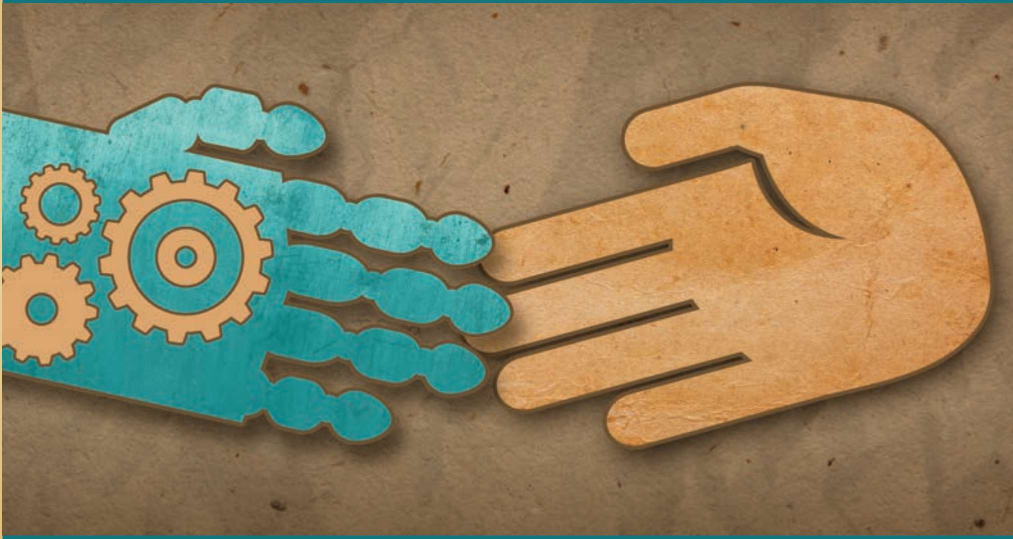
PONTIFICIAE ACADEMIAE SCIENTIARVM SCRIPTA VARIA 132

Power and Limits of Artificial Intelligence

Edited by
ANTONIO M. BATTRO, STANISLAS DEHAENE



Workshop | 30 November-1 December 2016
Casina Pio IV | Vatican City

LIBRERIA EDITRICE VATICANA
VATICAN CITY 2017

Power and Limits of Artificial Intelligence

Pontificiae Academiae Scientiarvm Scripta Varia 132

The Proceedings of the Workshop on

Power and Limits of Artificial Intelligence

30 November-1 December 2016

Edited by
Antonio M. Battro
Stanislas Dehaene

EX AEDIBVS ACADEMICIS
IN CIVITATE VATICANA • MMXVII

The Pontifical Academy of Sciences
Casina Pio IV, 00120 Vatican City
Tel: +39 0669883195 • Fax: +39 0669885218
Email: pas@pas.va • Website: www.pas.va

The opinions expressed with absolute freedom during the presentation of the papers of this meeting, although published by the Academy, represent only the points of view of the participants and not those of the Academy.

ISBN 978-88-7761-111-6

PONTIFICIA ACADEMIA SCIENTIARVM
LIBRERIA EDITRICE VATICANA

VATICAN CITY

In the Encyclical *Laudato Si'* I stated that "we are called to be instruments of God our Father, so that our planet might be what he desired when he created it and correspond with his plan for peace, beauty and fullness" (53). In our modern world, we have grown up thinking ourselves owners and masters of nature, authorized to plunder it without any consideration of its hidden potential and laws of development, as if subjecting inanimate matter to our whims, with the consequence of grave loss to biodiversity, among other ills. We are not custodians of a museum or of its major artefacts to be dusted each day, but rather co-operators in protecting and developing the life and biodiversity of the planet and of human life present there. An ecological conversion capable of supporting and promoting sustainable development includes, by its very nature, both the full assuming of our human responsibilities regarding creation and its resources, as well as the search for social justice and the overcoming of an immoral system that produces misery, inequality and exclusion.

Address of His Holiness Pope Francis to Participants in the Plenary Session of the Pontifical Academy of Sciences, Consistory Hall, Monday, 28 November 2016.

C. VILLANI

Contents

Preface

One of the key issues today concerns the place of the human person in a growing digital environment of increasing complexity that not only expands the range of his or her capacities, but also may compete with them or even replace them. Over the past fifty years, robots and computers have progressively supplemented humans, initially only in relatively simple computational and manipulation tasks, but more recently in higher cognitive tasks that used to be the prerogative of the human brain, including language, mathematics, probabilistic reasoning and decision making. A crucial question is how to enhance the productive interactions between humans and artificial intelligence (AI). As such interactions reach new orders of complexity, many researchers and philosophers feel that the outcome may defy our understanding and produce radical changes in our personal and social life in the near future.

Our Academy has already organized several meetings on the organization and functions of the human brain and mind (*The Educated Brain*, 2003; *Human Neuroplasticity and Education*, 2011; *Neurosciences and the Human Person*, 2012). We propose now to study the *Power and Limits of Artificial Intelligence*.

What is the state of the art in AI software and machine learning? Can all aspects of brain function be mimicked by artificial systems? Will machines soon surpass us in all domains of human competence? What is the proper form of mathematics that may capture the operation of minds and brains? What is consciousness? Could a machine be endowed with an artificial consciousness? What would it take for a machine to possess a sense of self? Will intelligent machines soon pose a danger to humanity? Is it possible to design and construct an intelligent robot endowed with an artificial sense of ethics? How can we enhance the humanitarian uses of artificial intelligence and robotics, in particular in the field of education, health and emergencies?

We know that all these questions are very difficult to answer today, but we want to open a discussion between experts of the different fields in order to map the new cognitive environment that humanity is creating for the first time in history.

Antonio M. Battro and Stanislas Dehaene

Programme

Wednesday, 30 November 2016

STATE OF THE ART IN ARTIFICIAL INTELLIGENCE, ROBOTICS, BRAIN MODELING, BRAIN-COMPUTER INTERFACES

09:00 Werner Arber, *Word of Welcome*
09:10 Marcelo Sánchez Sorondo, *Welcome Greetings*
09:20 Stanislas Dehaene, *Outline of the Workshop*
09:30 *Artificial Intelligence: A Survey of Achievements and Questions, Viewed from Mathematics*
Cédric Villani, Institut Henri Poincaré, PAS
09:50 Discussion
10:10 *The Evolutionary Success of Cerebral Cortex: Computing in High Dimensional Dynamic Space*
Wolf Singer, Strüngmann Institute, Frankfurt, PAS
10:30 Discussion
10:50 Coffee Break
11:20 *Breaking the Gap Between AI and Human Intelligence: What Are We Missing?*
Yann LeCun, Facebook
11:40 Discussion
12:00 Lunch at the Casina Pio IV
14:00 *Comment: The Ethics of Artificial Intelligence*
Stephen Hawking, University of Cambridge
14:05 *The Probabilistic Brain*
Alex Pouget, Université de Genève
14:25 Discussion
14:45 *Building Machines That Learn and Think Like People*
Josh Tenenbaum, MIT
15:05 Discussion
15:25 Coffee break
15:50 *Towards Artificial General Intelligence*
Demis Hassabis, Google DeepMind
16:10 Discussion
16:30 *Motivation and Evaluation are Computationally Messy*
Patricia Churchland, UCSD, California
16:50 Discussion
17:10 General Discussion

17:30 *Children and Robots*
Antonio Battro
18:00 Departure from the Casina Pio IV by bus for the visit to Palazzo Farnese
18:30 Dinner at the Casina Pio IV for those not attending the visit to Palazzo Farnese

Thursday, 1 December 2016

PUTATIVE PREROGATIVES OF THE HUMAN BRAIN: EDUCATION, REASONING, CREATIVITY, CONSCIOUSNESS, SENSE OF SELF, ETHICS...COULD THEY BE CAPTURED IN MACHINES?

09:00 *The Impact of Augmented Reality, Wearables and Robotics In Neuroscience and Neuropsychiatry*
Olaf Blanke, EPFL
9:20 Discussion
9:40 *What is Consciousness, and Could Machines Have It?*
Stanislas Dehaene, Collège de France, PAS
10:00 Discussion
10:20 Coffee break
10:50 *What Really Matters: Children's Inferences About Learning, Trying and Caring*
Laura Schulz, MIT
11:10 Discussion
11:30 Artificial Intelligence and Human Minds: Perspectives From Studies of Infants Elizabeth Spelke, Harvard
11:50 Discussion
12:10 Lunch at the Casina Pio IV
14:00 *The Limits and Potential for Brain Computer Interfaces*
John Donoghue EPFL, Lausanne
14:20 Discussion
14:40 *Collaborative Human-Robot Autonomy*
Manuela M.Veloso, Carnegie Mellon University
15:00 Discussion
15:20 *Who Am I?*
Laurie Paul, North Carolina Chapel Hill
15:40 Discussion
16:00 Coffee Break
16:30 General discussion and drafting of final statement by all participants (discussion led by Stanislas Dehaene)
18:30 Dinner at the Casina Pio IV

List of Participants

Werner Arber
President of the Pontifical
Academy of Sciences;
Biozentrum, Department of
Microbiology University of Basel
(Switzerland)

Antonio M. Battro
Pontifical Academy of Sciences,
Academia Nacional de Educación
(Argentina)

Olaf Blanke
Laboratory of Cognitive Neuroscience,
Brain-Mind Institute
Ecole Polytechnique Fédérale
de Lausanne (EPFL) Lausanne
(Switzerland)

Patricia Churchland
University of California,
San Diego, CA (USA)

Stanislas Dehaene
Pontifical Academy of Sciences;
Collège de France. Inserm-CEA,
Cognitive Neuroimaging Unit CEA/
SAC/DSV/DRM/NeuroSpin,
Gif sur Yvette (France)

John Donoghue
Director, Wyss Center for Bio
& Neuroengineering
Geneva (Switzerland)

Demis Hassabis
Google Deep Mind Technologies;
Computer Laboratory,
University of Cambridge (UK)

Stephen W. Hawking
Pontifical Academy of Sciences;
University of Cambridge, Department of
Applied Mathematics
and Theoretical Physics,
Cambridge (UK)

Yann LeCun
Director of AI Research, Facebook;
Computer Science, Neural Science, and
Electrical and Computer Engineering,
New York University, NY (USA)

Pierre Léna
Pontifical Academy of Sciences;
Fondation LAMAP La main à la Pâte,
Académie des Sciences,
Paris (France)

Laurie Ann Paul
University of North Carolina at Chapel
Hill, Department of Philosophy,
Chapel Hill, North Carolina (USA)

Alexandre Pouget
Université de Genève,
Department of Basic Neurosciences
(Switzerland)

H.E. Msgr. Marcelo Sánchez Sorondo
Chancellor, Pontifical Academy of
Sciences (Vatican City)

Laura Schulz
Massachusetts Institute of Technology,
Department of Brain
and Cognitive Sciences,
Cambridge, MA (USA)

Mariano Sigman
Universidad Torcuato di Tella;
Laboratorio de Neurociencia Cognitiva
(Argentina)

Wolf J. Singer
Pontifical Academy of Sciences;
Max-Planck-Institute for Brain Research,
Frankfurt am Main (Germany)

Elizabeth Spelke
Harvard University,
Department of Psychology,
Cambridge, MA (USA)

Josh Tenenbaum
Massachusetts Institute of Technology,
Department of Brain
and Cognitive Sciences,
Cambridge, MA (USA)

Manuela Veloso
Carnegie Mellon University, Head,
Machine Learning Department, School
of Computer Science,
Pittsburgh, PA (USA)

Cédric Villani
Pontifical Academy of Sciences;
Institut Henri Poincaré
(UPMC/CNRS),
Paris (France)

State of the Art in Artificial Intelligence, Robotics, Brain Modeling, Brain-Computer Interfaces

Artificial Intelligence – Big Achievements and Huge Questions Viewed from Mathematics

Cédric Villani

Introduction

Mathematical algorithms have probably been around for more than 4000 years, as suggested by the famous YBC7289 clay tablet (dated 1600 BC or earlier), displaying an amazing computation of $\sqrt{2}$. They have grown in scope, diversity and sophistication together with mathematical sciences. But the middle of the twentieth century marked an amazing new turn. On the one hand, arguably for the first time in history, the outcome of a major human event depended on the devising of a clever, mathematically sophisticated algorithm (this is the story of the work of Alan Turing's team during Second World War). On the other hand, within just a few years, the basis of modern computer technology and computer algorithms were laid with the discovery of transistors and the works of Turing, Shannon, Von Neumann, Wiener and others.

Progress had been slow, but then it accelerated. Fast-forward half a century, and here we are, with a world full of algorithms, and entire sectors of human activities have been revolutionized by algorithms. For instance, to get an idea of how it now looks in world finance, just read the books "6" and "5" by Alexandre Laumonier, providing an impressionistic but thoroughly documented of mysterious algorithms fighting against each other, fortunes evaporating in a fraction of a second, crazy race for speed of transmission and execution. Whether this vision, fascinating and frightening, will extend to all of society, has been the subject of considerable debate; but one sure thing is that algorithms will capture a more and more important role in our economies, our societies, our lives.

A chapter within this long rise of algorithms is the long rise of artificial intelligence. This field is old, by modern standards, since it started almost at the same time as computer science, with the works of Turing and Shannon. Actually, some of the most important methods and algorithms used nowadays in this field did originate from the fifties or even forties. A vision of the founding fathers was that artificial intelligence would help us understand our own intelligence. After some initial fascinating dreams and successes, crystallized in 1968s HAL computer in Kubrick's Odyssey of space, the field mostly stalled. Then it accelerated again recently, partly because of new methods, partly by taking advantage of the amazing new speeds and capacities of computers, partly by the exploitation of the huge databases which have appeared. And suddenly Artificial Intelligence has become an enormous hype, with speculations of superhuman intelligence, economic catastrophes; and any ambitious "global entrepreneur" has to keep artificial intelligence under his or her radar. Questions about artificial intelligence and machine learning are so frequent on Quora, appear in broad audience magazines, newspapers; they have also given rise to new directions of research and a renewed attraction for young scientists.

In this context, it is normal to be enthusiastic but to keep away from mystic overhype. It is also normal to remain cautious, and to try and point out questions which remain in the dark. So let us go for a nonexhaustive overview.

Disclaimer: I am not an expert on AI! But the field has been taking so much room that I could not leave it unexamined. Actually I have taken keen interest in the related field of Monte Carlo Markov Chain (MCMC) already for the past 15 years.

1. Basic principles

1.1. Optimization

An intelligent solution is one which tries to find the best analysis, best answer, best action in a certain context. So artificial intelligence will often be about optimizing. Linear optimization, in which the constraints and functions to optimize are all linear, has a rich theory with a lot of structure; but apart from that peculiar setting, not many methods are known for optimization when the setting is rather general.

By far the most popular general method of optimization is gradient descent: follow the gradient. For instance, to find the highest point in a landscape, just look for the direction in which altitude increases fastest, and continue this way. In nature, optimization is supposed to work in a different way, namely through competition (as in natural selection). Parallel to that, there are algorithmic methods based on competition, be it through mutations, auctions or other mechanisms.

Mutations introduce probability theory in the game, and huge progress was made when probabilistic and deterministic methods were mixed: these were, in particular, the MCMC methods, which go back to the forties but have become all the hype in the nineties. Consider again the problem of finding the highest point in a landscape: with the gradient method, you will in general get trapped in a local optimizer. But MCMC can get you out of the trap, by randomly allowing some motions which will get you down, thus leaving a possibility to get to the next hill and in the end to arrive at the true peak. And when arrived at that true highest peak, one will also, from time to time, get down the hill, so that the information is about the time spent in the various states. (And there are techniques to progressively make the exploration deterministic, so that one may eventually settle in the culminating point, or at least a very high peak).

Whatever the technique used, the field of artificial intelligence strongly depends on optimization.

1.2. Learning

But besides the notion of intelligence there is of course not just the notion of finding an optimal, or at least good, response. It should also adapt to the situation, and do things which it was not explicitly told to do. Or, to use a phrase by Samuel (1959), the program should have "the ability to learn without being explicitly programmed".

The field of machine learning is about letting an algorithm discover by itself a good way to handle a problem, through reviewing information and adapting to that information. One of the very first such systems was Shannon's electric mouse, Theseus (1952), which would explore a maze to find the best way out.

To continue with the analogy of finding the highest altitude, think that we don't want to only find the peak, but also to find the shortest path between the starting point and the peak, taking into account what we explored. Or, more ambitiously and more interestingly, that we wish to discover recipes, learnt from examples, that allow us to find the peak very fast, if we are put in a new environment which shares certain features with the previous environments that we explored.

A field of mathematics in which learning has always been at the core is Bayesian statistics. One wishes to evaluate the probability distribution of a certain set of parameters, and for that one starts with a prior distribution, then updates it with all the knowledge gained from successive information.

1.3. Classification

Imagine that you have a number of observations falling in several categories: maybe just two categories, A and B. You would like to describe the difference between A and B in the shortest possible way. In mathematical terms, it could be about separating the phase space in two regions, through an easily described interface, in such a way that all A observations lie on one side, and all B observations lie on the other side.

The best situation is when you can find a hyperplane to do this separation job; by the way there is a long tradition of separating hyperplanes in the context of convexity theory. But of course, most of the time you will not be able to do so. On the other hand, it might be that a change of parameterization gives rise to such a possibility. This is the principle of the method of linear classifiers (so strongly associated to machine learning that an icon about this principle was chosen as the logo of the machine learning Wiki!).

1.4. The curse of dimensionality

In practice the learning problem stumbles across the major problem that the phase space is huge. Already in the simple Theseus problem, the phase space is not the board on which the mouse crawls, but rather the set of all paths in this board, so there is a combinatorial increase of the complexity. But in any realistic problem things are way worse in terms: for combinatorial or complexity questions, problems have to be set in large dimension. Consider the problem of figure recognition: possible variations in shape of written numbers imply that the unknown lives in a space with dozens of dimensions. In some currently used modern models for semantic analysis, language representation occurs in a 300 dimensional space. In phylogenetic reconstruction, the number of possible trees is beyond imagination (500 taxa can be arranged in more than 10^{1275} trees!)

In the absence of specific information to reduce this high dimensionality, there is absolutely no hope to explore the set of possibilities via deterministic, systematic methods. Some guesses have to be made, and one has to resort to chance in a way or another. A good news is that random exploration will give, in many cases, surprisingly efficient methods. A bad news is that it is not really understood why. Another bad news is that there is in general no way to be completely sure that the method will achieve the desired goal.

1.5. The extraction of meaning

There is a classical distinction between information, knowledge and wisdom. How to get the wisdom from the knowledge, and the knowledge from the information, are longstanding multi-faceted questions. Henri Poincaré said it beautifully: An accumulation of facts is no more a science than a heap of stones is a house. The scientist has to order. But in our current era a no less pressing problem is to extract information from data. Indeed, the amount and size of data, which are considered, makes it impossible to just examine them with human senses and brains.

Some of the most important beliefs underlying current techniques are:

Belief 1. It all boils down to a "reasonable" number of parameters. That means, even if the data we are observing give information about a million sets of measurements in a space of several hundreds of dimensions, eventually we should be able to classify all this information according to a rather small number of parameters, say 10.

To give an example among a multitude, a case which became infamous recently in relation with automated election campaigning, is the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) which was used to re-present personality traits of users in social media. In this model, all the complexity of behaviours was summarized in a five-dimensional space.

Belief 2. We may let the algorithm discover (with or without precise instructions) these explanatory parameters.

For a long time statisticians have developed methods of "exploratory factor analysis" to identify the determining parameters in a multivariate set of data. Model selection is about finding the good compromise between a too precise and a too rough description of multiparameter problems. On the other hand, some of the new artificial intelligence methods perform such a task by very indirect ways.

While "understanding" is, in a way, about finding the best way to sort things out, this can serve a diversity of purposes, in particular: compress data, recognize data, react to a situation, or generalize data (either by interpolation or by extrapolation).

1.6. Parsimony and extrapolation

Implicit in the previous discussion is the notion of economic representation of the information. This underlying general principle is also made explicit in parsimony theory, in the form of a minimization problem, which has a taste of the entropy problem in statistical physics.

Parsimony could be summarized as follows: given an incomplete set of data, let us find the data, which complements the set in such a way as to achieve minimal "complexity". There is a certain art in choosing what complexity here means, and it can be based on phenomenology as well as on fundamental principles.

2. Some applications

Applications of artificial intelligence are now legion, and underlie a large part of the current innovation. Here is a very partial list.

- Pattern recognition: e.g. the algorithm will guess that the image which is presented to it is that of a Panda bear or a "priority road" sign or something else, and it will give a "confidence percentage".
- Prediction: For instance, it is through the sole use of data given by social users networks that the company QuantCube, specialized in data analysis, successfully predicted the election of Donald Trump at the Presidency of the USA, even though with a tiny margin of confidence.
- Preference guesses: from the behaviour of a list of users, and partial information about the behaviour of another user, guess what the latter prefers to see, hear or feel. The most famous example is the "Netflix problem": given an incomplete table of preferences for a large list of users (this user loved this movie and hated that one, etc.), and an incomplete set of preferences for another user, guess what he or she will love or hate. Accurate preference guesses are a holy grail in the fields of advertising, selling, but also for opinion campaigns.
- Translation: Google Translate rests on statistical methods of machine learning much more than on grammatical analysis.
- Composition: For instance, using artificial intelligence researchers have composed songs in the style of the Beatles, or automatically generated short movies from a selection of images, etc.
- Diagnosis: Expert systems such as IBM's Watson guess a likely disease based on symptoms, or evaluate a situation based on measurements; similarly, banks detect transactions with a high risk of being identity thefts.
- Clustering and sorting: For analysis of relations between words, languages, species, etc.
- Automation: self-driving cars, drones, etc.
- Interface man-machine: for instance through commands which adjust to the person's mentality or even thoughts; evolutive prosthetics which learn the morphology of the body.

Etc. Machine learning has become so routinely used, and with such a diversity of tricks, that competitions are now regularly organized between teams to select the most efficient method. The most famous of these platforms is the website www.kaggle.com

Implications of these methods are not only in technology: they also suggest new trends to scientists, even though often resting on somewhat shaky ground.

Let me comment on four examples which I came across recently.

- Inria in France has several projects based on robotics control: the motion of the eyes, or even just the thoughts (through recording of the brain activity), would be able to pilot a drone or robot.
- Microsoft Research in Asia developed an algorithm to generate avatars: it uses a collection of pairs (photograph of a face, artist's rendition) to develop its own recipe imitating the style of the artist. Thus when a new photograph is provided, the algorithm will suggest a translation in the style of the artist.
- A biology research paper published in September 2016 suggested that there were four species of giraffe in Africa, rather than one; it was based on learning from large samples of genetic data. (This is an example in which a scientific field is changed by the use of artificial intelligence).
- Riccardo Sabatini and his team showed how to teach a machine-learning algorithm to reconstruct the face of a human from the DNA sample. This is a typical case of application of machine learning: on the one hand the data is absolutely huge, on the other hand the correspondence between the data (genotype), and the expected outcome (phenotype) has famously remained a nightmare for geneticists.

3. The algorithms

3.1. Trends

Machine learning encompasses a diversity of techniques and tricks. Searching online it will be easy to find some lists and selections of them, such as "Top 10 machine learning algorithms" etc. Buzzwords evolve as new algorithms demonstrate their efficiency. A few years ago in lectures on the subject you would hear a lot about SVM (Support Vector Machines); now it is Deep Learning which gets the most credit.

The example of Deep Learning shows that it is important to retain a diversity of methods, and some unconventionality. Indeed, not so long ago, most of the renowned experts in the field would dismiss neural networks as inefficient and doomed; but the tenacity of Yann LeCun demonstrated that these methods can be amazingly efficient.

A beautiful reference about the many facets of artificial intelligence, at least up to a few years ago, is the book by Russell and Norwig. To get some glimpses of the state of the art in current artificial intelligence research, one may watch the online videos of the following emblematic events:

(a) The plenary lecture of Emmanuel Candes at the International Congress of Mathematicians in Seoul (2014), about parsimony methods applied to preference guessing as well as medical imagery; themes such as the right definition of "complexity" of an image, or the mathematical justification of the method, are enlightening.

(b) The course by LeCun at Collège de France in Paris, and the seminars given there, e.g. by Ollivier and Mallat, which will provide a diverse view as well as many questions.

A general remark is that, eventually, artificial intelligence algorithms boil down to technical keywords such as large matrix diagonalization, convex optimization, gradient flows etc. which to an outsider hardly evoke anything related to "intelligence".

3.2. Neural networks

Neural networks is just one of many fields in Artificial Intelligence. But it has become such a craze that it deserves a specific review. Let me just mention that

- Neural networks use a list of examples which we may write (x_i, h_i), and the goal is to produce a "rather simple" function h in such a way that $h(x_i) \simeq h_i$; in other words it is about guess an unknown function through examples;
- Neural networks are made of nodes (neurons) and links (synapses); while the general pattern is inspired from animal brains, the organization is quite different; nowadays the neurons are organized in a rather large number of layers (depth), with synapses joining neurons from one layer to the next one;
- Each synapse corresponds to an elementary nonlinear function, approximating a step function, and depending on some parameters; this mimics the fact that a synapse can transmit more or less information, and does so only at a certain level of excitation; so a function is a combination of elementary nonlinear functions;
- The number of neurons can be very large nowadays, with millions of parameters, and actually large neuron networks have achieved amazing results these past few years;
- The optimization procedure is based on a gradient flow method, here called "back-propagation".

I refer to the lecture ofYann LeCun for more information about networks and their use. I also refer to the lecture of Demis Hassabis for an in-depth discussion of the spectacular and instructive case of the AlphaGo program, which achieved world fame by demonstrating its super-human level at Go and showing at the same time what could be considered as creativity.

It is important to point out some conceptual differences between AlphaGo and a "classical" algorithmic approach to Go playing. AlphaGo achieved its super-human power by an extraordinarily intense training, taking primary examples from recorded Go games by humans. The particular rules which AlphaGo applies do depend on the particular examples that it was fed. But also, AlphaGo spent an enormous amount of time playing against himself and trying random departures from the sets of games it was given. Thus a good amount of randomness enters the making of AlphaGo, first through the selection of games it is primarily fed with, secondly through these random variations that it is trying.

4. Big Scientific Questions

The brilliance of programs such as AlphaGo, or the ever-increasing number of applications and programs which use modern AI, demonstrate the impact of these methods. But still this comes with big questions.

4.1. Why does it work so well?

This question, which is formulated verbatim in Mallat's contribution at the Collège de France, is on the lips of every researcher, especially since the surprise comeback of deep neuron networks. In fact, these methods are vexingly efficient, and took theoreticians off guard. For sure there is, among other things, an effect of the "Big" factor. It has been noted already some time ago that the most important asset of a database is its size; inaccuracies being washed out by the sheer number. With modern methods we also see that size does matter.

As Ollivier emphasizes in his own contribution, to better understand this question there is need for a much more conceptual modelling, with a geometric study of the phase space and the process.

A related question is: which problems can be solved? Mallat likes to formulate this in term of three keywords which are well known to harmonic analysts: complexity, regularity, and approximation theorems. Working in the particular context of parsimony methods, Candes insists on three ingredients for success which seem important: (i) structured solutions (the fact that only a few parameters really count; mathematically speaking this would be, typically, about a matrix having small rank, or being very close to having small rank); (ii) the ability to use convex programming for computational purposes; (iii) incoherence, that is, the fact that the missing information does not present any particular pattern in respect to the key parameters. With a proper mathematical formalization of these assumptions, Candes and Tao are able to prove a few neat mathematical theorems showing accuracy of the reconstruction for "most" samples.

Yet another related question, obviously, is: can one make those algorithms more efficient? There is real motivation for this, as modern algorithms are, by any rate, inefficient: they are very demanding in terms of storage, go through datasets dozens or hundreds of times, use up absurdly enormous power if we compare them to a human brain, do not yet adapt to quantum algorithmics...

Part of this inefficiency is also due to the use of randomness, which is typical (randomness is inefficient but in complex phase spaces all other methods are usually worse). Arguably, it may also be due to the poor incorporation of rules and semantics in the search for representation.

4.2. MCMC methods

I would like to recall that MCMC methods were all the buzz in the 1990s to magically solve problems with large phase spaces. The articles by Persi Diaconis on this MCMC Revolution are very instructive.

Arguably, MCMC was a particular case of machine learning, with a modelling (the probability distribution) which was improving with the amount of data, using randomness and gradient flow optimization.

But the analogy does not stop here: some of the same questions as above were also central: Why does it work so well? Which are the geometrical or structural conditions which make it work well, and so on?

Diaconis, who became famous for his discovery of the cut-off effect in the convergence of Markov chains (that is, the fact that the convergence often occurs very rapidly after a certain time, going in a small number of iterations to "hardly mixed" to "very much mixed"), has been fascinated by the problem of mathematically explaining the efficiency of MCMC methods. Together with Michel Lebeau and other collaborators, he worked on analysing this in controlled cases with very simple rules. The results, which appeared on the prestigious journal *Inventiones*, are fascinating: even if the model is oversimplified, they are based on an amazing level of mathematical sophistication, and the convergence estimates are quite conservative. While these authors have established admirable pioneer work, it is likely that there is still room for huge improvements.

By analogy, the following question is very natural: Is there a sharp cutoff effect for AI algorithms, in terms of the size of the data, or the number of parameters?

4.3. What about our intelligence?

The dream of the founding fathers was that artificial intelligence would lead us to a better understanding of our own intelligence. So far this has not borne so much fruit. On the contrary, some striking experiments suggest that our algorithms are very different from those which are used in AI. Maybe none is more spectacular than the correlated noise attacks performed by Christian Szegedy: from an image which is clearly recognized by the algorithm (say a truck), a tiny modification (invisible to a human mind) will fool the algorithm into recognizing an ostrich, with extremely high confidence.

Even without this, the inefficiency of artificial algorithms with respect to natural ones has been an elephant in the room: just a few observations are sufficient for a human to identify a pattern, where machine learning algorithms need huge numbers of them.

In such situations, however, comparison between natural and artificial mechanisms has helped suggesting new research directions, such as reinforcement by adversarial training (see LeCun's lecture), or the modelling of universes with categories and subcategories (see Tenenbaum's lecture).

Also, at qualitative level, some striking suggestions have been made by Dehaene, for instance about the encoding of numbers in the brain, based on artificial neuron networks.

It is likely that these features (strong discrepancy between natural and artificial algorithms, but mutual influence in their understanding) will continue, and that little by little we shall identify some ways to model some human intelligence features through AI.

4.4. Epistemological questions

Will AI be able one day to do science, to out-perform human scientists, or, more modestly, to help humans finding new science models, or science laws?

Mathematics has been a favourite science in this question, probably because (a) mathematics does not explicitly rely on experiments, (b) mathematics is the only science in which the rules of the game are fully known, (c) mathematics is both familiar to and admired by the (mostly geek-type) conceivers of AI programs. So the idea of a theorem-proving AI is a widely shared dream in AI.

Besides mathematics, one could hope for AI to identify patterns, formulas, or even equations, without "proving" them, but showing that this is how nature works.

One may object that mathematical proof requires exploration of an extraordinary combinatorics. Automated proof checking has gone a long way forward, but there is a whole world between proof checking and proof making.

One may also object that machine-learning methods, based on examples rather than models, will be poor at discovering new laws and reasons. But on the other hand, we have examples in science of laws which were first discovered through the examination of data and later turned into laws: one of the most famous is Kepler's law of elliptical orbits.

Still, so far the harvest is meagre. It is true that a computer program has been good at deriving the basic laws of Hamiltonian mechanics, and that some expert systems have managed to prove some nontrivial geometry theorems, but the whole of such achievements remains a tiny portion of science, and I am not aware of any novel law which has been found through AI. Let me also bring back the spectre of MCMC by recalling that it was originally used to discover the phenomenon of hard spheres transition; but that nobody has been able to justify or understand this phenomenon in more than half a century.

For the moment we may just say that time will tell!?

Now, one may for sure be more optimistic in the prospect of an AI-aided scientist, and there are already such examples, especially in biology. In his seminar, Mallat also shows how to use AI to derive the shape of an unknown energy in a mathematical physics problem.

However, these are not really about finding new laws, but about finding new ways to organize a complex given information. Here, for sure the most important themes revolve around genetics and related fields, such as phylogenetics and taxonomy.

An example of progress in the field of taxonomy is the recent work on giraffe genomes, which suggests that there are actually four species of giraffes (note that the notion of giraffe is no longer clearly defined!). As for the field of phylogenetics, which aims at identifying the "parenthood" relations between species, a recently debated issue was the respective places of Archaea, Bacteria and Eukaryots. In these fields MCMC and other machine learning methods have been used on large genomic samples. This is exciting, but leaves some big questions.

A first big question is how will researchers master these tools, and most importantly the safety rules for their use. Thinking again about MCMC, there is a well-known course by Alan Sokal warning users that results obtained by MCMC have no scientific value whatsoever if they do not come with justification of the convergence times and sampling rates through an estimate of features such as the autocorrelation times. It is likely that most of the published scientific literature based on these techniques does not perform such checks. Of course debates follow.

A second one is about the epistemological status of advances which have been obtained through AI algorithms. There is usually no proof of convergence of such algorithms, and thus no way to guarantee the accuracy of the method. Should we admit them as evidence, knowing that they use randomness and other black box features?

A third big question is about the meaning of"understanding". AI methods have made huge progress when we became more lenient in our demands for understanding the rules which produce the results. The good thing is that the algorithm does usually much better than what we could imagine, but the bad thing is that we don't understand the reasons for the outcome, even when we have it. To remedy this, one should work (and one already works) on the way to display and propose the results, singling out those parameters which played a most significant role in arriving at the result.

A final big question is the risk of seeing drops in the mastering of entire chunks of scientific skills, namely in the modelling. For instance, in mathematical finance, stochastic modelling is rapidly giving way to big data analysis. One certainly should rejoice about this diversification, but one can also worry that stochastic finance analysis, based on modelling, may soon be forgotten by younger generations of finance mathematicians. The same can be said about many fields. Whatever point of view one wishes to adopt, it is important to recall that in a classical scientific view, understanding always includes modelling, and it is certainly foolish to believe that data will get rid of that. Just think of the big difference between cause and correlation, that only a model can bring!

5. Big Societal questions

AI methods have invaded most fields of technology and will very likely be used more and more, for more and more tasks. This is a partly comforting, partly worrying trend.

5.1. How robust is AI?

Szegedy's experiments have shown that AI-based recognition may be fragile, and possibly subject to attacks exploiting the fact that it has been trained in a certain way. Currently, AI remains so far good for specialized tasks (like playing Go!) and this may lead to a lack of stability and robustness.

It is notable that one of the most promising directions of research in AI, namely adversarial learning, is precisely aimed at making learning algorithms improve by challenging them in situations of ambiguity (as when one is training a youngster by giving exercises with traps).

In the case of AlphaGo, Lee Sedok was able to fool the algorithm once by leading it into a highly non-comfortable zone that it had not explored enough. This is reminiscent of human strategies against chess programs. It certainly would be impossible with the current version of AlphaGo, which is way stronger than the one which Sedok played against. But it shows that it is not easy to ascertain the robustness of AI.

5.2. Who will take responsibility?

The achievements of AI are impressive, but the convergence is not guaranteed, the mechanisms remain mysterious. Who will take responsibility in case of legal battle or policy change? The question may be asked for an automatic driving car, but also in a number of other situations. For instance, it is now possible, through AI, to get a rather good reconstruction of the face through DNA sample: can this be used in a criminal case?

Then, there are discriminations which may come in AI programs. What happens if an AI program automatically leads to different rules and behaviours in front of different ethnical groups, or social groups? Such situations have already occurred.

5.3. Biases

The case of Tay, the chatbot by Microsoft, has been on the media: that AI was influenced and manipulated by users which transformed it quickly into a horrible racist (and, by the way, a worshipper of Donald Trump). This shows that AI, like humans, may inherit strong biases from their environment, impairing their tasks.

Currently, the method of building an AI, through example-intensive machine learning, leads to biases: the program will depend on the databases which we use. A 2016 paper by Caliskan-Islam, Bryson and Narayanan was pointing that Semantics derived automatically from language corpora necessarily contain human biases.

5.4. Myths and fears

AI has triggered a number of myths and fears already. One is the emergence of a super-human intelligence. The way some authors talk about it makes it closer to religion than science. By the way, this was the subject of a terrible movie, *Singularity* (translated in French by *Transcendence*, which had a very clear christic analogy). It is certainly a good advice to try and keep being objective.

Another fear is that humans will lose their thinking abilities relying too much on AI. Actually there has always been a transfer of tasks from humans to technology as the latter improves (for instance we are not so keen now about memorizing or computing since our books and computers do it so well for us). This debate was already going on at the time of Socrates with the technology of writing.

On the other hand, new technologies also come with new challenges, new ways to entertain. Computers have taken some abilities from humans, but also brought new abilities!

Also, a striking case in point about the AlphaGo experiment is that the new supremacy of algorithms on humans did not seem to deter humans from playing Go; actually, as Hassabis pointed out, the sales of Go games skyrocketed as a consequence of the competition. Playing with humans is a human activity after all.

5.5. Economics

In the field of economics and collective social affairs, things may be more worrying. So far the two main areas of concern are (a) robotization which may, to some extent and for some time, deprive humans of jobs, and (b) over personalization of interactions which may create bubbles and weaken the cohesion of society. Let us briefly discuss both.

The replacement of humans by robots and algorithms in certain tasks is a robust trend. In the case of AI it typically concerns medium skills (not the lowest, not the highest). As one example among many, in 2016 Foxconn has announced the replacement of 60,000 factory workers with robots.

The replacement of jobs may take place at the level of production sites (factories, companies) but it may also go in the interaction between humans and the task. Famous cases are those of photograph developers or travel agencies: those have mostly gone now, and users are handling it themselves with the technology. The same may occur with AI. Bankers, taxi drivers, generalist doctors, are all categories for which speculations are going on. Short-time traders are a famous category which was upset by algorithms, and there is hardly any doubt that finance will rely more and more on AI, jeopardizing classically trained finance officers.

In a Schumpeterian view, one may argue that these jobs will reappear in another format. But the characteristics of the present wave (so strong, so versatile, so quick, so globalized) make it possibly different from the previous ones; it may be that the "creative destruction" takes quite a long time by human life standards. The positive part of it is that there will be a whole chunk of new, high-level jobs devoted to interpreting, mastering or accompanying AI; so a bright future opens up for related jobs. In this context, statistician was elected CareerCast's "Best Job of 2016".

The second main concern is about the over efficiency of algorithms in personalizing the interaction. A few years ago the public diffuse fear with algorithms was mostly about a systematic uniform treatment for everybody, and it is ironic that now it is the other extreme which arises fears.

While personalization is certainly good news in certain fields (personal medicine for instance), it may be a problem in respect to solidarity. Insurance is about sharing risks, but if risks become tailor-made for each individual, the solidarity may effectively disappear. Profiled news will get citizens exactly the information which they wish to hear. Profiled politics will get the politicians to adapt their speech and convictions to exactly what their electors wish to receive; and it will help their campaign teams manipulating their opinions. Profiled suggestions will at the same time keep customers in their preferences and satisfy them (it was recently a sensation when it was found that Netflix used some 80,000 subcategories of users). Profiled advertisement will help trick customers into buying (while already trying to cover up the profiling by inserting some more random advertisement). All this is so efficient that it may have a destabilizing effect in a number of human affairs. Actually there is already ample clue that Big Data and AI methods played a significant role in both the controversial Brexit campaign and the controversial Trump campaign. In the same way as the narrative of Internet has been switching from freedom to mass surveillance and from sharing to bubble creation, it is possible that the narrative of AI will switch from fine support to manipulation. Already a notable book has appeared by the mathematician O'Neil with a strong title that says it all: *Weapons of Math Destruction: How Big Data increases inequality and threatens democracy.* This evolution, in conjunction with uncontrolled phenomena of fake news, "trolling", distortions and the like, may possibly be one of the most important problems facing humanity currently.

Let us also note that these topics are still controversial (as shown by Mark Zuckerberg's recent statement in disbelief of bubbles), that the environment is rapidly evolving, and that experiments are virtually impossible to do; so that it is not clear if the subject is amenable to science.

5.6. Human-AI interaction

One of the most fascinating features with AI is the interaction between humans and algorithms. This comes together with issues of human-human interactions.

It was already noted that human-AI matches do not deter human-human games. In medicine, algorithms may soon be able to outdo humans in diagnosis, but the patient-doctor relation is also one of human trust, and patients are certainly not ready to confide their emotions and fears to an algorithm (even though, on the contrary, some humans are much more comfortable to confide their traumatic experiments to neutral robots than to fellow humans). Because of this, or thanks to this, the paradigm of the medical doctor using AI is bound to be much more powerful than just the medical doctor or just AI.

Related to trust are subjects of responsibility and explanation procedure. It has already been widely debated that the responsibility issue for automatic car driving may be quite tricky. Also, even though automatic driving will certainly be more secure than human driving, some people object to, or fear putting their lives in the "hands" of an algorithm. In this case the job of driver cannot be considered similarly as the job of medical doctor: first because the affective bond between passenger and driver is much weaker than the bond between patient and doctor, but also because a human driver is very poor at securizing an automatic car (for taking the sequel of an automated procedure, or intervening in an emergency procedure, we humans are very bad). So the automatic car will basically have to be fully automated.

As a different example, consider the problem of detection of frauds by identity thefts. Already in the nineties, major companies were using insurance procedures and a certain number of rules to refuse transactions which were considered "fishy". In the absence of any explanation and any interaction, these gave rise to infuriating incidents. Nowadays banks are equipping themselves with AI-based recognition algorithms for what is fishy and what is not. When this is well designed, this comes with human arbitrage, explanation, and reaching out to the customer (through cell phone, for instance), so that responsibility is clear and interaction with the customer can take place. Of course budgetary issues, efficiency, trust to the consumer, and so on, will also be elements of choice for a bank company willing to improve in this direction.

It is certainly an interesting multidisciplinary subject to understand when the human-AI combination is an improvement and when it has to be fully human, or fully AI.

In any case, in such a context, the education of a wide audience to the basic principles of AI, with their powers and limitations, seems like a wise society option.

6. Bibliography

The bible of AI is the beautiful book by Stuart Russell and Peter Norvig, *Artificial Intelligence, A modern approach*. Even if it does not discuss so much the most modern algorithms, it is an amazing synthesis and a work of art.

Here are some interesting books about AI and its interaction with society: *Probably Approximately Correct* by Leslie Valiant (a milestone in the algorithmic approach to AI); *The Technological Singularity* by Murray Shanahan; *The Future of Machine Intelligence*, by David Beyer.

The Cerebral Cortex: An Evolutionary Breakthrough

Wolf Singer

The evaluation and encoding of perceptual relations

Living systems have to establish models of the world in which they evolve in order to be able to predict the outcome of actions and to thereby assure survival and reproduction. Establishing a good model of the world requires the detection of relevant and consistent relations between features of the environment and the efficient storage of these relations (rules). The simplest solution, found in virtually all neuronal systems, are relation encoding feed-forward circuits. Neurons tuned to respond to particular features of the environment converge on common target cells and these respond selectively to particular conjunctions of features provided that the gain of the input connections to these conjunction specific neurons are appropriately adjusted (Barlow, 1972). A particular relation among features gets encoded in the discharge rate of a neuron responding selectively to this relation. Because this neuron encodes always the same relation, one talks about a "labelled line code". In order to evaluate and encode combinations of relations (relations of higher order) this process of input recombination and gain adjustment is iterated across successive layers. This basic principle for the evaluation, encoding and classification of relational constructs has been implemented in numerous versions of artificial neuronal networks (Rosenblatt, 1958; Hopfield, 1987; DiCarlo and Cox, 2007; LeCun et al., 2015). The highly successful recent developments in the field of "deep learning" (LeCun et al., 2015), capitalize on the iteration of this principle in large multilayer architectures. As far as feed-forward connections are concerned, these artificial multilayer systems resemble the organization of sensory systems in the brain. Marked differences exist, however, with respect to other essential features. Feedback or recurrent lateral connections are prominent in brains (Markov et al., 2014; Bastos et al., 2015) but implementation of these architectural features is still rare in artificial systems. Moreover, the training mechanism used in technical systems for the supervised adjustment of the synaptic gain of connections, the so called "back-propagation algorithm" is biologically implausible and differs from both unsupervised and supervised learning mechanisms implemented in brains (Feldman, 2012; Singer, 2016).

A complementary way to detect and encode relations between signals is to evaluate temporal contingencies: If event A consistently precedes event B, event A is likely to be the cause of B, if A and B often coincide the two events most likely have a common cause and if A and B are uncorrelated in time they are most likely unrelated.

The learning rules implemented in nervous systems are adapted to evaluate such temporal relations and to translate them into lasting changes of coupling. Both the traditional Hebbian rules (Hebb, 1949) and the more recently discovered mechanisms (Stiefel et al., 2005; Holthoff et al., 2006; Carvalho and Buonomano, 2011; Grienberger et al., 2015) evaluate temporal relations among converging inputs as well as between pre- and postsynaptic activity (spike timing dependent plasticity – STDP), (Markram et al., 1997; Bi and Poo, 1998). The molecular mechanisms underlying these use-dependent synaptic modifications operate with a temporal precision in the millisecond range. This has two important implications: First, it implies that the precise timing of spikes in converging pathways matters in determining the occurrence and polarity of synaptic gain changes. Second, the mechanism subserving synaptic modifications not only evaluates simple covariations between pre-and postsynaptic firing rates, but also evaluates causal relations. It increases the gain of excitatory connections whose activity can be causally related to the activation of the postsynaptic neuron and it weakens connections whose activity could not have contributed to the postsynaptic response. Thus, temporal relations reflecting semantic relations among events are evaluated by time sensitive mechanisms and converted into lasting changes of the coupling strength of interacting neurons. In this way statistical contingencies among features of the sensory environment are translated into synaptic weight distributions in neuronal networks.

This time sensitivity of synaptic plasticity mechanisms has deep implications for signal processing. If the known plasticity mechanisms are used for the storage of relations in general, all relations eventually have to be expressed as temporal relations among distributed neuronal responses. Thus, for the association of responses that lack temporal structure or are offset in time by intervals longer than those bridgeable by the time constants of the molecular processes, mechanisms are required that endow neuronal responses with temporal structure and permit bridging temporal gaps. Otherwise rather different and still unknown mechanisms of synaptic plasticity have to be postulated.

Mechanisms for the generation of temporally structured activity

Results from an initially completely different line of research suggest the existence of mechanisms capable of imposing temporal structure on neuronal activity and of making perceptually related responses coherent in time. It had been discovered with multisite recordings from the visual cortex that cortical circuits have a high propensity to engage in oscillatory activity and that these intrinsically generated oscillations can become synchronized, leading to correlated firing of the synchronously oscillating neurons. Of particular importance is the fact that this temporal coordination is dynamically regulated. It is context sensitive and reflects meaningful relations among encoded features (Gray and Singer, 1989; Gray et al., 1989). One reason for the synchronization of neurons encoding features that should be bound together is that the reciprocal connections between the neurons are adaptive and undergo Hebbian modifications. As a consequence they strengthen between neurons that encode features which have a high probability of co-occurring in natural environments. As increased coupling among oscillators enhances the probability that they synchronize (Kuramoto et al., 1992), synchronization probability reflects the probability of feature contingencies. Thus, neurons encoding features that often co-occur, e.g. because they are constitutive for a particular object, have an increased likelihood to synchronize and to form a coherently active cell assembly. The saliency of their responses is enhanced jointly because synchronous discharges have a stronger impact on target neurons (Bruno and Sakmann, 2006). Thus, synchronously oscillating cells convey the message that the features they encode should be bound together because they have a high probability to be related in a meaningful way, e.g. because they are constitutive for a particular object and therefore have often co-occurred in the past (for review see Singer, 1999; Engel et al., 2001; Fries, 2009; Uhlhaas et al., 2009). Initially, the synchronization was seen as a relation defining mechanism mainly in the context of low-level visual processes such as feature binding and figure ground segregation. The reason was that synchronization probability reflected well the common Gestalt criteria for perceptual grouping and also reflected the architecture of the recurrent connections in the visual cortex that couple preferentially neurons coding for features which tend to be bound perceptually (Löwel and Singer, 1992). However, it soon turned out that synchronization of oscillatory activity is not confined to the visual system but a ubiquitous phenomenon (for review see Buzsáki et al., 2013). What makes these dynamic phenomena particularly interesting is the fact that they result from highly dynamic self-organizing processes that enable rapid reorganization of the temporal coherence of the responses of widely distributed groups of neurons. For this reason synchronization of oscillatory activity is now considered by many to serve a large number of different functions that have in common the requirement for temporal coordination of distributed neuronal responses. Examples are the enhancement of the saliency of responses (Fries et al., 1997; Biederlack et al., 2006), the dynamic formation of functional networks (Siegel et al., 2015; Fries, 2005; Deco and Kringelbach, 2016), the selection of responses by attention mechanisms (Fries et al., 2001a), the matching of top down signals with sensory input (Bastos et al., 2015), the parsing of speech signals (Ding et al., 2016) and the definition of relations in the context of learning and memory (Siapas et al., 2005; Fell et al., 2011; Yamamoto et al., 2014; for review see Singer, 2016).

Complex dynamics

As more laboratories engaged in multisite recordings, a prerequisite for the analysis of the correlation structure of neuronal dynamics, it became clear that oscillations with constant frequency sustained over long time intervals and synchronization of these oscillations with stable phase relations occur only under specific stimulation conditions. Especially the high frequency oscillations in the beta and gamma frequency range were found to exhibit a much more complex and variable dynamics than reported in the early days of their discovery. In the visual cortex, the frequency of stimulus-induced oscillations increases with the energy and the complexity of the stimuli and with their motion speed (Gray et al., 1990; Kayser et al., 2003; Ray and Maunsell, 2015; Lima et al., 2011). The amplitude of stimulus-induced oscillations decreases with the complexity of the inducing stimuli and increases with attention and expectancy (Lima et al., 2011; Fries et al., 2001a). Moreover, in awake behaving animals the oscillations are usually transient, occur as brief bursts (Pipa and Munk, 2011; Lundqvist et al., 2016) and are often coupled with the phase of other oscillations that have lower frequency (cross frequency coupling, Canolty et al., 2010). Accordingly the pairwise correlations between oscillating cell populations are also highly variable. They are transient and exhibit phase shifts that vary over time (for review see Fries et al., 2001b; Maris et al., 2016).

It has been argued that this high degree of variability of oscillations and synchrony is incompatible with a functional role of these dynamic phenomena (Ray and Maunsell, 2015). This critique concerns both the initial postulate that temporal coherence serves to encode relations and the formation of assemblies in distributed coding regimes (Singer, 1999) as well as the extension of this concept known as the Communication Through Coherence (CTC) hypothesis (Fries, 2005). However, others have argued that variability and non-stationarity of the dynamics are necessary properties for flexible processing in order to comply with the speed and versatility of cognitive operations (Roberts et al., 2013) and with the requirement to configure on the fly functional networks on the fixed backbone of the cortical connectome (Deco et al., 2016).

A unifying concept

These facts and arguments urge for a novel framework that attributes specific functions to the various manifestations of cortical dynamics and provides a cohesive interpretation of both low-dimensional states characterized by sustained frequency-stable oscillations and high-dimensional states with complex and rapidly changing correlation structure. The hypothesis proposed here is that the cortex exploits the high dimensional state space provided by the non-linear dynamics of recurrently coupled networks in order to perform flexible and efficient computation. In this framework, emphasis is placed on the characteristic parameters of self-organizing complex systems with non-linear dynamics. These parameters include changes in correlation structure, the entropy and dimensionality of distributed activity, network oscillations, synchronisation phenomena and phase shifts. The proposed computational strategy is likely to account for a number of hitherto poorly understood functions: The encoding of temporal sequences, the storage of vast amounts of information about the environment in the networks of sensory cortices, the ultrafast retrieval of information in processes requiring comparison between input signals and stored knowledge and the fast and effective classification of complex spatio-temporal input patterns.

Early theories of perception (von Helmholtz, 1867) have suggested that the brain interprets sparse input signals on the basis of an internal model of the world and these early ideas have received substantial support by studies on active sensing and predictive coding. The internal model is thought to build on inherited, genetically transmitted information and on knowledge acquired by experience. The information provided by this model is used to reduce redundancy in sensory signals and to facilitate perceptual grouping, feature binding, classification and identification. Because of the daunting complexity of the visual world, the store containing such an elaborate model must have an immense capacity. Moreover, read-out must be extremely fast to comply with behavioural evidence. When primates, including humans, scan their visual environment, they change the direction of their gaze on average four times a second. Thus, the prior knowledge required for the interpretation of a particular input pattern needs to be accessible within fractions of a second. The proposal is that these conditions can only be met if encoding, storage and processing of information take place in the *high-dimensional state space provided by a complex system with non-linear dynamics.*

The hypothesis

Neocortex, especially its supragranular compartment, is ideally suited to provide such a high-dimensional coding space. It is a recurrently coupled network (Gilbert and Wiesel, 1989; Stettler et al., 2002), whose nodes are feature selective and have a high propensity to oscillate (Gray and Singer, 1989). This network, so the assumption, provides the high-dimensional state space required for the storage of statistical priors, the fast integration with input signals and the representation of the results in a classifiable format. Statistical priors are supposed to be stored in the functional architecture of long-range horizontal connections and their synaptic weight distributions.

In this framework a number of experimentally testable predictions can be formulated. Spontaneous activity should reflect the dynamics of the structured network harbouring the entirety of latent internal priors and therefore exhibit very high dimensionality. Input signals are supposed to drive in a graded way the feature sensitive nodes *and* thereby constrain the network dynamics. If the evidence provided by the input patterns matches priors stored in the network architecture, the network dynamics will collapse to a specific substate, corresponding to a particular perceptual experience. Such a substate is expected to have a lower dimensionality than the resting activity, exhibit specific correlation structures and be metastable due to reverberation among nodes supporting the respective substate. Because these processes occur within a very high-dimensional state space, substates induced by different input patterns should be able to coexist (superposition of information), outlast the duration of the stimuli because of reverberation and be well segregated and therefore easy to classify. They can then either serve as input to the next cortical processing stage, where the same matching process is iterated, albeit with different, more global and abstract priors, or they can be classified by local readout units that directly feed into executive centres. According to this concept every cortical area has its own model of the world and these models differ in granularity and the degree of abstraction because of the mapping rules specifying the distribution and recombination of input signals across different processing stages.

Experimental evidence

Preliminary evidence is already available for some of these predictions. Developmental studies support the notion that the statistics of feature conjunctions in the outer world gets translated into cortical connectivity. Both feed forward as well as the reciprocal tangential connections in the visual cortex have been shaped during evolution and get further refined by experience dependent pruning to match the statistical properties of visual scenes (Hubel and Wiesel, 1962; Smith et al., 2015; Pecka et al., 2014; Eysel et al., 1998; Gilbert et al., 2009) according to a Hebbian mechanism (Singer and Tretter, 1976; Rauschecker and Singer, 1981; Löwel and Singer, 1992). In agreement with the hypothesis, the covariance structure of resting activity reflects the anisotropic layout of these connections (Kenet et al., 2003; Fries et al., 2001b; Bosking et al., 1997; Löwel and Singer, 1992; Gilbert and Wiesel, 1989), is modified by learning (Lewis et al., 2009; Kundu et al., 2013) and reveals hallmarks of an internal model of the environment (Berkes et al., 2011).

Ample evidence is also available for the ability of cortical circuits to engage in oscillatory activity in a wide range of frequencies and for stimulus dependent changes of correlations mediated by intracortical connections, both being hallmarks of recurrently coupled networks (for reviews see Singer, 1999; Buzsáki et al., 2013).

However, much less is known about how the ensuing oscillatory responses depend on the particular properties of natural stimuli, both in the spatial and temporal domain, how particular Gestalt principles of grouping translate into informative neuronal dynamics and how noise or ambiguity affect the efficiency of this encoding.

There are also indications that both sensory stimulation and top-down mechanisms related to attention induce changes in the dimensionality of states, because they can enhance synchronized oscillatory activity in distinct frequency bands (Gray et al., 1989; Lima et al., 2011; Churchland et al., 2010; Fries et al., 2001a). However, no direct analysis of dimensionality changes were performed in these studies. In Lima et al., 2011, the "attended" stimulus evoked gamma band oscillations of much higher amplitude than the "non-attended" stimulus. Thus, the expectancy of having to respond to a particular stimulus changed the correlation structure of the activity induced by this stimulus towards enhanced coherence. In other terms, anticipatory top down signals constrained the dynamics of an early visual area – most likely leading to a reduction of dimensionality.

Evidence is also available that cortical dynamics exhibit a fading memory for recent inputs. This is a hallmark of recurrent networks (Buonomano and Maass, 2009; Bertschinger and Natschläger, 2004; Lukoševičius and Jaeger, 2009) that greatly facilitates encoding and classification of sequences. As demonstrated in Nikolic et al. (2009) information about a briefly presented stimulus could persist for up to one second in the distributed responses of cortical neurons, could superimpose with information about subsequent stimuli and remain classifiable with high fidelity. We presented sequences of visual stimuli (letters and numbers) while recording from a large number of neurons in the visual cortex and trained a linear classifier on short segments (5-100 ms) of the high dimensional vector of responses obtained from a training set and then used the same classifier to identify the nature of the presented stimuli in a test set. We found that i) the information about a particular stimulus persists in the activity of the network for up to a second after the end of the stimulus, ii) the information about sequentially presented stimuli superimposes so that two subsequent stimuli can be correctly classified some time after the end of the second stimulus and iii) the information about stimulus identity is distributed across neurons and encoded both in the discharge rate of the neurons and in the precise timing of the spikes.

Finally, we have preliminary evidence that repeated exposure to a set of images changes the response properties of populations of neurons in the primary visual cortex, such that stimulus classification improves over time: we observe changes in the dynamics of the network through the state-space that favor the segregation of responses into stimulus specific substates. Hence the network "learns" in an unsupervised way about the statistics of feature constellations in frequently presented stimuli and this leads to enhanced segregation and classifiability of substates in the high-dimensional state space (Lazar and Singer, in preparation).

Concluding remarks

Despite considerable effort there is still no unifying theory of cortical processing and therefore numerous experimentally identified phenomena lack a cohesive theoretical framework. This is particularly true for the dynamic phenomena reviewed here because they cannot easily be accommodated in the prevailing concepts that emphasize serial feedforward processing and labelled line codes. Here we have proposed a concept that assigns specific functions to recurrent coupling and to features of the emerging dynamics. This concept is fully compatible with the robust evidence for labelled line codes and extends this notion by the proposal that precise temporal relations among the discharges of coupled neurons also serve as code for the definition of relational constructs both in signal processing and learning. We proposed a computational strategy that capitalizes on the high-dimensional coding space offered by reciprocally coupled networks. In this conceptual framework, information is distributed and encoded both in the discharge rate of individual nodes (labelled lines) and to a substantial degree also in the precise temporal relations among the discharge sequences of distributed nodes. The core of the hypothesis is that the dynamic interactions within recurrently coupled oscillator networks i) endow responses with the temporal structure required for the recoding of semantic relations into temporal relations, ii) exhibit complex, high dimensional correlation structures that reflect the signatures of an internal model stored in the weight distributions of the coupling connections and iii) permit fast convergence towards stimulus specific substates that are easy to classify because they occupy well segregated loci in the high-dimensional state space. The analysis of the correlation structure and consistency of these high-dimensional response vectors is still at the very beginning. However, methods are now available for massive parallel recording from large numbers of network nodes in behaving animals. It is to be expected, therefore, that many of the predictions formulated above will be amenable to experimental testing in the near future.

Irrespective of the outcome of these tests, available evidence suggests that nature – with the evolution of the cerebral cortex – succeeded to realize an extremely powerful, scalable and versatile computational strategy, that probably differs in some crucial aspects from algorithms implemented presently in artificial devices – and is probably not yet fully understood. As the intrinsic organization of cortical modules is strikingly similar across the whole cortical mantle, this strategy must be of a very general nature and capable of serving a wide spectrum of seemingly different cognitive and executive functions. This versatility is the likely reason for the tremendous evolutionary success of this structure. Its expansion is the hallmark of the evolutionary changes that distinguish the human species from its nearest neighbours, the great apes, and the cognitive functions resulting from the addition of cortical modules ultimately enabled humans to initiate cultural evolution.

Acknowledgments

This work was supported by the Max Planck Society and the Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, the DFG grant NI 708/51 and the Frankfurt Institute for Advanced Studies.

References

Barlow, H.B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* 1, 371-394.

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron 85*, 390-401.

Bi, G.Q., and Poo, M.M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464-10472.

Biederlack, J., Castelo-Branco, M., Neuenschwander, S., Wheeler, D.W., Singer, W., and Nikolic, D. (2006). Brightness induction: Rate enhancement and neuronal synchronization as complementary codes. *Neuron* 52, 1073-1083.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331, 83-87.

Bertschinger, N., and Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Comput.* 16, 1413-1436.

Bosking, W.H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* 17, 2112-2127.

Bruno, R.M., and Sakmann, B. (2006). Cortex is driven by weak but synchronously active thalamocortical synapses. *Science* 312, 1622-1627.

Buonomano, D.V., and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Rev. Neurosci.* 10, 113-125.

Buzsáki, G., Logothetis, N., and Singer, W. (2013). Scaling brain size, keeping timing: Evolutionary preservation of brain rhythms. *Neuron* 80, 751-764.

Canolty, R.T., Ganguly, K., Kennerley, S.W., Cadieu, C.F., Koepsell, K., Wallis, J.D., and Carmena, J.M. (2010). Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *Proc. Natl. Acad. Sci. USA* 107, 17356-17361.

Carvalho, T.P., and Buonomano, D.V. (2011). A novel learning rule for long-term plasticity of short-term synaptic plasticity enhances temporal processing. *Frontiers Integr. Neurosci.* 5, 20: 1-11.

Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B.*, et al.* (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neurosci.* 13, 369-378.

Deco, G., and Kringelbach, M.L. (2016). Metastability and coherence: Extending the communication through coherence hypothesis using a whole-brain computational perspective. *Trends Neurosci.* 39, 125-135.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333-341.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neurosci.* 19, 158-164.

Engel, A.K., Fries, P., and Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Rev. Neurosci.* 2, 704-716.

Eysel, U.T., Eyding, D., and Schweigart, G. (1998). Repetitive optical stimulation elicits fast receptive field changes in mature visual cortex. *NeuroReport* 9, 949-954.

Feldman, D.E. (2012). The spike-time dependence of plasticity. *Neuron* 75, 556-571.

Fell, J., Ludowig, E., Staresina, B.P., Wagner, T., Kranz, T., Elger, C.E., and Axmacher, N. (2011). Medial temporal theta/alpha power enhancement precedes succesful memory encoding: Evidence based on intracranial EEG. *J. Neurosci.* 31, 5392-5397.

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Rev. Neurosci.* 32, 209-224.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.* 9, 474-480.

Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001a). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291, 1560-1563.

Fries, P., Neuenschwander, S., Engel, A.K., Goebel, R., and Singer, W. (2001b). Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neurosci.* 4, 194-200.

Fries, P., Roelfsema, P.R., Engel, A.K., König, P., and Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proc. Natl. Acad. Sci. USA* 94, 12699-12704.

Gilbert, C.D., and Wiesel, T.N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* 9, 2432-2442.

Gilbert, C.D., Li, W., and Piech, V. (2009). Perceptual learning and adult cortical plasticity. *J. Physiol.* 587, 2743-2751.

Gray, C.M., and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. USA* 86, 1698-1702.

Gray, C.M., Engel, A.K., König, P., and Singer, W. (1990). Stimulus-dependent neuronal oscillations in cat visual cortex: Receptive field properties and feature dependence. *Eur. J. Neurosci.* 2, 607-619.

Gray, C.M., König, P., Engel, A.K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334-337.

Grienberger, C., Chen, X., and Konnerth, A. (2015). Dendritic function in vivo. *Trends Neurosci.* 38, 45-54.

Hebb, D.O. (1949). *The Organization of Behavior* (New York, John Wiley & Sons).

Holthoff, K., Kovalchuk, Y., and Konnerth, A. (2006). Dendritic spikes and activity-dependent synaptic plasticity. *Cell Tissue Res.* 326, 369-377.

Hopfield, J.J. (1987). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proc. Natl. Acad. Sci. USA* 84, 8429-8433.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* 160, 106-154.

Kayser, C., Salazar, R.F., and König, P. (2003). Responses to natural scenes in cat V1. *J. Neurophysiol.* 90, 1910-1920.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., and Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature* 425, 954-956.

Kundu, B., Sutterer, D.W., Emrich, S.M., and Postle, B.R. (2013). Strengthened effective connectivity underlies transfer of working memory training to tests of short-term memory and attention. *J. Neurosci.* 33, 8705-8715.

Kuramoto, Y., Aoyagi, T., Nishikawa, I., Chawanya, T., and Okuda, K. (1992). Neural network model carrying phase information with application to collective dynamics. *Prog. Theor. Phys.* 87, 1119-1126.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436-444.

Lewis, C.M., Baldassarre, A., Committeri, G., Romani, G.L., and Corbetta, M. (2009). Learning sculpts the spontaneous activity of the resting human brain. *Proc. Natl. Acad. Sci. USA* 106, 17558-17563.

Lima, B., Singer, W., and Neuenschwander, S. (2011). Gamma responses correlate with temporal expectation in monkey primary visual cortex. *J. Neurosci.* 31, 15919-15931.

Löwel, S., and Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science* 255, 209-212.

Lukoševičus, M., and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Sci. Rev.* 3, 127-149.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and beta bursts underlie working memory. *Neuron* 90, 152-164.

Maris, E., Fries, P., and van Ede, F. (2016). Diverse phase relations among neuronal rhythms and their potential function. *Trends Neurosci.* 39, 86-99.

Markov, N.T., Ercsey-Ravasz, M.M., Ribeiro Gomes, A.R., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M.A., Sallet, J., Gamanut, R., Huissoud, C., Clavagnier, S., Giroud, P., Sappey-Marinier, D., Barone, P., Dehay, C., Toroczkai, Z., Knoblauch, K., Van Essen, D.C., and Kennedy, H. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex* 24, 17-36.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275, 213-215.

Nikolic, D., Häusler, S., Singer, W., and Maass, W. (2009). Distributed fading memory for stimulus properties in the primary visual cortex. *PLoS Biol.* 7: *e1000260*, 1-19.

Pecka, M., Han, Y., Sader, E., and Mrsic-Flogel, T.D. (2014). Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron* 84, 457-469.

Pipa, G., and Munk, M.H.J. (2011). Higher order spike synchrony in prefrontal cortex during visual memory. *Frontiers Comput. Neurosci.* 5: 23, 1-13.

Rauschecker, J.P., and Singer, W. (1981). The effects of early visual experience on the cat's visual cortex and their possible explanation by Hebb synapses. *J. Physiol. (Lond.)* 310, 215-239.

Ray, S., and Maunsell, J.H.R. (2015). Do gamma oscillations play a role in cerebral cortex? *Trends Cogn. Sci.* 19, 78-85.

Roberts, M.J., Lowet, E., Brunet, N.M., Ter Wal, M., Tiesinga, P., Fries, P., and De Weerd, P. (2013). Robust gamma coherence between macaque V1 and V2 by dynamic frequency matching. *Neu-*

ron 78, 523-536.

Rosenblatt, F. (1958). The perceptron. A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386-408.

Siapas, A.G., Lubenov, E.V., and Wilson, M.A. (2005). Prefrontal phase locking to hippocampal theta oscillations. *Neuron* 46, 141-151.

Siegel, M., Buschman, T.J., and Miller, E.K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science* 348, 1352-1355.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron* 24, 49-65.

Singer, W. (2016). Synchronous oscillations and memory formation. In *Comprehensive Neurobiology of Learning and Memory*. Byrne, J.H., Menzel, R. (Eds.). Elsevier. New Edition (in print).

Singer, W., and Tretter, F. (1976). Unusually large receptive fields in cats with restricted visual experience. *Exp. Brain Res.* 26, 171-184.

Smith, G.B., Sederberg, A., Elyada, Y.M., Van Hooser, S.D., Kaschube, M., and Fitzpatrick, D. (2015). The development of cortical circuits for motion discrimination. *Nature Neurosci.* 18, 252-261.

Stettler, D.D., Das, A., and Bennett, J.G., C.D. (2002). Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron* 36, 739-750.

Stiefel, K.M., Tennigkeit, F., and Singer, W. (2005). Synaptic plasticity in the absence of backpropagating spikes of layer II inputs to layer V pyramidal cells in rat visual cortex. *Europ. J. Neurosci.* 21, 2605-2610.

Uhlhaas, P.J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D., and Singer, W. (2009). Neural synchrony in cortical networks: history, concept and current status. *Frontiers Integrat. Neurosci.* 3, 1-19.

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik* (Leipzig, Leopold Voss Verlag).

Yamamoto, J., Suh, J., Takeuchi, D., and Tonegawa, S. (2014). Successful execution of working memory linked to synchronized high-frequency gamma oscillations. *Cell* 157, 845-857.

Comments: The Ethics of Artificial Intelligence

Stephen Hawking

It is not clear whether intelligence has any long-term survival value. Bacteria multiply and flourish without it. However, intelligence is central to what it means to be human. It allows us to learn more about ourselves and our environment, and as a species it gives us competitive edge. Everything our civilized action has achieved is a product of human intelligence. I regard it a triumph that we, who are ourselves mere stardust, have come to such a detailed understanding of the universe in which we live.

The potential benefits of creating beneficial artificial intelligence are huge. Used as a toolkit, AI can augment our existing intelligence to open up advances in every area of science and society. However, it will also bring dangers. Governments around the world are already funding an AI arms race. And in the future, AI could develop a will of its own, a will that is in conflict with ours...

In short, AI will be either the best or the worst thing ever to happen to humanity. We do not yet know which. That is why in 2014 I and a few others, called for more research to be done in this area. I feel it is important to have this discussion now, in order that the research and its applications benefit society as a whole.

Optimal Strategies for Decision-Making and Their Neural Basis

Alexandre Pouget

Understanding how animals and humans make decisions is one of the key questions in neuroscience, economics and artificial intelligence. Decisions come in all sort of flavors but, in neuroscience, most of the work so far has focused on two types known as perceptual decision-making and value-based decision-making. In the case of perceptual decision-making, subjects must decide on the state of a stimulus based on sensory evidence. For instance, subjects might have to determine whether a set of dots is moving rightward or leftward based on a short movie [1]. In value-based decision-making, subjects have to choose between items with subjective values, such as choosing between two types of desserts [2]. In this case, and contrary to perceptual decision-making, there is no objectively correct answer since the value of an item is necessarily specific to the taste and preference of each subject.

The theory as well as the neural basis of binary perceptual decision-making are reasonably well understood [3]. A class of models known as drift diffusion model, or DDM for short, have been shown to predict remarkably well the percentage of correct responses as well as the reaction times as a function of the task difficulty. DDM are based on the assumption that subjects receive scalar samples at every time steps from their perceptual system, which serve as evidence for or against the two possible choices [4, 5]. For instance, in the case of leftward versus rightward motion, positive samples can be assigned to leftward motion, in which case negative samples would count as evidence for rightward motion. To be more specific, the samples are assumed to be drawn from a Gaussian distribution whose mean is proportional to the strength of the visual motion and its sign is related to the direction (positive for leftward motion in our example). The DDM simply takes the sum over time of the samples and stops whenever the accumulated evidence reaches one of two symmetric bounds. If the positive bound is hit first, the model 'chooses' left, while it chooses right if the negative bound is hit first. Critically, this simple strategy, and variations thereof, has been shown to optimize the number of correct answers per unit of time. Moreover, the response of neurons in several cortical areas suggests that they sum their momentary evidence, and stop integrating when their activity reaches a specific level, just as in a DDM. Therefore, it appears that, to a first approximation, neural circuits implement DDMs for binary perceptual decision-making.

Behavioral studies also suggest that humans and animals use a similar strategy for binary value-based decision-making [6]. In this case, it is assumed that the brain generates two samples at each time step drawn from two Gaussian distributions with means equal to the subjective values of the two items being considered. Specialized circuits compute the difference between the two samples at each time step and then sum this difference over time until an upper or lower bound is hit, with each bound associated with one particular choice. This strategy is appealing from a neural point of view since it requires the same circuits as for perceptual decision-making. However, unlike in the case of perceptual decision-making, it is unclear whether this strategy is optimal, i.e., whether it maximizes the number of rewards (or value) per unit of time across multiple trials. In fact, there are reasons to believe that this is not an optimal strategy.

Consider a choice between two items with nearly equal high values. This would be like choosing between your favorite ice cream and your favorite cake. In this case, the difference between the value samples at each time step will be very small on average, in which case the accumulation process will take a long time to hit either of the bounds. Therefore, this model predicts that, when confronted with two equally good choices, subjects should take a particular long time to decide, even though at the end of the decision, they are guaranteed to end up with a good choice. This is strange: it would make a lot more sense in this case to decide quickly rather than to procrastinate. A different class of models known as race models seems better suited to this type of situations. A race model uses two accumulators, one per choice, each summing the samples for one choice exclusively. The process stops whenever one of the accumulators reaches a preset bound. If both choices are highly valued, both accumulators will grow quickly and hit their bound in a short time.

Curiously, however, it is well known that subjects do take a very long time to decide between two items they like, a result consistent with the DDM, not the race model. In fact, we have all experienced this problem. If a restaurant menu contains two items you really like, you know you will agonize over the options for a long time. Other behaviors in value-based decision-making are just as puzzling. For instance, subjects have a particular hard time deciding between two items they really like if a third low-value choice is offered, even if the subjects never select this third choice. These strange interactions between options, and others results, have often been used to argue that humans rely on a suboptimal strategy for value-based decision-making. The problem with this conclusion is that, up until recently, the optimal strategy for value-based decision-making was unknown, making it difficult to determine whether a particular strategy is optimal or not. We have recently revisited this issue and used the theory of dynamic programming to derive the optimal strategy. In the case of binary decision-making, the answer was counterintuitive: DDM models do provide the optimal strategy in the sense that they optimize the reward rate [7]. Although it seems strange that the optimal decision policy involves waiting a long time when deciding between two good choices, this strategy has the advantage of leading to very fast responses when the difference in value between the two options is large, i.e., when the choice is easy, which increases the reward rate. As a result, DDMs work better than race models when the difficulty of the choices varies across trials.

For choices involving N options where N is greater than 3, the optimal solution requires N coupled accumulators, where the coupling comes from the fact that the mean across all the accumulators must be subtracted from each accumulator. As a result, the choice between two high-value items can be influenced by a third low-value item, because this item will contribute to the common mean term. As a result, the optimal strategy exhibits the same behavior as humans: it becomes hard to choose between two high-value items in the presence of third items even if it is never chosen.

Our work also shows that the neural implementation of the optimal strategy requires a very specific operation known as normalization with corresponds to the subtraction of the mean of the momentary evidence. Normalization has been reported in neural circuits involved in value-based decision-making but its role had remained obscure. Our analysis suggests that it is in fact a key operation that allows neural circuits to make near optimal decisions.

While it represents a significant step forward, this work only considers the simplest form of value-based decisions, such as choosing between two desserts. For much more complex cases, such as deciding a major in college, the decision involves very complex form of reasoning which cannot be captured by the simple DDM model we have explored here. Complex reasoning is believed to rely on probabilistic inference over rich data structures such as trees or graphs driven by a temporal stream of evidence. It remains to be seen how neural circuits could represent data structures of this type and implement efficient inference over such neural representations.

It is quite likely that this research could greatly benefit from the recent work in artificial intelligence on neural networks with long term memory ([8, 9]). Up until recently, learning algorithms were exclusively designed for networks with short-term memory but these algorithms have now been generalized to train networks composed of two sub-networks, one dedicated to long term-storage and one more specialized for online computation. A similar dichotomy appears to exist, to a first approximation, in the mammalian brain where the hippocampus is specialized in long-term storage while the cortex is more specifically focused on online processing. One can imagine storing knowledge about a particular domain in the long-term memory of the system and using the other network to integrate over time the information extracted from long-term memory. The current architectures used in AI lack full biologically plausibility but they provide an extremely promising starting point. Such a project would illustrate once again the extraordinary potential of artificial intelligence as a source of inspiration for research in Neuroscience.

References

1. Newsome, W.T., K.H. Britten, and J.A. Movshon, Neuronal correlates of a perceptual decision. *Nature*, 1989. 341(6237): p. 52-4.
2. Rangel, A., C. Camerer, and P.R. Montague, A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci*, 2008. 9(7): p. 545-56.
3. Shadlen, M.N. and R. Kiani, Decision making as a window on cognition. *Neuron*, 2013. 80(3): p. 791-806.
4. Kiani, R. and M.N. Shadlen, Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 2009. 324(5928): p. 759-64.
5. Gold, J.I. and M.N. Shadlen, The neural basis of decision making. *Annu Rev Neurosci*, 2007. 30: p. 535-74.
6. Krajbich, I., C. Armel, and A. Rangel, Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 2010. 13(10): p. 1292-8.
7. Tajima, S., J. Drugowitsch, and A. Pouget, Optimal policy for value-based decision-making. *Nat Commun*, 2016. 7: p. 12400.
8. Graves, A., et al., Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016. 538(7626): p. 471-476.
9. Henaff, M., et al., *Tracking the World State with Recurrent Entity Networks*. 2016.

Motivations and Drives Are Computationally Messy[1]

Patricia Smith Churchland

Deep learning strategies have achieved successes that have surprised those who favor a traditional write-a-program approach to artificial intelligence. The dramatic success of AlphaGo in defeating Lee Sodol, one of the very best Go players in the world, in 4 games out of 5, was a very public vindication for those who took advantage of increased computer power and steadfastly improved the performance of Artificial Neural Nets (ANNs).

Although insights relevant to neuroscience may emerge from how ANNs work, so far the accomplishments of deep learning are essentially confined to pattern recognition problems, and indeed, to pattern recognition in one single domain per machine. Striking as these pattern recognition feats are, any animal whose capacity was confined to one category of pattern recognition, even if it is brilliant pattern recognition for that category, would be an evolutionary casualty.

The success of ANNs notwithstanding, it must be acknowledged that the behavior is a far cry from what a rat or a human can do, as they live out their lives on the planet. But could we not just scale up pattern recognition so that it could be the equal of the accomplishments of a rat or a human? My judgment is that this is a lot harder than trumpeting the words "scale up pattern recognition" and confidently waving your hands.

All animals have the capacity to maintain homeostasis. Their inner milieu must stay within a restricted temperature range, the circuitry must organize the animal's movements so that it has sufficient energy, water, and oxygen. Homeostasis is anything but a simple business, and as endotherms emerged, a much narrower temperature range had to be maintained on pain of death. Maintaining homeostasis often involves competing values and competing opportunities, as well as trade-offs and priorities. While enthusiastic ANN designers might bet serious money that simple extensions to learning algorithms could easily handle these jobs, I would bet that a

[1] Special thanks to Anne Churchland and Adrienne Fairhall for their ideas and observations.

wholly new wrinkle is needed. To mimic what evolution discovered over many hundreds of millions of years may be much more difficult than scaling up pattern recognition in a really big ANN with a cobbled-up trick or two.

All vertebrate species are able to detect threats, and to behave appropriately in response to motivations to survive, thrive, and reproduce. In this domain, as well as maintaining homeostatic functions, there are typically competing values and competing opportunities: should I mate or hide from a predator, should I eat or mate, should I fight or flee or hide, should I back down in this fight or soldier on, should I find something to drink or sleep, and so forth. The underlying neural circuitry for essentially all of these decisions is understood if at all, then only in the barest outline. And they do involve some sense of "self", which is a sort of brain-construction, not a feat of pattern recognition. Biological evolution favors those whose values and decision allow them to survive long enough to pass on their genes.[2] But the dynamics of the neural business is poorly understood.

Not everything in the world is of equal interest across species. Dung beetles are highly motivated to seek dung; squirrels are not. On the other hand, squirrels are keen to find nuts and to distinguish between fresh and stale nuts, but dung beetles care not. Dogs are typically motivated to sniff the behinds of other dogs, humans are not. And so forth. Goals and plans to achieve them are internal to the animal. Often the stimuli are essentially neutral, but for the animal's goal.[3] So there are internal settings, some acquired but some not, that manipulate such pattern recognition functions as these animals deploy.

Mammals, at least, appear to build causal models of the world. Since *causality* is a stronger relation that *correlation*, the standard real-brain tactic for upgrading to causality involves intervention and manipulation. This may be easier for an animal that can move than a stationary, if deep, learning machine. The capacity for movement, especially if you have limbs, is anything but simple, that we do know.

So far as I can tell, no one has a genuinely workable plan concerning how to capture motivation and drives and motor control into a plausible

[2] Unless, for example, they are honeybees, where their decisions are geared to passing on the genes of the queen bee.

[3] Vikram Gadagkar et al., (Dec 8 2016), *Science*. Dopamine neurons encode performance error in singing birds.

pattern recognition regime.[4] The problem is not straightforward because motivations come in different packages – hunger is different from thirst, which is different from lust or fear or curiosity or joy. Temperament comes in different "colors on a spectrum" – introvert or less so, risk averse or less so, energetic or less so, and so forth. These factors change with age, with time of day, with sleep, with mood changes, and with diseases. These functions might be understood as the drivers of pattern recognition jobs in real animals, not as pattern recognition themselves.[5] The hypothalamus and brainstem, which are crucial in the nervous system of real animals for the managements of these functions, are not yet well understood in neuroscience, to put it mildly. The circuitry is ancient, and extremely complicated. It does not look like it is just doing pattern recognition, whatever that might mean in this context.

Why not just assign different numbers to different motivational forms, and add a plus or a minus to mark strength? Ditto for temperament, ditto for levels of arousal? One problem is that the "just assign numbers" idea blurs the differences between kinds of motivational states, and the relevance of such differences to decision-making.[6] Fear and lust both involve arousal, but they are different and have different trajectories in brain space. In any case, the idea needs to be fleshed out to show how such a system makes decisions to behave that are comparably suitable to those of a fruit fly or a rat.

There is a precise pattern of causality between kinds of functions that so far is not captured by the "pattern recognition" paradigm. Until we understand much more about the nervous systems of animals, we cannot specify with any precision the nature of the causal relationships managed by the hypothalamus, brainstem and basal ganglia, or how adequately to model what is going on.

What really are depression or exuberance or patience or tenacity or resilience? Not just pattern recognition, almost certainly. How do these phenomena interact with motivation, drive and desires? What is the role

[4] Though I should mention that Yann LeCun has some ideas about internal motivation, on a "happy or not happy" dimension. This could be a fruitful start.

[5] Sejnowski, T.J. Poizner, H. Lynch, G. Gepshtein, S. Greenspan, R. Prospective Optimization, Proceedings of the IEEE, 102, 799-811, 2014.

[6] Raposo, D., Kaufman M.T., and Churchland A.K. (2014) Nature Neuroscience: 17(12): 1784-92. A category free neural population supports evolving demands during decision-making.

of neuromodulators in these and other phenomena such as curiosity or wanderlust or sociality or aggression? Neuroscientists are indeed exploring these phenomena, one and all, but their neurobiological bases are not easy to plumb. That they are simply, at bottom, forms of pattern recognition seems unlikely at this point.

Neuromodulation more generally seems to affect what is learned, when something is learned, and how it is learned, yet so far, ANN modelers give it no role whatsoever, as though neuromodulation is a "mere biological by-product" – existing in us because we are the products of guess-and-by-golly evolution, but definitely not the crux of bit of engineering magic.

Perhaps they are right. But consider. Because so much is unknown about the relevant neurobiology, perhaps what are waved off as activities *incidental* to intelligence may turn out to be essential features that "scaffold" real intelligence. The analogy here is with early brain researchers who thought that all the cognitive action was in the ventricles, not in the brain itself.[7] The thing is, apart from biological intelligence, we have no understanding of what to count as real intelligence – we have no other criteria. For example, a person who is a great mathematician may be a dud in practical matters of health, finance, sex, and food. Mathematicians may say she is intelligent, but financiers or fighter pilots will not.

Go ahead and market something as "intelligent", but if it is brittle, lacks flexibility and "common sense"[8] and has nothing approximating motivation or drive or emotions or moods, it may be difficult to persuade the rest of us that it is intelligent in the way that biological entities can be. Redefine "intelligence" you may, but the redefinition *per se* will not make the machine intelligent in any generally recognizable sense.

At least some of the dystopian predictions concerning the eventual threat to humans of intelligent machines depend on the tendentious assumption that engineers have now cracked the problem of intelligence in a machine. However dramatic such predictions may be, they are not tethered by a biological understanding of what makes for intelligence, and they certainly are not grounded in a biological understanding of the nature of motivation and goals. Although it is always ticklish to downplay dystopian predictions lest one seem indifferent, it is nevertheless worth

[7] See for example, Hieronymous Brunschwig (1497, second edition 1525) *The Noble Experyence of the Vertuous Handy Werke of Surgerie.* Descartes seems to have thought along similar lines.

[8] See also Yann LeCun on this point.

balancing dystopian predictions by noting that our realistic time horizon is only about five to ten years out. Machines that care and desire control are unlikely within that time horizon.[9]

[9] Thanks to Terry Sejnowski for comments on an earlier draft.

Children and Robots

Antonio M. Battro and Magela Fuzatti*

One way to explore the relation between artificial intelligence and human consciousness is to look at the way children build robots and program them. It seems that when children construct a robot as a new toy or a new instrument not only are they putting together "atoms and bits" using physics and information technology, but they also attribute to their creation some "mental" properties. We will try to briefly analyze this phenomenon and the acquisition of the new robotics skills in our lives.

Animism, artificialism and roboticism

Jean Piaget described the combination of "animism" and "artificialism" in the cognitive development of young children some eighty years ago in his celebrated book *La répresentation du monde chez l'enfant,* 1938 (*The child's conception of the world*). For Piaget animism "is the tendency that the child has to ascribe life and consciousness to inanimate beings" and artificialism is the idea that "nature is directed by people or at least gravitates around people".[1]

Today, millions of children around the world have access to robots, and many acquire the skills to construct and program them since primary education. We can coin the term *roboticism* as *the belief that the robot is an autonomous object with liberty to make decisions.* As such, roboticism could be under-

*Director, Laboratorios Digitales, Plan Ceibal, Uruguay.

[1] Jean Piaget (Battro, 1973): Artificialism, 4 developmental periods: I) "nature is directed by people or at least gravitates around people", II) mythological artificialism "appears from the moment when the child asks questions about the origin of things or answers questions which we put to him", III) technical artificialism, "the child continues to attribute to man the general arrangement of things, but limiting his action to the operations which can be technically achieved", IV) immanent artificialism, "nature is the heir of man and manufacturer like a workman or artist … it considers things as the product of human manufacturing, much more than it attributes to the manufacturing activities".

Animism "is the tendency that the child has to ascribe life and consciousness to inanimate beings". "The child ascribes to things moral attributes rather than psychological": I) diffuse animism is the general tendency of children to confuse the living and the inert", II) systematic animism is the group of explicit beliefs which the child has. The clearest one of them is that children believe that the heavenly bodies follow them".

stood as a new synthesis of animism and artificialism. When children think that a robot is an "animated artifact" they are in fact putting together both beliefs. On the one hand, they have *constructed* an artifact that is working as they have predicted; on the other, they have given instructions that are followed *automatically* by the robot, without human control. This mixture of dependency (rules) and autonomy (freedom) is quite unique. At the time of Piaget, in the pre-digital era, it was impossible even to imagine such a combination in the hands of children. Children have now constructed an object that becomes in a certain sense "independent": a robot that works without human help. Of course, the child has prescribed the kind of work the robot performs by means of a set of rules in a well-defined environment; the robot's apparent freedom is limited by this particular environment and rules. Our thesis is that early hands-on experience in the construction and programming of robots may lead children to discover the real power and limits of artificial intelligence. However, we would need more field research and extended cognitive studies to disentangle the new composite of beliefs that may continue into adulthood in relation to robots.

A personal history: playing with turtles

My experience (AMB) with children and robots started in the nineteen-sixties when Seymour Papert promoted the revolutionary project of deploying computers in the classrooms and began to explore the way children learn to program and construct/control a robot. I met Papert in the early 60s at the Center of Genetic Epistemology directed by Jean Piaget in Geneva. At that time he was developing his cognitive theory of *constructionism,* as a complement to Piaget's theory of *constructivism*. In Piaget's words, *constructivism* is the "formal obligation of constantly transcending the systems already constructed to assure non-contradiction" (Piaget & Beth, 1961). In contrast, Papert's *constructionism* was more focused on the dynamics of developmental change than on the logical stability of mental structures or stages. Both authors were clearly opposed to instructionism in education. Papert left Geneva for MIT, where he became director of the AI Lab (1967) with Marvin Minsky. I would now like to pay a most sincere tribute to my dear friends Seymour and Marvin who passed away this year, we owe them so much.

With Wally Feurzeig, Papert created LOGO, a programming language inspired by LISP, and introduced it in schools in the 80s. He became professor at the MIT Media Lab, founded by Nicholas Negroponte, but, unfortunately, our Master Piaget died in 1980 and wasn't able to see his formidable breakthrough in education. Together with Horacio C. Reggini, we soon followed his example in Argentina, where we created "Asociación Amigos de Logo" to promote the practice of Logo in elementary and special education schools. From the very beginning Papert fully supported my work with disabled children with the help of computers. His seminal book, *Mindstorms* (1980) was followed by Reggini's *Alas para la mente. Logo: un lenguaje de computadoras y un estilo de pensar* (1982), a book that had a great impact in our Latin American region. Logo was used in many different school activities; one of the most popular was to draw on the computer screen using elementary geometric procedures to move a pointer, a small triangle that was called a "turtle".

The name "turtle" has an interesting history in cybernetics and was inspired by the (analog) robot created by neurophysiologist William Grey Walter (1910-1977) in the 1940s (http://www.rutherfordjournal.org/article020101.html). Grey Walter's "tortoise" had three wheels, light and touch sensors, steering and propulsion motors and two vacuum tube analog processors that allowed the robot to explore and avoid obstacles, and to simulate positive and negative phototropism. It was named *Machina Speculatrix* (http://www.extremenxt.com/walter.htm) and was used to simulate some brain mechanisms and simple behaviors. Gray Walter elaborated these ideas in an influential book *The living brain* (Norton, New York, 1963), that became a source of inspiration for many of us.

Following this trend the first robot programmed by children in the 1980s was a Logo "turtle". The turtle was a very simple and robust robotic vehicle, produced by Terrapin Co. (terrapins are small semi-aquatic turtles), equipped with two wheels, electric motors, a transparent shell, a ring as a contact sensor and the whole robot connected to a computer (https://www.terrapinlogo.com/). Children were taught to write modular and recursive Logo programs with a few simple commands such as forward (number), back (number), turn (degrees, left, right), pen down (to write the trajectory on the floor), pen up (stop drawing), etc. My early work with children and robots began with these charming Logo turtles in a variety of settings, working initially with disabled kids in a hospital and in a few elementary schools. Incidentally at that time very few physicians or clinical psychologists were using computers. One landmark event, perhaps the first of its kind in the world, was the communication by computer we managed to establish between deaf children in Argentina and the United States with Percival J. Denham in 1988. This ended the communication gap established by Graham Bell when he invented the telephone and excluded ipso facto all deaf people from the system, a cruel paradox because he was a dedicated teacher of the deaf (Battro & Denham, 1989). Another unforgettable experience in the early 1980s was to watch our turtle being moved in Buenos Aires by Logo commands from Boston via modem and telephones lines, well before the internet. Today remote controlled robots are very common, even in schools.

It is interesting that the robotic work with children started with turtles and not with androids/humanoid robots… Roboticism was not "about humans" in the early days of robotics. Today things are changing rapidly and androids capture the imagination of both children and adults. There are so many androids on the market today at child's reach. Nevertheless, there is an essential difference between *buying* a robot and *constructing* one. Both modalities can be used in the classroom or at home, but only constructing gives *transparency* to the inner organization of the machine, which is hidden in the manufactured robot.

Current research and implementations

A source of inspiration for all of us is the work of the Laboratory of Lifelong Kindergarten at the MIT Media Lab, directed by Mitchel Resnick, creator of Scratch, a very useful programming language to use in elementary robotics with children (http://scratch.mit.edu). MIT is one of the leading places that launched the LegoLogo equipment for children, where Lego blocks are provided with gears, motors and sensors connected to a computer (https://llk.media.mit.edu/press/).

A recent development by Mariana Umashi Bers, also a disciple of Papert, now at Tufts, is the ScratchJr software (available as a free download on iPad and Android tablets), which is making it possible to program robots without even knowing how to read or write. She calls it KinderLab robotics. The job is done using solid objects with various symbols for SPIN, SING, STOP, and so forth, that can be put together as a "solid sentence" that commands a small wireless robot called Kibo. It is all about "learning to code" through actions.[2] The same idea is at the core of the spirit of *La main à la pâte* foundation, which promotes inquiry-based learning, including learning by "doing robotics" (www.fondation-lamap.org, Calmet, Hitzig & Wilgenbus, 2016).

[2] https://www.youtube.com/watch?v=jOQ-9S3lOnM&list=PLXzFU_7W4n0t-5suyfWPX6R-zUpd1MQ876

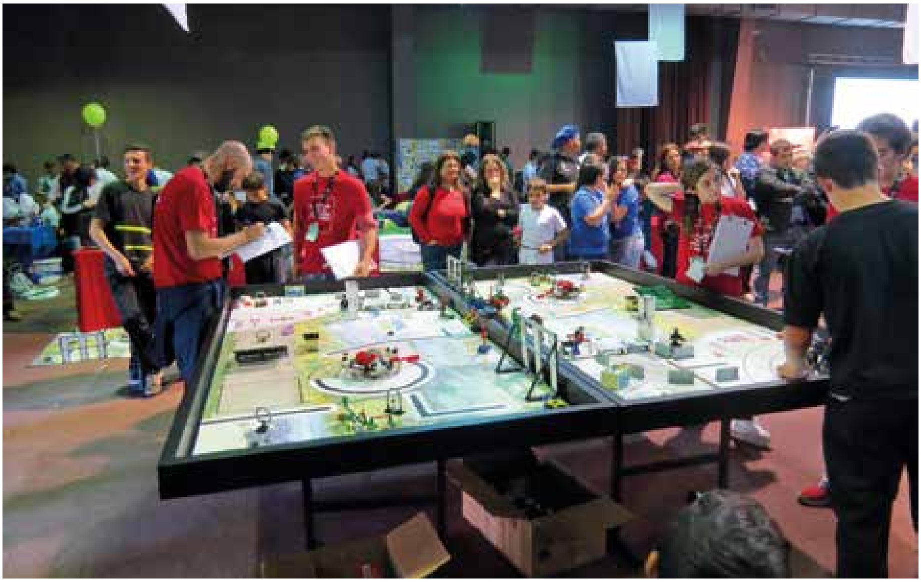


Figure 1. Robotics, Plan Ceibal, Uruguay, Nov. 2016.

Perhaps one of the most remarkable recent variations on robotic turtles is the "tortuga Butiá" made in 2012. It is essentially a "moving laptop", a laptop mounted on wheels, a clever invention of the School of Engineering in Uruguay that is easy to build and uses free software (Turtle Art) and free hardware.[3] Therefore, every device of the laptop, videos, photos, sounds, and a multiplicity of sensors, is already incorporated in the (laptop) robot and may be used freely without extra costs.

In Uruguay every student and teacher in public primary, secondary and technical schools owns his or her own laptop, the famous green XO produced by OLPC, *One laptop per child*, the program launched by Negroponte in 2005. Many children and adolescents are now capable of transforming their own XO into a robot that can compete with other robots and play all sort of games. Such a rapid transformation is bridging the technological gap between diverse socioeconomic populations, in particular in rural and urban deprived environments (Cobo and Mateu, 2016). Some 700,000 students today participate in the Plan Ceibal (www.ceibal.org) and there

[3] https://www.youtube.com/watch?v=6leWvweMEMc; https://www.youtube.com/watch?v=vP6DAdGnmaA; https://www.youtube.com/watch?v=fXRRd5M_Zzs

are over 1200 digital labs in public schools, well equipped with Lego/Logo (some 5000 kits) to construct robots, using Scratch to code. Among the very recent improvements we should mention the 3D printer that children are starting to use in order to produce solid pieces to build robots of the most diverse kinds. In November 2016 a national robotics competition (180 teams) demonstrated the students' creativity with these new tools.[4]

We could say that the Nobel Prize in Chemistry 2016 awarded jointly to Jean-Pierre Sauvage, J. Fraser Stoddart and Bernard L. Feringa "for the design and synthesis of molecular machines" also rewarded a nanoscale kind of robotic turtle to play with, this time at the frontier of molecular science. And when we play we learn, at all ages and in all fields.

In conclusion*, the construction of a robot* is certainly the main path to understand how the machine works, and creates enormous potential for invention, creativity and design starting with the early school years. It is a new cognitive skill that will have profound social, economical and moral consequences. This robotic experience, which today is available to millions of children, opens a new field of research for the neurocognitive and social sciences.

References

Battro, A. M. (1973). *Piaget: Dictionary of Terms* (Preface by Jean Piaget. Translated and Edited by Elizabeth Rutschi-Hermann and Sarah F. Campbell). New York: Pergamon.

Battro, A.M. & Denham, P.J. (1989). *Discomunicaciones: computación y niños sordos.* Fundación Navarro Viola: Buenos Aires, El Ateneo.

Piaget, J. (1938). *La représentation du monde chez l'enfant*. Paris: Presses Universitaires de France.

Piaget, J., Beth, E.W. (1961). *Epistémologie mathématique et psychologie. Essai sur les relations entre la logique formelle et la pensée réelle.* Études d' Epistémologie Génétique. Paris: Presses Universitaires de France. XIV (p. 324).

Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas*. Cambridge, Mass: MIT Press.

Reggini, H.C. (1982). *Alas para la mente. Logo: un lenguaje de computadoras un estilo de pensar.* Buenos Aires: Ediciones Galápago.

Calmet, C., Hitzig, M. & Wilgenbus, D. (2016). *1,2,3…Codez. Enseigner l'informatique à l'école et au collège (cycles 1, 2 et 3).* Paris: Le Pommier.

Cobo, C., & Mateu, M. (2016) A conceptual framework for the analysis and visualization of Uruguayan Internet for education. *Interactions*, 23(6), 70-73. dl.acm.org/citation.cfm?id=2998387&dl=ACM&coll=DL&CFID=860046993&CFTOKEN=15443708

[4] https://www.youtube.com/watch?v=hkERD8Oylzw; https://www.youtube.com/watch?v=ztM7_AwXJWs; https://www.youtube.com/watch?v=Na-lSM90oVA

Putative Prerogatives of the Human Brain: Education, Reasoning, Creativity, Consciousness, Sense of Self, Ethics...Could They be Captured in Machines?

Ghost In the Machine

Olaf Blanke

Neuroscience research has investigated some of the major mechanisms of conscious processing (i.e. Koch, 2004; Dehaene and Changeux, 2011). Influential data regarding the neural correlates of consciousness came from observations in neurological patients (i.e. Pöppel et al., 1973; Weiskrantz et al., 1974; Bisiach et al., 1979), extended by psychophysical research and brain imaging (i.e. Dehaene and Changeux, 2011; Dehaene, this issue). Although these studies have led to a better understanding of perceptual consciousness, they have mostly targeted visual consciousness, whereas conscious and unconscious perception for other senses has been underexplored, despite its importance for consciousness given its multisensory and integrated nature (Faivre et al., 2015, 2017).

Recently, consciousness research has targeted the observer, or subject of conscious experience, that was not accounted for in these models of visual-perceptual consciousness, although the self as the subject of conscious experience is a fundamental property of perceptual consciousness and some have even argued that a subject pervades all consciousness experience. Thus, conscious perception is not only a multisensory experience of external objects, but also includes the experience of a unitary subject. In what follows I summarize what is known about the brain mechanisms that are associated with the ghost in the machine, the feeling that the objects of conscious perception seem to be experienced by somebody, by a self. Many notions of self have been defined and studied in the neurosciences. Although many different classifications have been proposed for the self, I will here highlight only two kinds of self (*cognitive self*, *conscious self*). Both are of relevance for engineering and AI, but only one is fundamentally relevant for consciousness.

The multidimensional *cognitive self* includes self-related cognitions ranging from memory, to language, to imagery, theory of mind, and many other cognitive functions and overlaps with Dan Dennett's narrative self and Ulric Neisser's extended self (i.e. Neisser, 1988; Dennett, 1991; Gallagher, 2000). For example, the capacity to attribute mental states to self and others in order to predict and explain behavior (theory of mind) is often listed as part of the *cognitive self*. Such research has targeted brain mechanisms that may distinguish theory of mind for self and other, for example the different sensory inputs and cognitive systems processing signals relevant for attributing mental states to self and others. Contrary to our intuition that we know our own mind better than those of others there is, however, much evidence that brain mechanisms are actually quite similar and shared when attributing mental states to oneself or to other people (i.e. Gopnik and Meltzhoff, 1994). Another self-relevant cognitive function is memory. Already John Locke proposed that memory processes are a crucial building block for the self. Memory may ascertain continuity of the self across time (and space) and recent studies have defined the self-relevant brain mechanisms of mental time travel (i.e. Arzy et al., 2008; Schacter et al., 2007, 2012) or of autobiographical memory (Levine et al., 1998). Comparable to the systems dedicated to verbal or visuo-spatial memories, the remembered self provides humans with the capacity to store and recall past own life events, to imagine life events from one's past, and to imagine and predict future life events (Arzy et al., 2008; Schacter et al., 2007, 2012). Theory of mind and memory related aspects of the self are highly conscious self-representations and can be mentally accessed as any other cognitive operation. Future machines possessing capacities related to the *cognitive* self may thus be considered more likely as conscious as compared to machines not having such cognitive functions implemented. Advancing neuroscientific understanding of the *cognitive self* and implementing it in machines will probably make these machines more powerful. However, they are not likely to be conscious machines, because despite the importance of these systems for cognition and despite the mental access they may provide to such cognitive self-representations, they are distinct from a fundamental central processing system mediating the phenomenal self: the *conscious self*.

What kind of system should be implemented in a machine so that it is more likely to be phenomenological consciousness? Recent evidence, ranging from clinical to experimental data, suggests that the processing of specific bodily signals is what is needed to have the unitary experience of being the subject of conscious experience. This *conscious self* is based on the processing of multisensory bodily (and motor) signals. The *conscious* self is fundamentally based on the processing of trunk-centered multisensory signals, representing the person's body as a global and unitary entity (Blanke, 2012) and characterized by congruent self-identification, self-location, and first-person perspective. Experimental studies in healthy subjects used different visuo-tactile and visuo-vestibular stimulations for the induction of global changes in the *conscious self*, such as 'full-body', 'out-of-body', or 'body-swap' illusions (Ehrsson, 2007) (Lenggenhager et al., 2007) (Petkova and Ehrsson, 2008). Typically in these paradigms, tactile stimulations are repeatedly applied for several minutes to the back or chest of a participant who is being filmed and so simultaneously views (on a virtual reality headset) the stroking of a human body or avatar in real-time. When exposed to such stimulations, changes in the *conscious self* occur and participants self-identify with the seen virtual body and have changes in self-location towards the position of the virtual body (and thus not or less with their physical body). Additional visuo-vestibular conflicts may also lead to changes in the experienced direction of the subjective first-person perspective (i.e. Ionta et al., 2011; Pfeiffer et al., 2016). Similar effects on the *conscious self* have also been observed when integration of interoceptive bodily signals is tested (Aspell et al., 2013), linking the present concept of the *conscious self* to interoception-based self models (i.e. Craig, 2002; Park and Tallon-Baudry, 2014).

Several variants of such multisensory bodily illusions exist and were conceived to mimic alterations of the *conscious self* that have been reported by neurological patients. Two such clinical conditions that are both based on abnormal multisensory integration of trunk-centered signals are most relevant and are out-of-body experiences and heautoscopy. Out-of-body experiences are characterized by a first-person perspective that is not non body-centered (i.e. the conscious self is experienced as being outside one's bodily borders at an elevated position; Blanke et al., 2004; De Ridder et al., 2007). Heautoscopy is characterized by conscious bilocation and reduplication of the *conscious self* (i.e. the experience of two simultaneously conscious selves that are experienced at two distinct spatial locations; Heydrich and Blanke, 2014). Clearly, visual illusions (i.e. Ponzo or Ebbinghaus-Titchener illusion) are important tools to refine models of visual perception and consciousness (Eagleman, 2001). Likely, the highlighted multisensory bodily illusions will advance models of the *conscious self*. Computational implementations of the basic laws or constraints of the *conscious self* (i.e. proprioception, body-related visual information, peripersonal space, and embodiment; Blanke et al., 2015) may thus be systems with access to body-centered multisensory self-representations and may hence enable forms of phenomenal self-consciousness: a ghost in the machine with a tendency towards mind-body dualism.

The neuroscientific notions of *cognitive self* and *conscious self* relate differently to an old dichotomy, pursued since the dawn of philosophy, between the easy and the hard problem of consciousness (Chalmers, 1996) or between access and phenomenal consciousness (Block, 1995). It will be a fascinating neuroscience question to pursue whether the *conscious self* may allow us to make advances on phenomenal consciousness and the hard problem. Does the *conscious self* pervade all conscious experience (visual, auditory, cognitive, emotional) and underlies the integrated and unitary conscious experience of being a conscious self or subject as well as conscious experience of being a subject with a certain qualitative conscious experience (say the conscious experience of the blue sky). In other words, do the abovementioned brain mechanisms of the *conscious self* that are based on trunk-centered global bodily signals also play a role in visual and auditory consciousness? In my opinion too much speculation rather than experimentation and modeling has prevailed in the past on this topic. With many of the mechanisms of perceptual consciousness as well as those of the *conscious self* well established such studies seem possible. Some data suggest that changes in visual consciousness are tightly coupled with changes in bodily self-relevant signals (Park et al., 2014; Salomon et al., 2016; Faivre et al., 2016). Whether such phenomenal aspects of visual or auditory consciousness are mediated via a higher-order *cognitive self* representation or the first-order *conscious self* representation should open ground for fascinating research, including computational approaches. Such work in humans will require further experimental improvements as well as advances in virtual reality, augmented reality, and robotics/haptics. Such technology needs to be tailored to the needs of cognitive neuroscience and brain imaging, advancing towards a more systematic and fine-grained control of bodily states in humans (i.e. Rognini and Blanke, 2016). Whether machines one day will report out-of-body experiences and heautoscopy or be disposed to psychosis (hallucinations and delusions that have been linked to altered processing of the conscious self) remains to be seen, but may be likely.

References

Arzy S., Molnar-Szakacs I., Blanke O. (2008) Self in time: imagined self-location influences neural activity related to mental time travel. *J Neurosci.* 28(25): 6502-7.

Aspell, J.E., Heydrich, L., Herbelin, B., & Blanke, O. (2013). Turning body and self inside out. Cardio-visual illumination modulates bodily self consciousness and tactile perception. *Psychol Sci. 24, 2445-2453.*

Blanke, O., Landis, T., Spinelli, L., & Seeck, M. (2004). Out-of-body experience and autoscopy of neurological origin. *Brain, 127*, 243-258.

Blanke, O. (2012) Multisensory brain mechanisms of bodily self-consciousness. *Nat Rev Neurosci*, 13, 556-71.

Blanke O., Slater M., Serino A. (2015) Behavioral, Neural, and Computational

Principles of Bodily Self-Consciousness. *Neuron*. 88(1): 145-66.

Block, N. (1995) On a confusion about a function of consciousness. *Beh Brain Sci* 18: 227-287.

Bisiach, E., Luzziatti, C. & Perani, D. 1979. Unilateral neglect, representational schema and consciousness. *Brain: a journal of neurology,* 102, 609-618.

Chalmers, D. (1996). *The conscious mind*. MIT Press.

Craig, A.D. 2002. How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, 3, 655-666.

Dehaene, S., & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*, 200-227.

Dennett, D.C. (1991). *Consciousness explained*. New York: Penguin.

De Ridder D., Van Laere K., Dupont P., Menovsky T., Van de Heyning P. (2007) Visualizing out-of-body experience in the brain. *N Engl J Med.* 357(18): 1829-33.

Ehrsson, H.H. (2007). The experimental induction of out-of-body experiences. *Science, 317*, 1048.

Faivre, N., Salomon, R., & Blanke, O. (2015). Visual Consciousness and Bodily-Self Consciousness. *Current Opinion in Neurology*, 28(1), 23-28.

Faivre N., Arzi, A., Lunghi, S., & Salomon, S. (2017) Consciousness is more than meets the eye: a call for a multisensory study of subjective experience. *Neuroscience of Consciousness*. 3(1): nix003.

Faivre N., Dönz J., Scandola M., Dhanis H., Bello Ruiz J., Bernasconi F., Salomon R., Blanke O. (2017) Self-Grounded Vision: Hand Ownership Modulates Visual Location through Cortical β and γ Oscillations. *J Neurosci.* 2017 Jan 4;37(1):11-22.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*.

Heydrich, L., & Blanke, O. (2013). Distinct illusory own-body perceptions caused by damage to posterior insula and extrastriate cortex. *Brain,* 136, 790-803.

Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D., … Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70, 363-374.

Koch, C. (2004). *The quest of consciousness. A neurobiological approach*. Englewood, CO: Roberts.

Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science, 317*, 1096-1099.

Levine B., Black S.E., Cabeza R., Sinden M., Mcintosh A.R., Toth J.P., Tulving E., Stuss D.T. (1998) Episodic memory and the self in a case of isolated retrograde amnesia. *Brain*. 121:1951-73.

Park H.D., Correia S., Ducorps A., Tallon-Baudry C. (2014) Spontaneous fluctuations in neural responses to heartbeats predict visual detection. *Nat Neurosci*. 17(4):612-8.

Park HD, Tallon-Baudry C. (2014) The neural subjective frame: from bodily signals to perceptual consciousness. *Philos Trans R Soc Lond B Biol Sci*. 369 (1641): 20130208.

Petkova, V.I., & Ehrsson, H.H. (2008). If I were you: Perceptual illusion of body swapping. *PLoS ONE, 3*, e3832.

Pfeiffer, Christian; Grivaz, Petr; Herbelin, Bruno; Serino, Andrea; Blanke, Olaf (2016) Visual gravity contributes to subjective first-person perspective. Neuroscience of Consciousness. (1): niw006. doi: 10.1093/nc/niw006

Pöppel E., Held R., Frost D. (1973) Residual visual function after brain wounds involving the central visual pathways in man. *Nature*. 243(5405): 295-6.

Rognini G, Blanke O. (2016) Cognetics: Robotic Interfaces for the Conscious Mind. *Trends Cogn Sci*. 20(3): 162-4.

Salomon R., Ronchi R., Dönz J., Bello-Ruiz J., Herbelin B., Martet R., Faivre N., Schaller K., Blanke O. (2016) The Insula Mediates Access to Awareness of Visual Stimuli Presented Synchronously to the Heartbeat. *J Neurosci*. 2016 May 4;36(18):5115-27.

Schacter D.L., Addis D.R., Buckner R.L. (2007) Remembering the past to imagine the future: the prospective brain. *Nat Rev Neurosci*. 2007 Sep; 8(9):657-61.

Weiskrantz L., Warrington E.K., Sanders M.D., Marshall J. (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*. 97(4): 709-28.

What Is Consciousness, and Could Machines Have It?

Stanislas Dehaene

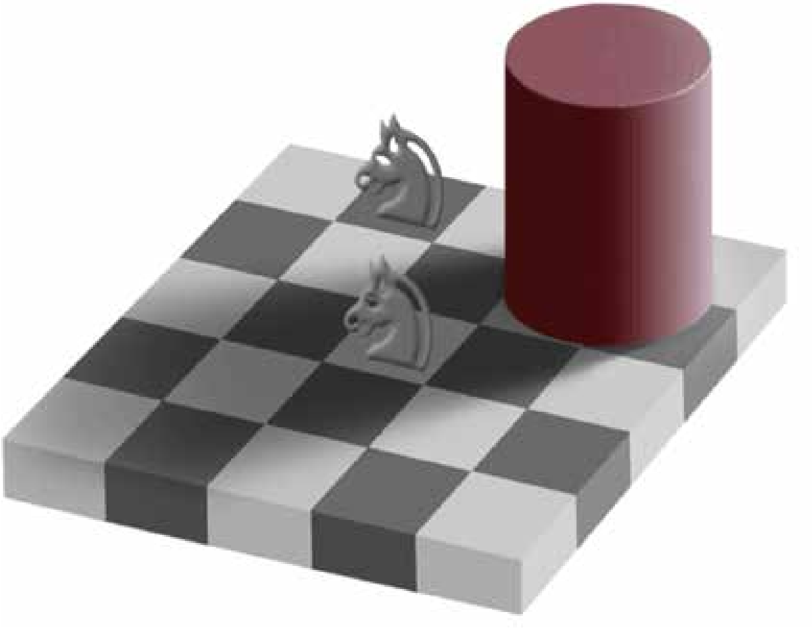
Although consciousness is frequently considered as the pinnacle of the brain, and something that is impossible to confer to machines, I would like to argue otherwise. In this report, I argue that, in fact, much is already known about how brains generate consciousness, and how those findings could be used in artificial intelligence. In the past twenty years, cognitive neuroscience has made great advances in understanding the "signatures of consciousness" – brain activity markers that occur only when the subject is aware of a certain content. Those findings support a theory of consciousness as a sharing device, a "global neuronal workspace" that allows us to share our knowledge, internally, between all of the specialized "processors" in our brain, and externally, with other people. On this basis, I make several tentative suggestions as to which functionalities should be added to present-day machines before they might be considered conscious.

The contemporary strategy to study consciousness

In the past 20 years, the problem of consciousness has ceased to appear insurmountable. Neuroscience has started to identify the objective brain mechanisms underlying subjective processes – what I call the "signatures" of conscious processing. Those discoveries have been reviewed in detail elsewhere (Dehaene & Changeux, 2011; Dehaene, 2014). Briefly, the first advance came with the advent of explicit theories of computation (Hilbert, Gödel, Turing, Von Neumann) and information representation (Shannon). Following those breakthroughs, consciousness could then be seen as a computational property associated with a certain level of information processing.

At present, three computational levels may be distinguished. At the lowest level, which we may call level 0, unconscious algorithms process symbols blindly and, obviously, without any awareness. For instance, our visual system blindly and unconsciously processes the following image (due to Adelson).

It lets us see a strictly normal checkerboard, although this is illusory – as you may check by masking the figure, the two knights and their squares seem to be black and white, but they are actually exactly the same shade of



By Mig after Edward H. Adelson.

grey. What is happening? Our visual system detects the presence of a dark zone in the image, which it interprets as a shadow, and it subtracts it from the image to let us see the "true" shade of grey of the pieces, thus making one knight look brighter than the other. Arguably, any machine whose aim would be to extract the genuine appearance of objects while getting rid of shadows and other defects of the image would have to go through the same inference process. In this sense, many of the brain's unconscious computations are rational computations. Paradoxically, any machine that strives towards objectivity would be submitted to similar human-like illusions.

Above the unconscious processing level, two higher levels of information processing may be defined, corresponding to what others have termed primary and secondary consciousness (Edelman, 1989).

- Level 1 is conscious access. At any given moment, although our brain is bombarded with stimuli and has a vast repertoire of possible sensory or memory states, only a single piece of information, selected for its relevance, is consciously accessed, amplified, and becomes the focus of additional processing. This selective attribution of higher-level computing resources is what we experience as "conscious access".

- Level 2 is conscious self-representation. At this level, the cognitive system entertains one or several representations of its own knowledge, for instance it may know what it is currently focusing on, that it made an error, etc. Thus, the system not only commits its resources to a specific piece of information (level 1), but also "knows that it knows" (level 2). The assumption is that this self-knowledge is represented in the same format as the knowledge of other people (also known as a "theory of mind"), thus allowing this information to be shared with others and to be used in social decision making (Bahrami et al., 2010; Frith, 2007).

Baars (1989) was one of the first cognitive scientists to realize that, given those simple definitions, it is quite possible to study consciousness experimentally. The experimental strategy proceeds in several steps (Dehaene, 2014):

1. Identify a minimal experimental paradigm (e.g. a visual illusion) that allows to contrast visible and invisible stimuli. My laboratory has used masking, whereby a flashed image can be made either subliminal or conscious (Kouider & Dehaene, 2007). Others have used binocular rivalry, whereby the competition between two images is used to render one of them conscious while the other is not (Logothetis, Leopold, & Sheinberg, 1996). Many other minimal contrasts are available, for instance sleep versus wakefulness; wakefulness versus anesthesia; vegetative-state versus minimally conscious patients, etc. (Baars, 1989).

2. Carefully quantify the subject's introspection, i.e. what he or she "knows that it knows". Introspection defines the very phenomenon that we want to study (conscious subjective perception) and must therefore be recorded alongside other objective measures of behavior and brain activity. The ideal situation consists in presenting a fixed stimulus closed to the threshold for awareness, and to sort the trials according to subjective reports, such that the very same stimulus is sometimes perceived consciously and sometimes remains unconscious.

3. As a consequence, focus on a particular and restricted sense of consciousness: the capacity to report a piece of information, to oneself or to others. Scientists have learned that this sense of consciousness, called *reportability*, is well-defined and differs from other concepts such as attention, vigilance, or self-consciousness.

4. Apply the panoply of modern neuro-imaging and neuroscience tools to compare the behaviors and brain activity patterns evoked by reportable and unreportable stimuli, thus uncovering the signatures of consciousness.

Current signatures of consciousness in the human brain

Experiments that have implemented this strategy have discovered that, although subliminal stimuli can induce considerable activity in many if not all circuits of the human brain, conscious perception is associated with a set of specific signatures:

- *Amplification and access to prefrontal cortex.* Compared to a subliminal image, a conscious image is amplified and gains access to higher levels of representation, particularly in prefrontal and parietal cortices.
- *Late global ignition and meta-stability.* Tracking the propagation of conscious and unconscious images shows that unconscious activity can be strong in early visual cortex, yet die out in a few hundreds of milliseconds within higher cortical areas. A conscious image, on the contrary, is amplified in a non-linear manner, an event called "global ignition". By about 300 milliseconds, brain activity becomes more stable when the stimulus is conscious than when it is not (Schurger, Sarigiannidis, Naccache, Sitt, & Dehaene, 2015).
- *Brain-scale diffusion of information.* Conscious ignition is accompanied by increased in bidirectional exchanges of information in the human brain. During a conscious episode, the cortex "talks to itself" at greater distances, and this is manifested by correlations of brain signals, particularly in the beta band (13-30 Hz) and theta band (3-8 Hz).
- *Global spontaneous activity.* Even in the absence of stimuli, the brain spontaneously generates its own patterns of distributed activity, which are constantly changing (Barttfeld et al., 2015). This resting state activity can partially predict the content of consciousness, for instance whether the subject currently experiences mental images or "mind wandering".
- *Late all-or-none firing of "concept cells".* Single-cell correlates of conscious ignition have been identified in human and non-human primates. Neurons in prefrontal and anterior temporal cortex fire to a specific concept (e.g. the Empire State building) and do so *only* when the corresponding word or image is presented consciously. Their late activity acts as a signature of conscious perception (Quiroga, Mukamel, Isham, Malach, & Fried, 2008).

Those findings are compatible with the Global Neuronal Workspace (GNW) hypothesis, a simple theory of consciousness (Baars, 1989; Dehaene & Changeux, 2011; Dehaene, 2014). Briefly, the hypothesis is that, while specialized subsystems of the brain ("modules") process information unconsciously, what we subjectively experience as consciousness is the global availability of information, which is made possible by a non-modular "global workspace". Consciousness is a computational device that evolved to break the modular organization of the brain. During conscious perception, a distributed parieto-frontal circuit, forming the global neuronal workspace, ignites to selectively amplify a relevant piece of information. Thanks to its long-distance connectivity, supported by giant neurons with long axons in layers 2-3, it stabilizes a selected piece of information and broadcasts it in a brain-wide manner to all other modules. The global workspace thus maintains information in an active state for as long as it is needed (meta-stability).

The GNW hypothesis addresses the classical question of the function of consciousness. Is consciousness a mere epiphenomenon, i.e. a useless side-effect of brain activity, similar to the whistle of the train? Theory and experiments suggest otherwise: consciousness appears to be required for specific operations. Thanks to the global workspace, we can reflect upon the information: subliminal information is evanescent, but conscious information is stabilized and available for long-term thinking. Consciousness is also helpful in order to discretize the incoming flux of information and reduce it to a few samples that can be reported or stored: while unconscious processes compute with an entire probability distribution, consciousness samples from it. Consciousness is also involved in routing information to other processing stages, thus allowing us to perform arbitrary chains of operations (for instance, computing 23x47; Sackur & Dehaene, 2009). Finally, consciousness plays a key role in monitoring our behavior and diagnosing our errors. We have found that a key component of the brain's error monitoring system, the "error negativity" that arises whenever we press the wrong button in a simple response-time task, only occurs on trials where subjects report seeing the stimulus (Charles, Van Opstal, Marti, & Dehaene, 2013). Only visible stimuli allow us to detect the occasional discrepancy between what we intended and what we did.

What machines are missing

In summary, cognitive neuroscientists are beginning to understand that the computations that we experience as "conscious processing" are useful aspects of brain function that are therefore likely to be equally useful to artificial-intelligence devices. Here, I list four computational features that conscious brains possess and that machines currently miss. My suggestion is that if those functions were implemented, the resulting machine would be likely to be considered conscious, or at least much closer to conscious than most machines currently are.

1. *A workspace for global information sharing.* In current computers and cellphones, computations are performed by special-purpose programs known as "apps". Each app possesses its own memory space and its specific knowledge base, carefully protected from others. Apps do not share their knowledge: it is frequent for one app to "know" a piece of information while others ignore it. In the human brain, this is characteristic of unconscious processing. According to the GNW hypothesis, consciousness evolved to break this modularity. The GNW can extract relevant information from virtually any brain module, and make it available to the entire organism. Machines may benefit from a similar architecture for flexible information sharing, capable of broadcasting to the entire system a potentially relevant piece of information. "Blackboard" architectures of this type were proposed in the 1970s. It would be interesting to pursue this idea in the context of present-day machine-learning algorithms, which are able to make the best use of the broadcasted information.
2. *A repertoire of self-knowledge.* To determine where the important information lies and where to route it, I believe that brains and machines alike must be endowed with a repertoire of self-knowledge. By this, I do not mean a bodily sense of self, as might be available for instance to a robot that would know the location of its limbs (in the human brain, the construction of this body map is, in fact, unconscious). What I have in mind is an internal representation of the machine's own abilities: a database that contains a list of its apps, the kind of knowledge they possess, what goals they can fulfill, how fast they can operate, how likely they are to be correct, etc. Even young children, when learning arithmetic, compile such a repertoire of the different strategies at their disposal (Siegler & Jenkins, 1989). In a machine endowed with learning algorithms, self-knowledge should be constantly updated, leading to a concrete implementation of the Socratic "know thyself".
3. *Confidence and "knowing that you don't know".*A conscious machine should know when it is wrong or when it is uncertain about something. In the human brain, this corresponds to meta-cognitive knowledge (cognition about cognition) which has been linked to prefrontal cortex. Even preverbal infants know that they don't know, as revealed by the fact that they turn to their mother for help whenever appropriate {Kouider}. There are several ways in which a computer could be equipped with a similar functionality. First, it could be endowed with statistical programs that do not just give an answer, but also compute the probability that

this answer is correct (according to Bayes' law or some approximation of it). Second, a computer could be endowed with an error-detection system, similar to the brain's error-negativity, which constantly compares ongoing activity with prior expectations and spontaneously reacts if the current behavior is likely to be wrong. Third, this error-detection device could be coupled to a corrective device, such that the system constantly looks for alternative ways to get the correct answer.

4. *Theory of mind and relevance*. One aspect of consciousness, which may be unique to humans, is the ability to represent self-knowledge in the same format as knowledge of others. Every human being holds distinct representations of what he knows; what others know; what he knows that others know; what he knows that others don't know; what others know that he doesn't know; etc. This faculty, called theory of mind, is what allows us to model other minds and to use this knowledge in order to maximize the usefulness of information that we can provide them (relevance, as defined by Sperber & Wilson, 1988). Current machines often lack such relevance. A machine that could simulate its user's mind would undoubtedly provide more relevant information. It would remember what it previously said, infer what its user knows, and avoid presenting trivial, useless, or otherwise contradictory information. Algorithms that handle such recursive representations of other minds are currently being developed (Baker, Saxe, & Tenenbaum, 2009; Daunizeau et al., 2010).

The above list is probably not exhaustive. However, I contend that conferring it to computers would arguably go a long way towards closing the consciousness gap. According to the present stance, consciousness is not an essence, but solely a functional property that can be progressively approximated. Humans seem to be quite generous in attributing consciousness to others, including animals, plants, and even inanimate objects such as clouds or storms. The steps that I outlined here should bring us closer to attributing consciousness to machines.

References

Baars, B. (1989). *A cognitive theory of consciousness*. Cambridge, Mass.: Cambridge University Press.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., & Frith, C.D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-5. https://doi.org/10.1126/science.1185718

Baker, C.L., Saxe, R., & Tenenbaum, J.B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-49. https://doi.org/10.1016/j.cognition.2009.07.005

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., & Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 887-892. https://doi.org/10.1073/pnas.1418031112

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80-94. https://doi.org/10.1016/j.neuroimage.2013.01.054

Daunizeau, J., den Ouden, H.E., Pessiglione, M., Kiebel, S.J., Stephan, K.E., & Friston, K.J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS One*, *5*(12), e15554. https://doi.org/10.1371/journal.pone.0015554

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Reprint edition). Penguin Books.

Dehaene, S., & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–27. https://doi.org/10.1016/j.neuron.2011.03.018

Edelman, G. (1989). *The remembered present.* Basic Books: New York.

Frith, C. (2007). *Making up the mind. How the brain creates our mental world.* London: Blackwell.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philos Trans R Soc Lond B Biol Sci*, *362*(1481), 857-75.

Logothetis, N.K., Leopold, D.A., & Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? *Nature*, *380*(6575), 621-4.

Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R., & Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proc Natl Acad Sci U S A*, *105*(9), 3599-604.

Sackur, J., & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. *Cognition*, *111*(2), 187-211.

Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J.D., & Dehaene, S. (2015). Cortical activity is more stable when sensory stimuli are consciously perceived. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(16), E2083-2092. https://doi.org/10.1073/pnas.1418730112

Siegler, R.S., & Jenkins, E.A. (1989). *How children discover new strategies.* Hillsdale N.J.: Lawrence Erlbaum Associates.

Sperber, D., & Wilson, D. (1988). Précis of relevance : Communication and cognition. *Behavioral and Brain Sciences*, *10*, 697-789.

Artificial Intelligence and Human Minds: Perspectives from Young Children

Elizabeth Spelke

Our species has a talent for technology: for imagining, crafting and using objects that extend our abilities and improve our lives. Although there is now much discussion, and some disquiet, over the prospect that future generations will live with autonomous machines that are more capable than we are, our talent for building such tools is not new. Even the earliest tools from human prehistory outperform us at their dedicated functions: arrowheads pierce animal skins better than fingernails, and bowls hold water better than cupped hands. From the beginning, moreover, our tools have functioned not only with us, like the stone flakes that early humans used to cut food, but autonomously and for our benefit, like the roofs that shelter us. Our ability to develop such objects testifies to our singular ability, as adults, to foresee how currently nonexistent objects, functions, and activities can transform our lives and experiences. It also testifies to the power of children to develop adaptively within the highly variable environments that human ingenuity creates, learning culture-specific skills that have come to include agriculture, reading, mathematics, and modern engineering (Dehaene, 2009, 2011).

Despite the ancient origins of our talent for technology, the emergence of machines that reason and learn prompts many questions, two of which pertain directly to the focus of my research, on the cognitive capacities of human infants and children. First, can the development of such machines shed light on the workings of young human minds and on the sources of our species' cognitive talents: insights that could deepen our understanding of human nature and improve children's education and welfare (Battro et al., 2011)? Second, will the presence of intelligent machines that interact with humans alter the ways in which children think and learn? If so, how can those machines best be structured to enhance children's development? To approach these questions, I begin by reviewing some pertinent findings from research on early human cognitive development.

Cognition in infancy

From birth, human infants perceive, act on, and make sense of their surroundings, anticipating its future states. Research provides evidence that infants both perceive inanimate objects when they are visible and track such objects when they are hidden, extrapolating object motions and mechanical interactions (Baillargeon, 1998; Stahl & Feigenson, 2015). Infants also perceive and reason about people and animals, predicting their future actions from their past behavior together with their powers to perceive accessible aspects of the environment (Gergely & Csibra, 2003; Luo & Johnson, 2009). And from the beginning, infants focus on people's social communications, using their speech, gaze, and coordinated actions to infer their engagement with the infant (Meltzoff & Moore, 1977; Kinzler, et al., 2007) and with one another (Hamlin et al., 2007; Powell & Spelke, 2013).

Inanimate objects, agents, and social beings behave in fundamentally different ways: objects are governed only by the laws of physics, whereas agents plan their actions to achieve valued goal states while minimizing costs, and social beings engage with one another so as to share information and experiences. Research provides evidence that infants are sensitive to these differences (Spelke & Kinzler, 2007). They perceive and interpret the behavior of inanimate objects primarily by analyzing objects' positions and motions, in accord with basic constraints that objects move as connected wholes on continuous paths and interact with one another on contact (Spelke, et al., 1995). Infants perceive and reason about the object-directed actions of people and animals by analyzing aspects of their shapes and motions (Bertenthal & Pinto, 1994), in accord with assumptions that agents perceive the world at a distance and act efficiently to transform it, in accord with their goals (Gergely et al, 1995; Woodward, 1998; Liu & Spelke, 2017). Finally, infants perceive and interpret people's social motives and relationships by analyzing their interactions with the infant and with one another. Recent research suggests that infants are especially sensitive to the asymmetrical relations that connect caregiving adults to their children (Johnson et al., 2007; Spokes et al., 2017), dominant individuals to their subordinates (Thomsen et al., 2011), and socially responsive imitators to the targets of their imitation (Powell & Spelke, 2013, in review).

These findings and others suggest that infants are endowed with core cognitive systems that form the foundation for the development of our common sense reasoning about the physical, living, and social worlds. These systems likely are connected, because agents' actions are constrained by physics and people's social bonds are conveyed by their actions. Nevertheless, each core cognitive system functions in accord with a distinct set of principles and operates with a high degree of independence from the other systems, especially in infancy. For example, infants likely can view their pet cat as a social being (a member of their family, with distinctive relations to other family members), an agent (that chases after butterflies and chews on house plants), and an object (that is heavy to lift), but they do not readily construe the cat in these three different ways at once. Young infants also do not appear to recognize a central property of tools and other artifacts: that they are *objects*, designed to foster the instrumental goals of *agents*, for use within a community of *social beings.*

Toward the end of the first year, infants' understanding of objects, agents, and social beings comes together: infants begin to conceive of objects as members of one or another *kind* – a body whose form affords dedicated functions for itself (if it is a person or animal) or for members of the infant's social world (if it is inanimate: Xu & Carey, 1996). This conception emerges as infants engage with others and thereby learn one of the earliest emerging and universal features of human language: noun phrases whose head nouns refer to kinds of animals ("dog"), natural objects ("stone", "tree"), or artifacts ("cup"). By nine months of age, infants expect each distinct noun to refer to a distinct kind of object with a characteristic form and function (Xu, 2007). Soon thereafter, infants begin to seek information about object kinds, asking of each thing that they encounter, "what is this?" and (if it is an artifact) "what is it *for*?" (Keil, 1989). There is wide consensus among psychologists that the capacity to view novel objects as individual members of novel artifact kinds is central to the child's developing mastery of culture in general and technology in particular. Moreover, this capacity is widely thought to depend on infants' predisposition to attend to their social partners, learning from their speech and actions (Tomasello, 2008; Csibra & Gergely, 2009). Because adults are apt to talk about things that matter to them, their language directs children to concepts that are socially useful. Because adults' actions on objects, such as drinking from a cup or turning the pages of a book, both exhibit the objects' functions and reveal aspects of their structure, those actions inform infants about the key properties of the things used in their culture. The artifact concepts that infants master at the end of the first year therefore serve as a basis for the prodigious cultural learning that distinguishes our species from others, and that sets humans on a path that leads toward the world we now are considering, in which humans interact with autonomous machines whose intelligence and action capacities, in some domains, equal or exceed our own.

Reverse engineering infant minds

Although research on human cognitive development has shed light both on what young infants know and on the fundamental changes in their knowledge that occur when one-year-old children begin to master artifacts, the psychological and brain sciences have not yet achieved a deep understanding of the mechanisms and processes that give rise to this knowledge. The content of infants' knowledge can be revealed by simple behavioral experiments, yet the most advanced investigations in experimental psychology and neuroscience have not yet revealed the basic computations of the human mind.

With the emergence of machine learning and artificial intelligence comes the promise of this deeper understanding. From its beginnings, computer scientists have aimed to build machines that learn as children do, the most capable learners on earth (Turing, 1950). Moreover, the most conspicuous recent successes in the field of artificial intelligence have centered on machines that are structured similarly to the brain's perceptual systems and that are built to learn (LeCun et al., 2015). Symmetrically, cognitive and developmental psychologists have looked to research in computer science and mathematics for guidance in studying the basic computations of mature and developing human minds (Tenenbaum, et al., 2011). Coordinated research across these fields, developing and testing computational models of human cognition and learning, could deepen understanding of human minds in general, and the minds of infants and young children in particular, while guiding the development of ever more intelligent machines.

For example, recent thinking about infants' "intuitive physics" – their grasp of the mechanical principles governing object motions and interactions – has benefited from the development, in computer science, of physics engines that simulate these motions and interactions (Battaglia et al., 2013). Physics engines are used in animated films and interactive video games to depict events in which objects collide, topple, or collapse on contact with other objects, surfaces, substances, and agents. The computational challenges solved by the designers of physics engines suggest insights into both the capacities of young infants and key limits to those capacities (Ullman et al., in review). For example, infants track moving objects over occlusion by taking account of their positions, motions, and approximate sizes but not their detailed shapes or surface texture (Baillargeon, 1998; Spelke, et al., 1995). For example, when young infants see a cup appear alternately on the opposite sides of one screen, they represent one persisting object in motion, but when they see a cup and a shoe appear in alternation on the screen's two sides, they fail to represent two distinct objects (Xu & Carey, 1996). Physics engines might behave similarly, for they use coarse representations of an object's position, mass, and motion in order to extrapolate its motion forward, and then call on stored, detailed representations of the object's appearance so that it can be rendered, by graphics programs, at places where it is visible. The use of a coarse representation in the computation of the object's changing position and motion is accurate enough to appear natural to adults, while sparing the computations that would be required if every detailed feature of the object were extrapolated forward. Infants' failure to track the detailed shapes of occluded objects may reflect a similarly efficient process for representing hidden object motion, and a division of labor between basic processes for representing objects' dynamic properties and their visual appearance.

Recent thinking about young children's psychological and social reasoning has benefited in similar ways from computational models of action understanding (e.g., Baker, et al., 2009, 2017) based on the assumption that agents plan actions that maximize their rewards while minimizing their costs (Gergely et al., 1995), and that social beings act as well to maximize the rewards of their valued social partners (Jara-Ettinger et al., 2016). Recent experiments provide evidence that representations of action plans guide young children's interpretation and evaluation of other agents' actions, motives, and mental states. Three-year-old children who see a social character refuse to help another character judge the first character more harshly if the requested helping action was easy to perform (Jara-Ettinger et al., 2015), and 10-month-old infants infer that an agent values one goal object more than another if he is willing to take a higher-cost action to obtain one of the objects, even if his behavior toward the two objects is otherwise the same (Liu et al., 2017).

Computational modeling of early cognitive development is still in its infancy, but these and other studies suggest that a deeper understanding of young human minds, and of our species' prodigious learning capacities, can emerge from coordinated research in machine learning, artificial intelligence, and human cognitive and brain development. Such an understanding may be critical to addressing key challenges posed by our rapidly changing technological landscape.

Protecting and enhancing children's development

As research on the nature of intelligence progresses, how will the development of increasingly intelligent machines affect the minds of those who use them, especially the children who learn with and from them? If artificial intelligence is to bring us new technologies that enhance our reasoning and benefit our lives, then this question looms large. Intelligent systems might extend our capacities by making useful information more accessible: for example, GPS-based navigation systems that display our current position in relation to our surroundings at multiple scales, and that bring us information about otherwise inaccessible events such as traffic accidents or roadblocks, have the potential to extend and enrich our representations of the environment. These same systems, however, could diminish our spatial cognitive capacities, if we use them to navigate for us, rather than to enrich and strengthen our spatial knowledge. Research in cognitive neuroscience reveals that the basic cognitive systems by which humans navigate are fundamental to human spatial reasoning and memory, and they are strengthened by exercise (Burgess, et al., 2002). Like visual and motor systems, these systems likely are weakened by disuse: thus, a person who moves solely at the direction of a GPS navigator may both fail to develop a spatial representation of her surroundings, and diminish her memory capacities more generally. These two contrasting uses of contemporary technology suggest a question and a challenge: How should navigation aids be crafted so as to enrich, rather than diminish, our ancient, autonomous capacities for spatial reasoning? Similar questions, calling for research, are raised by intelligent systems that help us plan our days, remember friends' birthdays, or select our music.

The advent of intelligent machines raises especially pointed questions concerning children's learning, including the learning that propels our talent for developing novel technology. Throughout history, children have learned both by acting and by observing the actions of their elders, who manipulated artifacts with perceptible structures and functions. In the second year, children become highly attentive to the manner in which adults act on objects, and highly predisposed to reproduce those actions exhibit (Tomasello, et al., 2005; Lyons et al., 2007). Young children also begin to attend to adults who copy their own detailed actions on objects (Agnetta & Rochat, 2004), and to the structural properties of the objects that adults manipulate (Booth & Waxman, 2002). These developments recruit infants' earlier developing sensitivity to object shapes and motions, to agents' detailed, multi-step actions, and to social beings' shared experience to propel a key feature of human cognition: the rapid development, in childhood, of encyclopedic knowledge of object kinds.

How will this development proceed for the current generation of infants, born into families using the tablets and smart phones that are now ubiquitous in many societies and constant companions to many parents? In contrast to the artifacts that smart phones replace, such as telephones, cameras, and books, smart phones have multiple functions. Neither the structures that permit their functions, nor the actions of their users, are perceptually accessible to the child (or, in most cases, to other adults): when a parent looks at and taps on a cell phone, he could be engaged in any of a multitude of diverse actions, undertaken to realize an even larger potential set of goals. His observable behavior does not reveal his action plans.

If multipurpose machines take on more and more of the functions that previously were performed by perceptually distinct objects, whose structure afforded specific actions that were diagnostic of their function, how will children develop the encyclopedic knowledge of object kinds that has long served as a foundation for cognitive development? Will future generations of children learn directly from smart machines, whose functioning has made the actions of other people less informative? Because the structures that support the behavior of these machines cannot be seen, and the behavior of adults who use them is only minimally informative about their goals, plans, and social relations, will children be less inclined to explore objects, or to use the object-directed actions of other people, so as to learn about the structure and functioning of the physical, living, and social world? If so, what will children learn in a world of smart, interactive machines, and how will their learning impact their social and cognitive growth? Because humans invent technologies for human benefit, we can combine and invert these questions: What kinds of intelligent machines should computer scientists aim to create, in order promote and support young children's cognitive development and well being?

Past research on cognitive development in infancy and early childhood does not answer this question. Although that research has taught us a great deal about *what* infants and young children know at different ages, it does not support strong predictions concerning children's learning in radically new or hypothetical environments. To make such predictions, the brain and cognitive sciences must achieve a deeper understanding of *how* infants and children reason and learn.

Fortunately, collaborative research in cognitive science, neuroscience and computer science promises to deepen our understanding, providing insights that can inform the development of new technologies to enhance children's lives. Side by side with our talents and propensities for transforming the world in ways that create both new opportunities and new problems, our species has a striking capacity for foreseeing the potential problems and addressing them. Thus, the development of physics and the atmospheric sciences has allowed its practitioners to anticipate, and devise ways to counter, the catastrophic consequences of massive climate change or global nuclear warfare – two challenges posed by human technological progress that now can be foreseen and countered, even though nothing in our history provides a precedent for them. Similarly, the development of computational cognitive science promises to bring knowledge that can support the design of thinking machines that act for the benefit of all people, and perhaps especially for the benefit of children, the most vulnerable and gifted human learners. I believe it will best do so if computer scientists and cognitive psychologists work together to achieve a better understanding of developing human minds.

References

Agnetta, B., & Rochat, P. (2004). Imitative games by 9-, 14-, and 18-month-old infants. *Infancy, 6*(1), 1-36.

Baillargeon, R. (1998). Infants' understanding of the physical world. In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in psychological science*, Vol. 2 (pp. 503- 529). London: Psychology Press.

Baker, C.L., Saxe, R., & Tenenbaum, J.B. (2009) Action understanding as inverse planning. *Cognition, 113*(3), 329-349.

Baker, C.L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behavior, 1 (0064).*

Battaglia, P.W., Hamrick, J.B., & Tenenbaum, J.B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 18327-32.

Battro, A., Dehaene, S. & Singer, W. (2011). *Human neuroplasticity and education: Scripta Varia, 117.* Pontifical Academy of Sciences.

Bertenthal, B.I., & Pinto, J. (1994). Global processing of biological motions. *Psychological Science, 5*(4), 221-225.

Booth, A.E. & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology, 38*(6), 948-957.

Burgess, N., Maguire, E.A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron, 35(4),* 625-641.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148-153.

Dehaene, S. (2009). *Reading in the brain.* Penguin.

Dehaene, S. (2011). *The number sense (second ed.).* Oxford.

Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7(7),* 287-292.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165-193.

Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Hamlin, J.K., Wynn, K., & Bloom, P.

(2007). Social evaluation in preverbal infants. *Nature, 450*(7169), 557-559.

Jara-Ettinger, J., Gweon, H., Schulz, L.E., & Tenenbaum, J.B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences.*

Jara-Ettinger, J., Tenenbaum, J.B., & Schulz, L.E. (2015). Not so innocent: Toddlers' reasoning about costs, competence, and culpability. *Psychological Science.*

Johnson, S.C., Dweck, C.S. & Chen, F.S. (2007). Evidence for infants' internal working models of attachment. *Psychological Science, 18(6)*, 501-502.

Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kinzler, K.D., Dupoux, E., & Spelke, E.S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences USA, 104*(30), 12577-12580.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Liu, S. & Spelke, E.S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition, 160,* 35-42.

Liu, S., Ullman, T., Tenenbaum, J., & Spelke, E.S. (2017). Origins of a naïve utility calculus: Infants infer the value of goals from the costs of actions. *Unpublished manuscript, Harvard University.*

Luo, Y., & Johnson, S.C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science, 12*(1), 142-149.

Lyons, D.E., Young, A.G., & Keil, F.C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences USA, 104(50), 19751-19756.*

Meltzoff, A.N., & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*(4312), 75-78.

Powell, L.J. & Spelke, E.S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences, 110,* 3965-3952.

Powell, L.J. & Spelke, E.S. (in review). *Infants' understanding of social imitation: Inferences of affiliation from third-party observation.* Manuscript submitted for publication.

Spokes, A.C. & Spelke, E.S. (2017). The cradle of social knowledge: Infants' reasoning about caregiving and affiliation. *Cognition, 159*, 102-116.

Spelke, E.S., Vishton, P., & von Hofsten, C. (1995). Object perception, object-directed action, and physical knowledge in infancy. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Spelke, E. & Kinzler, K. (2007). Core knowledge. *Developmental Science, 10,* 89-96.

Stahl, A.E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348*(6230), 91-94.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., & Goodman, N.D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

Thomsen, L., Frankenhuis, W., Ingold-Smith, M., & Carey, S. (2011). The big and the mighty: Preverbal infants represent social dominance. *Science, 331(6016)*, 477-480.

Tomasello, M. (2008). *Origins of human communication.* Cambridge, MA: MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28,* 675-735.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 49,* 433-460.

Ullman, T., Spelke, E., Battaglia, P. &

Tenenbaum, J. (in review). Mind games: Game engines as an architecture for intuitive physics. Manuscript submitted for publication.

Woodward, A.L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1-34.

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology, 30*(2), 111-153.

Xu, F. (2007). Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences, 11*(9), 400-406.

Neurotechnology For Human Benefit and the Impact of AI

John P. Donoghue

Neuroprosthetics is an emerging field that is beginning to provide a technological approach to restore lost sensory functions, restore movement for those with paralysis, or repair cognitive deficits produced by disordered brain circuits (Donoghue, 2015). Neural prosthetics are technologies (i.e. a system of devices) that can be placed in or on the body to partially recover lost sight, hearing, or movement, or repair brain circuits that affect mood, memory or movement are either already available commercially or in human clinical trials, and there is a growing pipeline of new neurotechnologies emerging from research laboratories. It is possible to use technology to repair and restore function both because of an impressive (but still very incomplete) body of neuroscience knowledge and the transformational technology and information processing achievements of the last decades.

Our sense organs provide electrical patterns of information about the state of the world. Neural machinery spread across the central nervous system uses those patterns to compute new representation patterns that nearly always make sense to us when processed properly in the brain. (How this remarkable process occurs is the driving force for a large fraction of neuroscientists). On the output side of the nervous system we are capable of an enormous repertoire of dexterous movements, like piano playing or ballet dancing. To generate skilled voluntary movement, the brain plans actions by assembling sensory signals and internally known information and outputs out electrical patterns that drive the coordinated muscle activity. Brain networks, in ways still quite unclear, also capture, store, organize, and generate memory, behavioral plans, and other cognitive functions. All this appears to include computing new information from internally generated activity patterns.

Fundamental knowledge about how the nervous system codes and computes information is now sufficient, and computing hardware and software good enough, to create neural prosthetic systems that can write-in or read-out neural codes to reproduce aspects of sensory and motor functions, and correct abnormal cognitive brain networks when these functions are lost or disrupted due to disease or injury. However, current neural prosthetics, which aim to emulate the function of neural circuits, still perform well below their biological counterparts, largely from two broad limitations. First, knowledge of brain function, particularly as an integrated information processing system, remains inadequate. Second, technology needed to replace neural function cannot adequately capture signal patterns and then copy computations that occur in most real neural networks – a software problem – and the devices capable of these computations are bulky, power-hungry and slow compared to their biological counterparts, and they are difficult to integrate into body – a hardware problem. Nevertheless, neuroscience knowledge is sufficient, technology adequate to create useful neural prosthetics, but there is great room for improvement. AI is one area that may be able to contribute to a major advance in the processing capabilities of neural prosthetics. Here, I will provide a high level overview of the current state of neural prosthetics from four examples of clinically motivated prosthetics, discuss the limitations faced now. In the spirit of this volume, I will comment on how AI could be a valuable approach to improve sensory, motor and brain circuit neural prosthetics (Fig. 1). AI is used specifically to refer to the deep learning approach (Hinton et al., 2006) (LeCun et al., 2015), because, as will be illustrated below, neural prosthetics suffer from a common challenges of detecting often incomplete activity patterns in raw complex and poorly characterized signals and transforming them into a new representation. The ability for Deep Learning DL nets to be able to generate reliable outputs in complex data is well suited to this class of problem, perhaps not surprisingly because DL nets are an attempt to copy the very processes neural prosthetics are trying to replicate. Figure 1 provides more detail and a schematic of this problem in the context of each of the examples that will be described next.

Sensory, Circuit and Motor Prosthetics

Four devices illustrate the current state of neurotechnology in helping humans: sensory neuroprosthetics for hearing and for vision restoration, deep brain stimulation (DBS) to modulate dysfunctional brain circuits, and (3) brain computer interfaces (BCIs) to restore movement. These neurotechnologies exemplify the forms, spectrum of developmental stages, and state of our ability to restore brain function with technology. Each could be enhanced through the methods available from the field of AI.

Sensory Neurotechnology

Each sense organ is an exquisite structure that provides the brain with spatially distributed and temporally changing patterns of electrical impuls-

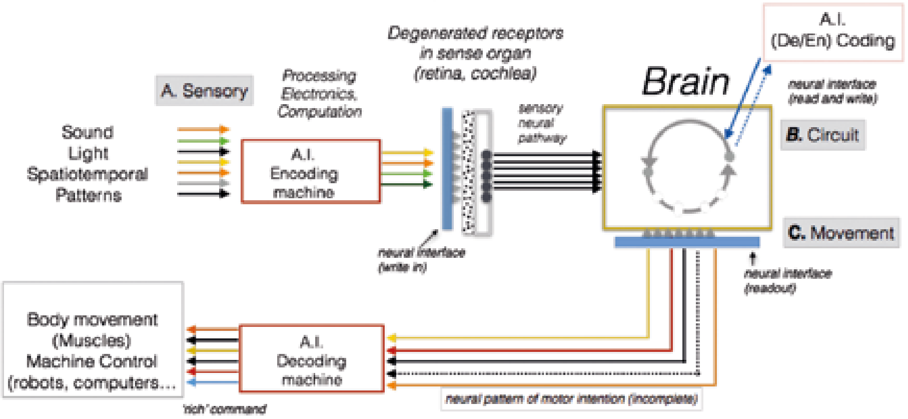


Figure 1. Role of AI algorithms in Sensory, Circuit and Motor Neural prosthetics. A. Sensory neural prosthetics to replace lost sensory receptors. Deep Learning neural nets would be used to learn the optimal match between natural sound or light patterns (colored arrow input to encoding 'machine') and patterns of stimulation (colored arrow output). The prosthetic device delivers this artificial pattern to brain pathways (black) via a stimulating (writing in) neural interface (blue) in sensory organ, after degeneration of sensory receptors (speckling). **B. Circuit prosthetics** modulate 'imbalanced' brain circuits (dashed arrow) to restore circuit function using stimulation targeted at a critical node in the pathway. Here deep learning could be used to search the stimulating pattern that best restores function. Here, a readout of either circuit activity or behavior would be needed to find optimal stimulation patterns. **C. Motor prosthetics** readout a limited sample of motor intentions (due to the restricted number of recording channels) from the brain through an electrode array (blue neural interface). Deep learning would be used to learn the best mapping between this incomplete neural signal pattern and a desired movement command signal pattern (decoder output) that would provide the best signal to operate devices like a computer for spelling, a robot arm to perform reach and grasping actions, or to activate muscles as a way to reconstruct the path from the brain to the body. Ideally, the implementation of AI (both the computational framework and computational power) could generate rich commands from these incomplete patterns to produce flexible, complex actions based on a limited sample of the desired control generated in neural activity patterns.

es emanating in large sets of neurons. These spatiotemporal neural activity patterns are a filtered version of various forms energy in the world – sound, light, chemicals, or mechanical forces. When activated through its specialized sense organ traveling initially through selective pathways, the brain interprets these patterns, for example, as sounds (air pressure changes transduced from the eardrum through the cochlea) or vision (electromagnetic radiation from 390 to 700 nm wavelength processed through the retina). Neural patterns are transformed and spread widely in brain networks, continuously 'computing' activity patterns (at least in the conscious state) that lead to the perception of an the object or the understanding of a spoken word (Fig. 1A). Thus, a large part of brain information processing appears to be the transformation of one pattern into another, put in a shamefully over-simplistic way. Damage to a sense organ disconnects the brain from that perceptual system (and more), limiting the use of that input to understand, remember or interact with the world. Most often, sensory receptor degeneration (e.g. inherited genetic disorders, mechanical damage) is the reason a sensory capability is lost, but the computing neural hardware remains without receiving patterned input needed to compute (Mysore et al., 2015). Available neural prosthetics can provide these lost patterns for both hearing and sight.

Hearing

The cochlear implant is the benchmark neurotechnology achievement for a human disorder. More that 250,000 devices have been implanted, allowing, for example, deaf children to attend standard educational programs. Nevertheless the understanding of sounds in the world is not at the level achieved by the intact biological interface between our acoustic world and the brain. This still crude neural prosthetic device has a profound personal and social impact (Bond et al., 2009).

The cochlea is a snail shell-shaped structure at the end of the middle ear where the mechanical motions of sound are converted by receptive hair cells lying a thin sheet along the length of the cochlea. Mechanical motion of the hairs atop these receptor cells result in electrical activity patterns in auditory nerve fibers, which connect to each hair cell. Sound generates patterns of activity across the auditory nerve. Many forms of deafness are the result of hair cell death but the auditory fibers typically still remain. A cochlear implant bypasses missing hair cells by delivering an artificial spatiotemporal electrical impulse patterns directly to the auditory nerve fibers in the cochlea. These electrical impulses, which at first are non-sense signals to the user when first supplied, over time become recognizable in brain auditory networks as meaningful, although not natural sound. Remarkably, comprehensible speech emerges when fewer than a dozen electrodes in the cochlea are used to replace thousands of lost hair cells. Thus, what is ordinarily a very rich spatiotemporal pattern of natural sound, can be replaced by impoverished pattern of electrical stimulation that the brain can still meaningfully use.

In the cochlear implant device, sounds are captured by an external microphone and processed using electronics housed in a small package worn behind the ear. The impulses are transmitted wirelessly through the skin to an implanted receiver-stimulator connected to a flexible, pencil-lead thin electrode that is threaded into the cochlea. In the intact ear, hairs atop different cells wiggle to different sound frequencies – different spots for different frequencies – which is in a simple sense a place code. Thus, the cochlea, in its simplest sense, has tonotopic map of sound in that frequency response is arranged spatially along the length of the cochlea. However, the actual transduction involves complex actions across the cochlea. In the healthy cochlea, hair cells chemically communicate their activation to auditory nerve fibers below them. The cochlear implant bypasses missing hair cells by directly activating auditory fibers, albeit with an electrode that probably activates hundreds of fibers at once because each stimulation site activates many fibers at once. Despite the impossibility of the current implant being able to recreate natural spatiotemporal auditory nerve fiber activity patterns, it is nevertheless quite successful in providing useful signal to the brain. In essence, the cochlear implant transforms sound in the world into a spatiotemporal pattern; this transformation is an attempt to copy the computation that the sound should have produced in the auditory nerve. A useful video of the system can be found at: www.youtube.com/watch?v=u8LpjkfvaSU.

Cochlear implants do not produce a natural sound in part because the technology cannot produce the correct spatiotemporal activity patterns in the auditory pathway nerve. We lack an understanding of how to compute natural neural patterns from sound. The inadequate transformation probably accounts for problems such as the difficulty users have understanding speech in noisy environments. Here is where AI strategies could improve function. Deep learning might provide an effective way not only in learning the optimal spatiotemporal pattern of stimulation to compute percepts from sound, but also help to learn and then generate missing components in the neural signals needed to correct for changing environments. Static sound processor algorithms aim to find the best algorithm, but deep learning approaches that identify relationships between input and output patterns to solve problems like filtering speech in crowds are already beginning to show promise (Healy et al., 2015). As will be repeated in the subsequent examples, these biological, computational and technological shortcomings of cochlear implants are shared by all current neurotechnologies and AI may be able to help improve pattern transformation across all of them.

Sensory Neurotechnology – vision

Retinal implants, another neural prosthetic device that has recently been approved, has the ability to restore a level of vision for people with blindness from retinal degenerative disorders. Several hundred have already been implanted. Like hearing, vision involves the transduction of a complex pattern light that falls on the retina into a neural activity pattern, that is further computed in brain networks and interpreted as form, structure or meaning. Photoreceptors and the ensuing circuitry at the back of the tissue-paper thin retina at the back of the eye produce spatiotemporal activity patterns. This pattern is transmitted to the brain via ganglion cells, which project their axons from the retina through the optic nerve to multiple sites in the brain. Vision, especially when in the service of a behavior, engages networks across vast extent of the nervous system, again in a continual re-computing of one pattern into another.

Vision loss is commonly the result of photoreceptor degeneration (e.g. macular degeneration (Mysore et al., 2015)), which stops light from engaging the first step of the retinal circuit that leads to ganglion cell activation, the obligatory path from eye to brain. Typically, ganglion cells and their brain connections remain, a parallel to auditory receptor (hair cell) degeneration with the auditory nerve remaining. Retinal implants, which were first approved in the US in 2013, have followed a design similar to the cochlear implant: A set of electrical stimulating electrodes is used to replace lost photoreceptors to activate intact visual projections, via the ganglion cells, to deliver spatiotemporal pattern of information to the brain (Lin et al., 2015). For a retinal implant, a two-dimensional sheet of stimulating electrodes is laid at the back of the eye (above or below the retina). Patterns of light detected on a camera (worn outside the body) are transformed into spatiotemporal electrical stimulation pattern on the array, which activate the ganglion cells. The activated ganglion cells carry this artificial pattern from the eye, through the optic nerve, to the brain. The user perceives this patterned electrical stimulation of the retina not like a typical visual scene, but instead the image is reportedly somewhat like a pattern of light flashes on a movie marquis made of many light bulbs. The number of artificial visual channels is low: dozens of electrodes are tapped into the roughly one million channels that go from the human eye to the brain. Importantly, retinal stimulation bypasses the complex intraretinal 'computational' neural machinery that transforms the light pattern falling on the photoreceptor sheet into a new pattern in the ganglion cells. 'Vision' provided by neural prosthetics can require significant time for the users to interpret, presumably as the brain mechanisms are used to interpret this unusual pattern of activation. An example showing the use of a retinal implant is available at: www.youtube.com/watch?v=DTiVKvs_lXg.

The fact that vision is possible with a neural prosthesis is remarkable, and of great impact for those seeking to see again. However, restoration of natural vision will ideally require much better hardware (to couple more channels across the full extent retina) and processing to recreate natural vision, including mapping computation that occurs in the retina itself. For vision, AI could help in learning and then computing a more effective representation of signal that are produced after light is processed by the eye and retinal circuitry, which could produce more natural vision and help correct for real world complexities like changing illumination that the brain 'expects' from the eye. AI based on deep learning of natural scenes, which can perform many human-like perceptual functions, appear not to have been implemented in retinal prosthetics yet, but should be able to help the still small number of channels activated by the neural interface to create more natural activity patterns, leading to more natural vision.

Brain Circuits – Neuromodulation prosthetics

Stimulating sensory neurons activates pathways that are eventually interpreted by brain circuits. These circuits store information in memory, or can immediately invoke action, or delay it for later actions (planning). Highly interconnected brain networks quickly and flexibly combine information from any input and various 'internal circuits' to engage almost any output in ways still poorly understood. Disorders that emerge from imbalanced activity of certain brain networks appear to lead to perceptual, cognitive (including affective), and movement disorders. Neural prosthetics to correct malfunctioning circuits through targeted electrical stimulation-termed *neuromodulation* are already in use, although they currently engage 'brute force' tactics that inject electrical impulses within a complex circuit without fully understanding how this injected 'information' modulates the computations produced by this network. The most remarkable and widely used example of neuromodulation success is the use of deep brain stimulation (DBS) in Parkinson's disease (PD), where the shaking, rigidity and tremor of the disorder is relieved by stimulating a particular point in a cortical-basal ganglia circuit at about 100 times per second. Parkinson's disease is the result of the loss of the neurotransmitter dopamine, which is essential for the normal operation of cortical-thalamic-basal ganglia networks that control movement planning and performance, as well as cognitive functions. Exactly how these circuits work, or depend on dopamine is not fully understood, but remarkably DBS stimulation at one node in this circuit overcomes the dopamine-induced deficit, readjusting the circuit so that it operates more normally, as long as stimulation is continued. With DBS, which has been applied in more 150,000 people symptoms are diminished often substantially (but not cured).

Deep brain stimulation (DBS) is a process of using a pattern of electrical stimulation through an electrode surgically inserted into a select region of a basal ganglia thalamo-cortical loop (i.e. subthalamic nucleus, STN) in order to alter the functional activity level, of a part of that circuit (Fig. 1B). Typical DBS electrodes consist of a spaghetti noodle-sized probe with four (or more) mm sized metal contacts near its end. The probe is inserted into the STN, a collection of neurons about the size of a lentil bean. Repeated electrical stimulation in STN – through an electronic pulse generator placed under the skin of the chest – modulates brain circuit function presumably modulating the circuit so that it computes properly. Typically one electrode is used, but multiple electrodes are being evaluated as a way to create more complex or accurately localized spatiotemporal patterns. The result in PD is impressive (one of many videos of the effect at: https://www.youtube.com/watch?v=17ch1guvoLA).

The safety profile for DBS implantation is quite good (DiLorenzo et al., 2014). But not surprisingly, DBS can have side effects, including effects on cognition (Wu et al., 2014), perhaps due to the large electrodes (>1mm), the difficulty in very exact placement requirements, the proximity of other circuits to the stimulation site or the type of stimulation pattern employed, but it remains impressive that crude stimulation works so well. Tuning stimulation parameters is complex. DBS is now open loop, in that it does not use information about activity patterns in the circuit to shape the best stimulation pattern. This is hindered in part by the difficulty in monitoring the state of the circuit, although at least basic macro recording is now beginning to be possible. When readout is available, AI might be useful in finding intelligent ways to adaptively learn the best timing or intensity of stimulation to optimize the effect, which now is a barrier to effective DBS therapy (Arlotti et al., 2016). Effectively this plan would produce a bioelectronics hybrid brain circuit.

DBS is also being investigated as a circuit neuroprosthetic for broad range of other disorders including affective disorders like depression (Choi et al., 2015), obsessive-compulsive disorder (repetitive habits) (Fayad et al., 2016) and memory loss in Alzheimer’s disease (Mirzadeh et al., 2016) involving various other frontal circuits. The potential for DBS success in affective or cognitive circuits still requires considerable further inquiry. Advances are currently limited by our poor understanding of computations occurring in these complex, dynamic brain circuits, difficulties associated with knowing how or where to intervene in these networks, and availability of technology to precisely deliver correct spatial and temporal patterns. DL approaches would potentially be helpful in repairing these circuit disorders, by learning from abnormal brain patterns (by *reading out* patterns of activity) and converting them into meaningful activity patterns for that circuit (*writing in* by stimulation of the correct circuit sites).

Changing brain circuits raises ethics issues. DBS provides a tool to shape virtually any brain circuits including those affecting personality or behavior and therefore must be judiciously monitored for ethical application. The concept of altering brain circuits, and potentially behavior, with electrical stimulation (or other forms of energy) is more immediately concerning as simpler, but much less precise brain stimulation devices are used. It is now possible to influence circuits with technology that can be applied outside the head, which has growing adoption in the public. Most specifically, Transcranial Direct Brain Stimulation TDCs is possible with everyday technology (batteries and saltwater soaked sponges on the head) and it is very cheap and easy to make. TDCS has a popular following and is being used for every imaginable issue, often with no valid scientific backing (Wexler, 2017) raising ethical, legal and social concerns (Kuersten and Hamilton, 2014).

Movement restoration – Brain computer interfaces

Voluntary movement also emerges from brain circuits and, not surprisingly, is elaborated by a vast network of central nervous system structures. However, the corticospinal pathway, a bundle of axons connecting neurons in cerebral motor areas to the spinal cord, is one critical path that provides a patterned input to the spinal cord (and many other structures) to generate skilled movement particularly of the fingers and hand. Paralysis results from a number of disorders, including stroke, spinal cord injury, or traumatic brain injury. When any of these disorders disrupts the corticospinal pathway anywhere along its route, paralysis of useful, skilled actions including hand motion, walking or speech may occur. A brain computer interface (BCI) is a system that is designed to bypass damaged brain structures and restore brain-controlled movement by using brain activity patterns as a source of movement commands. BCIs recreate action commands from limited samples of neural activity patterns from brain areas that have activity patterns related to movement intentions. These patterns can be read out and decoded into commands able to operate devices like a computer or a robot, or even the paralyzed muscles themselves. A BCI can be considered as the converse of devices discussed so far, in that a BCI is intended to *read out* brain activity (i.e. recording activity) so that intentions can become actions, rather than trying to *write in* signals into the brain or nerves. BCIs have been used in investigational studies in fewer than 20 people with severe paralysis to restore their ability to move or interact with the world. (for more comprehensive reviews see: (Donoghue et al., 2007; Hatsopoulos and Donoghue, 2009; Homer et al., 2013).

Movement intentions arise from a spatiotemporal pattern of activity in cortical neurons across a network of cerebral motor areas. The primary motor cortex (MI) is a major origin of the corticospinal pathway and a key node in a much larger cerebral motor network. Using an aspirin-sized bed of 100 hair-thin probes that are inserted just into the surface of the MI arm region, it is possible to record patterns of neural activity from a each of many individual neurons that reflect the coordinated motions of the arm, say to reach and grasp. Current, typically static, algorithms make it possible to convert that pattern from a small sample of neural activity into control signals that can allow a person who is fully paralyzed to control a multijoint robotic arm well enough to pick up a container of coffee, drink from it, and put it back on the table (for video see: https://www.youtube.com/watch?v=ogBX18maUiM). Other groups have extended this work and demonstrated even more dexterous control (Collinger et al, 2013). However, actions of computer cursors or robotic arms using BCIs are slow and far less dexterous than we effortlessly accomplish all the time. The computational power of AI, by potentially learning better mappings between limited, complex patterns of cortical neural activity and the requisite command structure, could provide a much richer, faster and complex control (Fig. 1C). As depicted in Fig. 1 for motor systems, current sensors only sample incompletely – they have access to a tiny fraction of the activity ongoing to coordinate even the simplest reach and grasp action. Optimized DL algorithms (and hardware) might be able to make up for the small sample and missing information to achieve the speed and dexterity achieved by able-bodied people. DL could also benefit from information about real world actions which help constrain the problem (Howard et al., 2009), but this has not yet been tried.

Conclusions

Neural prosthetics are remarkable mainly early stage attempts to replace missing neural structures, but they are not able to fully replicate the brain structures they are intended to replace. Their shortcoming emerges from technology limitations, namely the inability of current electrode interfaces to address (write in) or sample (read out) the full spectrum of channels, the size of computational technologies which can limit their processing power or portability, or their compatibility with the body. Importantly, successful neural prosthetics are limited by the inability to reproduce computations of biological circuits, which can be simplistically reduced to a problem of computing one spatiotemporal pattern from another. The problem is exacerbated by incomplete and noisy information inherent to neural data. Deep learning appears to be a framework very well suited to more faithfully mimic this computation compared to current approaches, because this approach is particularly good at finding high-level abstractions, (i.e., complex patterns) in large-scale data.

AI, ethics and the limitations and potential of Neuroprosthetics

The promise of neurotechnology to improve human health is substantial and not limited to the examples given above. There are many other neuroprosthetic technologies where advanced, intelligent computing can help improve the lives of those with neurological disorders or injury, such as creating a brain controlled artificial limb for people with limb loss. In my view, these technologies are likely to be realized, although it is very difficult to predict the timing and pace of success when fundamental research issues still are required. 'AI', neuroscience and engineering advances will all play a role in realizing more effective technology to restore vision, hearing, cognition, or movement to those disabled by nervous system disorders. AI offers neuroprosthetics a means to learn and then implement computations like those achieved by neural circuits, without full understanding how these computations are achieved. As they are integrated into neural functions, they may provide a framework to create more and more brain-like computing that could replicate or even exceed those capabilities. We should be aware of the impact of such advances on society and the individual.

Accelerating AI and in neuroprosthetics successes indicate that we are at an inflection point where the ability to augment human function with these bio-machine hybrids, though still a long way off, can be realized. Full restoration of humans who have lost critical functions of their nervous system would be an outstanding success for neuroengineering, the extension of this to ways to augment abilities in able-bodied individuals is easy to imagine as medical applications expand. One might envision retinal implants enabling night or infrared vision, or, more fancifully, memory circuit stimulation to double memory capacity. These speculations raise ethical and social challenges that need to be evaluated now in the scientific and legal communities so that we are prepared as these capabilities emerge. Lastly, it is important to recall that there are other big challenges to achieving the bionic human either to overcome disability or to augment function. High resolution communication with the nervous system, for the foreseeable future, will require surgical interventions that will slow adoption, due to cost and real or perceived risk, and will surely influence social views on using this type of neurotechnology.

References

Arlotti M, Rosa M, Marceglia S, Barbieri S, Priori A (2016) The adaptive deep brain stimulation challenge. *Parkinsonism Relat Disord* 28:12-17 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27079257 [Accessed February 26, 2017].

Bond M, Mealing S, Anderson R, Elston J, Weiner G, Taylor R, Hoyle M, Liu Z, Price A, Stein K (2009) The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model. *Health Technol Assess* (Rockv) 13:1-330 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19799825 [Accessed February 26, 2017].

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C.,

Weber, D.J., … Schwartz, A.B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet*, *381*(9866), 557-64. http://doi.org/10.1016/S0140-6736(12)61816-9

Choi KS, Riva-Posse P, Gross RE, Mayberg HS (2015) Mapping the "Depression Switch" During Intraoperative Testing of Subcallosal Cingulate Deep Brain Stimulation. *JAMA Neurol* 72:1252-1260 Available at: http://archneur.jamanetwork.com/article.aspx?doi=10.1001/jamaneurol.2015.2564 [Accessed February 26, 2017].

DiLorenzo DJ, Jankovic J, Simpson RK, Takei H, Powell SZ (2014) Neurohistopathological Findings at the Electrode-Tissue Interface in Long-Term Deep Brain Stimulation: Systematic Literature Review, Case Report, and Assessment of Stimulation Threshold Safety. *Neuromodulation Technol Neural Interface* 17:405-418 Available at: http://doi.wiley.com/10.1111/ner.12192 [Accessed February 25, 2017].

Donoghue JP (2015) Neurotechnology. In: *The future of the brain: essays by the world's leading neuroscientists* (Marcus GF (Gary F, Freeman J, eds), pp 219-233. Princeton University Press.

Donoghue JP, Nurmikko A, Black M, Hochberg LR (2007) Assistive technology and robotic control using motor cortex ensemble-based neural interface systems in humans with tetraplegia. *J Physiol*

579:603-611.

Fayad SM, Guzick AG, Reid AM, Mason DM, Bertone A, Foote KD, Okun MS, Goodman WK, Ward HE (2016) *Six-Nine Year Follow-Up of Deep Brain Stimulation for Obsessive-Compulsive Disorder*. Bankiewicz K, ed. PLoS One 11:e0167875 Available at: http://dx.plos.org/10.1371/journal.pone.0167875 [Accessed February 26, 2017].

Hatsopoulos NG, Donoghue JP (2009) The science of neural interface systems. *Annu Rev Neurosci* 32:249-266 Available at: http://www.annualreviews.org/doi/10.1146/annurev.neuro.051508.135241 [Accessed February 26, 2017].

Healy EW, Yoho SE, Chen J, Wang Y, Wang D (2015) An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *J Acoust Soc Am* 138:1660-1669 Available at: http://asa.scitation.org/doi/10.1121/1.4929493 [Accessed February 24, 2017].

Hinton GE, Osindero S, Teh Y-W (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput* 18:1527-1554 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16764513 [Accessed February 26, 2017].

Homer ML, Nurmikko A V, Donoghue JP, Hochberg LR (2013) Sensors and decoding for intracortical brain computer interfaces. *Annu Rev Biomed Eng* 15:383-405 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3985135&tool=pmcentrez&rendertype=abstract [Accessed July 14, 2014].

Howard IS, Ingram JN, Körding KP, Wolpert DM (2009) Statistics of natural movements are reflected in motor errors. *J Neurophysiol* 102:1902-1910 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19605616 [Accessed February 26, 2017].

Kuersten A, Hamilton RH (2014) The brain, cognitive enhancement devices, and European regulation. *J Law Biosci* 1:340-347 Available at: https://academic.oup.com/jlb/article-lookup/doi/10.1093/jlb/lsu019 [Accessed February 26, 2017].

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436-444 Available at: http://www.nature.com/doifinder/10.1038/nature14539 [Accessed February 26, 2017].

Lin T-C, Chang H-M, Hsu C-C, Hung K-H, Chen Y-T, Chen S-Y, Chen S-J (2015) Retinal prostheses in degenerative retinal diseases. *J Chinese Med Assoc* 78:501-505 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26142056 [Accessed February 26, 2017].

Mirzadeh Z, Bari A, Lozano AM (2016) The rationale for deep brain stimulation in Alzheimer's disease. *J Neural Transm* 123:775-783 Available at: http://link.springer.com/10.1007/s00702-015-1462-9 [Accessed February 26, 2017].

Mysore N, Koenekoop J, Li S, Ren H, Keser V, Lopez-Solache I, Koenekoop RK (2015) A Review of Secondary Photoreceptor Degenerations in Systemic Disease. *Cold Spring Harb Perspect Med* 5:a025825 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25475108 [Accessed February 26, 2017].

Wexler A (2017) Recurrent themes in the history of the home use of electrical stimulation: Transcranial direct current stimulation (tDCS) and the medical battery (1870-1920). *Brain Stimul* 10:187-195 Available at: http://linkinghub.elsevier.com/retrieve/pii/S1935861X16303825 [Accessed February 26, 2017].

Wu B, Han L, Sun B-M, Hu X-W, Wang X-P (2014) Influence of deep brain stimulation of the subthalamic nucleus on cognitive function in patients with Parkinson's disease. *Neurosci Bull* 30:153-161 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24338433 [Accessed February 26, 2017].

Who Am I? The Immersed First Personal View

Laurie Ann Paul

Case 1. On a train to a new destination in a foreign country, you are lulled to sleep by the gentle rocking of the carriage. Suddenly, you are startled awake by a sudden stop and the opening of the compartment doors. You realize you must have missed your stop! You leap up, gather your things, and jump off. You have no map and no phone. Disoriented, you wonder, *where am I?*

Case 2. You are using your virtual reality headset to explore a high mountain ridge in the Alps. As you walk along the thin edge of a precipice, you trip over a concrete block that your business partner, a known practical joker, put on the floor of the room you are in. A rush of fear combined with a disorienting return to external reality jerks you from your VR experience back into the room.

What is your mind doing when it reorients itself?

1. The self in experience and decision

Exploring these sorts of disorientation can help us to articulate the structure of what it is to be a self. Understanding the constituent features of one's self is highly relevant to questions of artificial intelligence and to designing a machine that could be a thinking self or that could think like a human.

A related philosophical connection is to recent work on the nature and structure of transformative decisions, experience, revisionary epistemic change, and self-change (Paul, 2014a). That work explores the deep epistemic structure of how we understand who we are, and how we re-construct ourselves through major epistemic upheavals. With transformative decision-making, the focus is on decision models for first personal decisions. A central idea involves the concept of transformative epistemic change: some decisions can lead to epistemic changes so profound that they create significant self-change. An example that brings out the idea involves a congenitally blind saxophonist.

Imagine a blind adult who makes his living playing the saxophone. One day, he is offered a one-time-only chance to have retinal surgery to become sighted. How does he assess the decision of whether to have the surgery? In this kind of case, before the epistemic change occurs, there is no way for him to imagine or represent the experiential nature of the change. What will it be like for him to become sighted? In effect, he has the chance to have a new kind of experience, an experience that he cannot assign a subjective, experiential value to. He cannot assign an experiential value because he lacks the capacity to imaginatively represent the nature of this lived experience.

The situation raises a distinctive set of decision problems, some of them associated with his inability to grasp his possible future self as a sighted adult. The problem illustrates ways we can lack the ability to imagine and model our future selves, and to assign values to possible lived experiences when our futures involve dramatically changed selves. By extension, it illustrates problems with formulating diachronic decision rules for radically incommensurable selves.

All of these ideas, at bottom, are founded on an understanding of the nature and structure of the experienced self, and on a picture where, from the first personal perspective, in many ordinary contexts, we are selves who plan, decide, and act as we evolve forward in time.

2. Constitutive features of selves

Today, I want to focus on identifying some of the foundational elements of the experiencing self. My interest in the examples of the train and the virtual Alps is in how they can be used to highlight constitutive features of thinking selves. Such features may be so basic that we don't explicitly attend to them in ordinary contexts. The examples also help us to take a distinctive perspective: they help us take the first personal view of the self. The view *of* the self, a view from an immersed perspective, is different from the view *on* the self.

To construct a machine that can think like a human, we want to find a way to represent a first personal point of view and capture the way that a self, from its perspective, deliberates and functions in the world. Ordinarily when people think about the self, they start by thinking about how an individual recognizes itself as a thinking self, usually in terms of what that self values or desires, its intentions, and how it has self-awareness. But this builds in a lot right away.

I am starting at a more fundamental level. To understand what a self is, and how an individual knows who she is, we need to understand her immersed perspective. We need to understand the view of the self in question.

This involves an exploration of the fundamental experiential structure of a first personal point of view.

So my first point is that the underpinning of an individual's understanding of who she is, of her self-conception and self-awareness, is structured by her consciously centered, experiential point of view.

What is a consciously centered, experiential point of view? An example, couched in terms of camera angles, can help to bring the idea out. Think of the sort of view that you get with a Go Pro, a type of digital camera designed for filming action while being immersed in it. Or take an immersive computer game. An "immersed first personal viewing angle" is a distinctive and important camera angle that you get from, metaphorically, occupying the boots of your character in a computer game. The view is as though you were looking out the eyes of the character, seeing the world as it sees the world. This captures the first personal visual perspective of the player. Computer games can add a further level of cognitive immersion from an action camera if they give you a certain amount of control over the character's visual perspective and actions.

The immersed first personal camera angle, set up as though you were looking out from the eyes of the character, gives us a visual analogue of an individual's consciously centered, experiential point of view. Note that the analogy is only partial, because the centering it represents is largely just visual and causal. A person's first personal perspective brings in more than this, as it is both a sensory and cognitive centering of the perspective.

Part of why I am emphasizing immersion here is that, when thinking about these issues in the abstract, we can miss details by moving too quickly. We can shift, almost without noticing it, into a third personal approach to the self. This shift is like shifting from the immersed visual angle where I am occupying the boots of my character to a third personal viewing angle using a "follow camera" to track my character. The importance of this difference is represented in how distinctive it can feel to make this visual shift in gameplay. Moreover, the shift in perspective can change the way the player is able to solve tasks in the game, aligning the perspectival shift from first personal to third personal with a functional shift.

When reasoning about the self, if we only explore the third personal angle, we miss the difference between, for example, a self being located in time and space and the experience of being located in time and space. A self may be located in space and time, *but it's the immersive or centered experience of being located* which is a constituent of the self. This is not the same thing as just having a location in space and time!

Further, if we miss the crucial difference between having a location and the immersive experience of being located, we can miss the deeper structure of how the immersive experience of being located is comprised of a sense of being here, now, along with coordination to external spatial and temporal cues.

My second main point is that these sorts of immersive experiential features are part of what make up the self. In the two examples I started with, you are disrupted along some dimension of your first personal orientation. In the train case, where you wake up and jump off the train, you are spatially disoriented, because your internal representation and monitoring of your spatial location (which involves keeping it correlated with the external facts) has been disrupted. To orient yourself, you need to recalibrate by finding your location on a map.

This disorientation highlights the immersive, first personal experience of *being located* which is different from knowing your location on a map. Ordinarily, I have an immersive experience of where I am created by constantly coordinating or updating my immersed first personal sense of being at a location with my third personal perspective or map view of where I am.

Time and temporal experience are the same. I engage in regular calibration and updating of my experienced temporal location by comparing my experienced sense of what's present or now, and my sense of how much time has passed, and coordinating it with my location as understood externally, using a clock.

There is an even deeper sort of temporal coordination involving the direction of time. My personal sense of time passing, and of the deep difference between the past and the future, are fundamental structural features of my point of view, and I orient myself in the world by coordinating this internal point of view with the external world. (Consideration of time travel cases can bring this out: imagine looking out the window of the time machine and watching the world running backwards as, inside, you live forwards). My internally directed, asymmetric sense of what counts as past and what counts as future are constituents of my first personal self. I find myself balanced in the nexus between the past and the future, and direct myself towards the future. (Relatedly, this gives me an internal sense of the direction of causation).

The structure of our immersed, centered temporal and spatial experience also includes another element, a more esoteric sense of who I am. I know I am here, I know I am here now, but I also know that I am *me*. When I anticipate, I am thinking of *my* future. When I remember, I recall *my* past. And the same is true for you. If you lose your memories, there is an important sense in which you no longer know who you are. If you had an accident where you lost your memories, you'd be disoriented with respect to who you are. And this is temporal and causal: you need to know that your experienced memories are representations of past experiences that played a role in creating who you are now.

Also note that, to recalibrate your sense of who you are, it isn't just a matter of thinking of past experiences. You need to recognize these thoughts as *your* memories. If you had the memories but somehow didn't recognize that the first personal experiences you are recalling are *your* experiences (perhaps you thought they were false experiences, or experiences of someone else's first personal perspective), you would not recover your sense of self.

So the point here is that an immersed first personal representational sense of *one's own* memories are essential elements of the centered conscious experiencer (the first personal self) (Paul, 2014b). At least one crucial way I know that I am me, and how I define who I am, is that I grasp my memories (as *my* memories). I sync and update my current experiences as causal outgrowths of my past experiences, and I recognize my memories as my past experiences.

Similarly for the temporal character of anticipation: I have to know what counts as a memory versus what counts as an anticipation. I must distinguish my past selves from my future selves, and recognize my future selves *as future*.

Very briefly: this can be important for rational deliberation and action, and it comes up in the discussion of transformative experience and decision. In the case with the congenitally blind saxophonist, the trouble is that he cannot, in the ordinary way, project himself forward into the shoes of his possible future self. What he wants to be able to do is consider future ways he could live, or future ways he could be, as a sighted individual. But in an essential sense he can't mentally project or evolve himself forward from his immersed perspective. He has to become sighted to know what it will be like to be sighted. Before the operation, he faces an epistemic wall that he can only get past by having the experience itself. The further implication is that this epistemic change, becoming sighted, will scale up into a change in who he is. The addition of sight fundamentally alters the structure of his immersed experience, and by extension alters the nature of his centered, conscious, experiencing self.

The point generalizes, especially because there are many other new kinds of experiences that can transform you, such as going to war, becoming a parent, or experiencing massive technological transformation. A profound epistemic change in the nature of an experiencer's first personal perspective can lead to a restructuring of his values or preferences, and thus can change, in a deep way, who he is. This again connects to AI, for the building blocks of AI include a conception of what a self is, how it is structured by its values, and how it makes decisions and updates itself in response to the external world (See Paul 2014a for further discussion).

3. Modality

What are some other features of the self? We can tease further elements out with more examples. What happens when you wake up from an intense dream, in an unfamiliar room? You are disoriented until you recall where you are and why you are there. Your immersed qualitative experience distinguishes between different realities, distinguishing what it takes to be real versus what it sees as merely the experience of the dream. The immersed self, then, wants to distinguish between the real world and other worlds, and needs to know what's real to know which features of its experience are part of who it is.

Now we've got several distinctive features that characterize what a self *is*: one's spatial and temporal immersive sense of being here and now, paired with regular updating to external spatial and temporal cues, and a directed difference between the past and the future. These blend with causal experience and the sense of having one's own memory, to give us a located, centered, and directed point of view that makes an implicit distinction between what's real and what's not. In addition, the updating and monitoring of location and other elements of my centered conscious experience seem to be an internal way of tracking and modeling myself, and in this way knowing myself: as I think of it, it forms part of my intuitive self. I use it to control, create, and know who I am.

It also defines a boundary between who I am and the rest of the world. (In the following sense: when I'm mentally coordinating my immersed perspective with external cues, I'm defining myself in juxtaposition to the rest of the world. Finding myself on a map, coordinating my sense of time's passing with the movement of the hands of the clock, and distinguishing reality from the dream world all help me know where I end and the rest of the world begins).

There are surely additional features of the self to explore. One important feature of the centered conscious self involves the nature and character of its experienced sensory information. Another very important one involves the self's relations to other people. Once we have the basic structure of a focused and centered first personal perspective in place, I'm inclined to think that another constituent of what a fully realized self is involves its relations to other people and things. It may be that a distinctive element of the self's relations to other selves is its representation of those other selves *as selves* or *as* conscious beings.

Now that we have all this in play, I'd like to go back to the virtual reality example. The case of the virtual Alps, where you stumble over a concrete block, is a case where you are disoriented because your immersed representation of the world, a virtual world defined by your visual immersion, has been disrupted.

You know where you are in the virtual reality, but you also need to know where you are in the external reality of the room. The concrete block disrupted your orientation in your visual (virtual) reality. To re-orient yourself, and to avoid tripping over the concrete block again, you have to recalibrate and coordinate your immersed visual perspective of your virtual reality with your tactile perspective on your external reality.

The example brings out how immersed experience in a kind of reality is a feature of the centered conscious experiencer: in the virtual Alps case, we can contrast your immersed visual experience of the virtual world with your immersed tactile experience of the external or real world. A VR user might even need to exploit her understanding of the contrasting modalities (virtual and real) for problem solving: imagine she has to find her way around a virtual boulder in her Alpine VR experience, but to do this she has to open a closed door in the external reality of the room she is in. Opening the door will move her around the virtual boulder.

What is the mind doing when it solves the task of the virtual boulder and closed door?

To successfully perform this task, she needs to clearly distinguish the two modalities she is working in: virtual and external, and she needs to coordinate between her visual VR's spatially, temporally, causally, immersed point of view and her tactile external spatially, temporally, causally, immersed point of view. Finally she must manage the interpretation between them.

This brings out how a centered understanding of the features of an agent's point of view can frame her actions and define her problem solving. Let's make the story a tiny bit more complicated: imagine that the virtual boulder has to be pushed aside with the help of other VR users, and the door is too heavy to be opened without their help as well. Together, you have to move the boulder by opening the door. You are the team leader.

What is your mind doing when you solve this joint action problem?

Here, part of what you need to solve the action puzzle is to clearly distinguish and represent the first personal features of your virtual and external realities, as well as represent the different virtual and external perspectives of others, and then coordinate between their represented modalities and your own modalities.

Again, computer gameplay has an analogue: in a multiplayer game where you can see the points of view of your teammates, you have your own first personal view *in addition to the first personal views of others* embedded into your screen. The structure of joint action can also be applied to cases where you are negotiating with or understanding yourself at different times: your self at a past time, at a future time, or even at a merely possible time (a merely possible location or situation) can be treated like another agent with its own first personal view. If so, the complicated virtual reality joint action case isn't just for acting and making decisions with other people. We do something similar when we are making decisions for our future selves or our merely possible selves. That is, we sometimes need to be able to represent and understand the points of view of our future selves, our past selves, and our merely possible selves in order to act rationally. The example of transformative experience I discussed above, where the saxophonist must rationally assess his possible future self, is another example. These are just some of the ways in which understanding the self connects deep philosophical issues to exciting questions in artificial intelligence.

References

Paul, L.A. (2014a). *Transformative Experience*. Oxford: Oxford University Press.

Paul, L.A. (2014b). "Experience and the Arrow". In *Asymmetries of Chance and Time*, 174-193, edited by Alistair Wilson. Oxford: Oxford University Press.

Statement on Artificial Intelligence

On November 30th and December 1st 2016, the Pontifical Academy of Sciences hosted an international symposium on *Power and Limits of Artificial Intelligence.*

In the past decade, computer science, robotics, and artificial intelligence (AI) have made remarkable progress. Those technologies hold great promise to address some of our most intractable social, economic and environmental problems, but they are also part of a long-term trend towards automatization, with consequences that may ultimately challenge the place of humans in society. This committee therefore reviewed the current trends of AI research, its potential utility and dangers, and formulated a number of recommendations.

Current trends

Major research is underway in areas that define us as humans, such as language, symbol processing, one-shot learning, self-evaluation, confidence judgment, program induction, conceiving goals, and integrating existing modules into an overarching, multi-purpose intelligent architecture. While progress is impressive, no evidence suggests the imminent emergence of a runaway intelligence with a will of its own. Artificial intelligence remains far from human and lacks an overarching mathematical framework.

Benefits

Used as a toolkit, AI has the potential to advance every area of science and society. It may help us overcome our cognitive limitations and solve complex problems such as energy management and ecology, where vast amounts of data present a challenge to human understanding. In combination with robotics and brain-computer interfaces, it may bring unique advances in medicine and care. By elucidating how we learn, it may bring dramatic changes in education. It may also help scientists shed light on the nature of intelligence, the organization of the universe, and our place in it.

Dangers

Unless channeled for public benefit, AI will soon raise important concerns for the economy and the stability of society. We are living in a drastic transition period where millions of jobs are being lost to computerized devices, with a resulting increase in income disparity and knowledge gaps.

With AI in the hands of companies, the revenues of intelligence may no longer be redistributed equitably. With AI in the military, we may witness a new and costly arm race. While intelligent assistants may benefit adults and children alike, they also carry risks because their impact on the developing brain is unknown, and because people may lose motivation in areas where AI is superior.

Recommendations

The effort to develop intelligent machines must remain continuously directed to the greater good, reducing the poverty gap and addressing general needs for health, education, happiness and sustainability. All governments should be alerted that a major industrial revolution is underway and must take new measures to manage it. Scientists and engineers, as the designers of AI devices, bear a primary responsibility in actively trying to ensure that their technologies are safe and used for good purposes. We welcome the initiatives of some companies to create in-house ethical and safety boards, and to join non-profit organizations that aim to establish best practices and standards for the beneficial deployment of AI. We also call for external civil boards to perform recurrent and transparent evaluation of all technologies including the military. The value functions that AI is asked to optimize require special attention, as they may have unexpected biases or inhuman consequences. Just like crash tests for transportation, the passing of ethical and safety tests, evaluating for instance social impact or racial prejudice, could become a prerequisite to the release of AI software.

Signatories

Werner Arber, Antonio M. Battro, Olaf Blanke, Patricia Churchland, Stanislas Dehaene, John Donoghue, Demis Hassabis, Stephen W. Hawking, Yann LeCun, Pierre Léna, Laurie Ann Paul, Alexandre Pouget, H.E. Msgr. Marcelo Sánchez Sorondo, Laura Schulz, Mariano Sigman, Wolf J. Singer, Elizabeth Spelke, Josh Tenenbaum, Manuela Veloso, Cédric Villani.

Declaración sobre inteligencia artificial

La Pontificia Academia de Ciencias celebró un simposio internacional del 30 de noviembre al primero de diciembre 2016 sobre *Poder y límites de la inteligencia artificial.*

En la década pasada, la ciencia de la computación, la robótica y la inteligencia artificial (IA) han realizado progresos considerables. Estas tecnologías son promisorias para encarar algunos de nuestros problemas sociales, económicos y medio-ambientales más acuciantes, pero también forman parte de una automatización a largo plazo cuyas consecuencias podrían comprometer el lugar que ocupa el ser humano en la sociedad. En consecuencia, nuestro comité pasó revista a las tendencias actuales de la investigación en IA, su utilidad potencial y sus peligros y formuló una serie de recomendaciones.

Tendencias actuales

Está en curso una considerable investigación en áreas que nos definen en tanto seres humanos como el lenguaje, el procesamiento de símbolos, el aprendizaje inmediato, la auto-evaluación, el juicio certero, la inducción de programas, proponer objetivos e integrar los módulos existentes en un arquitectura abarcativa y multipropósito. Aunque el progreso es impresionante no existe evidencia alguna sobre la emergencia inminente de una inteligencia descontrolada con una voluntad propia. La inteligencia artificial está muy lejos de la inteligencia humana y carece de un encuadre matemático abarcativo.

Beneficios

Cuando la IA se usa como instrumento tiene capacidad de hacer progresar todas las áreas de la ciencia y de la sociedad. Nos puede ayudar a superar nuestras limitaciones cognitivas y a resolver problemas complejos como la gestión de la energía y de la ecología, donde la enorme cantidad de datos representa un desafío para la comprensión humana. En combinación con la robótica y con interfaces cerebro-computadora podrá provocar avances considerables en medicina y en asistencia. Al elucidar cómo aprendemos podrá aportar cambios radicales en educación. También podrá ayudar a los científicos a comprender la naturaleza de la inteligencia, la organización del universo y nuestro lugar en él.

Peligros

Si la IA no se canaliza hacia el beneficio público traerá pronto problemas importantes para la economía y la sociedad. Estamos viviendo un drástico período de transición donde millones de trabajos se están perdiendo por causa de los equipos computarizados, lo que provoca un crecimiento en la disparidad de ingresos y en la brecha de conocimientos. Con el uso de la IA en las fuerzas armadas, podríamos asistir a una nueva y costosa carrera armamentística. Si bien la asistencia inteligente puede beneficiar tanto a adultos como a niños también ello puede ser riesgoso puesto que su impacto en el desarrollo cerebral no se conoce y podría hacer que las personas perdieran motivación en las áreas donde la IA es superior.

Recomendaciones

El esfuerzo para desarrollar máquinas inteligentes debe estar dirigido constantemente al bien mayor, reduciendo la brecha de pobreza y tratando las necesidades generales de salud, educación, felicidad y sustentabilidad. Se debe alertar a todos los gobiernos que estamos ante una revolución de gran magnitud y que debemos tomar nuevas medidas para gestionarla. Los científicos e ingenieros, en tanto diseñadores tienen una responsabilidad fundamental en asegurar que sus tecnologías sean seguras y se usen con buenos propósitos. Son bienvenidas las iniciativas de algunas compañías para crear comisiones de ética y de seguridad y para asociarse a organizaciones sin fines de lucro con el fin de establecer las mejores prácticas y medidas en la implementación beneficiosa de la IA. También recomendamos que comisiones civiles externas realicen evaluaciones periódicas de todas las tecnologías, incluyendo las militares. Se requiere prestar especial atención a aquellas funciones con valores que la IA debe optimizar en tanto pudiesen dar lugar a desviaciones inesperadas o a consecuencias inhumanas. De la misma forma que se realizan pruebas de colisiones en el transporte se deben aprobar pruebas éticas y de seguridad para evaluar el impacto social o el prejuicio racial como prerrequisitos para lanzar un software de IA.

Firmatarios

Werner Arber, Antonio M. Battro, Olaf Blanke, Patricia Churchland, Stanislas Dehaene, John Donoghue, Demis Hassabis, Stephen W. Hawking, Yann LeCun, Pierre Léna, Laurie Ann Paul, Alexandre Pouget, H.E. Msgr. Marcelo Sánchez Sorondo, Laura Schulz, Mariano Sigman, Wolf J. Singer, Elizabeth Spelke, Josh Tenenbaum, Manuela Veloso, Cédric Villani.