Consider, for example, the two actions *wave hand* and *scratch head*. The only part involved in the movement is the arm and the movement for both actions is the same, i.e. to raise the arm to the head. This could easily confuse the architecture to label both actions equally. The same reasoning can be applied to other actions that involve the movement of the same part of the body. Other considerations can be done for actions that involve movement of different parts of the body such as *kick* and *sit down*. In this case, the preprocessing operation such as subtraction of consecutive frames, gives rise to new sequences that sometimes can contain very similar frames, or frames that can be confused with each other, leading to a failed recognition of the observed action.

The promising experimental results show the potential of this hierarchical SOM based action recognition architecture. Potential future extensions include a more elaborate preprocessing procedure to enable a more potent view and size independence as well a explicit action segmentation.

Acknowledgements The authors gratefully acknowledge the support from the European Community's Seventh Framework Programme under grant agreement no: 612139.

References

1. Ahmad, M., Lee, S.W.: Human action recognition using shape and clg-motion flow from multi-view image sequences. Pattern Recognition **41**(7) (2008) 2237 – 2252
2. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104**(23) (2006) 249 – 257 Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.
3. Ahmad, M., Lee, S.W.: Hmm-based human action recognition using multiview image sequences. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Volume 1. (2006) 263–266
4. Buonamente, M., Dindo, H., Johnsson, M.: Recognizing actions with the associative self-organizing map. In: 2013 XXIV International Symposium on Information, Communication and Automation Technologies (ICAT), IEEE (2013) 1–5
5. Johnsson, M., Balkenius, C., Hesslow, G.: Associative self-organizing map. In: Proceedings of IJCCI. (2009) 363–370
6. Kohonen, T.: Self-Organization and Associative Memory. Springer Verlag (1988)
7. Buonamente, M., Dindo, H., Johnsson, M.: Simulating actions with the associative self-organizing map. In Lieto, A., Cruciani, M., eds.: AIC@AI*IA. Volume 1100 of CEUR Workshop Proceedings., CEUR-WS.org (2013) 13–24
8. Balkenius, C., Morén, J., Johansson, B., Johnsson, M.: Ikaros: Building cognitive models for robots. Advanced Engineering Informatics **24**(1) (2010) 40–48

Learning Graspability of Unknown Objects via Intrinsic Motivation

Ercin Temel[1], Beata J. Grzyb[2], and Sanem Sariel[1]

[1] Artificial Intelligence and Robotics Laboratory
Computer Engineering Department,
Istanbul Technical University, Istanbul, Turkey
{ercintemel,sariel}@itu.edu.tr
[2] Centre for Robotics and Neural Systems,
Plymouth University, Plymouth, United Kingdom
beata.grzyb@plymouth.ac.uk

Abstract. Interacting with unknown objects, and learning and producing effective grasping procedures in particular, are challenging problems for robots. This paper proposes an intrinsically motivated reinforcement learning mechanism for learning to grasp uknown objects. The mechanism uses frustration to determine when grasping of an object is not possible. The critical threshold of frustration is dynamically regulated by impulsiveness of the robot. Here, the artificial emotions regulate the learning rate according to the current task and performance of the robot. The proposed mechanism is tested in a real world scenario where the robot, using the grasp pairs generated in simulation, has to learn which objects are graspable. The results shows that the robot equipped with frustration and impulsiveness learns faster than the robot with standard action selection strategies providing some evidence that the use of artificial emotions can improve the learning time.

Keywords: Reinforcement Learning, Intrinsic motivation, Grasping unknown objects, Frustration, Impulsiveness, Visual scene representation, Vision-based grasping

1 Introduction

Robots need effective grasp procedures to interact with and manipulate unknown objects. In unstructured environments, challenges arise mainly due to uncertainties in sensing and control, and lack of prior knowledge and model of objects. Effective learning methods are essential to deal with these challenges. One classic approach here is to use reinforcement learning (RL) where an agent actively interacts with an environment and learns from the consequences of its actions, rather than from being explicitly taught. An agent selects its actions on basis of its past experiences (exploitation) and also by new choices (exploration). The goal of an agent is to maximize the global reward, therefore the agent needs to rely on actions that led to high rewards in the past. However, if the agent is too greedy and neglects exploration, it might never find the optimal strategy for the task. Hence, to find the best ways to perform an action they need to find a balance between exploitation of current knowledge and exploration to discover new knowledge that might lead to better performance in the future.

We propose a competence-based approach to reinforcement learning where exploration and exploitation is balanced while learning to grasp novel objects. In our approach, the dynamics of balancing between exploration and exploitation is tightly related to the level of frustration. The failures in obtaining a new goal may significantly increase the robot's level of frustration, and push it into searching new solutions in order to achieve its goal. However, a prolonged state of frustration, when no solution can been found, will lead to a state of learned helplessness, and the goal will be marked as unachievable at the current state (i.e., object not graspable). Simply speaking, an optimal level of frustration favours more explorative behaviour, whereas low or high level of frustration favours more exploitative behaviour. Additionally, we dynamically change the robot's impulsiveness that influences how fast the robot gets frustrated, and indirectly how much time it devotes to learning a particular task.

To demonstrate the advantages of our approach, we compare it with three other action selection methods: ε-greedy algorithm, softmax function with constant temperature parameter, softmax function with variable temperature depending on agent's overall frustration level. The results shows that the robot equipped with frustration and impulsiveness learns faster than the robot with standard action selection strategies providing some evidence that the use of artificial emotions can improve the learning time.

The rest of the paper is organized as follows. We first present related work in the area. Then, we give the details of the learning system including visual processing of objects, the RL framework and the proposed action selection strategies. In the next section, we present the experimental results and then conclude the paper.

2 Related Work

Our main focus is on learning graspability of objects. Previously, analytical methods are proposed for grasping objects [3], [8], [4]. These methods use contact point locations on objects and the gripper, and then find the friction coefficients by tactile sensors to compute force [15]. With these data, grasp stability values or promising grasp positions can be determined. Another approach for grasping is learning by exploration. In a recent work [6], grasp successes are associated with 3D object models which can lead algorithms to memorize object grasp coordination. According to their work, grasping unknown objects is a challenging problem and it varies in accordance with system complexity. This complexity depends on the chosen sensors, prior knowledge about environment and scene configuration. In [11], 2D contours are used for approximating the center of mass of objects for grasping.

In our work, we use reinforcement learning (RL) framework for learning and incorporate competence-based intrinsic motivation for guidance in search. The complexity of reinforcement learning is high in terms of the number of state-action pairs and the computations needed to determine utility values [14]. Approximate policy iteration methods can be used to alleviate this problem based on sampling [7]. Imitation learning before reinforcement learning [12] is one of the methods for decreasing the complexity in RL [5]. Furthermore, it is also used for robots learn crucial parameters in movement to accomplish the task.

In our work, we use a competence-based approach for intrinsic motivation for balancing exploration in RL. Frustration level of the robot is taken into account. We further extend this approach by adopting an adaptive frustration level depending on a task. Intrinsic motivation is investigated in earlier works. Lenat [13] propose a system considering "interestingness" and Schmidhuber introduce curiosity concept for reinforcement learning [19]. Uchibe and Doya [22] also consider intrinsic motivation as learning objective. Different from curiosity and reward functions, Wong [24] point out that ideal level of frustration is beneficial for exploration and faster learning. In addition, Baranes and Oudeyer [1] propose competence-based intrinsic motivation for learning. In our work, main difference is that impulsiveness [20] is adapted into the frustration rate in order to change the learning rate dynamically based on a task in real world environment for robots.

3 Learning to Grasp Unknown Objects

We propose an intrinsically motivated reinforcement learning system for robots to learn graspability of unknown objects. The system includes two main phases for determination of grasp points on objects and experimentation of them in the real world (Fig. 1). The first phase includes the required methods to determine candidate grasp point pairs in simulation. Note that a robot arm with a two-fingered end effector is selected as the target platform. For this reason, grasp points are determined as point pairs. In the second phase of the system, the grasp points determined in the first phase are experimented in the real world through reinforcement learning. The following subsections explain the details of these processes.

3.1 Visual Representation of Objects

In our system, objects are detected in the scene by using an ASUS Xtion Pro Live RGB-D camera mounted on a linear platform for interpreting the scene for tabletop manipulation scenarios by a robotic arm. We use a scene interpretation system that can both recognize known objects and detect unknown objects in the scene [9]. For unknown object detection, Organized Point Cloud Segmentation with Connected Components algorithm [21] from PCL [16] is used. This algorithm finds and marks connected pixels coming from the RGB-D camera

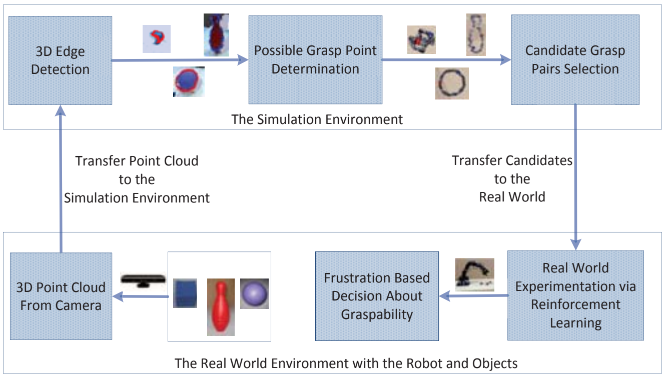

3D Edge Detection
Possible Grasp Point Determination
Candidate Grasp Pairs Selection
The Simulation Environment
Transfer Point Cloud to the Simulation Environment
Transfer Candidates to the Real World
3D Point Cloud From Camera
Frustration Based Decision About Graspability
Real World Experimentation via Reinforcement Learning
The Real World Environment with the Robot and Objects

Fig. 1. Overview of the intrinsically motivated reinforcement learning system.

and finds the outlier 3D edges by RANdom SAmple Consensus (RANSAC) algorithm [17]. Hence, the object's center of mass and its edges are detected to be used by the grasp point detection algorithm that finds candidate grasp point pairs for a two-fingered robotic hand.

3.2 Detection of Candidate Grasp Points in the Simulator

Objects are represented by their center of masses (μ) and 3D edges (H). Then candidate grasp point pairs ($\rho = [p_1, p_2]$) are determined as in Algorithm 1. In the algorithm, initially the reference points are determined. The center of mass, the upside and the bottom side center points are chosen as references. Based on these points, cross section points coplanar with the reference points and parallel to the table surface are determined. In the next step, the algorithm detects the closest point to the reference points on the same planar and draw a line crossing with reference points and closest to it. The second step is determining the opposite point to the closest one on the same line. This procedure continues until all points are tested. The algorithm produces the candidate grasp pairs (two grasp points with x,y,z values) and orientation of each pair according to (0,0) point in 2D (x,y) plane. These grasp points are tested in the simulator for finding out only the feasible ones.

In Fig. 2, the edges and sample grasp points for six different objects along with the number of grasp points are presented.

Algorithm 1 Grasp Point Detection (μ, H)

```
Input: Object Center Of Mass $\mu$, Edge Point Cloud $H$
Output: Grasp Pairs $P$
Detect $maxZ$, $minZ$ and $C$ as reference point $ref$.
for each reference point do
  $cPoints$ = findPointsOnTheSamePlane()
  $mPoint$ = findClosestPointToReferencePoint($cPoints$)
  $slope$ =findSlope($mPoint$,$ref$)
  for each $p \in cPoints$ do
    $Pslope$ =findSlope($mPoint$,$p$)
    if onTheSameLine($Pslope$,$slope$) then
      $P \leftarrow \{ p, mPoint \}$
    end if
  end for
end for
```

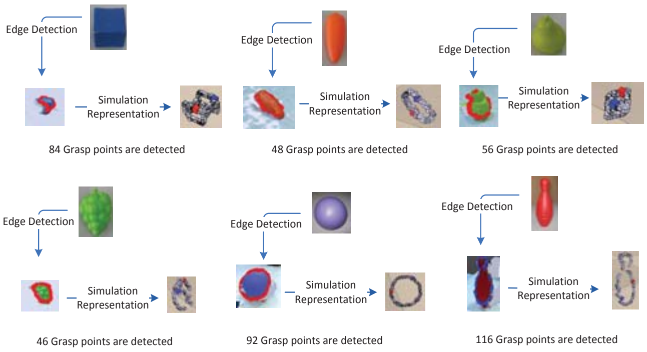

Edge Detection
Simulation Representation
84 Grasp points are detected
Edge Detection
Simulation Representation
48 Grasp points are detected
Edge Detection
Simulation Representation
56 Grasp points are detected
Edge Detection
Simulation Representation
46 Grasp points are detected
Edge Detection
Simulation Representation
92 Grasp points are detected
Edge Detection
Simulation Representation
116 Grasp points are detected

Fig. 2. Candidate grasp points on unknown objects are determined through a sequence of processes. Samples points for six objects are illustrated. The first step is 3D edge detection from 3D point cloud data. The second step is determination of candidate grasp point pairs for which samples are marked with red points on the 3D edges extracted from point clouds of objects. The number of feasible grasp points for each object is presented.

3.3 Learning When to Give Up Grasping

In the system, the output of the simulation environment is fed to the robotic arm to apply real-world experimentation. Intrinsic motivation with frustration level and new proposed impulsiveness method are evaluated to increase the learning process speed for the robot in order to give up quickly for the objects that are not graspable.

The main task of the robot is to learn which objects are graspable. We use a Reinforcement Learning (RL) framework with Q-learning [23] algorithm and softmax action selection [2] strategy. The state space here are all grasp point pairs generated during the simulation phase. A general state S is defined as:

$$S = [\mu, \rho, \phi, \omega, \boldsymbol{O_v}] \tag{1}$$

where, μ is the center of mass of the object, ρ is the selected set of two grasp points ρ =$[p_1, p_2]$, ϕ is the grasp orientation, ω is the approach direction of the gripper and $\boldsymbol{O_v}$ is the 3D translation vector for object during grasp trial. A collision between the robotic arm and the object may occur when there is a trajectory error that results in a non-zero vector.

Actions can be represented as follows,

$$A = [||\boldsymbol{R_v}||, \omega] \tag{2}$$

where, $||\boldsymbol{R_v}||$ is the slide amount on the x axis and ω represents the approach vector to the object of interest.

In our framework, the robot receives the reward value of 10 (R_{max}) when the grasp is successful and 0.1 (R_{min}) when the grasp is unsuccessful [10]. The Q-values are updated according to Eq. 3.

$$Q'(s,a) = Q(s,a) + \alpha * [R + (\gamma * maxQ(s',a)) - Q(s,a)] \tag{3}$$

where, $Q'(s,a)$ the next Q-value for state action pair (s,a), $Q'(s,a)$ is the current Q-value, α is the learning rate, R is the immediate reward after performing an action a in state s, γ is the discount factor, $maxQ(s',a)$ is the maximum estimate of optimal future value.

We investigate four action selection strategies. The first (and the simplest) one is the ε-greedy action selection method (M_1). This method most of the time selects the action with the highest estimated action value, but once in a while (with a small probability ε), selects an action at random, uniformly, independently of the action-value estimates.

The second one is the SoftMax Action selection (Eq. 4) method (M_2)with constant temperature value [2]:

$$P(a)_t = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \tag{4}$$

where, $P(a)_t$ is the probability of selecting an action a at the time step t, $Q_t(a)$ is the value function for an action a, and τ is the positive parameter called the temperature that controls the stochasticity of a decision. A high value of the temperature will cause the actions to be almost equiprobable and a low value will cause a greater difference in selection probability for actions that differ in their value estimates.

The third strategy (M_3) also uses the Softmax action selection rule. In this approach, however, the τ parameter is flexible and changes dynamically in relation to the robot's level of frustration and sense of control [10]. An optimal level of frustration favours more explorative behavior, whereas low or high level of frustration leads to a more exploitative behavior. For the purpose of our simulations, frustration was represented as a simple leaky integrator:

$$\frac{df}{dt} = -L * f + A_0 \tag{5}$$

where, f is the current level of frustration, A_0 is the outcome of the action ($success$ or $failure$) and L is the fixed rate of the 'Leak'.

In Eq. 5 the 'leak' rate (L) was fixed and kept at value 1 for all simulations [10]. Higher values of L cause the frustration rate to increase slower compared to smaller values of L. That means that the robot with a high value of L spends more time on exploration and possibly learns faster. Hence, we propose the forth method (M_4) that builds on this method and changes the value of L dynamically using an expected utilization motivation formula [20]:

$$L = \frac{\mathit{expectancy} * \mathit{value}}{Z + \mathit{\Gamma}(T - t)} \tag{6}$$

where *expectancy* represents the probability of getting the highest estimated action value (as in the greedy action selection method), *value* refers to the expected action reward (here $\mathit{value} = R_{max}$), Z is a constant derived from when rewards are immediate, $\mathit{\Gamma}$ indicates agent's sensitivity to delay (impulsiveness) and $(T - t)$ refers to the delay of the reward in terms of "time reward" minus "time now".

The impulsiveness is main focus of ours to develop interaction with frustration rate competence based motivation. According to triad "Frustration - Impulse - Temper", a person who has high impulsiveness is considered as "short tempered" and it means quickly get frustrated so that changes on frustration level for learning behavior. Our proposal with that, different values on impulsiveness directly affect rate of leak, L, on frustration formula so frustration rate of agent also will be dependent on impulsiveness.

The robot apart from learning how to grasp an object, also needs to learn whether the target object is graspable or not. The learning of a selected grasp pair ρ and action a finishes when overall frustration level becomes equal or greater than a certain threshold value. This value is determined based on a tolerance formula:

$$Tolerance = e^{-||\boldsymbol{O_v}|| * \varphi} \tag{7}$$

where, $||\boldsymbol{O_v}||$ denotes the translation of the object on the table because of the collision with the end effector and φ the number of trials from the beginning of learning.

Additionally, the online learning process may also end when the following criterium has been met:

$$FrustrationLimit = e^{1/\sqrt{n}} \tag{8}$$

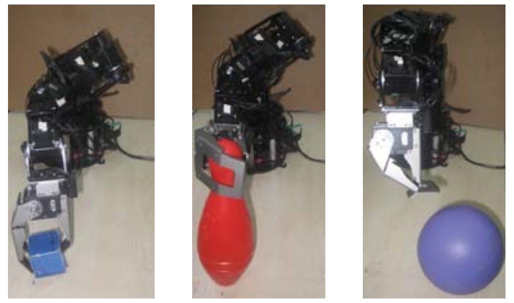
where, n refers to the number of grasp pairs.

3.4 Impulsiveness and Learning Rate

The main focus of the presented work is investigating an effect that impulsiveness has on frustration level and on learning. The learning rate and the speed of decision making is an important issue in human-robot interaction [18]. For example, when a robot plays a quick game with a human, it has to learn quickly. However, when the robot is alone, it can spend relatively more time on exploring different states. By changing the impulsiveness, the robot may dynamically control its level of frustration and therefore the time devoted for learning a particular task. Hence, the robot could behave differently in different environments and for different tasks.

4 Experimental Results

As mentioned before, the candidate grasping points are first determined in simulation, and then transferred to a robotic arm for real-word experimentation. V-REP simulator is used as the simulator and the Cyton-Veta Robotic 7-DOF robot arm by Robai (shown in Fig 3) is used as the experimental platform. The reachability of the arm is about 45 cm. Also in the experiments, we used three objects of different size and shape (i.e., a small cubic plastic block, a plastic bowling pin and a spherical plastic ball). We compare the performance of four different action selection methods discussed in the previous section. A high value



(a) Success of lengthwise grasp on the block.

(b) Success of transverse grasp on the pin.

(c) Failure of lengthwise grasp on the ball.

Fig. 3. Illustrative examples for grasping three different objects. (a)A cubic block which is relatively easy to grasp (b) a plastic bowling pin which can be grasped from top but not for all prasp points (c) a plastic ball which cannot be grasped as it is solid and too large.

of impulsiveness results in a faster increase in a frustration level (in other words, in a “short-tempered” agent). For comparison reasons, we use here two different values of impulsiveness: a low value of 0.01 and a high value of 100. The results of our experiments support our proposed hypothesis. An agent with low impulsiveness spends more time on exploration, testing more grasp pair possibilities than an agent with a higher value of impulsiveness. For demonstration purposes, we chose three different objects that vary in their graspability properties: a cube that is relatively easy to grasp, a plastic bowling pin that is easily graspable but it is liable of toppling down, and finally, a ball that is not graspable at all. We compare the decision and learning rate of the robot that uses our proposed strategy (M_4) with the one based only on frustration (M_3). Fig. 4 shows robot’s level of frustration for each learning epoch while the robot was learning how to grasp the block. The 84 possible grasp pairs generated in simulation were used in a real world scenario. Since the robot can easily grasp the cube, the frustration level is kept low and the learning process terminates before it reaches its limit value, 1.115 (i.e., according to Eq. 8.). In case of the pin (see Fig. 5), the simulation generated 116 possible grasp pair candidates that were subsequently used by the robotic arm. Since the pin is quite light, the arm pulls it down for some grasp pairs. When the pin fells down, the frustration threshold is decreased for the related grasp pairs according to the Eq. 7. Hence, the robot learns that these grasp pairs should be eliminated from the set and immediately proceeds to test another grasp pair. While for some grasp pairs grasping of the pin was possible, the robot was not able to grasp the ball for any of grasp pairs. The ball was made of a hard plastic material and quite light, so every robot’s attempt to grasp it resulted in a ball rolling over on the scene Fig. 3(c). After each trial, the robot’s tolerance for frustration decreased rapidly resulting in that the robot switches to another grasp pair. With each failure, the overall frustration level was raising and quickly exceeded the tolerance threshold (that at the same time was being decreased). Although 92 grasp pairs were transferred to the real world scenario, only after a few steps the robot learned that the object is not graspable. Fig.

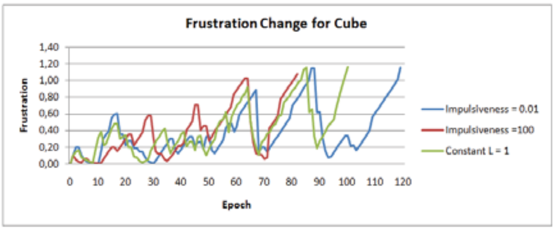


Fig. 4. Frustration Rate Changes For Block Grasping with Methods M_3 and M_4.

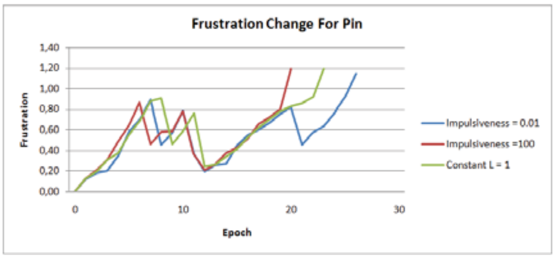


Fig. 5. Frustration Rate Changes For Pin Grasping with Methods M_3 and M_4.

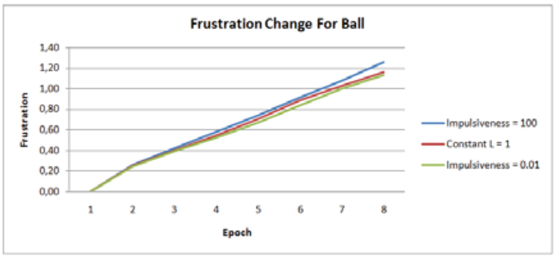


Fig. 6. Frustration Rate Changes For Ball Grasping with Methods M_3 and M_4.

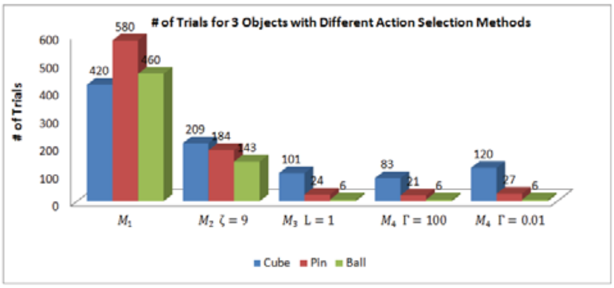


Fig. 7. Trial Count for Three Selected Object and Action Selection Method.

7 shows the comparison of the results for all four strategies of action selection. The frustration-based action selection methods require a lower number of trials

to learn the graspability of the objects compared to the standard softmax action selection with fixed temperature parameter and ε-greedy action selection. The agent with higher value of impulsiveness performs slightly better than the agent with low value.

5 Conclusion

We have presented our intrinsically motivated reinforcement learning system for learning graspability of novel objects. Intrinsic motivation is provided by frustration-based action selection methods during learning, and tolerance values are determined based on impulsiveness of the robot. Our claim is that impulsiveness can be adjusted based on the task that the robot is executing. We have analyzed this mechanism on a robotic arm to learn graspability of different-shaped objects. Our results reveal that the intrinsic motivation helps the robot learn faster. Furthermore, the decision on graspability is made earlier by taking impulsiveness into account. Our future work includes extending the experiment set and investigating impulsiveness parameters in detail for different domains with varying time constraints.

Acknowledgment

This research is funded by a grant from the Scientific and Technological Research Council of Turkey (TUBITAK), Grant No. 111E-286. TUBITAK's support is gratefully acknowledged. We thank Burak Topal for his contribution for robotic arm movement and also thank Mehmet Biberci for his effort on vision algorithms.

References

1. Baranes, A., Oudeyer, P.Y.: Maturationally-constrained competence-based intrinsically motivated learning. In: Development and Learning (ICDL), 2010 IEEE 9th International Conference on. pp. 197–203. IEEE (2010)
2. Barto, A.G.: Reinforcement learning: An introduction. MIT press (1998)
3. Bicchi, A.: On the closure properties of robotic grasping. The International Journal of Robotics Research 14(4), 319–334 (1995)
4. Buss, M., Hashimoto, H., Moore, J.B.: Dextrous hand grasping force optimization. Robotics and Automation, IEEE Transactions on 12(3), 406–418 (1996)
5. Chebotar, Y., Kroemer, O., Peters, J.: Learning robot tactile sensing for object manipulation
6. Detry, R., Baseski, E., Popovic, M., Touati, Y., Kruger, N., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on. pp. 1–7. IEEE (2009)
7. Dimitrakakis, C., Lagoudakis, M.G.: Rollout sampling approximate policy iteration. Machine Learning 72(3), 157–171 (2008)
8. Ding, D., Liu, Y.H., Wang, S.: The synthesis of 3-d form-closure grasps. Robotica 18(01), 51–58 (2000)

9. Ersen, M., Ozturk, M.D., Biberci, M., Sariel, S., Yalcin, H.: Scene interpretation for lifelong robot learning. In: The 9th International Workshop on Cognitive Robotics (CogRob 2014) held in conjunction with ECAI-2014. Prague, Czech Republic (2014)
10. Grzyb, B., Boedecker, J., Asada, M., Del Pobil, A.P., Smith, L.B.: Between frustration and elation: Sense of control regulates the lntrinsic motivation for motor learning. In: Lifelong Learning (2011)
11. Huebner, K., Ruthotto, S., Kragic, D.: Minimum volume bounding box decomposition for shape approximation in robot grasping. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on. pp. 1628–1633. IEEE (2008)
12. Kober, J., Peters, J.: Learning motor primitives for robotics. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. pp. 2112–2118. IEEE (2009)
13. Lenat, D.B.: Am: An artificial intelligence approach to discovery in mathematics as heuristic search. Tech. rep., DTIC Document (1976)
14. Peters: Machine learning of motor skills for robotics (2007)
15. Platt, R.: Learning grasp strategies composed of contact relative motions. In: Humanoid Robots, 2007 7th IEEE-RAS International Conference on. pp. 49–56. IEEE (2007)
16. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China (May 9-13 2011)
17. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: Robotics and Automation (ICRA), 2011 IEEE International Conference on. pp. 1–4. IEEE (2011)
18. Sauser, E.L., Billard, A.G.: Biologically inspired multimodal integration: Interferences in a human-robot interaction game. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. pp. 5619–5624. IEEE (2006)
19. iirgen Schmidhuber, J.: A possibility for lmplementing curiosity and boredom in model-building neural controllers (1991)
20. Steel, P., König, C.J.: Integrating theories of motivation. Academy of Management Review 31(4), 889–913 (2006)
21. Trevor, A., Gedikli, S., Rusu, R., Christensen, H.: Efficient organized point cloud segmentation with connected components. Semantic Perception Mapping and Exploration (SPME) (2013)
22. Uchibe, E., Doya, K.: Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. Neural Networks 21(10), 1447–1455 (2008)
23. Watkins, C.J., Dayan, P.: Q-learning. Machine learning 8(3-4), 279–292 (1992)
24. Wong, P.T.: Frustration, exploration, and learning. Canadian Psychological Review/Psychologie canadienne 20(3), 133 (1979)

On the cognitive and logical role of image schemas in computational conceptual blending

Maria Hedblom, Oliver Kutz, and Fabian Neuhaus

Institute of Knowledge and Language Engineering
Otto-von-Guericke University of Magdeburg, Germany

Abstract. In cognitive science, image schemas are identified as the fundamental patterns for the cognition of objects, which are perceived, conceptualised and manipulated in space and time. In this paper, we discuss a role for image schemas in computational concept creation. We propose to build a library of formalised image schemas, and illustrate how they can guide the search for a base space in the concept invention workflow.

1 Introduction

The notion that human cognition should guide the advancement of AI is as old as computer science itself [35, 39]. In this paper we apply this idea to computational creativity, in particular to computational concept invention.

In cognitive science, *image schemas* are identified as the fundamental patterns for the cognition of objects, which are perceived, conceptualised and manipulated in space and time [25]. Further, *conceptual blending* is considered as the cognitive engine for generating novel concepts [36]. In this paper we investigate how these two theories can be utilised in the context of computational concept invention and creativity [34, 17].

Within the European FP7 project COINVENT [34], a major effort is currently underway trying to fill the gap between the solid evidence from cognitive psychology and linguistics for the importance of the ideas of conceptual blending and image schema, and the lack of a computational and formal theory. The computational realisation of conceptual blending here is grounded on the basic formalisation ideas of Joseph Goguen [6]. In this paper, we address a particular piece of the puzzle to put together the various components of such a concept invention platform, namely to study the cognitive and logical role of image schema in concept invention.

The paper is structured as follows: first we introduce the notion of image schema and the basics of conceptual blending theory. This is followed by a discussion on how conceptual blending can be computationally modelled and implemented. As we will see, one critical step in the computational model for blending is the identification of shared structure across different domains. This is where image schemas can play a critical role to reduce the potential search space. We finish the paper with an extended example and a discussion of future work.

2 Image schema

Embodied theories of cognition [1] emphasise bodily experiences as the prime source for concept formation about of the world. Based on this view, the theory of image schemas suggests the perceptive spatial relationships between objects to constitute the foundation of our conceptual world. Typical examples of image schemas are SUPPORT[1], CONTAINMENT, LINK, and SOURCE_PATH_GOAL.

Both embodied theories and the image schema theory have support from both neuroscience [32], developmental psychology [23], and linguistic research in which image schemas can be observed in language development [22] and in the use of metaphoric information transfer and abstract thought [12].

As research on image schema is performed in several disciplines there is some incoherence on the terminology surrounding image schema, and the relationship between socio-cultural aspects and the neurobiology of embodied cognition is heavily disputed [9]. In order to proceed with our findings we follow the definition introduced by Mark Johnson [12], one of the founding theorists:

> An image schema is a recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience. [p. xiv]

We follow Johnson's footsteps and the further specialisations made by Kuhn [15] according to which image schemas are pre-linguistic structures of object relations in time and space.[2]

We also take into account the attempt of a hierarchical structuring of these phenomena as recently presented by Mandler and Pagán Cánovas [25] in which image schemas are explained as "simple spatial stories" using certain spatial primitives. We therefore build our approach from the view that image schemas are the abstract cognitive patterns that are obtained after repeated perceptual experience.

As an infant experiences similar perceptual events repeatedly – e.g., plates and other objects being placed on a table – an image schema is learnt based on this particular stimulation. This image schema represents the relationships between the objects in the event; in the mentioned example the image schema of SUPPORT is learnt.

Another basic example of an image schema is the notion of CONTAINMENT. This involves the understanding that an object can be within a border, or inside a container, including the events of entering and exiting. The CONTAINMENT schema is one of the most investigated image schemas [11] as it is one of the very first to be developed [23]. Perhaps unsurprisingly, this results in a complex relationship between spatial situations, learnt spatial concepts, and a corresponding use of natural language. Bennett and Cialone [2], in this connection, distinguish eight different spatial relationships and their mappings to natural language constructs, illustrated in Figure 1.

[1] All image schema concepts are printed in upper case letters.

[2] In particular, in [15] image schemas are hypothesised to capture the needed abstractions to model affordances related to spatio-temporal processes.

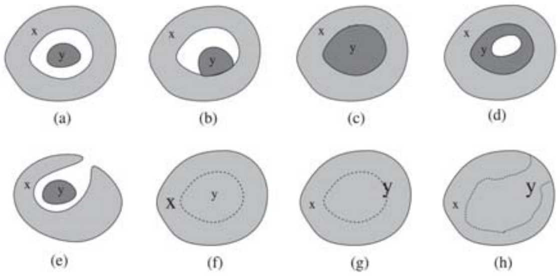


Fig. 1. Eight variations of containment as discussed in Bennett and Cialone [2].

When an image schema has formed, it can be generalised upon and can be transferred through analogical reasoning into other domains with similar characteristics in which the relationship is not yet known [23].

Following the cognitive development of the image schema of Containment, it would seem that it is the movement in and out of containers that inspires the learning of this particular structure [24]. One explanation may be that moving objects hold an increased perceptual value and that the surprise of objects disappearing in a container might trigger the mind to fast build theories in order to explain the feeling of surprise.

It is thought that image schemas develop systematically from perception and become more fine-tuned as the child is exposed to more experience of the same, or similar, relations. Mandler and Pagán Cánovas [25] made a hierarchical division of the umbrella term image schema into *spatial primitives*, *image schemas* and *conceptual integrations*. This follows the psychological research on the development of pre-linguistic concept formation. Spatial primitives are defined as the basic spatial relationships such as Path and Link. Image schemas are the spatial stories that can be built from these spatial primitives, and conceptual integrations are combinations of either spatial primitives or image schemas combined with non-spatial elements such as emotion or force. This is particularly interesting for research attempting to combine image schemas with conceptual blending, discussed below in more detail. It suggests that the operation of conceptual blending is already part of the most fundamental conceptualisation: the formation of complex image schemas.

A core idea is that image schemas provide a 'cognitive benefit' in information transfer. That is, an image schema structure may be used as a shortcut utilised in an analogical transfer from the spatial domain of the image schema to more abstract concepts, including concepts involving force, time and emotions. Traces of this can often be viewed in how language is used to explain concepts such as affection; we say that we are *in* love using the Containment schema, marriage can be explained with a LINK combined with a temporal PATH, and much of our metaphorical language is based on sensory-motor experiences.

The basic conceptual structures that image schemas provide for language acquisition and cognitive development are not only an important topic in spatial semantics and developmental psychology. A formalisation of image schemas could become a valuable asset and powerful tool for computational concept generation, as has been stressed by [14, 26, 6, 17]. A more systematic formalisation of image schemas could be used to aid computational creativity by supporting the generation of novel concepts following the conceptual blending approach, as outlined in more detail below.

3 Conceptual blending

The theory of *Conceptual Blending* was introduced during the 1990s as the cognitive machinery that helps us generate novel concepts, cf. e.g. Fauconnier and Turner's [3]. The theory has strong support from the cognitive psychology and linguistics domains [13, 8, 40] as well as in more computational areas [38] in which conceptual blending often is used to explain creativity and approach concept generation.

A central idea in conceptual blending theory is that the generation of novel concepts may happen via the combination of already existing ideas and knowledge. It is furthermore suggested that such novel concepts are selective and 'compressed' combinations, or blends, of previously formed concepts. This cognitive process is thought to happen as two, or more, input domains (or information sources) are combined into a new domain, the blended domain, see figure 2. The blend here inherits some of the attributes and relationships from the source domains and at the same time the unique mix allows the blends to have emergent properties that are unique to each particular blend.

Veale [38] captures the nature of conceptual blending as follows:

> "...conceptual blending combines the smoothness of metaphor with the structural complexity and organizing power of analogy. We can think of blending as a cognitive operation in which conceptual ingredients do not flow in a single direction, but are thoroughly stirred together, to create a new structure with its own emergent meanings." (p. 1)

As Veale points out, conceptual blending differs from analogical transfer in the following way: in analogical transfer information flows from a source domain to a target domain. In contrast, in conceptual blending knowledge is transferred from two source domains to a third, newly created blended space. However, similarly to the search for common structure in the source and target domain in analogy, conceptual blending looks for structural pattern that can be found in both of the input domains; these shared structural patterns – the so-called base, or generic space – are identified and provide the core for the blended conceptual space.

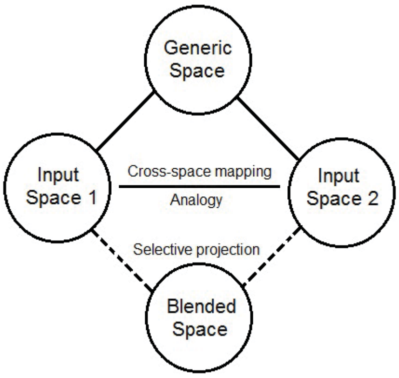


Fig. 2. The blending process as described by Fauconnier and Turner [3].

4 Formalising conceptual blending

Goguen defines an approach that he terms *algebraic semiotics* in which certain structural aspects of semiotic systems are logically formalised in terms of algebraic theories, sign systems, and their mappings in [4].

In [6] algebraic semiotics has been applied to user interface design and conceptual blending. Algebraic semiotics does not claim to provide a comprehensive formal theory of blending – indeed, Goguen and Harrell admit that many aspects of blending, in particular concerning the meaning of the involved notions, as well as the optimality principles for blending, cannot be captured formally. However, the structural aspects *can* be formalised and provide insights into the space of possible blends. The formalisation of these blends can be formulated using languages from the area of algebraic specification, e.g. OBJ3 [7].

In [10, 18, 20], we have presented an approach to computational conceptual blending, which is in the tradition of Goguen's proposal. In these earlier papers, we suggested to represent the input spaces as ontologies (e.g., in the OWL Web Ontology Language[3]). The structure that is shared across the input spaces is also represented as an ontology, which is linked by mappings to the input spaces. As proposed by Goguen, the blending process is modelled by a colimit computation, a construction that abstracts the operation of disjoint unions modulo the identification of certain parts specified by the base and the interpretations, as discussed in detail in [5, 19, 18].

We moreover presented how the Distributed Ontology Language (DOL) can be used to specify conceptual blends with the help of *blending diagrams*. These diagrams encode the relationships between the base space and the (two or more)

[3] With 'OWL' we refer to OWL 2 DL, see http://www.w3.org/TR/owl2-overview/

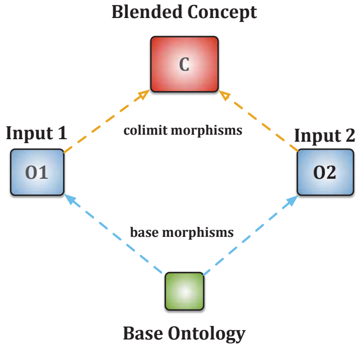


Fig. 3. The blending process as described by Goguen [6].

input spaces. These *blending diagrams* can be executed by Hets, a proof management system. Hets is integrated into Ontohub,[4] an ontology repository which allows users to manage and collaboratively work on ontologies. DOL, Hets, and Ontohub provide a powerful set of tools, which make it easy to specify and computationally execute conceptual blends, as seen in [29].

A critical step in the blending process is the identification of the base space and its mapping to the input spaces. One approach to computationally implement this step consists of applying techniques of finding generalisations of two input spaces, which have already been pursuit by analogy-making engines such as Heuristic Driven Theory Projection, HDTP [33]. HDTP computes a common generalisation B of two input spaces $O1$ and $O2$. This is done by anti-unification to find common structures in both input spaces $O1$ and $O2$. HDTP's algorithm for anti-unification is, analogously to unification, a purely syntactical approach that is based on finding matching substitutions.[5]

While this is an interesting approach, it has a major disadvantage. Typically, for any two ontologies there exists a large number of potential generalisations. Thus, the search space for potential base spaces and, therefore, potential conceptual blends is vast. HDTP implements heuristics to identify interesting anti-unifiers; e.g., it prefers anti-unifiers that contain rich theories over anti-unifiers that contain weak theories. However, since anti-unification is a purely syntactical approach, there is no way to distinguish cognitively relevant from irrelevant information. As a result, the combinatorial possibilities for anti-unification of axioms in the two input ontologies explodes.

[4] www.ontohub.org

[5] There are several other methods for finding generalisations. One example is the Analogical Thesaurus [37] which uses WordNet to identify common categories for the source and target spaces.

5 Blending with image schemas

Instead of relying on a purely syntactical approach as was illustrated in the example above using HDTP, we propose to guide the search for base spaces by a library of formalised image schemas.

Here, a (formalisation of) an image schema is searched for within two input theories $O1$ and $O2$ by a simultaneous theory-interpretation search. Computational support for this operation has already been investigated in [30], and a prototypical system has been developed that was tested as an add-on to the *Heterogeneous Tool Set* Hets [28]. Experiments carried out in [31, 21] showed that this works particularly well with more complex axiomatisations in first-order logic, rather than with simple taxonomies expressed in OWL, for the simple reason that in the latter cases there is simply too little structure to control the combinatorial explosion of such a search task. From the point of view of embedding image schemas into non-trivial concepts, we may see this as an encouraging fact, as image schemas are, despite their foundational nature, complex objects to axiomatise.

We now discuss in more detail an example for concept invention where an image schema plays an essential role in the construction of the newly blended concept. Consider the two concepts "space ship" and "mother". Both are associated with a multitude of concepts. Space ships travel through space, they visit space stations, and they are used to move cargo. Mothers give birth, they provide guidance for their children and have authority over them. There are many ways how these concepts can be blended. E.g., one blend would be a space ship that provides guidance and has authority over other, smaller ships – in other words, a flag ship. For other potential blends it is less obvious whether they would be useful; e.g., the concept of a mother that travels trough space.

Our thesis is that shared image schemas provide a useful heuristic to identify interesting blends.

To capture these ideas formally we first need to represent Containment in some formal language. For the sake of illustrating the basic ideas, we choose here a simplified representation in OWL (see Fig. 4). Containers are defined as material objects that have a cavity as a proper part. A container contains an object if and only if the object is located in the cavity that is part of the container.

```
Class: Container
        SubClassOf: MaterialObject
        EquivalentTo: has_capability ContainerCapability
        EquivalentTo: has_proper_part Cavity

ObjectProperty: contains
        SubPropertyChain: has_proper_part o is_location_of
        DisjointWith: has_proper_part
```

Fig. 4. A (partial) representation of Containment in OWL

Mothers realise the Containment schema, since before birth their children are contained within their wombs. Similarly, ships realise Containment since they may be used to transport goods and passengers. Of course, in almost any other aspect mothers and ships are completely different; in Fig. 5 we only represent that mothers are female humans with children and that space ships are capable of space travel.

```
Class: Mother
        EquivalentTo: Female and Human and parent_of some (Small and Human)
        SubClassOf: has_proper_part UterineCavity

Class: SpaceShip
        EquivalentTo: Vessel and has_capability some SpaceTravel
        SubClassOf: has_proper_part some CargoSpace
```

Fig. 5. Mothers and space ships

During the blending of "Mother" and "Ship" the Containment schema structure of both input spaces is preserved, forming the concept of "Mother ship" (see Fig. 6). In this case, the uterine cavity and the cargo space are mapped to the docking space. This concept inherits some features from both input spaces, while others are dropped. Obviously, a mother ship is a space travelling vessel. But like a mother, it is a 'parent' to some smaller entities of the same type. These smaller vessels can be contained within the mother ship, they may leave its hull (a process analogous to a birth) and are supported and under the authority of the larger vessel.[6]

```
Class: MotherShip
    EquivalentTo: Vessel and has_capability some SpaceTravel
        SubClassOf: has_proper_part DockingStation
        SubClassOf: has_proper_part some CargoSpace
        SubClassOf: parent_of some (Small and Vessel)
```

Fig. 6. Mother ship

To summarise, in our example we try to blend the input spaces of "Mother" and "Space ship". Instead of trying to utilise a syntactic approach like anti-unification to search for a base space, we recognise that both input spaces have cavities and, thus, are containers. Using the base space Containment in the blending process yields a blended concept of "Mother ship". Here, the precise mappings from the base space axiomatisation of Containment to the two input

[6] To represent dynamic aspects like birth and vessels leaving a docking bay adequately, one needs a more expressive language than OWL.

spaces regulate the various properties of the blended concept. Fig. 7 illustrates this blend by populating the generic blending schema shown in Fig. 3.

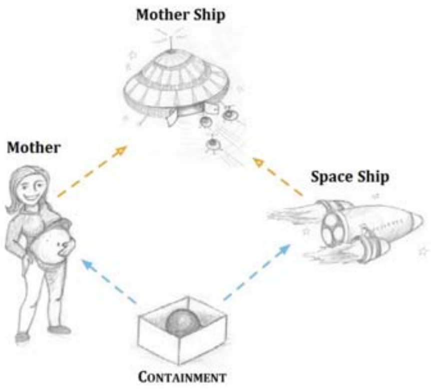


Fig. 7. The blending of mother ship

6 Outlook

The work on systematically formalising and ontologically structuring image schemas is largely unexplored ground. Our idea of using the cognitive structure of image schemas as the driving force behind the creation of the base space and the mappings in computational conceptual blending has yet to be fully explored, but similar work can be seen in analogy engines like HDTP.

Although several scattered formalisation attempts of image schemas may be found in the literature on conceptual blending and common sense reasoning [26, 14, 6], these attempts are directed at particular blends or common sense problems, without much systematicity. The most looked at image schema, by far, is the notion of Containment. Here, the work of [2], with its distinction of eight cases of Containment and its systematic mapping to natural language meanings, provides a fresh new perspective and a valuable starting point for our enterprise. Exploring the fruitfulness of these distinctions in future blending experiments will be of great interest.

Our main roadmap for developing the theory of image schemas formally is as follows: we plan specifically to

- design a formal ontology of image schemas, building on the work of [25];
- specify blending templates in the Distributed Ontology Language DOL [27];

- perform blending experiments with basic image schemas;
- create complex integration templates from basic image schemas via blending.

All this work will be directed towards the goal of building a library of basic, formalised image schemas, as discussed earlier. The most important, and arguably hardest, problem is to further investigate the interplay between dynamic and static aspects of image schemas, that is, the relationship between their embodied nature, i.e. 'simulating' an image schema in a particular scenario, and related 'static' logical formalisations. The late Joseph Goguen proposed to employ dynamical systems theory to address this aspect [16]. To evaluate this and related approaches to the formalisation problem of image schemas will be an important future task.

Acknowledgments. The project COINVENT acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant number: 611553.

References

1. Lawrence W. Barsalou. Grounded cognition. *Annual review of psychology*, 59:617–645, 2008.
2. Brandon Bennett and Claudia Cialone. Corpus Guided Sense Cluster Analysis: a methodology for ontology development (with examples from the spatial domain). In Pawel Garbacz and Oliver Kutz, editors, *8th International Conference on Formal Ontology in Information Systems (FOIS)*, volume 267 of *Frontiers in Artificial Intelligence and Applications*, pages 213–226. IOS Press, 2014.
3. Gilles Fauconnier and Mark Turner. Conceptual integration networks. *Cognitive Science*, 22(2):133–187, 1998.
4. Joseph A. Goguen. An Introduction to Algebraic Semiotics, with Applications to User Interface Design. In *Computation for Metaphors, Analogy and Agents*, number 1562 in LNCS, pages 242–291. Springer, 1999.
5. Joseph. A. Goguen. Semiotic Morphisms, Representations and Blending for Interface Design. In *Proc. of the AMAST Workshop on Algebraic Methods in Language Processing*, pages 1–15. AMAST Press, 2003.
6. Joseph A. Goguen and D. Fox Harrell. Style: A Computational and Conceptual Blending-Based Approach. In Shlomo Argamon and Shlomo Dubnov, editors, *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, pages 147–170. Springer, Berlin, 2010.
7. Joseph A. Goguen and Grant Malcolm. *Algebraic Semantics of Imperative Programs*. MIT, 1996.
8. Joseph E. Grady. Cognitive mechanisms of conceptual integration. *Cognitive Linguistics*, 11(3-4):335–345, 2001.
9. Beate Hampe. Image schemas in cognitive linguistics: Introduction. In Beate Hampe and Joseph E Grady, editors, *From perception to meaning: Image schemas in cognitive linguistics*, pages 1–14. Walter de Gruyter, 2005.

10. Joana Hois, Oliver Kutz, Till Mossakowski, and John Bateman. Towards Ontological Blending. In *Proc. of the The 14th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA-2010)*, Varna, Bulgaria, September 8th–10th, 2010.
11. Megan Johanson and Anna Papafragou. What does children's spatial language reveal about spatial concepts? Evidence from the use of containment expressions. *Cognitive science*, 38(5):881–910, June 2014.
12. Mark Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* The University of Chicago Press, Chicago and London, 1987.
13. Raymond W. Gibbs Jr. Making good psychology out of blending theory. *Cognitive Linguistics*, 11(3-4):347–358, 2001.
14. Werner Kuhn. Modeling the Semantics of Geographic Categories through Conceptual Integration. In *Proc. of GIScience 2002*, pages 108–118. Springer, 2002.
15. Werner Kuhn. An Image-Schematic Account of Spatial Categories. In Stephan Winter, Matt Duckham, Lars Kulik, and Ben Kuipers, editors, *Spatial Information Theory*, volume 4736 of *Lecture Notes in Computer Science*, pages 152–168. Springer, 2007.
16. Werner Kuhn, Martin Raubal, and Peter Gärdenfors. Editorial: Cognitive Semantics and Spatio-Temporal Ontologies. *Spatial Cognition and Computation*, 7(1):3–12, 2007.
17. Oliver Kutz, John Bateman, Fabian Neuhaus, Till Mossakowski, and Mehul Bhatt. E pluribus unum: Formalisation, Use-Cases, and Computational Support for Conceptual Blending. In Tarek R. Besold, Marco Schorlemmer, and Allain Smaill, editors, *Computational Creativity Research: Towards Creative Machines*, Thinking Machines. Atlantis/Springer, 2014.
18. Oliver Kutz, Till Mossakowski, Joana Hois, Mehul Bhatt, and John Bateman. Ontological Blending in DOL. In Tarek Besold, Kai-Uwe Kühnberger, Marco Schorlemmer, and Alan Smaill, editors, *Computational Creativity, Concept Invention, and General Intelligence, Proc. of the 1st Int. Workshop C3GI@ECAI*, volume 01-2012, Montpellier, France, August 27 2012. Publications of the Institute of Cognitive Science, Osnabrück.
19. Oliver Kutz, Till Mossakowski, and Dominik Lücke. Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis*, 4(2):255–333, 2010. Special Issue on 'Is Logic Universal?'.
20. Oliver Kutz, Fabian Neuhaus, Till Mossakowski, and Mihai Codescu. Blending in the Hub—Towards a collaborative concept invention platform. In *Proc. of the 5th International Conference on Computational Creativity*, Ljubljana, Slovenia, June 10–13 2014.
21. Oliver Kutz and Immanuel Normann. Context Discovery via Theory Interpretation. In *Proc. of the IJCAI Workshop on Automated Reasoning about Context and Ontology Evolution, ARCOE-09, Pasadena, California*, 2009.
22. Jean M. Mandler. The foundations of conceptual thought in infancy. *Cognitive Development*, 7(3):273 – 285, 1992.
23. Jean M. Mandler. How to build a baby: II. Conceptual primitives. *Psychological review*, 99(4):587–604, October 1992.
24. Jean M. Mandler. *The Foundations of Mind : Origins of Conceptual Thought: Origins of Conceptual Though.* Oxford University Press, New York, 2004.
25. Jean M. Mandler and Cristóbal Pagán Cánovas. On defining image schemas. *Language and Cognition*, 0:1–23, May 2014.

26. Leora Morgenstern. Mid-Sized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking. *Studia Logica*, 67:333–384, 2001.
27. Till Mossakowski, Christoph Lange, and Oliver Kutz. Three Semantics for the Core of the Distributed Ontology Language. In Michael Grüninger, editor, *7th International Conference on Formal Ontology in Information Systems (FOIS)*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2012.
28. Till Mossakowski, Christian Maeder, and Klaus Lüttich. The Heterogeneous Tool Set. In Orna Grumberg and Michael Huth, editors, *TACAS 2007*, volume 4424 of *Lecture Notes in Computer Science*, pages 519–522. Springer-Verlag Heidelberg, 2007.
29. Fabian Neuhaus, Oliver Kutz, Mihai Codescu, and Till Mossakowski. Fabricating Monsters is Hard - Towards the Automation of Conceptual Blending. In *Proc. of Computational Creativity, Concept Invention, and General Intelligence (C3GI-14)*, volume 1-2014, pages 2–5, Prague, 2014. Publications of the Institute of Cognitive Science, Osnabrück.
30. Immanuel Normann. *Automated Theory Interpretation*. PhD thesis, Department of Computer Science, Jacobs University, Bremen, 2008.
31. Immanuel Normann and Oliver Kutz. Ontology Correspondence via Theory Interpretation. In *Workshop on Matching and Meaning, AISB-09, Edinburgh, UK*, 2009.
32. Tim Rohrer. Image schemata in the brain. In Beate Hampe and Joseph E Grady, editors, *From perception to meaning: Image schemas in cognitive linguistics*, volume 29 of *Cognitive Linguistics Research*, pages 165–196. Walter de Gruyter, 2005.
33. Martin Schmidt, Ulf Krumnack, Helmar Gust, and Kai-uwe Kühnberger. *Computational Approaches to Analogical Reasoning: Current Trends*, volume 548 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
34. Marco Schorlemmer, Alan Smaill, Kai-Uwe Kühnberger, Oliver Kutz, Simon Colton, Emilios Cambouropoulos, and Alison Pease. COINVENT: Towards a Computational Concept Invention Theory. In *Proc. of the 5th International Conference on Computational Creativity*, Ljubljana, Slovenia, June 10–13 2014.
35. Alan M. Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.
36. Mark Turner. *The Origin of Ideas: Blending, Creativity, and the Human Spark*. Oxford University Press, 2014.
37. Tony Veale. The analogical thesaurus. In J. Riedl and R. Hill, editors, *Proceedings of the Fifteenth Innovative Applications of Artificial Intelligence Conference*, pages 137–142. AAAI Press, 2003.
38. Tony Veale. From conceptual mash-ups to bad-ass blends: A robust computational model of conceptual blending. In Mary Lou Maher, Kristian Hammond, Alison Pease, Rafael Prez y Prez, Dan Ventura, and Geraint Wiggins, editors, *Proceedings of the Third International Conference on Computational Creativity*, pages 1–8, May 2012.
39. John von Neumann. *The Computer and the Brain*. Yale University Press, London, 1958.
40. Fan-Pei Gloria Yang, Kailyn Bradley, Madiha Huq, Dai-Lin Wu, and Daniel C. Krawczyk. Contextual effects on conceptual blending in metaphors: An event-related potential study. *Journal of Neurolinguistics*, 26:312–326, 2012.

Monoidal logics: How to avoid paradoxes

Clayton Peterson*

Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität München
clayton.peterson@outlook.com

Abstract. Monoidal logics are logics that can be seen as specific instances of monoidal categories. They are constructed using specific rules and axiom schemata that allow to make explicit the monoidal structure of the logics. Among monoidal logics, we find *Cartesian* logics, which are instances of Cartesian categories. As it happens, many paradoxes in epistemic, deontic and action logics can be related to the Cartesian structure of the logics that are used. While in epistemic and deontic logics the source of the paradoxes is often found within the principles that govern the modal operators, our framework enables us to show that many problems can be avoided by adopting a proper monoidal structure. Thus, the usual modal rules and axiom schemata do not necessarily need to be discarded to avoid the paradoxes. In this respect, monoidal logics offer an alternative way to model knowledge, actions and normative reasoning. Furthermore, it provides us with new avenues to analyze modalities.

Keywords: Nonmonotonic reasoning, Conditional normative reasoning, Category theory, Categorical logic, Logical omniscience

1 Introduction

Monoidal logics were recently introduced by Peterson [28] as a framework to classify logical systems through their categorical structure.[1] Inspired by the work of Lambek (see for instance [22]), the idea is to use category theory as a foundational framework for logic and make explicit the relations between the categorical structure of the logics and the rules and axiom schemata that are used.

In the present paper, we show how monoidal logics are relevant to artificial intelligence given that they enable us to expose and solve some problems that are related to epistemic, deontic and action logics. While these kinds of logic are often formalized as different variations of modal logics, we begin in section 2 by summarizing the framework we adopt for modal logics (section 2.1) and monoidal logics (section 2.2). That being done, we present and discuss some paradoxes in section 3 and analyze them in light of our framework. We conclude in section 4 with avenues for future research.

* The author would like to thank Jean-Pierre Marquis for his time and support. This research was financially supported by the Social Sciences and Humanities Research Council of Canada.

[1] See also [30,29].

2 Framework

2.1 Modal logics

Following Chellas [8], let Δ contain the axiom schemata and rules of propositional logic. Assume the usual definition for the $\Diamond$ operator (i.e., $\Diamond\varphi =_{df} \neg\Box\neg\varphi$) together with the language $\mathcal{L} = \{Prop, (,), \wedge, \top, \supset, \vee, \bot, \Box\}$, where $Prop$ is a collection of atomic propositions. The $\Box$ operator is a modality that can represent necessity (e.g., alethic logic), knowledge (e.g., epistemic logic), obligation (e.g., deontic logic), past/future (e.g., temporal logic) or the execution of an action or a computer program (e.g., dynamic logic). The connectives of the language are the usual classical connectives (i.e., conjunction, implication and disjunction). Negation is defined by $\neg\varphi =_{df} \varphi \supset \bot$ and well-formed formulas are defined recursively as follows[2]:

$$\varphi := p_i \mid \bot \mid \top \mid \varphi \wedge \psi \mid \varphi \supset \psi \mid \varphi \vee \psi \mid \Box\varphi$$

The interest of Chellas's approach is that it clearly relates the rules governing the modalities to the consequence relation of classical logic. Using the following inference rules, we can adopt the following definitions[3]:

- Δ is *classical* if it is closed under (RE);
- Δ is *monotonic* if it is closed under (RM);
- Δ is *normal* if it is closed under (RK).

$$\frac{\varphi \equiv \psi}{\Box\varphi \equiv \Box\psi}\ \text{(RE)} \qquad \frac{\varphi \supset \psi}{\Box\varphi \supset \Box\psi}\ \text{(RM)} \qquad \frac{\varphi}{\Box\varphi}\ \text{(RN)}$$

$$\frac{(\varphi_1 \wedge \cdots \wedge \varphi_n) \supset \psi}{(\Box\varphi_1 \wedge \cdots \wedge \Box\varphi_n) \supset \Box\psi}\ \text{(RK)} \quad \text{with } n \geq 0$$

While a classical system preserves logical equivalences under $\Box$, a monotonic system insures that $\Box$ preserves consequences. The relations between these systems is as follows: if Δ is normal, then it is monotonic, and furthermore if it is monotonic, then it is classical. A classical system E is usually defined by $LPC +$ (RE), a monotonic system M by $LPC +$ (RM) and a normal system K by $LPC +$ (RK). In addition to the usual definition of these systems, one can also have alternative formulations using the following axiom schemata:

$$\Box(\varphi \wedge \psi) \supset (\Box\varphi \wedge \Box\psi)\ \text{(M)} \qquad \Box(\varphi \supset \psi) \supset (\Box\varphi \supset \Box\psi)\ \text{(K)} \qquad \Box\top\ \text{(N)}$$

Using these axioms, monotonic and normal systems can alternatively be defined:

$$\begin{aligned} K &= LPC + \text{(K)} + \text{(RN)} \\ &= M + \text{(K)} + \text{(N)} \end{aligned} \qquad M = E + \text{(M)}$$

[2] That is, atoms, $\top$ and $\bot$ are formulas, and if φ and ψ are formulas, then so are $\varphi \wedge \psi$, $\varphi \supset \psi$, $\varphi \vee \psi$ and $\Box\varphi$.

[3] Note that there are other types of modal systems, such as *regular* systems, but we leave them aside for the purpose of the present paper.

Many extensions can be constructed from these systems. The usual modal axioms are D, T, $\mathsf{4}$ and $\mathsf{5}$.[4]

$$\Box\varphi \supset \Diamond\varphi \ (\mathsf{D}) \qquad \Box\varphi \supset \varphi \ (\mathsf{T}) \qquad \Box\varphi \supset \Box\Box\varphi \ (\mathsf{4}) \qquad \varphi \supset \Box\Diamond\varphi \ (\mathsf{5})$$

D is usually considered as a deontic axiom, which means that if φ is obligatory, then it is also permitted. T is usually used as an axiom for necessity, meaning that if 'it is necessary that φ is true' is true, then φ is true. $\mathsf{4}$ and $\mathsf{5}$ are usually used for epistemic modalities, the former meaning that if an agent knows φ, then he knows that he knows φ and the latter meaning that if φ is true, then an agent knows that it is possible for φ to be true.

2.2 Monoidal logics

The rationale behind monoidal logics is to use category theory to analyze the proof theory of logical systems. By doing so, one can expose the categorical structure of different logics and classify these systems accordingly. Consider the language $\mathcal{L} = \{Prop, (,), \otimes, 1, \multimap, \oplus, 0\}$, where $Prop$ is a collection of atomic propositions. The connective $\otimes$ is understood as some form of conjunction (although not necessarily $\wedge$), $\multimap$ is an implication and $\oplus$ a disjunction (but not necessarily $\vee$). Negation and well-formed formulas are defined as usual ($\neg\varphi =_{df} \varphi \multimap 0$).

$$\varphi := p_i \mid 0 \mid 1 \mid \varphi \otimes \psi \mid \varphi \multimap \psi \mid \varphi \oplus \psi$$

To define monoidal logics, we need to first define the consequence relation (see the rules and axiom schemata in figure 1).[5] To do so, we define a deductive system and we require that proofs are reflexive and transitive.

Definition 1. *A* deductive system $\mathcal{D}$ *is a collection of formulas and (equivalence classes of) proofs (deductions). It has to satisfy (1) and (cut).*

Then, one can introduce a conjunction $\otimes$ with a unit 1 using a monoidal deductive system. This conjunction is minimally associative but is not necessarily commutative. The unit 1 can be absorbed by $\otimes$ from (r) and (l).

Definition 2. *A* monoidal deductive system M *is a deductive system satisfying (r), (l), (t) and (a).*

When this is done, one can do either one of two things. Either one keeps the monoidal structure and adds an implication, and then perhaps classical negations, or one adds some structure to the conjunction by requiring that it be commutative.[6] In the latter case, one can define a symmetric deductive system,

[4] Note that there are other axioms, see [8,19,10].

[5] A double line means that the rule can be applied both top-down and bottom-up.

[6] Given space limitations, we will not expose the whole plethora of monoidal logics that can be defined. For instance, we will not elaborate on monoidal closed deductive systems or monoidal closed deductive systems with classical negations. For a thorough presentation and further explanations, we refer the reader to [30,29].

where the conjunction satisfies a braiding rule (b). That said, at this stage, it is also possible to keep the symmetric structure and introduce an implication by defining a closed deductive system, and then adding classical negation by defining a closed deductive system with classical negation.

Definition 3. *A* symmetric monoidal deductive system **S** *is a monoidal deductive system satisfying (b).*

3.1 *A* symmetric closed *deductive system* **SC** *satisfies (cl).*
3.2 *A* symmetric closed deductive system with classical negation **SCC** *satisfies* $(\neg\neg)$*.*

From a symmetric deductive system, one can add some more structure to the conjunction and define a Cartesian deductive system. In such a case, $\otimes$ is the usual conjunction $\wedge$ of classical or intuitionistic logics. The rule (Cart) allows us to introduce and eliminate the conjunction, while (!) means that anything implies the truth. As it was the case for symmetric deductive system, one can also add an implication and classical negation.

Definition 4. *A* Cartesian deductive system **C** *is a deductive system satisfying (Cart) and (!).*

4.1 *A* Cartesian closed *deductive system* **CC** *satisfies (cl).*
4.2 *A* Cartesian closed deductive system with classical negation **CCC** *satisfies* $(\neg\neg)$*.*

The relationship between these deductive systems is as follows: if $\mathcal{D}$ is Cartesian, then it is symmetric, and furthermore if it is symmetric, then it is monoidal. As a notational convention, we use the symbols $\{\otimes, 1, \multimap, \oplus, 0\}$ for non-Cartesian deductive systems and $\{\wedge, \top, \supset, \vee, \bot\}$ for Cartesian ones.

$$\begin{array}{c} \\ \hline \varphi \longrightarrow \varphi \end{array}\ (1) \qquad \begin{array}{c} \\ \hline \neg\neg\varphi \longrightarrow \varphi \end{array}\ (\neg\neg)$$

$$\begin{array}{c} \varphi \longrightarrow \psi \qquad \psi \longrightarrow \rho \\ \hline \varphi \longrightarrow \rho \end{array}\ (\text{cut}) \qquad \begin{array}{c} \varphi \longrightarrow \psi \otimes 1 \\ \hline\hline \varphi \longrightarrow \psi \end{array}\ (\text{r}) \qquad \begin{array}{c} \varphi \longrightarrow 1 \otimes \psi \\ \hline\hline \varphi \longrightarrow \psi \end{array}\ (\text{l})$$

$$\begin{array}{c} \varphi \longrightarrow \psi \qquad \rho \longrightarrow \tau \\ \hline \varphi \otimes \rho \longrightarrow \psi \otimes \tau \end{array}\ (\text{t}) \qquad \begin{array}{c} \tau \longrightarrow (\varphi \otimes \psi) \otimes \rho \\ \hline\hline \tau \longrightarrow \varphi \otimes (\psi \otimes \rho) \end{array}\ (\text{a}) \qquad \begin{array}{c} \varphi \longrightarrow \psi \otimes \tau \\ \hline\hline \varphi \longrightarrow \tau \otimes \psi \end{array}\ (\text{b})$$

$$\begin{array}{c} \varphi \otimes \psi \longrightarrow \rho \\ \hline\hline \varphi \longrightarrow \psi \multimap \rho \end{array}\ (\text{cl}) \qquad \begin{array}{c} \\ \hline \varphi \longrightarrow 1 \end{array}\ (!) \qquad \begin{array}{c} \varphi \longrightarrow \psi \qquad \varphi \longrightarrow \rho \\ \hline\hline \varphi \longrightarrow \psi \otimes \rho \end{array}\ (\text{Cart})$$

Fig. 1. Rules and axiom schema

The co-tensor $\oplus$ can be axiomatized through a deductive system defined as an opposite deductive system $\mathcal{D}^{op}$ where the formulas remain the same but the deduction arrows are reversed and $\otimes/1$ are respectively replaced by $\oplus/0$. Hence, we obtain the co-versions of the aforementioned rules and we can define co-monoidal (co**M**), co-symmetric (co**S**) and co-Cartesian (co**C**) deductive systems.

The interest of this approach is that deductive systems can be classified according to their categorical structure (cf. [29]). For instance, **M** is an instance of a monoidal category, **SC** is an instance of a (monoidal) symmetric closed category and **C** is an instance of a Cartesian category (cf. [23] for the definitions). Using this framework, we can classify existing logical systems and create new ones. For example, classical logic is an instance of a **CCC**co**C**, intuitionistic logic is an instance of a **CC**co**C**, the multiplicative fragment of linear logic (cf. [11]) is an instance of a **SCC** satisfying $\varphi \oplus \psi =_{df} \neg\varphi \multimap \psi$ and the additive fragment of linear logic is an instance of a **C**co**C**.

On the semantical level, monoidal logics can be interpreted within the framework of partially-ordered residuated monoids (see [30,29]).[7] While it is well-known that **CC**s and **CCC**s are sound and complete with respect to Heyting and Boolean algebras, **SCC**s can be shown to be sound and complete with respect to partially-ordered commutative residuated involutive monoids.[8]

3 Some paradoxes

3.1 Logical omniscience

Epistemic logics are usually defined as normal K45-, KD45-, KT4- or KT5-systems. Notwithstanding these different axiomatizations, the problem of logical omniscience is linked to the basic structure of a normal system and can be related to many rules and axioms. While it is usually attributed to the K-axiom for distribution (e.g., [14]), it can also be attributed to the rule RK (e.g., [16]) or even RN. As we noted earlier, these rules and this axiom are all derivable in a normal system.

The rule RN expresses a weak form of logical omniscience. It means that an agent knows each and every theorem of the system. Combined with the K-axiom for distribution, this implies a stronger form of omniscience. Indeed, K is logically equivalent to the following formula, which states that knowledge (or belief) is closed under implications that are known (or believed).

$$(\Box\varphi \land \Box(\varphi \supset \psi)) \supset \Box\psi$$

Considered together with RN, the K-axiom implies that an agent knows every logical consequence of his prior knowledge. This is the usual presentation of the problem of logical omniscience, which amounts to attribute the problem to RK (which, as we know, is logically equivalent to K+RN). Hence, even though 'full' logical omniscience happens when RK is satisfied, it should be emphasized that some weaker form of logical omniscience can also happen in non-normal modal logics that satisfy either K or RN (but not both).

In addition to these three forms of logical omniscience, others are also present in some non-normal modal logics. For instance, the rule RM entails that if an agent knows something, then he knows all tautologies. That does not imply that the agent knows per se every tautology, but only that as soon as he knows, say, φ, *then* he knows every tautology. This is a consequence of the following instance of RM (with $\top$ some tautology).

[7] This semantical framework is inspired by the work of [9] on residuated lattices.

[8] They are also sound and complete with respect to a specific string diagrammatic language (see [31]).

$$\frac{\varphi \supset \top}{\Box\varphi \supset \Box\top}\ (\mathsf{RM})$$

Furthermore, another weaker form of omniscience can be related to RE. Although he specifies that this does not reduce to logical omniscience per se, Stalnaker [41] points out that RE also poses a problem given that as soon as one knows (or believes) a trivial tautology, such as $\neg(\varphi \wedge \neg\varphi)$, then one knows (or believes) all tautologies. As such, given that tautologies are logically equivalent, it follows that if one knows some tautology, then he knows them all.

Consequently, it appears that even classical systems are not completely immune to the objection of logical omniscience. But still, modal logics are widely relevant to the analysis of epistemic modalities, and thus an important question is whether or not it is possible to *utterly* avoid logical omniscience while keeping other relevant principles of modal logics. Fortunately, the answer to that question is *yes*, and the solution is to look at this problem from the perspective of monoidal logics.

Despite all the modal rules and axioms that were used in the presentation of the problem of logical omniscience, it should be noted that there were also two propositional principles at play. On the one hand, in the case of RM, it is the fact that $\varphi \supset \top$ is a theorem that allows us to conclude that if an agent knows that φ, then he knows every tautology. From a categorical perspective, this amounts to the fact that $\top$ is a terminal object. On the other hand, in the case of RE, it is the fact that tautologies are logically equivalent that leads to a weaker form of logical omniscience. Although this might not be explicit at first glance, it happens that this is also related to the fact that $\top$ is terminal.

$$\begin{tikzcd}
\varphi \arrow[r, dotted, "!"] \arrow[dr] & \top \arrow[d] \\
 & \varphi
\end{tikzcd}$$

As it is shown in the diagram above, if a formula φ is a theorem, then we know that there is a proof $\top \longrightarrow \varphi$. This is standard for any monoidal logic. That being said, it is the arrow ! that entails the logical equivalence between any tautology and $\top$ (hence between each and every tautologies).

From a categorical perspective, this arrow is related to the Cartesian structure of classical modal logics, which follows from the fact that they are extensions of (classical) propositional logic. It is however possible to define propositional logics that still have a classical negation but that do not have this Cartesian structure. Indeed, the closest alternative system would be a symmetric monoidal closed deductive system with classical negation **SCC**. In such a system, $\top$ is not terminal and tautologies (resp. contradictions) are not logically equivalent. Therefore, one could easily add the rules RE or RM to a **SCC** without facing the weaker forms of logical omniscience related to these rules. Note, however, that RN would still imply that the agent knows every tautology and, moreover, that K would still mean that knowledge is closed under known implications.

3.2 Ross's paradox

Ross's paradox [37,38] concerns deontic logic and the logic of imperatives. It aims to show that normative propositions (or imperatives) and descriptive propositions are not satisfied in the same conditions. In the standard system (i.e., in a normal KD-system), it amounts to say that the following is derivable: If Peter ought to mail a letter, then he ought to either mail it or burn it.

Despite the fact that Ross's paradox is ususally objected to the standard system (i.e., a normal system), it should be noted that it is actually derivable in monotonic systems. Indeed, it is a special case of RM:

$$\dfrac{\varphi \supset (\varphi \vee \psi)}{\Box\varphi \supset \Box(\varphi \vee \psi)}\ (\mathsf{RM})$$

But even though Ross's paradox appears as a consequence of monotonic systems, it happens that RM does not necessarily leads to it. As we can see in the instance of RM above, the formula that allows us to derive the undesired consequence is $\varphi \supset (\varphi \vee \psi)$. This formula is a specific instance (on the right below) of the co-version of the rule for Cartesian systems (on the left).[9]

$$\dfrac{\varphi \longrightarrow \rho \qquad \psi \longrightarrow \rho}{\varphi \vee \psi \longrightarrow \rho}\ (\text{co-Cart}) \qquad\qquad \dfrac{\dfrac{}{\varphi \vee \psi \longrightarrow \varphi \vee \psi}\ (1)}{\varphi \longrightarrow \varphi \vee \psi}\ (\text{co-Cart})$$

It is noteworthy that the arrow $\varphi \longrightarrow \varphi \vee \psi$ expresses a fundamental property of the disjunction $\vee$. Indeed, this arrow is an injection map that allows us to define $\vee$ as a categorical co-product. Put differently, it is a fundamental property of co-Cartesian deductive systems, which is not derivable in non-Cartesian ones. Hence, in the presence of RM, Ross's paradox is derivable as soon as the co-tensor is axiomatized within a co**C**.

As a result, it is possible to keep some desired principles governing the $\Box$ operator and add RM or RK to a **SCC**co**S** while avoiding Ross's paradox.

3.3 Prior's paradox

Prior's paradox [35] of derived obligations aims to show that von Wright's [45] notion of commitment was not adequately modeled by his initial approach. While von Wright interpreted $\Box(\varphi \supset \psi)$ as 'φ *commits* us to ψ', Prior showed that this leads to paradoxical results given that the following formula is derivable within von Wright's system.

$$\Box\neg\varphi \supset \Box(\varphi \supset \psi)$$

In words, this means that if φ is forbidden, then carrying out φ commits us to any ψ. This is obviously an undesirable principle. As in the case of Ross's paradox, this is actually a consequence of RM.

$$\dfrac{\neg\varphi \supset (\varphi \supset \psi)}{\Box\neg\varphi \supset \Box(\varphi \supset \psi)}\ (\mathsf{RM})$$

[9] Note that $\varphi \longrightarrow \psi$ if and only if $\top \longrightarrow \varphi \supset \psi$.

Yet, although Prior's paradox might be seen as an instance of RM, it is still possible to have a modal system satisfying that rule without enabling the derivation of the paradox. If we consider the logical equivalence between $\varphi \supset \psi$ and $\neg\varphi \vee \psi$, then $\neg\varphi \supset (\varphi \supset \psi)$ can also be seen as a special instance of (co-Cart). That being said, it is noteworthy that Prior's paradox is deeply related to the (co)-Cartesian structure of the logic. Indeed, the formula $\neg\varphi \supset (\varphi \supset \psi)$ actually hides the fact that $\bot$ is initial, which is also a fundamental characteristic of (co)-Cartesian deductive systems.

$$\dfrac{\dfrac{\dfrac{}{\neg\varphi \longrightarrow \neg\varphi}\,(1)}{\varphi \wedge \neg\varphi \longrightarrow \bot}\,(\mathrm{cl}) \qquad \dfrac{}{\bot \longrightarrow \psi}\,(\bot)}{\dfrac{\varphi \wedge \neg\varphi \longrightarrow \psi}{\neg\varphi \longrightarrow \varphi \supset \psi}\,(\mathrm{cl})}\,(\mathrm{cut})$$

As it is shown in the proof above, the derivation of Prior's paradox requires the axiom schema stating that $\bot$ is initial.[10] In this respect, the paradox can be correlated to the (co-)Cartesian structure of the logic. Therefore, it is possible to avoid Prior's paradox while keeping RM or RK, for instance if we add RM or RK to a **SCC**co**S**.

3.4 Idempotent action

In the philosophy literature, the two main action logics that are used are *stit* and dynamic logics (cf. [40]). On the one hand, the building blocks of *stit* logics can be found within the work of Kanger [21] and Pörn [36], but the explicit *stit* frameworks were introduced by Belnap and Perloff [5] and further developed by Xu [46] (see also Horty [18]).[11] Actions within *stit* frameworks[12] are modeled using a normal K-system and further axioms, depending on the desired structure of the model.[13] In this respect, the structure of *stit* logics is essentially Cartesian. On the other hand, dynamic logics where developed by Pratt [33,34] and where introduced within the context of deontic logic by Meyer [24,25]. Dynamic logics also use a normal K-system, which expresses that after the execution of some action (or computer program), a description of the state holds. In dynamic logics, however, there is a distinction between *actions* and *propositions*. As such, the 'action logics' inherent to these approaches are not expressed via the structure of the normal K-system. Instead, actions are modeled using a Kleene algebra in the standard formulation of dynamic logic (cf. [15]) and with a Boolean algebra in the case of deontic dynamic logic (cf. [28]). In addition to dynamic and *stit* logics, there are also other approaches that explicitly use Boolean algebras to model actions, for instance [39,42,7].

Apart from dynamic logics based on Kleene algebras, all the aforementioned approaches share a common structure, namely that of a Cartesian deductive

[10] Note that the axiom $\bot$ is actually co-!.

[11] The acronym *stit* stands for 'seeing to it that'.

[12] More precisely, consequences of actions (intended or not).

[13] See for example [17,26,6].

system. While it is trivial in the case of *stit* logics since they are normal modal logics, it is also a direct consequence of using Boolean algebras to model actions. Indeed, the syntactical equivalence between classical propositional logic and Boolean algebras is well-known, notwithstanding the fact that Boolean algebras can be seen as instances of Cartesian closed categories (cf. [2,13]).

Now, an interesting property of Cartesian deductive systems is that they satisfy idempotence of conjunction (i.e., φ is logically equivalent to $\varphi \wedge \varphi$). This follows from the derivations below.

$$\dfrac{\dfrac{}{\varphi \longrightarrow \varphi}\,(1) \qquad \dfrac{}{\varphi \longrightarrow \varphi}\,(1)}{\varphi \longrightarrow \varphi \wedge \varphi}\,(\text{Cart}) \qquad\qquad \dfrac{\dfrac{}{\varphi \wedge \varphi \longrightarrow \varphi \wedge \varphi}\,(1)}{\varphi \wedge \varphi \longrightarrow \varphi}\,(\text{Cart})$$

Although it was not formulated in these terms when he introduced linear logic, Girard [11] presented the backbone of what we might call the 'paradox of idempotent action'. Let φ stands for 'giving one dollar'. Clearly, giving one dollar is not logically equivalent to giving one dollar *and* giving one dollar. Consequently, the paradox of idempotent action can be objected to action logics that have a Cartesian structure given that they trivially satisfy idempotence of conjunction.

From the perspective of monoidal logics, we can see that this paradox affects **CCC**s, and thus the closest alternative to model action while avoiding the paradox is to use an instance of a **SCC**.

3.5 Contrary-to-duty reasoning

Contrary-to-duty reasoning is deeply relevant to artificial intelligence. As it stands, the three main problems one faces when trying to model contrary-to-duty reasoning are augmentation, factual detachment and deontic explosion.

Augmentation (cf. [20]), also known as the problem of strengthening the antecedent of a deontic conditional (cf. [1]), arises when a logic satisfies the following inference pattern.

$$\dfrac{\varphi \supset \Box\psi}{(\varphi \wedge \rho) \supset \Box\psi}\,(\text{aug})$$

Modeling a deontic conditional using $\varphi \supset \Box\psi$, this implies that whenever there is an obligation $\Box\psi$ conditional to a context φ, then this obligation is also conditional to the augmented context $\varphi \wedge \rho$ for any ρ. This is undesirable given that the extra conditions ρ might be such that the obligation does not hold anymore.[14]

The problem of factual detachment (cf. [44]) can be analyzed in similar terms. It arises when a system satisfies the following inference pattern (i.e., weakening):

$$\dfrac{(\varphi \wedge (\varphi \supset \Box\psi)) \supset \Box\psi}{(\rho \wedge (\varphi \wedge (\varphi \supset \Box\psi))) \supset \Box\psi}\,(\text{wk})$$

[14] The obligation can also be overridden or canceled (cf. [43]).

In a nutshell, the problem of factual detachment can be formulated as follows: even though one might want to detach the obligation $\Box\psi$ from the context φ and the deontic conditional $\varphi \supset \Box\psi$, there might be other conditions ρ that will thwart the detachment of $\Box\psi$. Thus the problem: detachment is desired but only when we can insure that nothing else will thwart the detached obligation. However, if a logic satisfies the aforementioned inference pattern, then it allows for unrestricted detachment.

Finally, the problem of deontic explosion (see for instance [12]) amounts to the fact that from a conflict of obligations one can deduce that anything is obligatory within a normal system.[15] Indeed, normal systems validate the formula $(\Box\varphi \wedge \Box\neg\varphi) \supset \Box\psi$ for any ψ.

These issues have been thoroughly analyzed in [32] and we showed that these three problems are actually related to the Cartesian structure of the logics that are used to model contrary-to-duty reasoning. While the proof of the weakening and the augmentation inference patterns depend on (Cart), deontic explosion actually comes from the fact that $\bot$ is initial in a **CCC**.

$$\dfrac{\dfrac{\dfrac{}{\rho \wedge (\varphi \wedge (\varphi \supset \Box\psi)) \longrightarrow \rho \wedge (\varphi \wedge (\varphi \supset \Box\psi))}\,(1)}{\rho \wedge (\varphi \wedge (\varphi \supset \Box\psi)) \longrightarrow \varphi \wedge (\varphi \supset \Box\psi)}\,(\text{Cart}) \qquad \dfrac{\dfrac{}{\varphi \supset \Box\psi \longrightarrow \varphi \supset \Box\psi}\,(1)}{\varphi \wedge (\varphi \supset \Box\psi) \longrightarrow \Box\psi}\,(\text{cl})}{\rho \wedge (\varphi \wedge (\varphi \supset \Box\psi)) \longrightarrow \Box\psi}\,(\text{cut})$$

$$\dfrac{\dfrac{(\text{Cart})\,\dfrac{(1)\,\dfrac{}{(\varphi \wedge \rho) \wedge (\varphi \supset \Box\psi) \longrightarrow (\varphi \wedge \rho) \wedge (\varphi \supset \Box\psi)}}{(\varphi \wedge \rho) \wedge (\varphi \supset \Box\psi) \longrightarrow \varphi \wedge (\varphi \supset \Box\psi)} \qquad \dfrac{\dfrac{}{\varphi \supset \Box\psi \longrightarrow \varphi \supset \Box\psi}\,(1)}{\varphi \wedge (\varphi \supset \Box\psi) \longrightarrow \Box\psi}\,(\text{cl})}{(\varphi \wedge \rho) \wedge (\varphi \supset \Box\psi) \longrightarrow \Box\psi}\,(\text{cut})}{\varphi \supset \Box\psi \longrightarrow (\varphi \wedge \rho) \supset \Box\psi}\,(\text{cl})$$

$$\dfrac{\dfrac{\vdots}{\Box\varphi \wedge \Box\neg\varphi \longrightarrow \Box\bot} \qquad \dfrac{\dfrac{}{\bot \longrightarrow \psi}\,(\bot)}{\Box\bot \longrightarrow \Box\psi}\,(\mathsf{RM})}{\Box\varphi \wedge \Box\neg\varphi \longrightarrow \Box\psi}\,(\text{cut})$$

In this respect, it can be argued that the three major problems one faces when trying to model contrary-to-duty reasoning are related to the Cartesian structure of the logic that is used. To avoid these problems, one must therefore use a logic that has a weaker structure to model contrary-to-duty reasoning. As such, we developed a logic for conditional normative reasoning on the grounds of a monoidal logic (precisely, an instance of a **SCC**co**S**) in [30].

4 Conclusion

Summing up, we showed using the framework of monoidal logics that many paradoxes in epistemic, deontic and actions logics are related to the Cartesian structure of the logic that are used. While the source of some paradoxes in epistemic and deontic logic is usually attributed to the rules and axiom schemata

[15] Or within a regular system.

that govern the modalities, we showed that the source of these problem is actually the Cartesian structure of the logic. As a result, it is possible to keep some desired modal rules and axiom schemata while avoiding the paradoxes by using a logic that has a monoidal structure rather than a Cartesian one.

For future research, it remains to explore the logical properties of the monoidal modal logics that can be constructed from the rules and axiom schemata of classical modal logics. We will need to properly study the relations between the different rules and axioms and determine how accessibility relations can be defined within the framework of partially-ordered residuated monoids. We also intend to explore how monoidal modal logics can be used to model artificial agents with the help of monoidal computers (cf. [27]).

References

1. Carlos Alchourrón, *Detachment and defeasibility in deontic logic*, Studia Logica **57** (1996), no. 1, 5–18.
2. Steve Awodey, *Category theory*, 2nd ed., Oxford University Press, 2006.
3. Micheal Barr, *∗-autonomous categories*, Lecture Notes in Mathematics, vol. 752, Springer, 1979.
4. ______, *∗-autonomous categories and linear logic*, Mathematical Structures in Computer Science **1** (1991), no. 2, 159–178.
5. Nuel Belnap and Michael Perloff, *Seeing to it that: A canonical form for agentives*, Theoria **54** (1988), no. 3, 175–199.
6. Jan Broersen, *Deontic epistemic stit logic distinguishing modes of mens rea*, Journal of Applied Logic **9** (2011), no. 2, 137–152.
7. Pablo F. Castro and Tom S. E. Maibaum, *Deontic action logic, atomic Boolean algebras and fault-tolerance*, Journal of Applied Logic **7** (2009), no. 4, 441–466.
8. Brian F. Chellas, *Modal logic: An introduction*, Cambridge University Press, 1980.
9. N. Galatos, P. Jipsen, T. Kowalski, and H. Ono (eds.), *Residuated lattices: An algebraic glimpse at substructural logics*, Studies in Logic and the Foundations of Mathematics, vol. 151, Elsevier, 2007.
10. James Garson, *Modal logic for philosophers*, Cambridge University Press, 2006.
11. Jean-Yves Girard, *Linear Logic*, Theoretical Computer Science **50** (1987), no. 1, 1–102.
12. Lou Goble, *Normative conflicts and the logic of 'ought'*, Noûs **43** (2009), no. 3, 450–489.
13. Robert Goldblatt, *Topoi: The categorical analysis of logic*, Dover Publications, 2006.
14. Joseph Y. Halpern and Riccardo Pucella, *Dealing with logical omniscience: Expressiveness and pragmatics*, Artificial Intelligence **175** (2011), no. 1, 220–235.
15. David Harel, Dexter Kozen, and Jerzy Tiuryn, *Dynamic logic*, MIT Press, 2000.
16. Wesley H. Holliday, *Epistemic logic and epistemology*, Handbook of Formal Philosophy (S. O. Hansson and V. F. Hendricks, eds.), Springer, 2014, forthcoming, pp. 1–35.
17. John Horty, *Agency and deontic logic*, Oxford University Press, 2001.
18. John Horty and Nuel Belnap, *The deliberative stit: A study of action, omission, ability and obligation*, Journal of Philosophical Logic **24** (1995), no. 6, 583–644.
19. George Edward Hughes and Maxwell John Cresswell, *A new introduction to modal logic*, Routledge, 1996.

20. Andrew J. I. Jones, *On the logic of deontic conditionals*, Ratio Juris **4** (1991), no. 3, 355–366.
21. Stig Kanger, *New foundations for ethical theory*, Stockholm, 1957.
22. Joachim Lambek and Philip Scott, *Introduction to higher order categorical logic*, Cambridge University Press, 1986.
23. Saunders Mac Lane, *Categories for the working mathematician*, 2nd ed., Springer, 1971.
24. John-Jules Ch. Meyer, *A simple solution to the "deepest" paradox in deontic logic*, Logique et Analyse **30** (1987), no. 117-118, 81–90.
25. John-Jules Ch. Meyer, *A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic*, Notre Dame Journal of Formal Logic **29** (1988), no. 1, 109–136.
26. Olga Pacheco and José Carmo, *A role based model for the normative specification of organized collective agency and agents interaction*, Autonomous Agents and Multi-Agent Systems **6** (2003), no. 2, 145–184.
27. Dusko Pavlovic, *Monoidal computer I: Basic computability by string diagrams*, Information and Computation **226** (2013), 94–116.
28. Clayton Peterson, *Analyse de la structure logique des inférences légales et modélisation du discours juridique*, Ph.D. thesis, Université de Montréal, 2014.
29. ______, *Categorical foundations for logic: Lambek's legacy revisited*, Manuscript submitted for publication, 2014.
30. ______, *The categorical imperative: Category theory as a foundation for deontic logic*, Journal of Applied Logic **12** (2014), no. 4, 417–461.
31. ______, *Conditional reasoning with string diagrams*, Manuscript submitted for publication, 2014.
32. ______, *Contrary-to-duty reasoning: A categorical approach*, Manuscript submitted for publication, 2014.
33. Vaughan Pratt, *Semantical considerations of Floyd-Hoare Logic*, Tech. Report MIT/LCS/TR-168, 1976.
34. ______, *Application of modal logic to programming*, Studia Logica **39** (1980), no. 2-3, 257–274.
35. Arthur Prior, *The paradoxes of derived obligation*, Mind **63** (1954), no. 249, 64–65.
36. Ingmar Pörn, *The logic of power*, Basil Blackwell, 1970.
37. Alf Ross, *Imperatives and logic*, Theoria **7** (1941), no. 1, 53–71.
38. ______, *Imperatives and logic*, Philosophy of Science **11** (1944), no. 1, 30–46.
39. Krister Segerberg, *A deontic logic of action*, Studia Logica **41** (1982), no. 2, 269–282.
40. Krister Segerberg, John-Jules Meyer, and Marcus Kracht, *The logic of action*, The Stanford Encyclopedia of Philosophy (E. N. Zalta, ed.), 2009.
41. Robert Stalnaker, *The problem of logical omniscience (i)*, Synthese **89** (1991), no. 3, 425–440.
42. Robert Trypuz and Piotr Kulicki, *A systematics of deontic action logics based on Boolean algebra*, Logic and Logical Philosophy **18** (2009), 253–270.
43. Leendert van der Torre and Yao-Hua Tan, *The many faces of defeasibility in defeasible deontic logic*, Defeasible Deontic Logic (D. Nute, ed.), Kluwer Academic Publishers, 1997, pp. 79–121.
44. Job van Eck, *A system of temporally relative modal and deontic predicate logic and it's philosophical applications*, Logique et Analyse **25** (1982), no. 99, 249–290.
45. Georg H. von Wright, *Deontic logic*, Mind **60** (1951), no. 237, 1–15.
46. Ming Xu, *On the basic logic of stit with a single agent*, The Journal of Symbolic Logic **60** (1995), no. 2, 459–483.

Mathematical Patterns and Cognitive Architectures

Agnese Augello[1], Salvatore Gaglio[1,2], Gianluigi Oliveri[2,3], and Giovanni Pilato[1]

[1] ICAR - Italian National Research Council
Viale delle Scienze - Edificio 11 - 90128 Palermo, Italy
[2] DICGIM- Università di Palermo
Viale delle Scienze, Edificio 6 - 90128, Palermo - ITALY
[3] Dipartimento di Scienze Umanistiche - Università di Palermo
Viale delle Scienze, Edificio 12 - 90128, Palermo - ITALY
{agnese.augello,giovanni.pilato}@cnr.it
{salvatore.gaglio,gianluigi.oliveri}@unipa.it

Abstract. Mathematical patterns are an important subclass of the class of patterns. The main task of this paper is examining a particular proposal concerning the nature of mathematical patterns and some elements of the cognitive structure an agent should have to recognize them.

1 Introduction

As is well known, the main aim of pattern recognition is to determine whether, and to what extent, what we call 'pattern recognition' can be accounted for in terms of automatic processes. From this it follows that two of its central problems are how to: (i) describe and explain the way humans, and other biological systems, produce/discover and characterize patterns; and how to (ii) develop automatic systems capable of performing pattern recognition behaviour.

Having stated these important facts, we need to point out that at the foundations of pattern recognition there are two more basic questions which we can formulate in the following way: (a) what is a pattern? (b) how do we come to know patterns? And it is clear that, if we intend to develop a science of pattern recognition able to provide a rigorous way of achieving its main aim, and of pursuing its central objects of study, it is very important to address questions (a) and (b).

What we intend to do in this paper is tackling questions (a) and (b) not in their full generality, but in the privileged context provided by mathematics, where there exists a consolidated tradition which regards it as a science of patterns,[4] connecting the results of our enquiries to the appropriate levels of the cognitive architecture we propose for a cognitive agent.

[4] See on this [Oliveri, 1997], [Shapiro, 2000], [Resnik, 2001], [Oliveri, 2007], [Oliveri, 2012], [Bombieri, 2013].

2 A case study

If we are presented with the two following objects **a** and **b**, it is very difficult to see what interesting mathematical feature they might have in common, if any, let alone that they exemplify the same mathematical pattern:

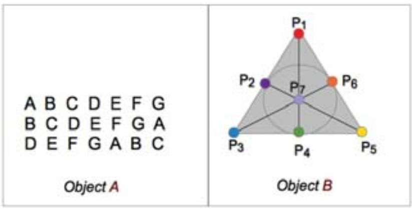


Fig. 1. Different Information Processing for two different Cognitive Agents

Indeed, whereas object **a** is a 3×7 matrix whose elements are the first seven letters of the Italian alphabet, object **b** is an equilateral triangle in which we have inscribed a circle, drawn three bisecting segments, and singled out the points of intersection of three curves.

However, the situation radically changes if we introduce the following formal system T with the appropriate interpretations.

Let T be a formal system such that the language of T contains a primitive binary relation 'x belongs to a set X' ($x \in X$), and its inverse 'X contains an element x' ($X \ni x$).

Furthermore, let us assume that D, the domain, is a set of countably many undefined elements $a_1, a_2, \ldots$; call 'm-set' a subset X of D; and consider the following as the axioms of T:

Axiom 1 If x and y are distinct elements of D there is at least one m-set containing x and y;

Axiom 2 If x and y are distinct elements of D there is not more than one m-set containing x and y;

Axiom 3 Any two m-sets have at least one element of D in common;

Axiom 4 There exists at least one m-set.

Axiom 5 Every m-set contains at least three elements of D;

Axiom 6 All the elements of D do not belong to the same m-set;

Axiom 7 No m-set contains more than three elements of D.[5]

[5] The case study discussed in this section has been taken from [Oliveri, 2012], §3, pp. 410-414. These axioms have been taken, with some minor alterations, from [Tuller, 1967], §2.10, p. 30.

At this point, if we put $I_1(a_1) = A, \ldots, I_1(a_7) = G$, we find that, under this interpretation, what corresponds to the m-sets are the columns of the matrix in fig. 1, and that object **a** is a model of T.

On the other hand, if we put $I_2(a_1) = P_1, \ldots, I_2(a_7) = P_7$, we find that, under this interpretation, what corresponds to the m-sets are the curves in fig. 1, and that object **b** is a model of T. But the surprises do not end up here, because we can now prove that the two models of T mentioned above are isomorphic to one another (see on this [Oliveri, 2012], §3, p. 413, footnote 12).

Several are the things that interest us in this example. First of all, the expression 'the pattern described by T' appears to refer to the mathematical structure which is realized/instantiated in objects **a** and **b**. What this seems to suggest is that, in the mathematical case, the concept of pattern coincides with that of mathematical structure.

Secondly, in the absence of our formal system T, we cannot see the pattern/structure instantiated by **a** and **b** because we are in no position for making the relevant observations concerning the salient features of the pattern/structure in question as is shown by the fact that, in particular, we are unable to make a number of fundamental distinctions such as that between part and whole, etc. etc.

Thirdly, the mathematical structure which becomes salient when we observe objects **a** and **b** *through* T depends not only on T, but also on **a** and **b**. In fact, given that we can prove in T that there exist exactly seven elements in D and seven m-sets if, for instance, the number of letters of the Italian alphabet we considered as elements of our matrix were different from seven, the matrix could not be a model of T (the same applies *mutatis mutandis* to the number of points of intersection of three curves in **b**).

Taking stock of some of the main points made in this section in our study of the mathematical case, we need to say that: (i) we must distinguish between object and structure; (ii) there are strong reasons for identifying mathematical patterns with structures; (iii) necessary conditions for pattern recognition in mathematics are the existence of (1) an observer O; (2) a domain of objects D; and (3) a system of representation Σ, i.e. $(\mathrm{O}, \mathrm{D}, \Sigma)$.[6]

With regard to the problem of how we come to know mathematical patterns, given that mathematical patterns are neither sensible objects nor properties of sensible objects, e.g., what in our example we saw *as* a Euclidean equilateral triangle is not a perfect Euclidean equilateral triangle, because its sides do not have exactly the same length, do not contain an infinite number of points, are not breadthless, etc. (see on this [Oliveri, 2012], §§3 and 4, pp. 410-417), it follows that they are not given to us as a consequence of abstraction or induction/generalization carried out on pure observations. But, on the other hand, if mathematical patterns are (also) dependent on objects, as in the case of **a** and **b**, they cannot simply be in the eyes of the beholder either. They are given to

[6] Actually, the system of representation Σ is an ordered pair $\Sigma = (\mathrm{T}, I)$, where T is a set containing (as a subset) a recursive set of axioms $\mathcal{A}$ and all the logical consequences of $\mathcal{A}$, and I is an interpretation of T on to D.

us as a consequence of our activity of representing entities like **a** and **b** within a given system of representation Σ.

3 Patterns and conceptual spaces

Conceptual spaces (CS) were originally introduced by Gärdenfors as a bridge between symbolic and associationist models of information representation. This was part of an attempt to describe what he calls the 'geometry of thought'.

In [Gärdenfors, 2004] and [Gärdenfors, 2004a] we find a description of a cognitive architecture for modelling representations. The cognitive architecture is composed by three levels of representation: a *subconceptual level*, in which data coming from the environment (sensory input) are processed by means of a neural network based system; a *conceptual level*, where data are represented and conceptualized independently of language; and, finally, a *symbolic level* which makes it possible to manage the information produced at the conceptual level at a higher level through symbolic computations.

Gärdenfors' proposal of a way of representing information *via* his conceptual spaces exploits geometrical structures rather than symbols or connections between neurons. This geometrical representation is based on the existence/construction of a space endowed with a number of what Gärdenfors calls 'quality dimensions' whose main function is to represent different qualities of objects such as brightness, temperature, height, width, depth.

Moreover, for Gärdenfors, judgments of similarity play a crucial role in cognitive processes and, according to him, it is possible to associate the concept of distance to many kinds of quality dimensions. This idea naturally leads to the conjecture that the smaller is the distance between the representations of two given objects in a conceptual space the more similar to each other the objects represented are.

According to Gärdenfors, objects can be represented as points in a conceptual space, points which we are going to call 'knoxels',[7] and concepts as regions within a conceptual space. These regions may have various shapes, although to some concepts — those which refer to natural kinds or natural properties — correspond regions which are characterized by convexity.[8]

Of course, at this point a whole host of important questions come to the forefront, questions like how could a cognitive agent: (1) learn the appropriate conceptual spaces? (2) select between different spaces that could fill the data? (3) determine the possible dimensions for representing objects? etc. etc. And although all such questions are central to our attempt to use Gärdenfors conceptual spaces as part of the cognitive architecture of a conceptual agent — we have addressed some of them in [Augello et al., 2013a] and [Augello et al., 2013b] —

[7] The term 'knoxel' originates from [Gaglio, 1988] by the analogy with "pixel". A knoxel k is a point in Conceptual Space and it represents the epistemologically primitive element at the considered level of analysis.

[8] A set S is *convex* if and only if whenever $a, b \in S$ and c is between a and b then $c \in S$.

what we aim to do in this paper is: (α) showing the existence of at least three different pattern recognition procedures; and (β) individuating which of the 3 corresponding levels of the cognitive architecture of our cognitive agent is involved in the processing of mathematical patterns.

To do this consider the case study discussed in §2 (taken from [Oliveri, 2012], §3, pp. 410-414), and imagine we have before us a cognitive agent A endowed with level 1 information processing system. In this case A (its neural network) can be trained to recognize letters A, B, ..., G and distinguish them from one another; and do the same thing for the coloured round objects $\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_7$.

Furthermore, suppose that the letters and the coloured round objects are presented to A exactly as they are in fig. 1. Once more A, exploiting its level 2 information processing system, i.e. the conceptual spaces of letters and of colours, is able to give a correct representation of **a** and **b**, for example, by representing **a** and **b** in an appropriate finite-dimensional vector space using rigid motions and some operations which act on the spaces.

However, what A cannot do, if the formal system T (see §2) is absent from its symbolic level 3 information processing system, is recognizing that **a** and **b** exemplify/realize the same pattern. Therefore, if what we have argued so far is correct, it follows that in the dawning of a mathematical pattern all the three levels of information processing systems we mentioned above are involved

4 Conclusion

In this work we have revisited a three levels cognitive architecture as a foundational approach to pattern recognition for an agent. We have illustrated this possibility by exploiting a mathematical domain. We have also highlighted the relevance of a linguistic, symbolic, level in order to produce abstractions and see deeper mathematical patterns.

Acknowledgements

This work has been partially supported by the PON01_01687 - SINTESYS (Security and INTElligence SYSstem) Research Project.

References

[Augello et al., 2013a] Augello, A., Gaglio, S., Oliveri, G., Pilato, G., 'An Algebra for the Manipulation of Conceptual Spaces in Cognitive Agents'. *Biologically Inspired Cognitive Architectures*, **6**, 23-29, 2013.

[Augello et al., 2013b] Augello, A., Gaglio, S., Oliveri, G., Pilato, G., 'Acting on Conceptual Spaces in Cognitive Agents'. in [Lieto et al., 2013], pp. 25-32, 2013.

[Bombieri, 2013] Bombieri, E.: 2013, 'The shifting aspects of truth in mathematics', *Euresis*, **vol. 5**, pp. 249–272.

[Chella, 1997] A. Chella, M. Frixione, and S. Gaglio. A cognitive architecture for artificial vision. Artif. Intell., 89:73111, 1997.

[Gaglio, 1988] S. Gaglio, P. P. Puliafito, M. Paolucci, and P. P. Perotto. 1988. Some problems on uncertain knowledge acquisition for rule based systems. Decis. Support Syst. 4, 3 (September 1988), 307-312. DOI=10.1016/0167-9236(88)90018-8 http://dx.doi.org/10.1016/0167-9236(88)90018-8

[Gärdenfors, 2004] Gärdenfors, P.: 2004, *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, Massachusetts.

[Gärdenfors, 2004a] Gärdenfors, P.: 2004. 'Conceptual spaces as a framework for knowledge representation'. *Mind and Matter* 2 (2):9-27.

[Lieto et al., 2013] Lieto A. e Cruciani M. (ed.s): Proceedings of the First International Workshop on Artificial Intelligence and Cognition (AIC 2013). An official workshop of the 13th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Torino, Italy, December 3, 2013 on Conceptual spaces as a framework for knowledge representation. CEUR Workshop Proceedings, vol. 1100.

[Oliveri, 1997] Oliveri, G.: 1997, 'Mathematics. A Science of Patterns?', *Synthese*, **vol. 112**, issue 3, pp. 379–402.

[Oliveri, 2007] Oliveri, G.: 2007, *A Realist Philosophy of Mathematics*, College Publications, London.

[Oliveri, 2012] Oliveri, G.: 2012, 'Object, Structure, and Form', *Logique & Analyse*, **vol. 219**, pp. 401-442.

[Resnik, 2001] Resnik, M.D.: 2001, *Mathematics as a Science of Patterns*, Clarendon Press, Oxford.

[Scholkopf, 2001] Bernhard Scholkopf and Alexander J. Smola. 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

[Shapiro, 2000] Shapiro, S.: 2000, *Philosophy of Mathematics. Structure and Ontology*, Oxford University Press, Oxford.

[Tuller, 1967] Tuller, A.: 1967, *A Modern Introduction to Geometries*, D. Van Nostrand Company, Inc., Princeton, New Jersey.

Romeo2 Project: Humanoid Robot Assistant and Companion for Everyday Life: I. Situation Assessment for Social Intelligence [1]

Amit Kumar Pandey[1], Rodolphe Gelin[1], Rachid Alami[2], Renaud Viry[2], Axel Buendia[3], Roland Meertens[3], Mohamed Chetouani[4], Laurence Devillers[5], Marie Tahon[5], David Filliat[6], Yves Grenier[7], Mounira Maazaoui[7], Abderrahmane Kheddar[8], Frédéric Lerasle[2], and Laurent Fitte Duval[2]

[1]Aldebaran, A-Lab, France, *akpandey@aldebaran.com; rgelin@aldebaran.com*
[2]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France;
rachid.alami@laas.fr; frederic.lerasle@laas.fr; renaud.viry@laas.fr; lfittedu@laas.fr
[3]Spirops/CNAM (CEDRIC), Paris; *axel.buendia@cnam.fr; rolandmeertens@gmail.com*
[4]ISIR, UPMC, France; *mohamed.chetouani@upmc.fr*
[5]LIMSI-CNRS University Paris-Sorbonne; *devil@limsi.fr; marie.tahon@limsi.fr*
[6]ENSTA ParisTech - INRIA FLOWERS; *david.filliat@ensta-paristech.fr*
[7]Inst. Mines-Télécom; Télécom ParisTech; CNRS LTCI;
yves.grenier@telecom-paristech.fr; maazaoui@telecom-paristech.fr
[8]CNRS-UM2 LIRMM IDH; *kheddar@gmail.com*

Abstract. For a socially intelligent robot, different levels of situation assessment are required, ranging from basic processing of sensor input to high-level analysis of semantics and intention. However, the attempt to combine them all prompts new research challenges and the need of a coherent framework and architecture.
This paper presents the situation assessment aspect of Romeo2, a unique project aiming to bring multi-modal and multi-layered perception on a single system and targeting for a unified theoretical and functional framework for a robot companion for everyday life. It also discusses some of the innovation potentials, which the combination of these various perception abilities adds into the robot's socio-cognitive capabilities.

Keywords: Situation Assessment, Socially Intelligent Robot, Human Robot Interaction, Robot Companion

1 Introduction

As robots started to co-exist in a human-centered environment, the human awareness capabilities must be considered. With safety being a basic requirement, such robots should be able to behave in a socially accepted and expected manner. This requires robots to reason about the situation, not only from the perspective of physical locations of objects, but also from that of 'mental' and 'physical' states of the human partner. Further, such reasoning should build knowledge with the human understandable attributes, to facilitate natural human-robot interaction.

The Romeo2 project (website [1]), the focus of this paper, is unique in that it brings together different perception components in a unified framework for real-life personal assistant and companion robot in an everyday scenario. This paper outlines our perception architecture, the categorization of basic requirements, the key elements to perceive, and the innovation advantages such a system provides.

[1] This work is funded by Romeo2 project, (http://www.projetromeo.com/), BPIFrance in the framework of the Structuring Projects of Competitiveness Clusters (PSPC)

Fig. 1 shows the Romeo robot and its sensors. It is a *40kg* and *1.4m* tall humanoid robot with *41 degrees-of-freedom*, *vertebral column*, *exoskeleton on legs*, *partially soft torso* and *mobile eyes*.

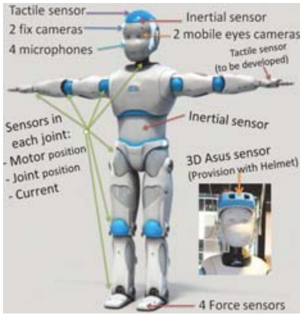


Fig. 1. Romeo robot and sensors.

1.1 An Example Scenario

Mr. Smith lives alone (with his Romeo robot companion). He is elderly and visually impaired. Romeo understands his speech, emotion and gestures, assists him in his daily life. It provides physical support by bringing the 'desired' items, and cognitive support by reminding about medicine, items to add in to-buy list, playing memory games, etc. It monitors Mr. Smith's activities and calls for assistance if abnormalities are detected in his behaviors. As a social inhabitant, it plays with Mr. Smith's grandchildren visiting him.

This outlined partial target scenario of Romeo2 project (also illustrated in fig. 2), depicts that being aware about human, his/her activities, the environment and the situation are the key aspects towards practical achievement of the project's objective.

Fig. 2. Romeo2 Project scenario: A Humanoid Robot Assistant and Companion for Everyday Life.

1.2 Related Works and the main Contributions

Situation awareness is the ability to perceive and abstract information from the environment [2]. It is an important aspect of day-to-day interaction, decision-making, and planning, so as important is the domain-based identification of the *elements* and *attributes*, constituting the state of the environment. In this paper, we will identify and present such elements from companion robot domain perspective, sec. 2.2. Further, three levels of it have been identified (Endsley *et al.* [9]): ***Level 1 situation awareness***: To perceive the state of the *elements* composing the surrounding environment. ***Level 2 situation awareness***: To build a goal oriented understanding of the situation. Experience and comprehension of the meaning are important. ***Level 3 situation awareness***: To project on the future. Sec. 2.1 will present our sense-interact perception loop and map these levels.

Further, there have been efforts to develop integrated architecture to utilize multiple components of situation assessment. However, most of them are specific for a particular task like navigating [21], intention detection [16], robot's self-perception [5], spatial and temporal situation assessment for robot passing through a narrow passage [1], laser data based human-robot-location situation assessment, e.g. human entering, coming closer, etc. [12]. Therefore, they are either limited by the variety of perception attributes, sensors or restricted to a particular perception-action scenario loop. On the other hand, various projects on Human Robot Interaction try to overcome perception limitations by different means and focus on high-level semantic and decision-making. Such as, the detection of objects is simplified by putting tags/markers on the objects, in the detection of people no audio information is used, [6], [14], etc. In [10], different layers of perception have

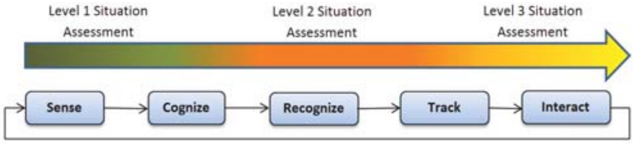

Level 1 Situation Assessment
Level 2 Situation Assessment
Level 3 Situation Assessment
Sense
Cognize
Recognize
Track
Interact

Fig. 3. A generalized perception system for sense-interact in Romeo2 project, with five layers functioning in a closed loop.

been analyzed to build representations of the 3D space, but focused on eye-hand coordination for active perception and not on high-level semantics and perception of the human.

In the Romeo2 project, we are making effort to bring a range of multi-sensor perception components within a unified framework (Naoqi, [18]), at the same time making the entire multi-modal perception system independent from a very specific scenario or task, and explicitly incorporating reasoning about human, towards realizing effective and more natural multi-modal human robot interaction. In this regard, to the best of our knowledge, Romeo2 project is the first effort of its kind for a real world companion robot. In this paper, we do not provide the details of each component. Instead, we give an overview of the entire situation assessment system in Romeo2 project (sec. 2.1). Interested readers can find the details in documentation of the system [18] and in dedicated publications for individual components, such as [4], [11], [19], [15], [3], [24], [17], [23], etc. (see the complete list of publications [1]). Further, the combined effort to bring different components together helps us to identify some of the innovation potentials and to develop them, as discussed in section 3.

2 Perceiving Situation in Romeo2 Project

2.1 A Generalized Sense-Interact Perception Architecture for HRI

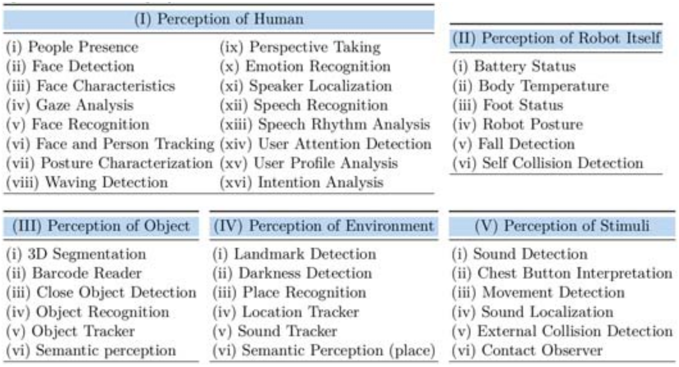
We have adapted a simple yet meaningful, sensing-interaction oriented perception architecture, by carefully identifying various requirements and their interdependencies, as shown in fig. 3. The roles of the five identified layers are:

(i) Sense: To receive signals/data from various sensors. Depending upon the sensors and their fusion. This layer can build 3D point cloud world; sense stimuli like touch, sound; know about the robot's internal states such as joint, heat; record speech signals; etc. Therefore, it belongs to *level 1* of situation assessment.

(ii) Cognize: Corresponds to the 'meaningful' (human-understandable level) and relevant information extraction, e.g. learning shapes of objects; learning to extract the semantics from 3D point cloud, the meaningful words from speech, the meaningful parameters in demonstration, etc. In most of the perception-action systems, this *cognize* part is provided a priori to the system. However, in Romeo2 projects we are taking steps to make *cognize* layer more visible by bringing together different learning modules, such as to learn objects, learn faces, learn the meaning of instructions, learn to categorize emotions, etc. This layer lies across *level 1* and *level 2* of situation assessment, as it is building knowledge in terms of attributes and their values and also extracting some meaning for future use and interaction.

(iii) Recognize: Dedicated to recognizing what has been 'cognized' earlier by the system, e.g. a place, face, word, meaning, emotion, etc. This mostly belongs to *level 2* of situation assessment, as it is more on utilizing the knowledge either learned or provided a priori, hence 'experience' becomes the dominating factor.

Table 1. Identification and Classification of the key situation assessment components



(I) Perception of Human	
(i) People Presence	(ix) Perspective Taking
(ii) Face Detection	(x) Emotion Recognition
(iii) Face Characteristics	(xi) Speaker Localization
(iv) Gaze Analysis	(xii) Speech Recognition
(v) Face Recognition	(xiii) Speech Rhythm Analysis
(vi) Face and Person Tracking	(xiv) User Attention Detection
(vii) Posture Characterization	(xv) User Profile Analysis
(viii) Waving Detection	(xvi) Intention Analysis

(II) Perception of Robot Itself
(i) Battery Status
(ii) Body Temperature
(iii) Foot Status
(iv) Robot Posture
(v) Fall Detection
(vi) Self Collision Detection

(III) Perception of Object	(IV) Perception of Environment	(V) Perception of Stimuli
(i) 3D Segmentation	(i) Landmark Detection	(i) Sound Detection
(ii) Barcode Reader	(ii) Darkness Detection	(ii) Chest Button Interpretation
(iii) Close Object Detection	(iii) Place Recognition	(iii) Movement Detection
(iv) Object Recognition	(iv) Location Tracker	(iv) Sound Localization
(v) Object Tracker	(v) Sound Tracker	(v) External Collision Detection
(vi) Semantic perception	(vi) Semantic Perception (place)	(vi) Contact Observer

(iv) Track: This layer corresponds to the requirement to track something (sound, object, person, etc.) during the course of interaction. From this layer, *level 3* of situation assessment begins, as tracking allows to update in time the state of the beforehand entity (person, object, etc.), hence involves a kind of 'projection'.

(v) Interact: This corresponds to the high-level perception requirements for interaction with the human and the environment. E.g. activity, action and intention prediction, perspective taking, social signal and gaze analyses, semantic and affordance prediction (e.g. pushable objects, sitable objects, etc.). It mainly belongs to *level 3* of situation assessment, as involves 'predicting' side of perception.

Sometimes, practically there are some intermediate loops and bridges among these layers, for example a kind of loop between tracking and recognition. Those are not shown for the sake of making main idea of the architecture better visible.

Note the *closed loop* aspect of the architecture from *interaction* to *sense*. As shown in some preliminary examples in section 3, such as *Ex1*, we are able to practically achieve this, which is important to facilitate natural human-robot interaction process, which can be viewed as: *Sense → Build knowledge for interaction → Interact → Decide what to sense → Sense →...*

2.2 Basic Requirements, Key Attributes and Developments

In Romeo2 project, we have identified the key attributes and elements of situation assessment, to be perceived from companion robotics domain perspective, and categorized along five basic requirements as summarized in table 1. In this section, we describe some of those modules. See Naoqi [18] for details of all the modules.

I. Perception of Human

People presence: Perceives presence of people, assign unique ID to each detected person. **Face characteristics**: To predict age, gender and degree of smile on a detected face. **Posture characterization (human)**: To find position and orientation of different body parts of the human, shoulder, hand, etc. **Perspective taking**: To perceive reachable and visible places and objects from the human's perspective, with the level of effort required to see and reach. **Emotion recognition**: For basic emotions of anxiety, anger, sadness, joy, etc. based on multi-modal audio-video signal analysis. **Speaker localization**: Localizes spatially the speaking person. **Speech rhythm analysis**: Analyzing the characterization of speech rhythm by using acoustic or prosodic anchoring, to extract social signals such as engagement, etc. **User profile**: To generate emotional and interactional profile of the interacting user. Used to dynamically interpret the emotional behavior as well as to build behavioral model of the individual over a longer period of time. **Intention analysis**: To interpret the intention and desire of the user through conversation in order to provide context, and switch among different topics to talk. The context also helps other perception components about what to perceive and where to focus. Thus, facilitates closing the interaction-sense loop of fig. 3.

II. Perception of Robot Itself

Fall detection: To detect if the robot is falling and to take some human user and self-protection measures with its arms before touching the ground.

Other modules in this category are self-descriptive. However it is worth to mention that, such modules also provide symbolic level information, such as *battery nearly empty*, *getting charged*, *foot touching ground*, symbolic posture *sitting*, *standing*, *standing in init pose*, etc. All these help in achieving one of the aims of Romeo2 project: sensing for natural interaction with human.

III. Perception of Object

Object Tracker: It consists of different aspects of tracking, such as moving to track, tracking a moving object and tracking while the robot is moving. **Semantic perception (object)**: Extracts high-level meaningful information, such as object *type* (*chair*, *table*, etc.), *categories and affordances* (*sitable*, *pushable*, etc.)

IV. Perception of Environment

Darkness detection: Estimates based on the lighting conditions of the environment around the robot. **Semantic perception (place)**: Extracts meaningful information from the environment about places and landmarks (a kitchen, corridor, etc.), and builds topological maps.

V. Perception of Stimuli

Contact observer: To be aware of desired or non-desired contacts when they occur, by interpreting information from various embedded sensors, such as accelerometers, gyro, inclinometers, joints, IMU and motor torques'.

3 Results and Discussion on Innovation Potentials

We will not go in detail of the individual modules and the results, as those can be found online [18]. Instead, we will discuss some of the advantages and innovation potentials, which such modules functioning on a unified platform could bring.

Ex1: The capability of multi-modal perception, combining input from the interacting user, the events triggered by other perception components, and the centralized memorization mechanism of robot, help to achieve the goal of closing the interact-sense loop and dynamically shaping the interaction.

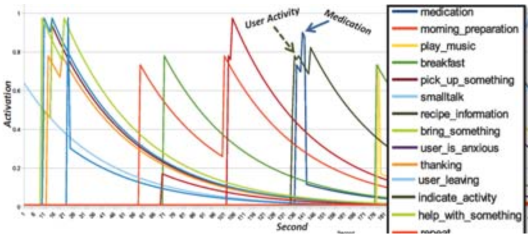
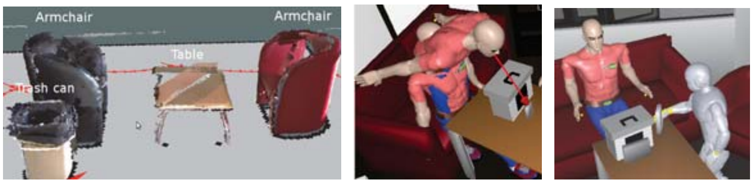


Fig. 4. Subset of interaction topics (right), and their dynamic activation levels based on multi-modal perception and events.



(a) (b) (c)

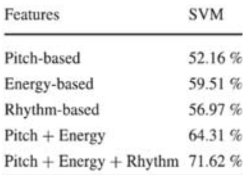
Fig. 5. High-level situation assessment. (a) The semantics map of the environment. (b) Effort and Perspective taking based situation assessment. (c) Combining (a) and (b), the robot will be able to make the object accessible to the human.

To demonstrate, we programmed an extensive dialogue with 26 topics that shows the capabilities of the Romeo robot. During this dialogue the user often interrupts Romeo to quickly ask a question, this leads to several 'conflicting' topics in the dialogue manager. The activation of different topics during an interaction over a period is shown in fig. 4. The plot shows that around *136th* second the user has to take his medicine, but the situation assessment based memory indicates that the user has ignored and not yet taken the medicine. Eventually, the system results the robot urging the user to take his medication (pointed by blue arrow), making it more important than the activity indicated by the user during the conversation (to engage in reading a book, pointed by dotted arrow in dark green). Hence, a close loop between the perception and interaction is getting achieved in a real time, dynamic and interactive manner.

Ex2: Fig. 5(a) shows situation assessment of the environment and objects at the level of semantics and affordances, such as there is a 'table' recognized at position X, and this belongs to an affordance category on which something can be put. Fig. 5(b) shows situation assessment by perspective taking, in terms of abilities and effort of the human. This enables the robot to infer that the sitting human (as shown in fig. 5(c)) will be required to stand up and lean forward to see and take the object behind the box. Thanks to the combined reasoning of (a) and (b), the robot will be able to make the object accessible to the human by placing it on the table (knowing that something can be put on it), at a place reachable and visible by the human with least effort (through the perspective taking mechanism), as shown in fig. 5(c).

In Romeo2 we also aim to use this combined reasoning about abilities and efforts of agents, and affordances of the environment, for autonomous human-level understanding of task semantics through interactive demonstration, for the development of robot's proactive behaviors, etc. as suggested the feasibility and advantages in some of our complementary studies in those directions, [19], [20].

Ex3: Analyzing verbal and non-verbal behaviors such as head direction (e.g. on-view or off-view detection) [15], speech rhythm (e.g. on-talk or self-talk) [22], laugh detection [8], emotion detection [24], attention detection [23], and their dynamics (e.g. synchrony [7]), combined with acoustic analysis (e.g. spectrum) and prosodic analysis altogether greatly allows to improve social engagement characterization of the human during interaction.



Features	SVM
Pitch-based	52.16 %
Energy-based	59.51 %
Rhythm-based	56.97 %
Pitch + Energy	64.31 %
Pitch + Energy + Rhythm	71.62 %

Fig. 6. Self-talk detection

To demonstrate, we collected a database of human-robot interaction during sessions of cognitive stimulation. The preliminary result with *14* users shows that on a *7* level evaluation scheme, the average scores for questions, ”*Did robot show any empathy?*”, ”*Was it nice to you?*” and ”*Was it polite?*” were *6.3*, *6.2* and *6.4* respectively. In addition, the multi-modality combination of the rhythmic, energy and pitch characteristics seems to be elevating the detection of self-talk (known to reflect the cognitive load of the user, especially for elderly) as shown in table of fig. 6.

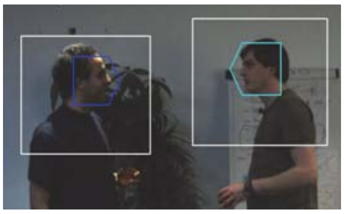


Fig. 7. Face, shoulder and face orientation detection of two interacting people.

Ex4: Inferring face gaze (as illustrated in fig. 7), combined with sound localization and object detection, altogether provides enhanced knowledge about who might be speaking in a multi-people human-robot interaction, and further facilitates analyzing the attention and intention. To demonstrate this, we conducted an experiment with two speakers, initially speaking at the different sides of the robot and then slowly moving towards each other and eventually separate away. Fig. 8 shows the preliminary result for the sound source separation by the system based on beamforming. The left part (BF-SS) shows when only the audio signal is used. When the system uses the visual information combined with the audio signals, the performance is better (AVBF-SS) in all the three types of analyses: signal-to-interference ratio (SIR), signal-to-distortion ratio(SDR) and signal-to-artifact (SAR) ratio.

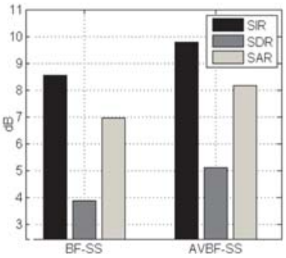


Fig. 8. Sound source separation, only audio based (BF-SS) and audio-video based (AVBF-SS).

Ex5: The fusion of rich information about visual clues, audio speech rhythm, lexical content and the user profile is also opening doors for automated context extraction, helping for better interaction and emotion grounding and making the interaction interesting, like doing humor [13].

4 Conclusion and Future Work

In this paper, we have provided an overview of the rich multi-modal perception and situation assessment system within the scope of Romeo2 project. We have presented our sensing-interaction perception architecture and identified the key perception components requirements for companion robot. The main novelty lies in the provision for rich reasoning about the human and practically closing the sensing-interaction loop. We have pointed towards some of the work in progress innovation potentials, achievable when different situation assessment components are working on a unified theoretical and functional framework. It would be interesting to see how it could serve as guideline in different context than companion robot, such as robot co-worker.

References

1. Beck, A., Risager, C., Andersen, N., Ravn, O.: Spacio-temporal situation assessment for mobile robots. In: Int. Conf. on Information Fusion (FUSION) (2011)

2. Bolstad, C.A.: Situation awareness: Does it change with age. vol. 45, pp. 272–276. Human Factors and Ergonomics Society (2001)
3. Buendia, A., Devillers, L.: From informative cooperative dialogues to long-term social relation with a robot. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 135–151 (2014)
4. Caron, L.C., Song, Y., Filliat, D., Gepperth, A.: Neural network based 2d/3d fusion for robotic object recognition. In: Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2014)
5. Chella, A.: A robot architecture based on higher order perception loop. In: Brain Inspired Cognitive Systems 2008, pp. 267–283. Springer (2010)
6. CHRIS-Project: Cooperative human robot interaction systems. http://www.chrisfp7.eu/
7. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. Affective Computing, IEEE Transactions on 3(3), 349–365 (July 2012)
8. Devillers, L.Y., Soury, M.: A social interaction system for studying humor with the robot nao. In: ICMI. pp. 313–314 (2013)
9. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors: Journal of the Human Factors and Ergonomics Society 37(1), 32–64 (1995)
10. EYESHOTS-Project: Heterogeneous 3-d perception across visual fragments. http://www.eyeshots.it/
11. Filliat, D., Battesti, E., Bazeille, S., Duceux, G., Gepperth, A., Harrath, L., Jebari, I., Pereira, R., Tapus, A., Meyer, C., Ieng, S., Benosman, R., Cizeron, E., Mamanna, J.C., Pothier, B.: Rgbd object recognition and visual texture classification for indoor semantic mapping. In: Technologies for Practical Robot Applications (2012)
12. Jensen, B., Philippsen, R., Siegwart, R.: Narrative situation assessment for human-robot interaction. In: IEEE ICRA. vol. 1, pp. 1503–1508 vol.1 (Sept 2003)
13. JOKER-Project: Joke and empathy of a robot/eca: Towards social and affective relations with a robot. http://www.chistera.eu/projects/joker
14. Lallee, S., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., Skachek, S., van Der Zant, T., Warneken, F., Dominey, P.F.: Towards a platform-independent cooperative human-robot interaction system: I. perception. In: IEEE/RSJ IROS. pp. 4444–4451 (Oct 2010)
15. Le Maitre, J., Chetouani, M.: Self-talk discrimination in human-robot interaction situations for supporting social awareness. J. of Social Robotics 5(2), 277–289 (2013)
16. Lee, S., Baek, S.M., Lee, J.: Cognitive robotic engine: Behavioral perception architecture for human-robot interaction. In: Human Robot Interaction (2007)
17. Mekonnen, A.A., Lerasle, F., Herbulot, A., Briand, C.: People detection with heterogeneous features and explicit optimization on computation time. In: ICPR (2014)
18. NAOqi-Documentation: https://community.aldebaran-robotics.com/doc/2-00/naoqi/index.html/
19. Pandey, A.K., Alami, R.: Towards human-level semantics understanding of human-centered object manipulation tasks for hri: Reasoning about effect, ability, effort and perspective taking. Int. J. of Social Robotics pp. 1–28 (2014)
20. Pandey, A.K., Ali, M., Alami, R.: Towards a task-aware proactive sociable robot based on multi-state perspective-taking. J. of Social Robotics 5(2), 215–236 (2013)
21. Pomerleau, D.A.: Neural network perception for mobile robot guidance. Tech. rep., DTIC Document (1992)
22. Ringeval, F., Chetouani, M., Schuller, B.: Novel metrics of speech rhythm for the assessment of emotion. Interspeech pp. 2763–2766 (2012)
23. Sehili, M., Yang, F., Devillers, L.: Attention detection in elderly people-robot spoken interaction. In: ICMI WS on Multimodal Multiparty real-world HRI (2014)
24. Tahon, M., Delaborde, A., Devillers, L.: Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices. In: INTERSPEECH. pp. 3121–3124 (2011)

Information for Cognitive Agents

Nir Fresco

The Edelstein Centre,
The Hebrew University of
Jerusalem, Israel
fresco.nir@gmail.com

Abstract. Humans use information in everyday activities, including learning, planning, reasoning and decision-making. There is broad agreement that, in some sense, human cognition involves the processing of information, and, indeed, many psychological and neuroscientific theories explain cognitive phenomena in information-theoretic terms. However, it is not always clear which of the many concepts of 'information' is the one relevant to understanding the nature of human cognition. Here, I suggest that information should be understood pragmatically. Whatever the criteria for information are, what makes some *x* informational has to do with how an agent either processes or can process *x*. Information is defined as meaningful structured representations of perceptual data. Their meaningfulness is determined by their behavioural effect on the agent.

Keywords: Cognition; Information; Behavioural effect; Data; Cognitive Science.

1 Introduction

There is broad agreement that, in some sense, human cognition involves the processing of information. Humans regularly use information in learning, planning, reasoning and decision-making. Many theories in cognitive science explain cognitive phenomena in information-theoretic terms. Yet, 'information' means many things to many people. So, it is not always clear which of the many concepts of 'information' is the one relevant to understanding the nature of human cognition. C. Shannon and W. Weaver defined information-content as the probability of a message being selected from a finite set of messages with any selection being equally probable [1]. R. V. L. Hartley before them had developed measures for the capacities of different types of information systems to transmit information [2]. More recently, Kolmogorov Complexity has defined the information-content in a binary string *s* as the length of the shortest program that produces *s* on a universal Turing machine [3, 4].

However, all these offer quantitative analyses of information for measuring the information-content in a message, rather than a theory of information as the *thing* that is to be measured. As noted by Hartley, Shannon and Weaver, their theories focused on physical features of signal communication, rather than the psychological or semantic features of information. Whilst quantitative aspects of information-content are clearly of importance to an information-theoretic analysis of cognition, it seems crucial to fix the concept of semantic 'information' that is used by information theories of cognition in artificial intelligence and cognitive science broadly. In the next section, I survey a few of the well-known theories of semantic information and point out their deficiencies as the basis for informational theories of cognition.

In this paper, I suggest that information should be understood pragmatically first and foremost, if we are to understand human cognition information-theoretically. Whatever the criteria for information are, what makes data informational (for an agent) has to do with how the agent either processes or can process these data. (Here, I adopt L. Floridi's *data-oriented* definition of information [5] with important modifications as is discussed below.) Information should be best understood as meaningful structured representations of perceptual data as is discussed in Section 3. The meaningfulness of perceived data is determined by their behavioural effect on the agent as a triadic, rather than dyadic, relation involving a physical object (or event or property or state of affairs), the agent's neural state and the behavioural effect on the agent. The account sketched here resembles other neo-Peircean analyses of representation [6, 7] as well as more recent accounts of information [8, 9]. The relationships between the present account and other neo-Peircean analyses are discussed in Section 4. Section 5 concludes the paper with some general reflections.

2 A brief survey of accounts of semantic information

An important principle underlying many probabilistic accounts of semantic information had been originally formulated by K. Popper. "[T]he amount of empirical information conveyed by a [set of sentences...] increases with its degree of falsifiability" [10]. This principle was later coined the *Inverse Relationship Principle* (IRP): the less likely a message is, the more informative (or rather *informational*) it is [11]. The first systematic theory of semantic information based on IRP was formulated by Y. Bar-Hillel and R. Carnap [12]. According to this theory, the thing that carries information or has informational content is *sentences*. The *meaningfulness* of information is relative to some logical probability space. Information is assigned to messages about events and the selected information measure depends on the logical probability of events or some properties of an object the message is about. Logical probability is defined in this context as a function of the set of possible worlds a sentence rules out.

Some have argued that this theory (and any other IRP-based theory) leads to a paradoxical result [5, 13]. If all the consequences of known sentences are known, any *logically* true sentence (that is, a tautology) does not increase knowledge and, hence, does not contain information. A tautology excludes no possible worlds and its logical probability is 1. At the same time, a self-contradictory sentence excludes all possible worlds and its logical probability is 0. Counter intuitively it contains *maximal* information. I return to this so-called paradox below, but for now, it should be noted that the Bar-Hillel/Carnap theory cannot serve as a basis for human cognition broadly. For it is defined in terms of sentences, and the domain of cognition is broader than language processing alone.

A more recent theory of information was offered by F. Dretske [14]. His theory is premised on the idea that information can be used as part of a reductive analysis of knowledge and cognition. On his view, the information carried by a message is relative to the epistemic state of the agent receiving that message. He was motivated by the central observation in the Shannon/Weaver theory that the receipt of information should reduce the agent's uncertainty. By applying the underlying communication model in the Shannon/Weaver theory to knowledge, the source of messages is the physical world and the receiver is a would-be knower. For Dretske, perceptual knowledge can (and should) be understood in terms of information. "*K* knows that *s* is *F* = *K*'s belief that *s* is *F* is caused (or causally sustained) by the information that *s* is *F*" [14]. The information that *s* is *F* affects *K*'s belief in such a way that the information suffices for the formation of the belief absent other contributing (or conflicting) factors. *K* must discern physical events in the world that carry the particular information, and those events have to cause (or causally sustain) *K*'s belief that *s* is *F*. Moreover, the informational content of a message is *also* conditional on what *K* already knows when receiving the message. Importantly, Dretske maintained that information must be *truthful*. "Information is what is capable of yielding knowledge, and since knowledge requires truth, information requires it also" [14]. Other supporters of the idea that information must be truthful include P. Grice [15], J. Barwise [11] and P. Allo [13].

Floridi has adopted some of Dretske's main ideas (including the idea that information cannot be false), whilst rejecting IRP and insisting on a stronger constraint on semantic content. His two main motivations for adopting the Veridicality Thesis (i.e., that information must be truthful) are (a) to provide a link between information and *knowledge*, and (b) to avoid the Bar-Hillel/Carnap paradox concerning the alleged informativeness of contradictions [5]. The first motivation is similar in spirit to Dretske's in establishing a close link between knowledge and information. The second motivation – being that tautologies contain *no* information, whereas contradictions contain *maximum* information (an underlying principle of classical logic) – has led him to deny IRP and suggest a stronger constraint that is based on closeness to truth. According to Floridi, "the amount of informativeness of each [message] can be evaluated *absolutely*, as a function of (a) […] the alethic value possessed by [the message] and (b) the degree of discrepancy […] between [the message] and a given state of the world" [5]. (Note the difference from Dretske's approach where information is conditional on the epistemic state of the receiver.)

Yet, besides the veridicality constraint, he proposes to understand information as *meaningful* and *structured* data. Unlike the Bar-Hillel/Carnap theory, information carriers are understood as *data* rather than *sentences* only. What is a datum? In its simplest form, it is the *lack of uniformity* in the real world. Examples of a datum include a black dot on a white page, the presence of some noise, a light in the dark or a logical 0 as opposed to a 1. A datum is defined as two distinct uninterpreted variables in a domain that is left open to further interpretation [5]. Data are *structured* when they are "rightly put together, according to the rules (syntax) that govern the chosen system, code or language being used. Syntax here must be understood broadly, not just linguistically" [16]. That they are meaningful means that the data "must comply with the meanings (semantics) of the chosen system, code or language in question. […] The data constituting information can be meaningful independently of an informee [and need not be] necessarily linguistic" [16].

There are clearly other important theories of information that are worth exploring, but this exceeds the scope of this paper. For example, D. MacKay offered a quantitative theory of semantic information based on the receiver's increase in knowledge. "[W]e have gained information when we know something now that we didn't know before; when 'what we know' has changed" [17]. Another example is B. Skyrms' analysis of information – grounded in signalling games – where senders of signals observe states of the world and communicate with receivers that in turn choose an act in response to receiving signals [18]. For him, information is correlated with states of the world as well as with actions.

3 Towards a theory of semantic information as meaningful structured representations of data

Space only permits a few, brief remarks regarding the adequacy of the theories of information outlined in Section 2. (This is discussed elsewhere [19].) The Bar-Hillel/Carnap theory of information is defined in terms of sentences, and, thus, is unable to account for many non-linguistic informational aspects of cognition. Dretske and Floridi's accounts of information aim specifically at explaining knowledge. Yet, that objective has led them to adopt the Veridicality thesis that restricts the applicability of information to other cognitive phenomena. Cognitive agents cannot always ascertain the veracity of the information they process and one of the most important methods of learning is by trial and error that clearly involves making mistakes (or false information). The processing of information in cognitive agents is insensitive to the veridicality of the information. Belief change models, for example, explain rationality is terms of *justified* doxastic commitments that are *consistent*. These models are underpinned by the principle that all information, even veridical information, is defeasible and subject to revision under the right conditions. Besides, on standard frameworks of belief change, false perceptual information can actually lead to truth approximation via belief revision and increase the agent's overall knowledge base.

To underscore the *pragmatic* value of information for the receiving agent consider a simple example. Suppose that the *same* message is sent twice by the same information source. The two messages clearly carry the *same* information-content. Nevertheless, only the message that is *successfully received* by the receiver *first* is informative. Of course, receiving the second message – with the very same information content – can still be useful, for example, in the presence of noise: the first message could have been distorted during transmission. Moreover, in some contexts, each of the messages, arguably, carries additional meta-information that is its temporal indexing: message one was sent (or rather received) at T_x and the second at T_y. This temporal indexing might also be pragmatically significant: it may tell the receiver that some state of the information source has remained unchanged. Nevertheless, all this is meta-information in addition to the information-content of each of the individual messages (e.g., if each message includes a timestamp as part of its content, the information-content of the two messages is different).

Crucial to the new theory sketched herein is the triadic basis of information. Rather than taking information to be a dyadic relation that obtains between signs and objects (or states of affairs) in the world, information requires a third element: its receiver. On Floridi's theory, for example, some information (i.e., *environmental* information) supposedly exists in the world independently of any receivers (e.g., concentric rings in the trunk of a tree that can be used to calculate the tree's age qualify as information even in the absence of any perceiver) [5]. But as argued by Dretske, the informativeness of a message is relative to the epistemic state of the receiving agent. Smoke in the forest (usually reliably) signifies there being fire to receivers of information that interpret the signals (smoke particles or combustion aerosols) as a potential imminent danger nearby. This triadic relation can already be found in the works of C. S. Peirce: something is a *sign* (also "*representamen*") only if it signifies an *object* with respect to an "*interpretant*" (i.e., a mediating representation in the mind of some agent) [20]. Whilst there is a causal correlation between smoke and fire based on natural regularities, the receiver of the signals (smoke particles) plays a key role in the formation of the information (*there being fire in the forest*). The receiver may know that smoke machines are used in the forest (for some bizarre reason) and, consequently, may not interpret the signals received as there being fire in the forest.

The theory proposed here uses Floridi's data-oriented definition of information with some important modifications. Objects, events or states of affairs in the world are sources of physical *signals* or *data* with which they are causally correlated. Physical data as *discontinuities in the world* exist "out there" *un*structured. Their structuring is an ongoing dynamic interaction between the *receiving agent* and her *environment*. But data need not always originate externally to the receiver. An organism, for example, can receive pain signals from one of its limbs. Further, the structure of the data in the wild is determined by an agent-environment function. If either of these two contributing factors is missing, there *is no* information just data. In that sense, the physical data "out there" constrain the information that can be formed by the receiver on their basis. Unless the agent is hallucinating in a void or dreaming, her perceptions are formed on the basis of stimuli (understood as data) from the world to which she is sensitive. Our cognitive apparatus only allows us to discriminate *some*, but *not all*, physical discontinuities and nomic regularities in the world. (Whilst elephants, for example, are sensitive to infrasound, humans are not readily sensitive to infrasound signals.) Only those data to which we are cognitively sensitive can give rise to the formation of information. Any perceived physical data "out there" are encoded, or represented, as some form of neural patterns (e.g., as action potentials or activation patterns). The precise form of representation is a further empirical question.

The *meaningfulness* of the perceptually structured data is determined by their behavioural effect (either positive or negative) on the receiver. Such behavioural effect is broadly construed to encompass more than just *observable* behaviour. It amounts to, roughly, the change produced in the receiver's action(s), belief(s) or goal(s) resulting from the data perceived (e.g., leaving the forest immediately when smelling or seeing smoke on a very hot day). In that sense, the state of the world – as signified by the perceived data – and the receiver are connected. This change implies, as argued in [21], that there exists a requisite flexibility of behaviour in the receiver, such that the perceived data *can* yield some change in the receiver. It makes little or no sense to describe a rigid system *S* as being *informed* by something if *S* cannot somehow behave differently upon receiving these data. Further, any consequence of the perceived data is the result of how the receiver interprets the data and behaves in the world accordingly [22]. However, for the perceived data to be *meaningful* there need not be any necessary dependence on a kind of coordination system amongst *senders* and receivers. Data need not be communicated amongst agents in order to *be meaningful*, and can flow directly from the world to the receiver [21]. Indeed, the world does not *communicate* with agents. It is rather the sensitivity of the receiver to particular regularities or physical discontinuities in the world that "flow" to the receiver.

Moreover, the effect concerned need not be necessarily positive (e.g., the receiver being informed about a nearby reservoir of water); it can often be negative (e.g., drawing a false conclusion regarding the distance of the reservoir). The distinction between negative and positive effects is what determines the *relevance* of information, as argued by D. Wilson and D. Sperber [23], not whether the meaningful structured data *qualify as information*. On their view, information is relevant to the agent when it (1) relates to her *background* information to derive conclusions that matter for her beliefs or actions, and (2) requires less processing effort by the agent. Others define the relevance of information relative to goals. A piece of information is relevant (for a goal) *iff* "it is a candidate for a belief that supports the processing of that goal" [24]. But either way, the relevance of information can only be determined once we have established what qualifies as information. The meaningfulness of the perceptual data is a prerequisite for the information *being* relevant. Understood this way, there is clearly room for mistakes (as a negative effect) in the agent forming information. An agent may mistake smoke particles for indicators of fire nearby, where, as a matter of fact, that smoke may be produced by smoke machines. Her escape from the forest would be rationally justified absent other overriding factors, despite there being no fire or imminent danger.

The theory proposed herein postulates that there is an important distinction to be made between *information-that* and *information-how* on the basis of the role information plays in cognitive processing. Information-how (e.g., 'In case of fire, break the glass and press the button') is prescriptive and informs an agent about which action has to be performed to achieve a particular result. As such, for cognitive agents it expresses an expectation for some goal-directed action on the part of the receiver in a given context. Information-that (e.g., 'Not all birds can fly') is descriptive and is about events, objects and states of affairs in the world. Cognitive agents use information-that to represent and form beliefs about, rather than merely externally react to, their environment. Both types of information play an important role in the way cognitive agents negotiate with their environment in terms of acting and believing. Neither information-how nor information-that need be restricted to *sentences*.

Lastly, why is this particular view of information considered apt to capture the kind of information processing often invoked in cognitive science? First, understanding information as being carried by data allows a broader applicability of the theory beyond linguistic aspects of cognition alone. To understand cognitive agency, what we want is a theory that focuses on *physical* information, and in that regard data-centred theories fare better. Sentences convey information, but so do sunlight and smoke, for example. Yet, unlike the Floridian data-centred theory of information, the present theory does not insist on the Veridicality thesis. Cognitive agents all too often make mistakes in interpreting perceptual data. Such mistakes should also be accounted for in explaining cognition. Second, information in cognitive science provides a naturalistic foundation for the explanation of cognition and behaviour. Humans and other organisms survive and reproduce by tuning themselves to reliable but imperfect cues that represent correlations between internal variables and environmental stimuli as well as between environmental stimuli and opportunities and threats [25]. The *meaningfulness* of perceived data described above is determined precisely by such "reliable but imperfect cues" the agent is sensitive to.

Third, the theory is neither too narrow nor too broad for our purposes. It is not too narrow in either imposing strict conditions that only few cognitive processes satisfy (e.g., the veridicality of the data for knowledge) or being limited to a subset of cognitive phenomena (e.g., language processing). It is compatible with the contemporary cognitive scientific view that "the brain reveals itself proactive in its interface with external reality" being an *interpreter* rather than a mirror of that reality [26]. "[R]esearch [...] has shown how signals coding predictions about [...] simple features of relevant events can influence several stages of neural processing" [26]. The proposed theory is equally compatible, for example, with a recent, and contentious, view of the brain as an hypothesis-testing mechanism that attempts to minimise the error of its predictions about perceptual data from the world [27]. Both "bottom-up" signals (perceptual input data) and "top-down" signals embodying predictions about the probable causes of the perceptual input data can qualify as information according to our theory. At the same time, the theory is not too broad so as to make information vacuous. Information can come at degrees. Some data do not give rise to information, since the receiver is not sensitive to them. Other data are simply not meaningful to their receiver. And although both a tautology and a contradiction, for example, can be informational, they are less or more *useful* and/or *relevant* in a given context.

4 A comparison with other neo-Peircean theories

In this section, the relationships between the proposed theory and other neo-Peircean analyses of representation and information are discussed. To begin with, consider B. von Eckardt's analysis of non-mental representation. In [6] she adapts Peirce's triadic relation that obtains amongst the represented object, the representing vehicle (*representamen*) and the mental effect in the mind of the interpreter of the sign (*interpretant*). The represented object could be a physical object, a relation, a state of affairs or a property. The representing vehicle – what she calls the *representation bearer* – such as a map, a photo or a spoken word, can be individuated in terms of its nonrepresentational (or material) properties. Both the represented object and the representation bearer are, at least in principle, objectively verifiable. von Eckardt claims that in order for *R* to be an *actual* – rather than merely a possible – representation there must currently exist an actual interpreter bearing the right relation to *R*. The resemblance to the proposed theory of information should be clear. Information is understood pragmatically and in a manner that requires an actual consumer of physical data (that can be upgraded to information under the right conditions). On the other hand, data need not be communicated by *senders*. Physical data "out there" can at best be classified as *potential* information in the absence of consumers.

G. O'Brien and J. Opie build on von Eckardt analysis of non-mental representation and add that the vehicles of mental representation should be understood as some kind of neural states [7]. Given their commitment to a naturalistic account of cognition, they seek to explain the act of interpretation in *naturalistic* terms in order to avoid a vicious circle. They claim that the only viable alternative is treating interpretation in terms of some modification of the cognitive agent's behavioural dispositions towards the represented object. Here, too, the similarity is clear. The proposed theory of information suggests that the meaningfulness of perceived data (and, therefore, their being *informational*) is determined by their behavioural effect on the agent. It is suggested that on receiving new information some effect in the agent triggers an action or a response (e.g., forming/changing a belief-state).

On E. Jablonka's functional-evolutionary analysis of semantic information, the distinction suggested above between information-that and information-how becomes very blurry. That is the case, for example, when 'functional' means that signals received by either a human- or natural-selection designed system play a causal role that "usually contributes to the goal-oriented behavior of this system" [9]. An apple pie recipe and a piece of software are instances of functional-evolutionary information for a cook and a computer, respectively, in a manner akin to the appearance of black cloudy sky leading to the shelter-seeking action of an observing ape. Nevertheless, insofar as we seek to understand the role information processing plays in cognitive tasks in the *lifetime* of an agent, rather than over evolutionary time, the information-how/information-that distinction seems worth preserving.

Lastly, on J. Queiroz, et al. neo-Peircean theory, information has the nature of a process of communicating a "form" to the interpretant [8]. That process constrains the possible patterns of behaviour of the interpreter. Information is taken typically as an interpreter-dependent "objective" process. Accordingly, it cannot be dissociated from a situated agent. On their view, it is only as a result of the interpretation process that information triadically connects the sign, object(s), and an effect on the interpreter. A sign (somehow) effectively communicates a form from the (represented) object to the interpretant, whilst changing the state of the interpreter. This account raises some interesting questions, which are not tackled here, about the objectivity of this process when it is dependent on a particular agent and about the communication of the form of an object to the interpreter (the world does not talk to us…). Nevertheless, it can be seen again that information is not simply "out there" in the world independently of a perceiver. Information is a dynamic construct that results from an ongoing interaction between the agent and its environment.

5 Concluding remarks

This short paper contributes to a long-standing and much-debated question of what concept of 'information' is suitable for understanding human cognition in terms of information processing. It is often argued, in cognitive science, that cognition is an information processing system. The literature contains many diverse theories of information (of which I have surveyed but a few here) pulling in different directions, thereby leading to disparate definitions of 'information'. Information, so I have suggested whilst adapting a neo-Peircean approach, should be understood pragmatically. Whatever the criteria for information are, what makes *x* a piece of information has to do with the way the agent either processes or can process *x* in actively engaging with her environment. Of course, it does not follow that a unified theory of information is either forthcoming or even possible. In different contexts, such as game theory or economics, information may be defined differently. The theory proposed herein is motivated by doing justice to the cognitive sciences. However, much more work is required to fully develop it.

Acknowledgements. This paper has benefited from the insightful comments and suggestions of two anonymous referees for the 2014 International Workshop on Artificial Intelligence and Cognition. Thanks are due to Aditya Ghose, Patrick McGivern, Michaelis Michael and Joel Pearson for useful discussions on topics related to the paper. This research has been supported by a fellowship from the Edelstein Centre for the History and Philosophy of Science, Technology and Medicine and a research grant from the Israeli Ministry of Aliyah and Immigrant Absorption. My attendance at the workshop was funded in part by the Rosselli Foundation.

References

1. Shannon, C.E., Weaver, W.: The mathematical theory of communication. University of Illinois Press, Urbana (1949).
2. Hartley, R.V.L.: Transmission of Information. Bell Syst. Tech. J. 7, 535–563 (1928).
3. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Probl. Inf. Transm. 1, 1–7 (1965).
4. Chaitin, G.J.: Algorithmic information theory. Cambridge University Press, Cambridge, UK (2004).
5. Floridi, L.: The philosophy of information. Oxford University Press, Oxford (2011).
6. Von Eckardt, B.: What is cognitive science? MIT Press, Cambridge, Mass. (1993).
7. O'Brien, G., Opie, J.: Notes toward a structuralist theory of mental representation. In: Staines, P.J., Clapin, H., and Slezak, P.P. (eds.) Representation in mind : new approaches to mental representation. pp. 1–20. Elsevier: Morgan Kaufmann, Amsterdam (2004).
8. Queiroz, J., Emmeche, C., El-Hani, C.N.: A Peircean Approach to "Information" and

its Relationship with Bateson's and Jablonka's Ideas. Am. J. Semiot. 24, 75 (2008).

9. Jablonka, E.: Information: Its Interpretation, Its Inheritance, and Its Sharing. Philos. Sci. 69, 578–605 (2002).

10. Popper, K.R.: The Logic Of Scientific Discovery. Routledge, London (2002).

11. Barwise, J.: Information flow: the logic of distributed systems. Cambridge University Press, Cambridge (1997).

12. Bar-Hillel, Y., Carnap, R.: Semantic Information. Br. J. Philos. Sci. 4, 147–157 (1953).

13. Allo, P.: A Classical Prejudice? Knowl. Technol. Policy. 23, 25–40 (2010).

14. Dretske, F.I.: Knowledge & the flow of information. MIT Press, Cambridge, Mass. (1981).

15. Grice, H.P.: Studies in the way of words. Harvard University Press, Cambridge, Mass. (1989).

16. Floridi, L.: Information: a very short introduction. Oxford University Press, Oxford ; New York (2010).

17. MacKay, D.M.: Information, mechanism and meaning. MIT Press, Cambridge, MA (1969).

18. Skyrms, B.: Signals: evolution, learning, & information. Oxford University Press, Oxford (2010).

19. Fresco, N., Pearson, J.: How theories of cognition define information. (Unpublished).

20. Peirce, C.S.: On a New List of Categories. Proc. Am. Acad. Arts Sci. 7, 287–298 (1868).

21. Cao, R.: A teleosemantic approach to information in the brain. Biol. Philos. 27, 49–71 (2012).

22. Millikan, R.G.: Biosemantics. J. Philos. 86, 281–297 (1989).

23. Wilson, D., Sperber, D.: Relevance Theory. In: Horn, L.R. and Ward, G.L. (eds.) The handbook of pragmatics. pp. 607–632. Blackwell, Malden, MA (2005).

24. Paglieri, F., Cristiano Castelfranchi: Trust in Relevance. In: Ossowski, S., Toni, F., and Vouros, G. (eds.) Proceedings of AT 2012 - First International Conference on Agreement Technologies. pp. 332 – 346. CEUR-WS.org, Dubrovnik (2012).

25. Scarantino, A., Piccinini, G.: Information without truth. Metaphilosophy. 41, 313–330 (2010).

26. Nobre, A.C., Correa, A., Coull, J.: The hazards of time. Curr. Opin. Neurobiol. 17, 465–470 (2007).

27. Hohwy, J.: The predictive mind. Oxford University Press, Oxford, United Kingdom (2013).

Mining and Visualizing Uncertain Data Objects and Named Data Networking Traffics by Fuzzy Self-Organizing Map

Amin Karami[1,2] and Manel Guerrero-Zapata[1,2]

[1] Computer Architecture Department (DAC), Universitat Politècnica de Catalunya (UPC), Campus Nord, C. Jordi Girona 1-3. 08034 Barcelona, Spain
[2] amin@ac.upc.edu and guerrero@ac.upc.edu

Abstract. Uncertainty is widely spread in real-world data. Uncertain data -in computer science- is typically found in the area of sensor networks where the sensors sense the environment with certain error. Mining and visualizing uncertain data is one of the new challenges that face uncertain databases. This paper presents a new intelligent hybrid algorithm that applies fuzzy set theory into the context of the Self-Organizing Map to mine and visualize uncertain objects. The algorithm is tested in some benchmark problems and the uncertain traffics in Named Data Networking (NDN). Experimental results indicate that the proposed algorithm is precise and effective in terms of the applied performance criteria.

1 Introduction

Uncertainty is a frequent issue in data analysis. The various factors that lead to data uncertainty include: approximate measurement, data sampling fault, transmission error or latency, data integration with noise, data acquisition by device error, and so on [1] [2]. These factors produce vague and imprecise data. Visualizing uncertain data is one of the new challenges in the uncertain databases [3]. Among the many visualization techniques, the Self-Organizing Map (SOM) [4] is widely and successfully applied due to its good result. SOM is a very popular unsupervised learning algorithm based on the classical set theory. An important application of SOM is discovering the topological relationship among multidimensional input vectors and mapping them to a low dimensional output which is easy for further analysis by experts [5] [6]. The process of SOM training requires a certain and an unambiguous input data either belongs or not belong to a weight vector (cluster), where the membership evaluation is boolean. In contrast, uncertain and vague input vectors are not either entirely belong or not belong to a weight vector. A data may be considered vague and imprecise where some things are not either entirely true nor entirely false and where the some things are somehow ambiguous. For instance, fuzzy location in the right side of Fig. 1 is a way to represent the item of vague information: the object is *approximately* at position (4, 3), in which the grey levels indicate membership values with white representing 0 and black representing 1. In contrast, the left side of Fig. 1 shows the exact position of a certain data where the membership evaluation of centers (weights) is boolean. There has been a lot of research in the application of Fuzzy sets theory to model vague and uncertain information [7]. The Fuzzy set (FS) theory introduced by Zadeh [8] is a more flexible approach than classical set theory, where objects belong to sets (clusters) with certain degree of membership ranging [0..1]. This makes FS theory suitable for representing and visualizing uncertain data [9]. Therefore, a combination of SOM and FS is able to illustrate dependencies in the uncertain data sets in a very intuitive manner. SOM is indeed originally intended as a classification method, not a visualization method so there are a few additions to apply SOM for visualization. Li et al. [3] proposed a mining and visualizing algorithm for uncertain data, called USOM which combines fuzzy distance function and SOM. In this paper, we employ the FS theory in the context of SOM algorithm to mine and visualize the uncertain objects in the uncertain databases. Experimental results over four classic benchmark problems and a new network architecture as Named Data Networking (NDN) show that the proposed method outperforms standalone SOM and USOM [3] in terms of the applied performance metrics. The remainder of the paper is organized as follows: Section 2 presents self-organizing map. Section 3 presents our contribution. Section 4 evaluates the new approach experimentally. Section 5 is the conclusion and future work.

2 Self-Organizing Map (SOM)

SOM (also known as Kohonen SOM) is a very popular algorithm based on competitive and unsupervised learning [4]. The SOM projects and represents higher dimensional data in a lower dimension, typically 2-D, while preserving the relationships among the input data. The main process of SOM is generally introduced in three main phases: competition, cooperation and adaptation which are described in detail in [4].

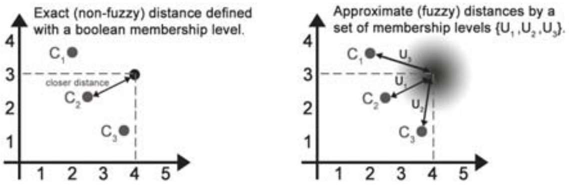

Exact (non-fuzzy) distance defined with a boolean membership level.
C1
C2
C3
closer distance
1 2 3 4 5
1 2 3 4
Approximate (fuzzy) distances by a set of membership levels {U1 ,U2 ,U3}.
C1
C2
C3
U1
U1
U2
1 2 3 4 5
1 2 3 4

Fig. 1: An example of exact (non-fuzzy) and approximate (fuzzy) distances in a 2-D space for a certain and vague data.

3 The Proposed Method

The procedure of the proposed method, application of fuzzy set theory in the context of SOM for mining and visualizing uncertainties is as follows. A diagram of the proposed method is shown in Fig. 2.

1. Fuzzy competition: in *hard competition*, the input vector is divided into distinct weights (clusters), where each input element belongs to exactly one weight. In *fuzzy competition*, input vector can belong to more than one weight, and associated with each element by a set of membership levels. Fuzzy c-means (FCM) [10] method allows one piece of input data to belong to two or more clusters (weights). The standard function is:

$$U_x = \frac{1}{\sum_j \left(\frac{d(weight_k,x)}{d(weight_j,x)}\right)^{\frac{2}{m-1}}} \tag{1}$$

   Where, U_x is the membership value of each input vector x to all weights, $j = 1, 2, ..., w$, and m is the level of cluster fuzziness which is commonly set to 2. By the fuzzy competition all the neurons are wining neurons with the membership degree ranging [0..1].
2. Fuzzy cooperation: in fuzzy cooperation, all wining neurons cooperate with their neighboring neurons in terms of the membership degree by Eq. 2. For the size of the neighborhood, we employed the Gaussian function that shrinks on each iteration until eventually the neighborhood is just the BMU itself.

$$h(j,i) = U_{xi} \times exp(\frac{-{d_{j,i}}^2}{2\sigma^2}) \quad i,j = 1,2,..,n; \;\; i \neq j \tag{2}$$

   Where, i is the number of the wining neurons including all the neurons with different membership degrees, j is the number of the cooperating neighbor neurons. U_{xi} is the membership value of input vector x from i^{th} wining neuron. $h(j,i)$ is the topological area centered around the wining neuron i and the cooperating neuron j. The size σ of the neighborhood needs to decrease with time. A popular time dependence is an exponential decay by:

$$\sigma(t) = \sigma_0 exp(\frac{-t}{\lambda}) \tag{3}$$

   Where, $\sigma(t)$ is the width of the lattice at time t, σ_0 is the width of the lattice at time t_0, and λ is the time constant.
3. Fuzzy adaption: the adaption phase is the weight update by:

$$w_j = w_j + U_j \times (\eta h(j,i) \times (x - w_j)) \quad i,j = 1,2,..,n; \;\; i \neq j \tag{4}$$

   Where, U_j is the membership value of input x from neuron j.

These three phases are repeated, until the maximum number of iterations is reached or the changes become smaller than a predefined threshold.

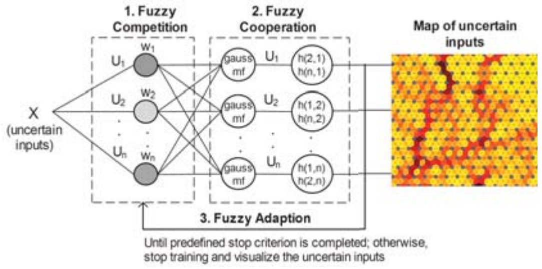


Fig. 2: The proposed method for mining and visualizing uncertainties.

4 Experimental Results

The proposed method, USOM and SOM were implemented by the MATLAB on an Intel Pentium 2.13 GHz CPU, 4 GB RAM running Windows 7 Ultimate.

4.1 Uncertain data modeling

To assess the accuracy and performance of the proposed method, four classic benchmark problems from the UCI machine learning repository [11] are applied. The selected data sets are Iris (4-D), Glass (9-D), Wine (13-D), and Zoo (17-D). In practice, uncertainties are usually modeled in the form of Gaussian distributions [2]. For some attributes in data sets, we add a Gaussian noise with a zero mean and the standard deviation with the normal distribution $[0, f]$, where, f is an integer parameter from the set of $\{1, 2, 3\}$ to define different uncertain levels.

4.2 Assessing the quality of visualizations

To assess the quality of the proposed method, several measures have been applied, including Quantization Error (QE), Topographic Error (TE), Trustworthiness of a Visualization, and Continuity of the Neighborhoods [12].

4.3 Visualization results

The experiments on each method were repeated 10 times independently. We evaluate the several SOM network structures on applied uncertain data sets which the optimal ones are Iris with 16x16 nodes, Glass with 16x16 nodes, Wine with 17x17 nodes, and Zoo with 15x15 nodes.

Table 1 shows that our proposed method outperforms SOM and USOM methods in terms of the Quantization Error (QE) and Topographic Error (TE). The proposed method seems to be more time consuming (with Exec.) than the other

Table 1: Performance improvements achieved by the proposed scheme

Data	SOM				USOM				Proposed Method			
			Time				Time				Time	
	QE	TE	Exe.	Inc.	QE	TE	Exe.	Inc.	QE	TE	Exe.	Inc.
Iris (16x16)	0.024	0.0404	9.6	5.45	0.023	0.034	11.34	4.16	0.02	0.0267	16.43	2.17
Glass (16x16)	0.066	0.0312	21.17	15.45	0.042	0.02	23.11	14.23	0.028	0.0174	26.82	10.1
Wine (17x17)	0.072	0.0381	18.87	12.05	0.06	0.022	19.12	10.16	0.049	0.0102	22.07	7.74
Zoo (15x15)	0.067	0.0215	15.54	11.23	0.046	0.016	18.51	10.36	0.039	0.0103	21.34	8.52

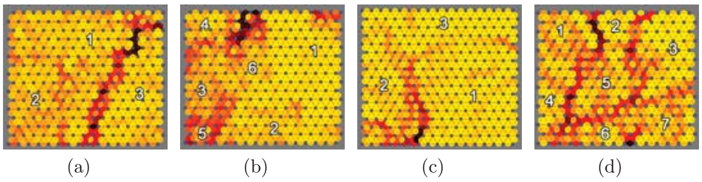
Table 2: The quality measurement by Trustworthiness

Data	SOM			USOM			Proposed Method		
	K=1	K=10	K=20	K=1	K=10	K=20	K=1	K=10	K=20
Iris (16x16)	0.937	0.94	0.93	0.95	0.962	0.968	0.962	0.968	0.974
Glass (16x16)	0.913	0.903	0.898	0.914	0.921	0.933	0.915	0.93	0.939
Wine (17x17)	0.917	0.921	0.904	0.924	0.941	0.953	0.925	0.951	0.962
Zoo (15x15)	0.957	0.96	0.96	0.962	0.963	0.968	0.963	0.97	0.978

Table 3: The quality measurement by Continuity

Data	SOM			USOM			Proposed Method		
	K=1	K=10	K=20	K=1	K=10	K=20	K=1	K=10	K=20
Iris (16x16)	0.945	0.901	0.892	0.961	0.964	0.966	0.97	0.974	0.982
Glass (16x16)	0.911	0.898	0.873	0.914	0.916	0.92	0.92	0.931	0.937
Wine (17x17)	0.921	0.892	0.883	0.93	0.931	0.935	0.935	0.939	0.941
Zoo (15x15)	0.86	0.841	0.812	0.89	0.898	0.902	0.91	0.918	0.927

methods due to the application of fuzzy set theories in the context of the SOM, in which all the neurons are winner with different membership grading. However, the proposed method can find a better solution with less times of increment on computational time (with Inc.) than the other methods due to its fast convergence speed. The trustworthiness and continuity values for K={1, 10, 20} are shown in Tables 2 and 3, respectively. The trustworthiness and continuity measures show that the proposed method obtains the better results as compared to SOM and USOM. The results show that the proposed method with the application of fuzzy set theory in the context of the SOM yields high accuracy as compared to other methods without very much computational cost. Since our proposed method performs well as compared to SOM and USOM, we visualize uncertainties in the applied uncertain data sets. To facilitate the interpretation of results, we use the U-Matrix (unified distance matrix) where visualize the high-dimensional uncertain data into a 2-D space in Fig. 3. In this figure, the blue hexagons represent the neurons (weights). The darker colors in the regions between neurons represent larger distance, while the lighter colors represent smaller distances. Fig. 3(a) shows that the constructed 4-D uncertain Iris SOM network has been clearly clustered into three distinct groups. The Glass SOM



(a) (b) (c) (d)

Fig. 3: U-Matrix of the applied benchmark problems: 3(a) Iris 16x16 SOM, 3(b) Glass 16x16 SOM, 3(c) Wine 17x17 SOM, and 3(d) Zoo 15x15 SOM.

Table 4: NDN traffic generation

Type of traffic		Frequency	Pattern
Normal (526 records)		[100..500]	Exponential
Attack (211 records)	Cache pollution	[200..800]	Locality-disruption attacks uniformly
	DoS attacks	[400..1500]	Interest flooding attacks for non-existent and existent content uniformly and exponentially

network (Fig. 3(b)) has been apparently classified 9-D uncertain data objects into six distinct types of glass. Figs. 3(c) and 3(d) show the three and the seven distinct groups of 13-D and 17-D uncertain data from Wine and Zoo data sets, respectively. The results confirm that the proposed method performs well in mining and visualizing uncertain data into somewhat expected distinct groups.

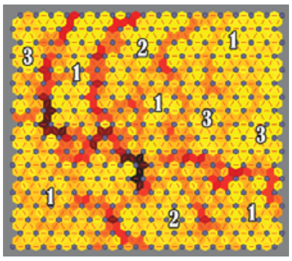


Fig. 4: U-Matrix of the NDN traffic. 1: normal, 2: DoS attack, 3: cache pollution attack.

4.4 Visualizing uncertain traffics in Named Data Networking

After evaluating the robustness and the accuracy of our proposed method with some benchmark problems, we apply the proposed method for visualizing uncertain traffics in Named Data Networking (NDN). NDN [13] is a promising network architecture being considered as a possible replacement to overcome the fundamental limitations of the current IP-based Internet. Traffic uncertainty refers to traffic volumes belong to more than one pattern, and associated with each pattern by a set of membership levels. Fuzzy approach can reduce the false positive rate with higher reliability in identifying the pattern of traffic volumes, due to any uncertain attack data may be similar to some normal patterns [14]. We conduct the same testbed configuration from papers [14] [15]. The employed features for traffic generation come from paper [14] as well as the ratio of (1) cache hit, (2) dropped Interest packet, (3) dropped data packets, (4) satisfied Interest packet, and (5) timed-out Interest packets in each 1 sec time interval. The structure of the traffic generated is shown in Table 4. We modeled uncertainties for some attributes in NDN traffic samples in the form of Gaussian distributions similar to Section 4.1. Fig. 4 maps the 11-D uncertain traffic samples to the 2-D space through our proposed method. This figure shows that the our proposed method performs somewhat well in mining and visualizing uncertainties into predefined distinct groups. Fig. 4 illustrates that there are some small groups of clustered data points with the lighter regions. These small clusters may contain some normal or attack data that try to be incorrectly placed in the neighboring regions, due to their uncertain nature. The results in Table 5 show that our proposed method offers the best performance and outperforms sufficiently other preexisting methods.

Table 5: Comparing results of visualizing NDN traffic samples

Criteria		Methods		
		SOM	USOM	Proposed Method
Quantization Error		0.042	0.029	0.0125
Topographic Error		0.074	0.053	0.031
Trustworthiness	K=1	0.91	0.95	0.968
	K=15	0.905	0.943	0.954
	K=30	0.877	0.925	0.942
Continuity	K=1	0.914	0.922	0.954
	K=15	0.893	0.931	0.941
	K=30	0.867	0.917	0.936

5 Conclusion

In this paper, we propose a new hybrid algorithm for mining and visualizing uncertain data. We investigate the implementation of fuzzy set theory in the design of SOM neural network in order to improve the accuracy of visualizing uncertain data bases. The experimental results over the uncertain benchmarking data sets and the uncertain traffics in Named Data Networking show that the proposed method is effective and precise in terms of the applied performance criteria. We plan to improve the proposed method for various uncertain models and big uncertain network traffic data in the future.

6 Acknowledgment

This work was partially supported by projects TIN2013-47272-C2-2 and SGR-2014-881.

References

1. Pavle Milošević, Bratislav Petrović, Dragan Radojević, and Darko Kovačević. A software tool for uncertainty modeling using interpolative boolean algebra. *Knowledge-Based Systems*, 62:1 – 10, 2014.
2. Jiaqi Ge, Yuni Xia, and Yicheng Tu. A discretization algorithm for uncertain data. In *Proceedings of the 21st International Conference on Database and Expert Systems Applications: Part II*, pages 485 – 499, 2010.
3. Le Li, Xiaohang Zhang, Zhiwen Yu, Zijian Feng, and Ruiping Wei. Usom: Mining and visualizing uncertain data based on self-organizing maps. In *International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 804 – 809, 2011.
4. T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995.
5. Mohammed Khalilia and Mihail Popescu. Topology preservation in fuzzy self-organizing maps. *Advance Trends in Soft Computing*, 312:105 – 114, 2014.
6. Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586 – 600, 2000.
7. D. Herrero-Pérez, H. Martnez-Barberá, K. LeBlanc, and A. Saffiotti. Fuzzy uncertainty modeling for grid based localization of mobile robots. *International Journal of Approximate Reasoning*, 51(8):912 – 932, 2010.
8. Lotfi A. Zadeh. Fuzzy sets. *Information Control*, 8:338 – 353, 1965.
9. Feng Qi and A-Xing Zhu. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. *Computers & Geosciences*, 37(9):1425 – 1436, 2011.
10. J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York, 1981.
11. A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.
12. Gonzalo A. Ruz and Duc Truong Pham. Nbsom: The naive bayes self-organizing map. *Neural Comput. Appl.*, 21(6):1319 – 1330, 2012.
13. Diana K. Smetters James D. Thornton Michael F. Plass Nicholas H. Briggs Jacobson, Van and Rebecca L. Braynard. Networking named content. In *In Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 1 – 12, 2009.
14. Amin Karami and Manel Guerrero-Zapata. A fuzzy anomaly detection system based on hybrid pso-kmeans algorithm in content-centric networks. *Neurocomputing*, 2014.
15. Amin Karami. Data clustering for anomaly detection in content-centric networks. *International Journal of Computer Applications*, 81(7):1 – 8, 2013.

Implementation of Evolutionary Algorithms for Deep Architectures

Sreenivas Sremath Tirumala

Abstract. Deep learning is becoming an increasingly interesting and powerful machine learning method with successful applications in many domains, such as natural language processing, image recognition, and hand-written character recognition. Despite of its eminent success, limitations of traditional learning approach may still prevent deep learning from achieving a wide range of realistic learning tasks. Due to the flexibility and proven effectiveness of evolutionary learning techniques, they may therefore play a crucial role towards unleashing the full potential of deep learning in practice. Unfortunately, many researchers with a strong background on evolutionary computation are not fully aware of the state-of-the-art research on deep learning. To close this knowledge gap and to promote the research on evolutionary inspired deep learning techniques, this paper presents a comprehensive review of the latest deep architectures and surveys important evolutionary algorithms that can potentially be explored for training these deep architectures.

Index terms — Deep Architectures, Deep Learning, Evolutionary Algorithms

1 Introduction

Deep Learning is a topic of high interest with its extensive application in natural language processing, image recognition [1] [2] and computer vision. Corporate giants like Google, Microsoft, Apple, Facebook, Yahoo etc. established their deep learning research groups for implementing this concept in their products. Applications based on deep learning have won numerous machine learning competitions in ICML and NIPS with considerable margins which were earlier dominated by other machine learning approaches like Support Vector Machines. In 2013 it has topped in Chinese Handwriting Recognition Competition, Galaxy Zoo Competition, MICCAI 2013 Challenge, Merck Drug Discovery Competition, Dogs versus Cats Competition etc. Deep Learning is rated as the most interesting topic of research interests by Massachusetts Institute of Technology (MIT).

The importance of studying deep architectures is motivated from the deep architecture found in human brain. It is a common practice to reduce a high level problem into a set of low level problems in a hierarchical manner with easiest problem at the bottom. Interestingly, deep architectures based systems can achieve the learning that a shallow architecture can, but the vice versa is not feasible [3]. A Deep Neural Network (DNN) is an Artificial Neural Network

(ANN) with multiple hidden layers. One of major problems of DNNs is overfitting which was unaddressed till 2014 [4]. Further, due to the extensive use of gradient descent based learning techniques, DNNs may easily be trapped into local optima, resulting in undesirable learning performance. Moreover, the initial topology of DNN is often determined through a seemingly arbitrary trial and error process. However, the fixed topology thus created may seriously affect the learning flexibility and practical applicability of DNNs. Deep learning has been applied on other machine learning paradigms like Support Vector Machines and Reinforcement Learning.

In this paper, we argue that Evolutionary Computation (EC) techniques can, to a large extent, present satisfactory and effective solutions to above mentioned problems. In fact, several Neuroevolutoinary systems have been successfully developed to solve various challenging learning tasks with remarkably better performance than traditional learning techniques. Unfortunately, many researchers with a strong background in evolutionary computation are still not fully aware of the state-of-the-art research on deep learning. To meet this knowledge gap and to promote the research on evolutionary inspired deep learning techniques, this paper presents a review of latest deep architectures and surveys important evolutionary algorithms that can potentially be explored for training these deep architectures. This paper is divided into 5 sections. Section 1 details the history of deep architectures. Section 2 provides a detailed study on various deep architectures. Recent implementations of evolutionary algorithms on deep architectures are explored in section 3. Section 4 summarizes the paper with outcomes and conclusion.

2 Deep Architectures

Deep architecture is a hierarchical structure of multiple layers with each layer being self-trained to learn from the output of its preceding layer. This learning process i.e., 'deep learning' is based on distributed representation learning with multiple levels of representation for various layers. In simple terms, each layer learns a new feature from its preceding layer which makes the learning process concrete. Thus, the learning process is hierarchical with low level feature at the bottom and very high level feature at the top with intermediate features in the middle that can also be utilized. From these features, greedy-layer-wise training mechanism enables to extract only those features that are useful for learning. Along with this, a pre-unsupervised training with unlabelled data makes deep learning more effective.

Shallow architectures have only two levels of computation and learning elements which makes them inefficient to handle training data [5]. Deep architectures require fewer computational units that allow non-local generalization which result in increased comprehensibility and efficiency that has been proved with its success in Natural Language Processing (NLP) and image processing. According to complexity theory of circuits, deep architectures can be exponentially more efficient than traditional narrow architectures in terms of functional representation for problem solving [5]. Traditional Artificial Neural Networks (ANNs) are considered to be most suitable for implementing deep architectures.

In 1980 Fukushima proposed Neocognition using Convolutional Neural Networks (ConvNets) [6] which served as a successful model for later works on deep architectures which later been improved by Lecun [7]. The theoretical concepts of deep architecture were proposed in 1998 by Lecun [8]. The Breakthrough in the research of training deep architectures was achieved in 2006 when Lecun, G.E. Hinton and Yoshua Bengio proposed 3 different types of deep architectures with efficient training mechanism. Lecun implemented efficient training mechanism for ConvNets [9] in which he was not successful earlier. Hinton implemented Deep Belief Networks (DBNs) [10] and Yoshua Bengio proposed Stacked Autoencoders [11].

A simple form of deep architecture implementation is DNNs, feed-forward ANNs with more than one hidden layer units that make them more efficient than a normal ANNs [12]. DNNs are trained with BP by discriminative probabilistic models that calculate the difference between target outputs and actual outputs. The weights in the DNNs are updated using stochastic gradient descent defined as $\Delta w_{ij}(t+1) = \Delta w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}}$, where η represents the learning rate, C is the associated cost function and w_{ij} represents weight. For larger training sets, DNNs may be trained in multiple batches of small sizes without losing the efficiency [13]. However it is very complex to train DNNs with many layers and many hidden units since the number of parameters to be optimized are very high.

2.1 Convolutional Neural Networks (ConvNets)

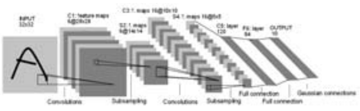


Fig. 1. ConvNets Structure proposed by Lecun [9]

ConvNets are a special type of feed-forward ANNs that performs feature extraction by applying convolution and sub sampling. The principle application of ConvNets is feature identification. ConvNets are biologically inspired MLPs based on virtual cortex principle [14] and the earliest implementation is by Fukushima in 1980 [6] for pattern recognition followed by Lecun in 1998 [8]. ConvNets diversify by applying local connections, sub sampling and sharing the weights which is similar to the principle approach of ANNs in early 60s. In ConvNets each unit in the layer receives input from set of units in small groups from its neighbouring layer which is similar to earlier MLP model. Using local connections for feature extraction has been proven successful, especially for extracting edges, end points and corners. These features extracted at the initial layer will be combined subsequently at the later layers to achieve higher or better features. The features that are detected at the initial stages may also be used at the subsequent stages. The training procedure of the ConvNets is shown in Fig. 1. The first layer takes a raw pixel with 32 x 32 from the input image. The second layer consists of 6 kernels with 5 x 5 local window. From this, a sub sampling is done in the 3rd layer (sub sampling) layer. For the 4th layer, another ConvNets with 16 kernels was exploited with the same 5 x 5 windows. Then the 5th layer is also constructed using sub sampling. This procedure continues till the last layer and the entire structure is developed as Gaussian connections.

2.2 Deep Belief Networks

Deep Belief Network (DBN) is a type of DNN proposed by Hinton in 2006 [15]. DBN is based on MLP model with greedy layer-wise training. DBN consists of multiple interconnected hidden layers with each layer acting as an input to the next layer and is visible only to the next layer. Each layer in a DBN has no lateral connection between its nodes present in that layer. The nodes of DBN are probabilistic logic nodes thus allowing the possibility of using activation function. Restricted Boltzmann machine (RBM) is stochastic ANN with input and hidden units with each and every connection connecting a hidden and visible unit. RBMs act as the building blocks of DBNs because of their capability of learning probabilistic distributions on their inputs. Initially the first layer of the DBNs is trained as RBM that transforms input into output. The output thus received is used as data for the second layer which is treated as a RBM for the next level of training and the process continues. Similarly the output of the second layer will be the input for the third layer and the process continues as shown in Fig. 2 .The transformation of data is done using activation function or sampling. In this way the subsequent hidden layer becomes a visible layer for current hidden layer so as to train it as a RBM. An RBM with two layers, a visible layer as layer 1 and a hidden layer as layer 2 is the simplest form of DBN. The units of the visible layer are used to represent data and the units (hidden with no connection between them) will learn to represent features. If a hidden layer 3 is added to this, then layer 2 will be visible to only layer 3 (still hidden to layer 1) and now the RBM will transform the data from layer 2 to layer 3. This process is illustrated in Fig. 2.

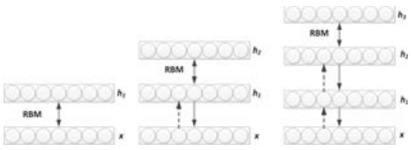


Fig. 2. Structure of Deep Belief Networks [15]

In 2006, Hinton proposed a greedy layer-wise unsupervised pre-learning algorithm for training that addresses the problem of training multilayer ANNs [10]. In DBNs, the lower level features of the input are extracted as lower layers and an abstract representation (high level features) of the input is performed at the higher layers. The training procedure of DBNs is carried out in three phrases. Each layer of the DBN is pre-trained with greedy layer wise training followed by unsupervised learning for each layer and finally training the entire network with supervised training. The significance of this training procedure is determined by the generative weights. After learning, the values of the latent variables in every layer can be inferred by a single, bottom-up pass that starts with observed data vector in the bottom layer using generative weights in the reverse direction. DBNs proved to be the most efficient in image recognition [10], Face Recognition [16], Character Recognition [11] and various other applications.

2.3 Stacked Auto-encoders

The idea of auto-encoders is evolved from the process of reducing dimensionality of data by identifying efficient method to transform complex high dimensional data into lower dimensional code using an encoding multilayer ANN. A decoder network will be used to recover the data from the code. Initially both encoder and decoder networks are assigned with random weights and trained by observing the discrepancy between original data and output obtained from encoding and decoding. After this the error is back propagated first through the decoder network followed by encoder network and this entire system is named as auto-encoders [15].

An auto-encoder with input $x \in R^d$ is ”encoded” as $h \in R^{d^1}$ using deterministic function defined as $f_\theta = \sigma(Wx + b), \theta = W, b$. To ”decode”, a reverse mapping of $f : y = f_{\theta^1}(h) = \sigma W^1 h + b^1$ with $\theta = (W^1,\ b^1)$ and $W^1 = W^T$ with encoding and decoding with the same inputs. This process continues for every training patten. For i training x_i is mapped to h_i with a reconstruction y_i. Parameter optimization is achieved by minimizing the cost function over the training set. However, optimizing an auto-encoder network with multiple hidden layers is difficult. Being similar to DBN greedy layer wise training procedure, this approach replaces RBMs by auto-encoders that perform learning by reproducing every data vector from its own feature activation [5]. The considerable change that has been applied in this model by Yoshua Bengio is changing the un supervised training procedure to supervised in order to identify the significance of training paradigm.

The process of greedy layer wise training is as follows. In the entire ANN, three layers are considered at one instance with the middle layer being the hidden layer. In the next instance, the middle layer becomes input layer and the output layer of the previous instance become hidden layer (the parameters from the output becomes the training parameters) and the layer next to it will be the new output layer. This process continues for the entire network. However, the results were not efficient since the network becomes too greedy [5]. It can be concluded that, the performance of stacked auto-encoders with unsupervised training was almost similar to that of RBNs with similar type of training whereas stacked auto-encoders with supervised pre-training is less efficient. Stacked auto-encoders were not successful in ignoring random noise in its training data due to which its performance is slightly less (almost equal performance but not same) than RBM based deep architectures. However, this gap is narrowed by stacked de-noising auto-encoder algorithm proposed in 2010 [17].

3 Applying Evolutionary Algorithms on Deep Architectures

3.1 Generative Neuroevolution for Deep Learning

In 2013 Phillip Verbancsics and Josh Harguess proposed Generative Neuroevolution for Deep Learning by implementing HyperNEAT as a feature learner on a ANN similar to ConvNets [18]. Compositional pattern producing network (CPPN) is an indirect encoding procedure of HyperNEAT that encodes weight patterns of ANN using composite functions. The topology and weights required for CPNN is evolved by HyperNEAT. In HyperNEAT process, CPPN defines an ANN as a solution for required problem. CPNNs fitness score is determined by evaluating the ANNs performance for the task for which it is evolved. Diverging from traditional methods, this approach trains ANN to learn features by transforming input into features. Then these features are evaluated by Machine Learning (ML) approach thus defining the fitness of CPNN. Therefore, this process will maximize the performance of the learned solution since HyperNEAT determines the best features out of other ML approach. ConvNets can be represented in a graph like structure with coordinates of the nodes associated with each other which are similar to HyperNEAT structure. This similarity enables to apply HyperNEAT on ConvNets based architectures.

For the experiment, an eight dimensional Hypercube representation of CPNN is used with f-axis as feature axis, x-axis as neuron constellation of each feature and y-axis being pixel locations. HyperNEAT topology is a multilayer neural network with layers traveling along z-axis with CPPN representing the points in an eight-dimensional Hyper-cube that corresponds to connections in the four dimensional substrate. The location of each neuron can be identified using (x,y,f,z) coordinate and each layer can be represented with a trait constituting number of features(F) with X and Y dimensions. HyperNEAT is applied to the LeNet-5 [8]. The experiment is conducted on MNIST database with a population size of 250 with 30 runs for 2500 generations. With this comparative results its been concluded that HyperNEAT with ANN architectures is overthrown by HyperNEAT with CNN architecture.

3.2 Deep Learning using Genetic Algorithm

In 2012, Joshua proposed a learning method for deep architectures using genetic algorithm [19]. A DNN for image classification is implemented using a genetic algorithm and training each layer using generic algorithm. Further this study tries to justify the possibility of using genetic algorithms to train non trivial DNNs for feature extraction. Initially a matrix representing the DNN is generated with Sparse Network Design with most of the values being close to zero, whereas the ideal solution in this case is an identity matrix. The genetic sequence of individuals with non-zero elements (which is considered as a gene) is kept and computed instead of re-generating the complete matrix which will reduce the amount of data required to store in the matrix and the process complexity. The position of the gene in the matrix can be determined by row and column and every gene has a magnitude.

The proposed algorithms are tested on image data normalized in the range of 0.0 and 1.0. Apart from applying to image data, the algorithm has been applied to handwriting, face image (small and large) and cat image identification. The experimental results section shows the reconstruction (of input) error rate for each experiment. Another experiment for reconstruction of faces with noisy data claim to prove that this algorithm is not just copying blocks of data, but generating the connections in the data and reconstructing the initial image. The theoretical limitations of the algorithm is not addressed. The cost of reconstruction becomes 0 for a single training image as it will be efficient only with a large set of data.

4 Conclusion

This paper provides a theoretical review of standard deep architectures and study the possibilities of implementing evolutionary computation principles on deep architectures. Apart from introducing various types of deep architecture, this paper provides a detailed explanation of their training procedure and implementations. Further, this paper analyses the implications of applying evolutionary algorithms on deep architectures with details of two such implementations and a critical review on their achievement. The Neuroevolution approach for deep architectures that is discussed in previous section is with respect to the application of HyperNEAT on deep architectures. The success of this proposed method cannot be determined since CNN holds the best classification for MNIST database. But, this drives a way of implementing Neuroevolution algorithms on deep architectures. Similarly, the second work of using genetic algorithms for training DNNs, justifies the possibility of using genetic algorithms for training deep architectures but does not show any signs of comparative studies of its efficiency with respect to speed or quality.

It is noteworthy that evolutionary algorithms may not be a complete replacement for deep learning algorithms at least not at this stage. However, the successful application of evolutionary techniques on deep architectures will lead to an improved learning mechanism for deep architectures. This might result in reducing the training time which is the main drawback for deep architectures. Future direction in this research could be evolving an optimized deep architecture based neural networks using Neuroevolutonary principles. This could provide a warm start to the deep learning process and could improve the performance of the deep learning algorithms.

References

1. Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 253–256, May 2010.

2. J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *In NIPS*, 2012.
3. Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009. Also published as a book. Now Publishers, 2009.
4. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
5. Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," in *Large Scale Kernel Machines* (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), MIT Press, 2007.
6. K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.
7. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems (NIPS 1989)* (D. Touretzky, ed.), vol. 2, (Denver, CO), Morgan Kaufman, 1990.
8. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, pp. 2278–2324, 1998.
9. M. A. Ranzato, C. S. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *NIPS*, pp. 1137–1144, 2006.
10. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, July 2006.
11. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. D. Montral, and M. Qubec, "Greedy layer-wise training of deep networks," in *In NIPS*, MIT Press, 2007.
12. G. Tesauro, "Practical issues in temporal difference learning," in *Machine Learning*, pp. 257–277, 1992.
13. G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.
14. D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *Journal of Physiology (London)*, vol. 195, pp. 215–243, 1968.
15. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, 2006.
16. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
17. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.
18. P. Verbancsics and J. Harguess, "Generative neuroevolution for deep learning," *CoRR*, vol. abs/1312.5355, 2013.
19. J. Lamos-Sweeney, "Deep learning using genetic algorithms. Master thesis, Institute Thomas Golisano College of Computing and Information Sciences," 2012. Advisor: Gaborski, Roger.

Author Index