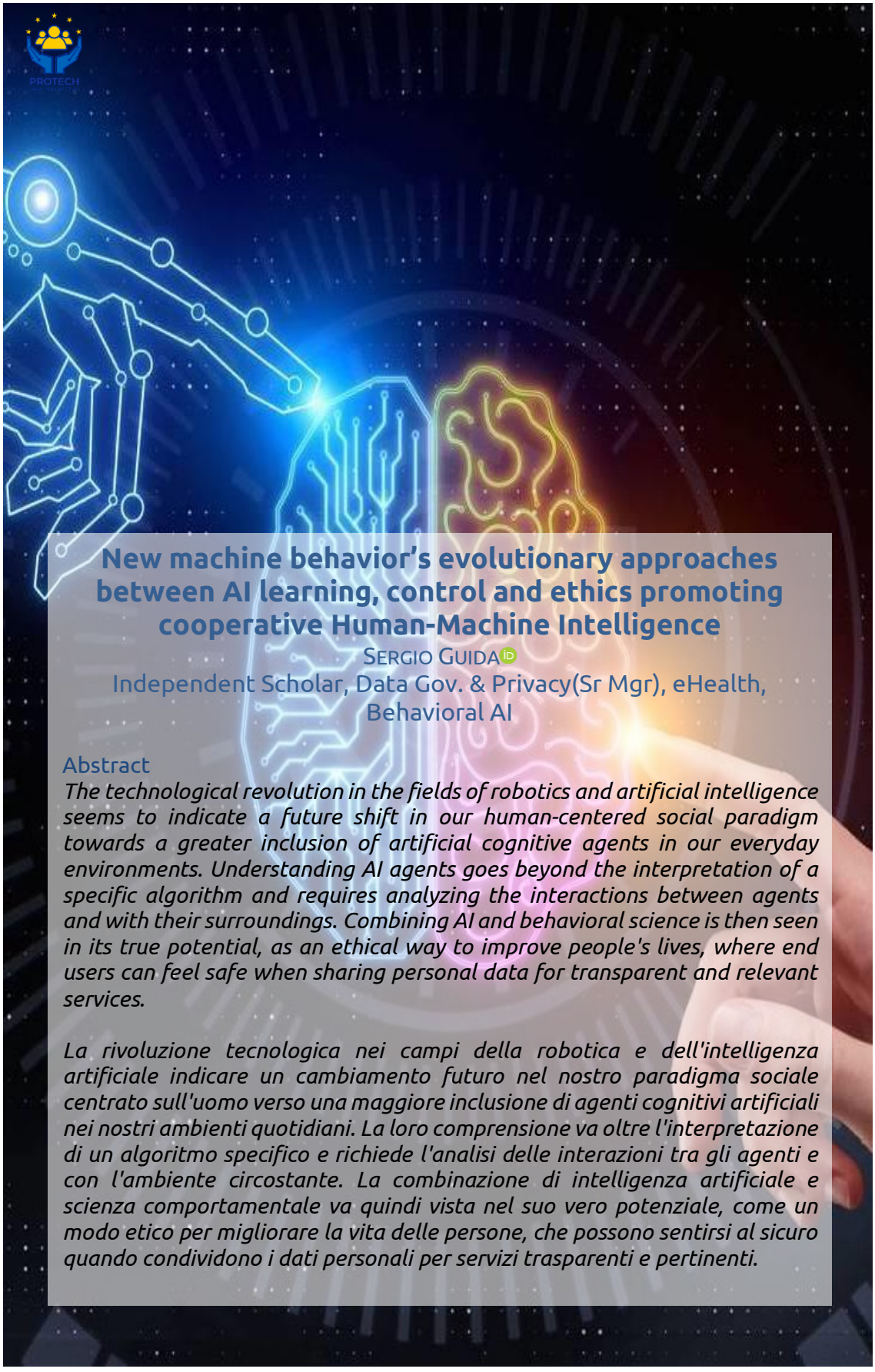
Guida S, 'New machine behavior's evolutionary approaches between AI learning, control and ethics promoting cooperative Human-Machine Intelligence' (2022) 1 EJPLT, 315-335.
DOI: https://doi.org/10.57230/EJPLT221SG

EJPLT
ISSN 2704-8012



New machine behavior's evolutionary approaches between AI learning, control and ethics promoting cooperative Human-Machine Intelligence

SERGIO GUIDA
Independent Scholar, Data Gov. & Privacy(Sr Mgr), eHealth, Behavioral AI

Abstract

The technological revolution in the fields of robotics and artificial intelligence seems to indicate a future shift in our human-centered social paradigm towards a greater inclusion of artificial cognitive agents in our everyday environments. Understanding AI agents goes beyond the interpretation of a specific algorithm and requires analyzing the interactions between agents and with their surroundings. Combining AI and behavioral science is then seen in its true potential, as an ethical way to improve people's lives, where end users can feel safe when sharing personal data for transparent and relevant services.

La rivoluzione tecnologica nei campi della robotica e dell'intelligenza artificiale indicare un cambiamento futuro nel nostro paradigma sociale centrato sull'uomo verso una maggiore inclusione di agenti cognitivi artificiali nei nostri ambienti quotidiani. La loro comprensione va oltre l'interpretazione di un algoritmo specifico e richiede l'analisi delle interazioni tra gli agenti e con l'ambiente circostante. La combinazione di intelligenza artificiale e scienza comportamentale va quindi vista nel suo vero potenziale, come un modo etico per migliorare la vita delle persone, che possono sentirsi al sicuro quando condividono i dati personali per servizi trasparenti e pertinenti.

© The author(s) 2022, published by Suor Orsola Benincasa Università Editrice. This contribution is licensed under a Creative Common Attribution 4.0 International Licence CC-BY-NC-ND, all the details on the license are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Keywords: Artificial agents, evolutionary robotics, machine behavior, hybrid intelligence, saliency maps, adaptive robots, neuroevolution, collaborative behavior.

Summary: 1. Introduction. – 2. The behavior of the machines. – 3. New evolutionary approaches. – 4. The most recent developments. – 5. Conclusions.

1. Introduction.

The technological revolution taking place in the fields of robotics and artificial intelligence seems to indicate a future shift in our human-centered social paradigm towards a greater inclusion of artificial cognitive agents in our everyday environments. Collaborative scenarios between humans and robots will become more frequent and have a deeper impact on everyday life.[1]

In one of the best examples of the evolutionary robotics' potential, the authors[2] used physical robots equipped with motors and sensors to conduct various evolutionary models and fitness goals. They concluded that 'these examples of experimental evolution with robots verify the power of evolution by mutation, recombination and natural selection. In all cases, the robots initially exhibited completely uncoordinated behavior because their genomes had random values'.[3] In summary, the study concluded that 'a few hundred generations of random mutations and selective reproduction were sufficient to promote the evolution of efficient behaviors in a wide range of environmental conditions'.[4]

This kind of evolution is also illustrated by Alphabet's recent release of over 100 daily robot prototypes in Google's offices, which is still under development.[5]

[1] Cf. S Vinanzi, M Patacchiola, A Chella, A Cangelosi, 'Would a robot trust you? Developmental robotics model of trust and theory of mind' (2019) Phil. Trans. R. Soc. B 374: 20180032, in https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2018.0032,1.

[2] Cf. D Floreano, L Keller, 'Evolution of Adaptive Behavior in Robots by Means of Darwinian Selection' (2010) PLoS Biol 8 (1): e1000292, in https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000292.

[3] *Ibid.*

[4] *Ibid.*

[5] 'Google's parent company Alphabet announced today that its *Everyday Robots Project* — a team within its experimental X labs dedicated to creating 'a general-purpose learning robot' — has moved some of its prototype machines out of the lab and into Google's Bay Area campuses to carry out some light custodial tasks. 'We are now operating a fleet of more than 100 robot prototypes that are autonomously performing a range of useful tasks around our offices' said Everyday Robot's chief robot officer HP Brøndmo in a blog post (https: // x. company / blog / posts / everyday-robots). (..) The big promise that's being made by the company is that machine learning will finally enable robots to operate in 'unstructured' environments like homes and offices. (..) Think about it: you may have seen robots from Boston Dynamics performing backflips and dancing to The Rolling Stones, but have you ever seen one take out the trash? It's because getting a machine to manipulate never-before-seen objects in a novel setting (something humans do every day) is extremely difficult. This is the problem Alphabet wants to solve', in J Vincent, 'Alphabet is putting its prototype robots to work cleaning up around Google's offices' (2021) The Verge, Nov 19, in https://www.theverge.com/2021/11/19/ 22791267/alphabet-google -everyday-robot-project-cleaning-office-prototype accessed 4 February 2022.

In truth, evolutionary robotics is not a new concept[6], but the tools needed to put the concept into action have become available with technological innovations that have grown exponentially in recent years.

For the first time in modern history, we have all the building blocks needed to facilitate evolutionary robotics: rapid prototyping and physical reproduction using 3D printing, neural networks for learning and training, improved battery life, cheaper materials, and much more other.

NASA has already used artificial evolution to develop antennas for satellites[7], for example. For their part, scientists from the University of Vermont and Tufts University in 2020 unveiled '*xenobots'*, which are 'small biological machines first designed in computer simulations using evolutionary robotics techniques'[8]. These self-healing biological machines were built using frog stem cells and have shown the ability to move and push payloads; one purpose is that these *nanobots* could one day be used to deliver drugs after being injected into the bloodstream.

Indeed, evolutionary robotics is the only way to create robots capable of complex and autonomous interactions in the real world. The benefits of such robots are too long to list, but use cases can range from robotic firefighters and search and rescue robots to nuclear waste cleaning robots, home care robots and more (see *infra*).

We may also gain a better understanding of organic evolution. A more nuanced understanding of evolution could have such broad applications that it is difficult to understand. We could gain incredible insights into the best ways to cure disease and build immunity, improve our lifespan, reduce our impact on the ecological world, and otherwise gain a better understanding of our future on this planet.

Recognizing the growing need to effectively identify and manage artificial intelligence risks, COSO published '*Realize the Full Potential of Artificial Intelligence'*[9] : this new guide exploits the principles of Enterprise Risk Management(ERM) and serves as a guide to help organizations align risk management with the strategy and execution of their artificial intelligence (AI) initiatives and to realize the potential of humans who collaborate with AI.[10]

In 2019, a group of 23 researchers from MIT's Media Lab announced on their blog the urgent need for a new scientific discipline to study the behavior of

---

[6] 'The first pioneering attempt to design adaptive robots were carried by cyberneticists in the 50s, before the digital computers era and before the development of the deliberative approach. The first systematic studies, instead, were carried in the 90s after the development of evolutionary, reinforcement learning, and regression learning algorithms' in S Nolfi, *Behavioral and Cognitive Robotics: An Adaptive Perspective*. (Institute of Cognitive Sciences and Technologies, National Research Council 2021, ISBN 9791220082372) inhttps://bacrobotics.com/Behavioral%20and%20Cognitive%20 Robotics %20An%20Adaptive%20Perspective%20 - %20Stefano%20Nolfi.pdf,15.

[7] See GS Hornby and Al Globus, 'Automated Antenna Design with Evolutionary Algorithms', American Institute of Aeronautics and Astronautics, in https://ti.arc.nasa.gov/m/pub-archive/1244h/ 1244 %20%28Hornby%29.pdf accessed 4 February 2022.

[8] See J Brown, 'Scientists Create the Next Generation of Living Robots', March 31, 2021, in https://www.uvm.edu/news/story/ scientists-create-next-generation-living-robots accessed 4 February 2022.

[9] Committee of Sponsoring Organizations of the Treadway Commission (COSO), 'Realize the Full Potential of Artificial Intelligence', 9/15/2021, in https://www.coso.org/Documents/COSO-News-Release-Realize-the-Full-Potential-of-Artificial-Intelligence.pdfaccessed 4 February 2022.

[10] As reported in the Press release, *ibid*.

machines. Iyad Rahwan, Director and Principal Investigator of the Media Lab's Scalable Cooperation Group, wrote: 'We are seeing the rise of *agency* [11]machines, machines that are actors who make decisions and take action on their own. This requires a new field of scientific studies that look at them not only as products of engineering and computer science, but also as a new class of actors with their own behavioral models and ecology'.[12]

The day after, the group published the seminal article 'Machine Behavior'[13], which describes a broad scientific research agenda to study the behavior of machines in order to integrate computer science and the sciences that study the behavior of biological agents. The researchers point out that just as animal and human behavior cannot be understood without the context in which they occur, the behavior of machines also requires a coordinated study of algorithms and the social environments in which they frequently occur.

Moreover, the current prevalence of different algorithms in society is unprecedented (Fig. 1).

[11] 'In very general terms, an agent is a being with the capacity to act, and *agency* denotes the exercise or manifestation of this capacity. The philosophy of action provides us with a standard conception and a standard theory of action. The former construes action in terms of intentionality, the latter explains the intentionality of action in terms of causation by the agent's mental states and events. From this, we obtain a standard conception and a standard theory of agency. (.). Examples of human action include instances of habitual actions, such as actions taken while driving a car, and instances in which the agent is engaged in a reactive flow of interaction, such as in jazz improvisation or verbal exchanges. Examples from robotics include skills such as coordination of limb movements and the ability to navigate through new environments', in M Schlosser, 'Agency', *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), https://plato.stanford.edu/archives/win2019/entries/agency/ accessed 8 March 2022.

[12] See M Castelluccio, 'Understanding machine behavior', May 8, 2019, in https://sfmagazine.com/technotes/may-2019-understanding-machine-behavior/ accessed 5 February 2022.

[13] I Rahwan, M Cebrian, N Obradovich et al., 'Machine behavior' (2019) Nature 568, 477–486 inhttps://www.nature.com/articles/s41586-019-1138-y.pdf .

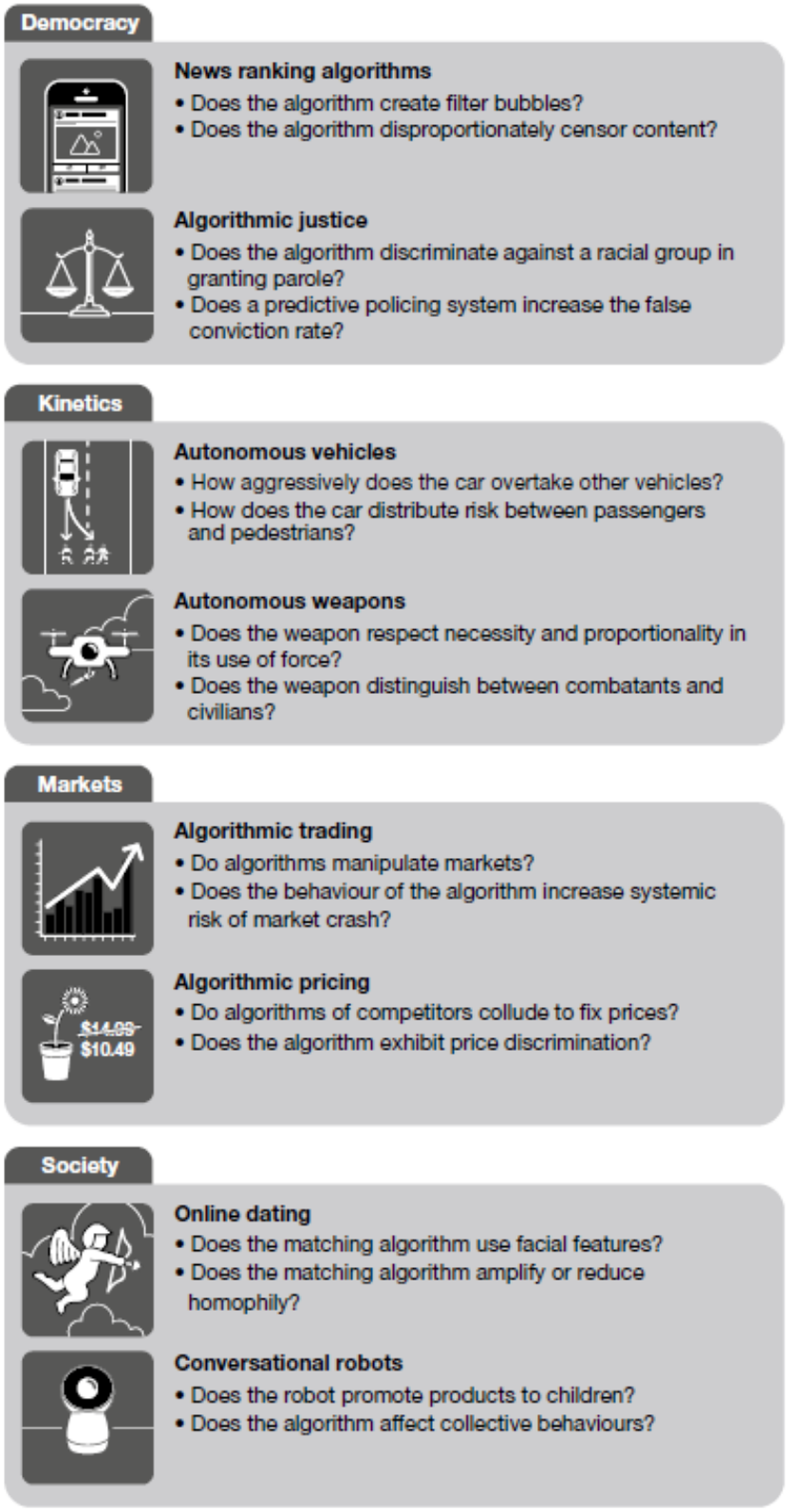


Figure 1 - Issues affecting the behavior of machines span a wide variety of scientific disciplines and topics. (Source: 'Machine Behavior', MIT Media Lab, cit.).

2. The behavior of the machines.

Understanding the behavior of AI agents is clearly one of the crucial challenges of the next decade of AI. *Interpretability* or *explainability* are some of the terms often used to describe methods that provide insights into the behavior of AI programs, although most have focused on exploring the internal structure of deep neural networks.

Recently, a group of artificial intelligence researchers from the Massachusetts Institute of Technology (MIT) are exploring a radical approach that attempts to explain the behavior of artificial intelligence by observing them in the same way we study human or animal behavior. They group ideas in this area under the captivating name of *machine behavior* which promises to be one of the most interesting fields of AI in the coming years.

The ideas behind the machine's behavior may be transformative, but its principles are relatively simple. The behavior of machines relies more on observations than engineering knowledge to understand the behavior of artificial intelligence agents. Think about how we observe and draw conclusions from the behavior of animals in a natural environment: most of the conclusions we get from observations are not related to our knowledge of biology, but rather to our understanding of social interactions. Understanding AI agents goes beyond the interpretation of a specific algorithm and requires analyzing the interactions between agents and with their surroundings. To achieve this, *behavioral analysis*[14] using simple observations can be a powerful tool.

So *machine behavior* is a field that leverages behavioral sciences to understand the behavior of artificial intelligence agents: experimental methodology, population-based statistics and sampling paradigms, observational causal inference, neuroscience, collective behavior, or social theory.

From this point of view, the behavior of machines places itself at the intersection of computer science and engineering and behavioral sciences in order to obtain a *holistic* (see also *infra*) understanding[15] of the behavior of artificial intelligence agents (fig. 2).

---

[14] Behavioral analysis 'is a natural science that seeks to understand the behavior of individuals. That is, behavior analysts study how biological, pharmacological, and experiential factors influence the behavior of humans and nonhuman animals. Recognizing that behavior is something that individuals do, behavior analysts place special emphasis on studying factors that reliably influence the behavior of individuals, an emphasis that works well when the goal is to acquire adaptive behavior or ameliorate problem behavior', in The Association for Behavior Analysis International, 'What Is Behavior Analysis?' in https://www.abainternational.org/about-us/ behavior-analysis.aspx, Copyright 2022 accessed 10 February 2022.

[15] "In terms of psychology, the holistic view suggests that it is important to view the mind as a unit, rather than trying to break it down into its individual parts. Each individual part plays its own important role, but it also works within an integrated system. Essentially, holism suggests that people are more than simply the sum of their parts. In order to understand how people think, the holistic perspective stresses that you need to do more than simply focus on how each individual component functions in isolation. Instead, psychologists who take this approach believe that it is more important to look at how all the parts work together. (..) One key phrase that summarizes the key idea behind the holistic approach is that 'the whole is more than the sum of its parts.' The field of holistic medicine, for example, focuses on treating all aspects of a person's health including physical symptoms, psychological factors, and societal influences. One reason why it is so important to consider the entire being is that the whole may possess emergent properties. These are qualities or characteristics that are present in the whole but cannot be observed by looking at the individual pieces", as we read in K Cherry, A Morin, 'What Is Holism? How psychologists use holism to

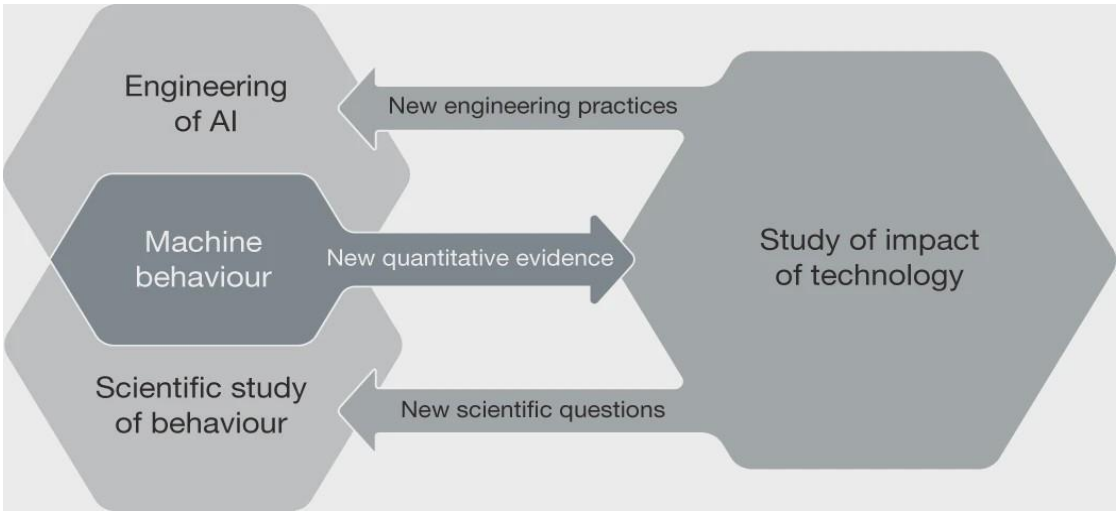


Figure 2 - The interdisciplinary nature of the machine behavior. The behavior of machines lies at the intersection between the fields that design and engineer artificial intelligence systems and the fields that traditionally use scientific methods to study the behavior of biological agents. Insights from machine behavior studies provide quantitative evidence that can help inform those fields studying the potential effects of technology on social and technological systems. In turn, these fields can provide useful engineering practices and scientific questions to fields examining the behaviors of machines. Finally, the scientific study of behavior helps AI scholars make more accurate claims about what AI systems can and cannot do. (Source: 'Machine Behavior', MIT Media Lab, cit.).

As AI agents become more sophisticated, analyzing their behavior will be a combination of understanding their internal architecture and their interaction with other agents and their environment. While the former will be a function of deep learning optimization techniques, the latter will rely in part on behavioral sciences.

Ethology is the field of biology that focuses on the study of animal behavior under natural conditions and as a result of evolutionary traits. One of the fathers of ethology was Nikolaas Tinbergen, who in 1973 won the Nobel Prize in Physiology or Medicine for his work identifying key dimensions of animal behavior. Tinbergen's thesis was that there were four complementary dimensions to understanding animal and human behavior: function, mechanism, development, and evolutionary history.

Despite the fundamental differences between artificial intelligence and animals, machine behavior borrows some of Tinbergen's ideas to delineate major behavior blocks in AI agents. Machines have mechanisms that produce behavior, undergo development that integrates environmental information into behavior, produce functional consequences that cause specific machines to become more or less common in specific environments, and embody evolutionary histories through which past environments and human decisions continue to influence the behavior of machines. An adaptation of the Tinbergen framework to the behavior of the machine can be seen in the following figure:

understand behavior', updated on April 19, 2020, in https://www.verywellmind.com/what-is-holism-4685432 accessed 10 February 2022.

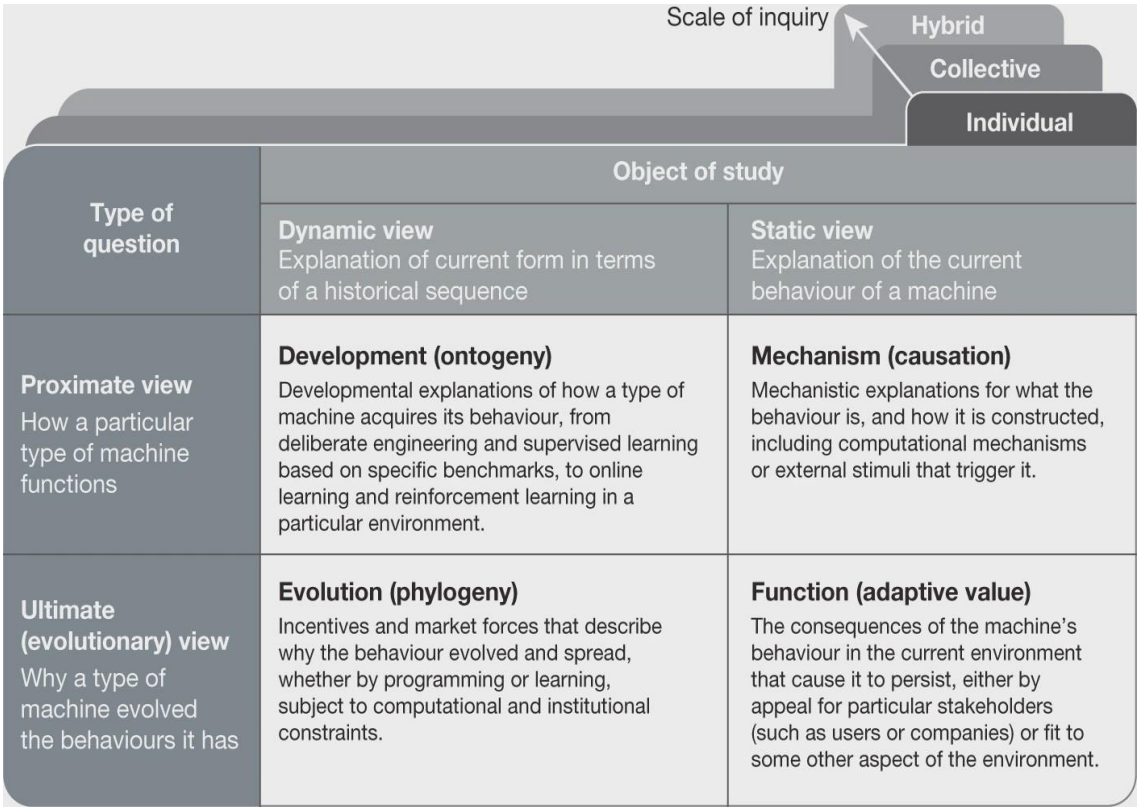


Figure 3 - The four categories proposed by Tinbergen for the study of animal behavior can be adapted to the study of the behavior of machines. Tinbergen's framework proposes two types of question, as well as two points of view of these questions. Each question can be examined on three scales of investigation: individual machines, collective machines and hybrid human-machine systems. (Source: 'Machine Behavior', MIT Media Lab, cit.).

In this framework, the study of the behavior of the machine focuses on four fundamental areas:

1. **Mechanism**: the mechanisms for generating the behavior of AI agents are based on its algorithms and on the characteristics of the execution environment. At its most basic level, machine behavior exploits interpretability techniques to understand the specific mechanisms underlying a given behavioral model.
2. **Development**: The behavior of AI agents is not something that happens in one fell swoop, but rather evolves over time. *Machine behavior* studies how machines acquire (develop) a specific individual or collective behavior. Behavioral development could be the result of engineering choices as well as the agent's experiences.
3. **Function**: An interesting aspect of behavioral analysis is understanding how a specific behavior affects the life function of an AI agent and how those functions can be copied or optimized on other AI agents.
4. **Evolution**: In addition to functions, AI agents are also vulnerable to evolutionary history and interactions with other agents. In the course of its evolution, aspects of the algorithms of artificial intelligence agents are reused in new contexts, both by constraining future behavior and by making further innovations possible. From this point of view, the behavior of the machine also studies the evolutionary aspects of artificial intelligence agents.

The four aspects provide a *holistic* model for understanding the behavior of AI agents.

However, these four elements do not apply equally when we are evaluating a single agent classification model versus a self-driving car environment with hundreds of vehicles.In this sense, the behavior of the machine applies the four previous aspects on three different scales (fig. 4):

1) **Individual Machine Behavior**: there are two general approaches to studying the behavior of individual machines. The first focuses on profiling the set of behaviors of any specific machine agent using an *in-machine* approach, comparing the behavior of a particular machine under different conditions. The second, an *inter-machine* approach, examines how a variety of individual machine agents behave in the same condition.

2) **Collective behavior of the machine**: unlike the individual dimension, this area seeks to understand the behavior of AI agents by studying the interactions in a group. The collective dimension attempts to identify behaviors that do not emerge at the individual level.

3) Hybrid Human-Machine Behavior: there are many scenarios where the behavioral patterns in AI agents are triggered by interacting with humans.

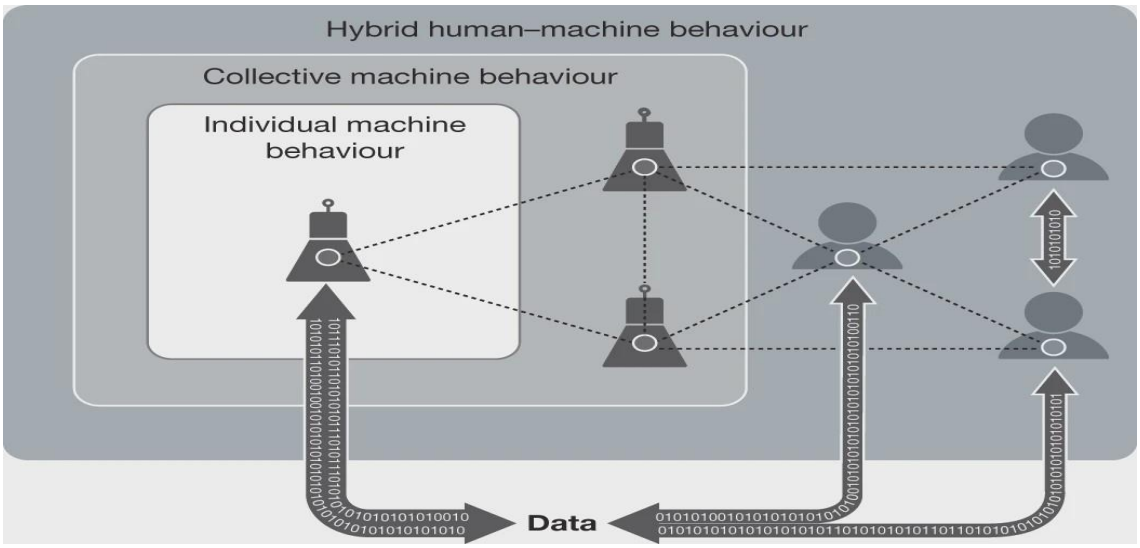


Figure 4 - Scale of investigation in the ecosystem of the machine behavior. Artificial intelligence systems represent the amalgamation of human beings, data and algorithms. Each of these domains influences the other in ways both well understood and unknown. The data, filtered through man-made algorithms, influence the individual and collective behavior of machines. AI systems are trained on data, which in turn affects how humans generate data. AI intelligence systems interact and influence each other. Human interactions can be altered by the introduction of these systems. (Source: 'Machine Behavior', MIT Media Lab, cit.).

Indeed, as interactions between humans and artificial intelligence become more sophisticated, the behavior of machines plays a critical role in enabling

the next level of 'hybrid' intelligence. This is why machine behavior is one of the most intriguing and nascent fields of behavioral sciences.

3. New evolutionary approaches.

Interpretability remains one of the greatest challenges of modern AI applications. Recent advances in computational models and deep learning research have enabled the creation of highly sophisticated models that can include thousands of hidden layers and tens of millions of neurons. While it is relatively simple to create incredibly advanced deep neural network models, understanding how such models create and use knowledge remains a challenge.

Recently, researchers from the Google Brain team published a paper proposing a new method called *Concept Activation Vectors* (CAVs)[16] that opens a new perspective to the interpretability of AI models.[17]

In the current generation of technologies the friction between *interpretability* and *accuracy* is the friction between being able to perform complex knowledge tasks and understanding how those tasks were performed: in essence, knowledge vs. control, performance vs. responsibility, efficiency vs. simplicity, all these antinomies can be explained by balancing the trade-offs between accuracy and interpretability.

Are you interested in getting the best results or are you interested in understanding how those results were produced? This is a question data scientists must answer in any deep learning scenario. Many techniques are complex in nature and, although they are very accurate in many scenarios, they can become incredibly difficult to get interpreted (Fig. 5).

[16] 'The interpretation of deep learning models is a challenge due to their size, complexity, and often opaque internal state. In addition, many systems, such as image classifiers, operate on low-level features rather than high-level concepts. To address these challenges, we introduce Concept Activation Vectors (CAVs), which provide an interpretation of a neural net's internal state in terms of human-friendly concepts. The key idea is to view the high-dimensional internal state of a neural net as an aid, not an obstacle. We show how to use CAVs as part of a technique, *Testing with CAVs* (TCAV), that uses directional derivatives to quantify the degree to which a user-defined concept is important to a classification result–for example, how sensitive a prediction of zebra is to the presence of stripes. Using the domain of image classification as a testing ground, we describe how CAVs may be used to explore hypotheses and generate insights for a standard image classification network as well as a medical application' in B Kim, M Wattenberg, J Gilmer, C Cai, J Wexler, F Viegas, 'Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)' in International conference on machine learning (2018) PMLR 2668-2677, in https://arxiv.org/pdf/1711.11279.pdf,1.

[17] See J Rodriguez, 'This New Google Technique Help Us Understand How Neural Networks are Thinking', 24 Jul 2019, in https://www.kdnuggets.com/2019/07/google-technique-understand-neural-networks-thinking .html accessed 15 February 2022.

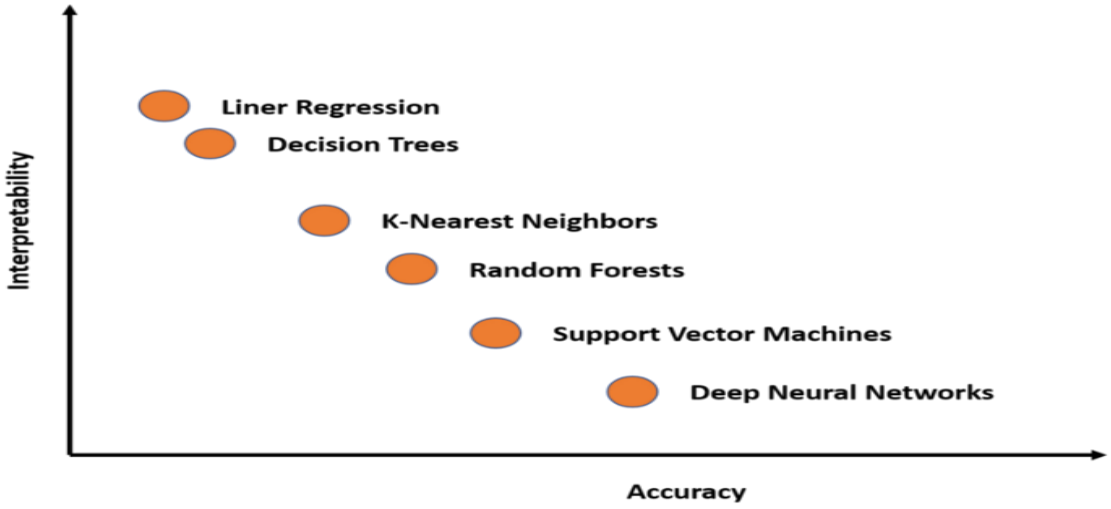


Figure 5 - The main deep learning models in a graph correlating accuracy and interpretability. (Source: 'This New Google Technique Help Us Understand How Neural Networks are Thinking Machine Behavior', cit.).

Furthermore, interpretability in such models is not a single concept and can be seen on multiple levels (fig. 6).

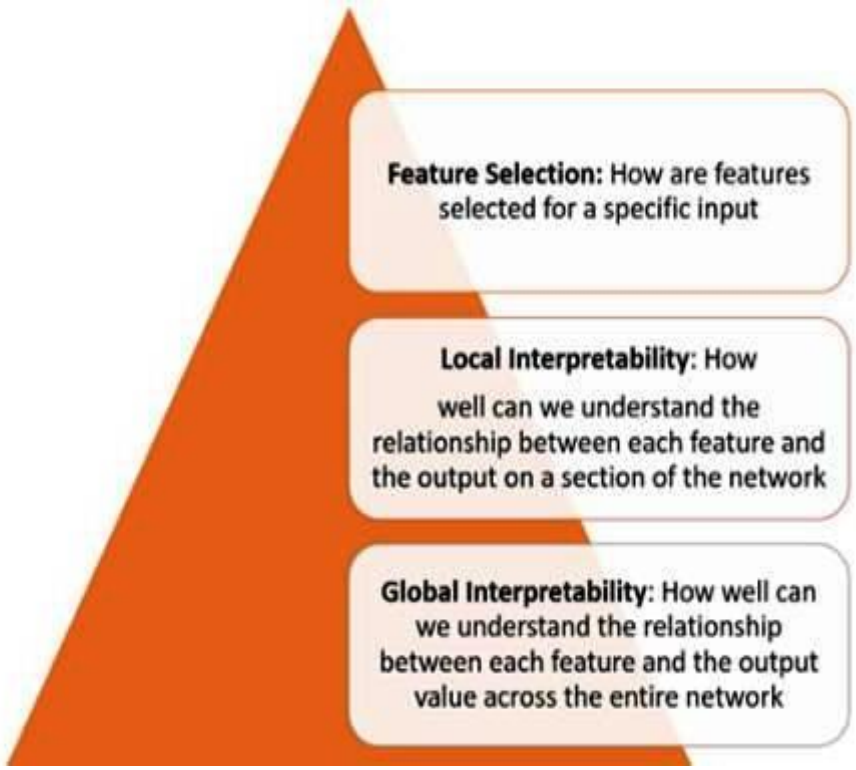


Figure 6 - The levels of interpretability in deep learning models. (Source: 'This New Google Technique Help Us Understand How Neural Networks are Thinking Machine Behavior', cit.).

Achieving interpretability through each of the levels defined in the previous figure requires several fundamental building blocks[18]:

- ✓ Understanding What Hidden Layers Do: most of the knowledge in a DL modelis formed in the hidden layers, so understanding the functionality of the different hidden layers at the macro level is essential to be able to interpret the model.
- ✓ Understanding How Nodes Are Activated: the key to interpretability is not understanding the functionality of individual neurons in a network, but rather groups of interconnected neurons firing together in the same spatial location. Segmentation of a network by groups of interconnected neurons will provide a simpler level of abstraction to understand its functionality.
- ✓ Understanding How Concepts Are Formed: understanding how the deep neural network forms individual concepts that can then be assembled into the final output is another key element.

These principles are the theoretical foundation behind Google's new CAV technique, based on *saliency maps*[19], to measure the relevance of a concept in the model's outputs (fig. 7).

[18] Cf. C Olah, et al., 'The Building Blocks of Interpretability', Distill, 2018.Doi: 10.23915/distill.00010, in https://distill.pub/2018/building-blocks/ accessed 15 February 2022.

[19]Deep learningsalience maps were first seen in the article 'Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps' (https://arxiv.org/pdf/1312.6034.pdf). The paper presented by researchers from the *Visual Geometry Group* at the University of Oxford highlighted visualization techniques for calculating images, including salience maps. (..) Saliency refers to unique characteristics (pixels, resolution, etc.) of the image in the context of visual processing. These unique features describe the visually alluring positions in an image. The salience map is a topographical representation of this. These maps, first proposed by neuroscientists Itti and Koch in their study on image feature extraction (*'A saliency-based search mechanism for overt and covert shifts of visual attention'* (2000) Vision Research 40 1489-1506, in http://ilab.usc.edu/publications/doc/ Itti_Koch00vr.pdf), give a detailed description: 'The purpose of the saliency map is to represent the conspicuity— or 'saliency'—at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map, modelled as a dynamical neural network. Saliency maps process images to differentiate visual features in images. For example, coloured images are converted to black-and-white images in order toanalyse the strongest colours present in them. Other instances would be using infrared to detect temperature (red colour is hot and blue is cold) and night vision to detect light sources (green is bright and black is dark). (..) The aggressive developments in saliency detection have almost achieved a human-like precision when it comes to recognising features. Be it with respect to datasets, learning models or with performance, saliency maps is the next big thing for computer vision and image processing projects.", as reported in A Sharma, 'What Are Saliency Maps In Deep Learning?',July 11, 2018, in https://analyticsindiamag.com/what-are-saliency-maps-in-deep-learning/ accessed 16 February 2022.

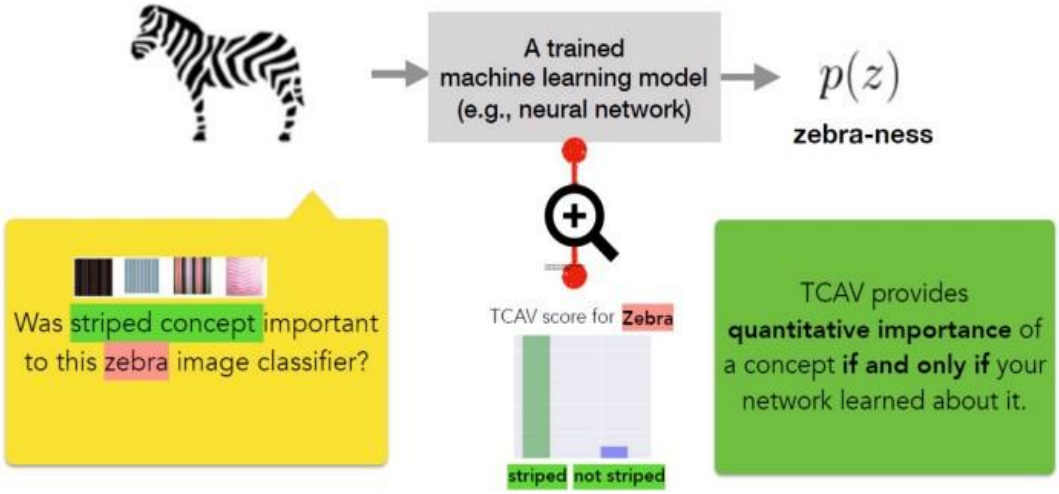


Figure 7 - TCAV measures the relevance of a concept in the model's outputs (Source: 'This New Google Technique Help Us Understand How Neural Networks are Thinking Machine Behavior', cit.).

A CAV for a concept is simply a vector in the direction of the values (e.g. activations) of the set of examples of that concept. In its article, Google's research team outlines a new linear interpretability method called '*Testing with CAV*' (TCAV) that uses directional derivatives to quantify the sensitivity of the model's prediction to a learned, high-level underlying concept from a CAV.

The methodology is defined in three basic steps (fig. 8):

1) Define concepts relevant to a model.
TCAV achieves this by simply choosing a set of examples that represent this concept or find an independent dataset with the concept labeled. CAVs are learned by training a linear classifier to distinguish between the activations produced by examples of a concept and by examples at any level.
2) Understand the sensitivity of the forecast to those concepts.
The second step is to generate a TCAV score that quantifies the sensitivity of the predictions to a specific concept. TCAV achieves this by using directional derivatives that measure the sensitivity of ML predictions to changes in inputs towards the direction of a concept, at the level of neural activation.
3) Extract an overall quantitative explanation of the relative importance of each concept for each prediction class of the model.

The final step seeks to assess the overall relevance of learned CAVs to avoid relying on irrelevant CAVs. To address this challenge, TCAV introduces a statistical significance test that evaluates a CAV against a random number of training sessions (typically 500), because a meaningful concept should lead to TCAV scores that behave consistently throughout training sessions.

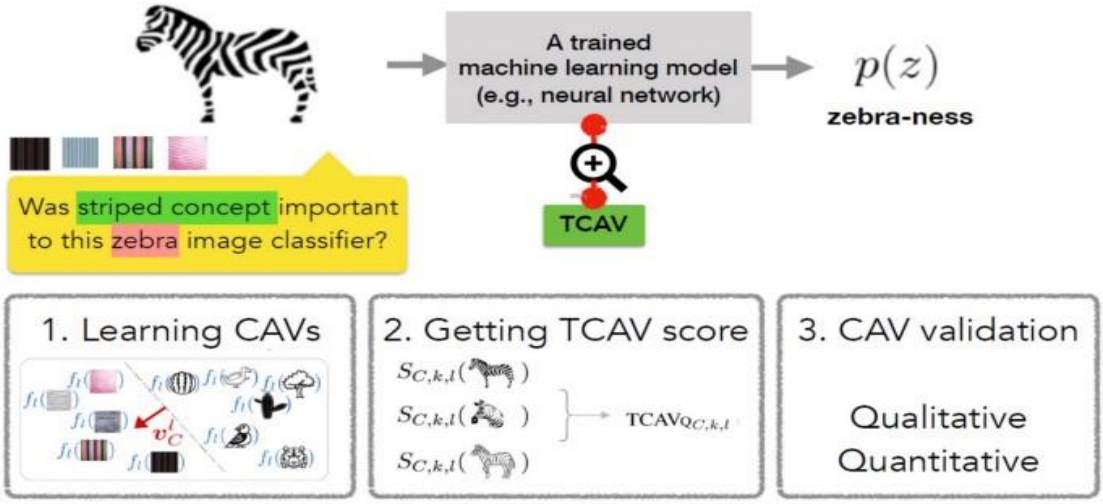


Figure 8 - The three fundamental steps of the TCAV methodology (Source: 'This New Google Technique Help Us Understand How Neural Networks are Thinking Machine Behavior', cit.).

The Google Brain team conducted several experiments to evaluate the efficiency of TCAV compared to other methods of interpretability. In one of the most extraordinary tests, the team used a salience map that attempts to predict the relevance of a caption or image to understand the concept of a taxi (fig.9).

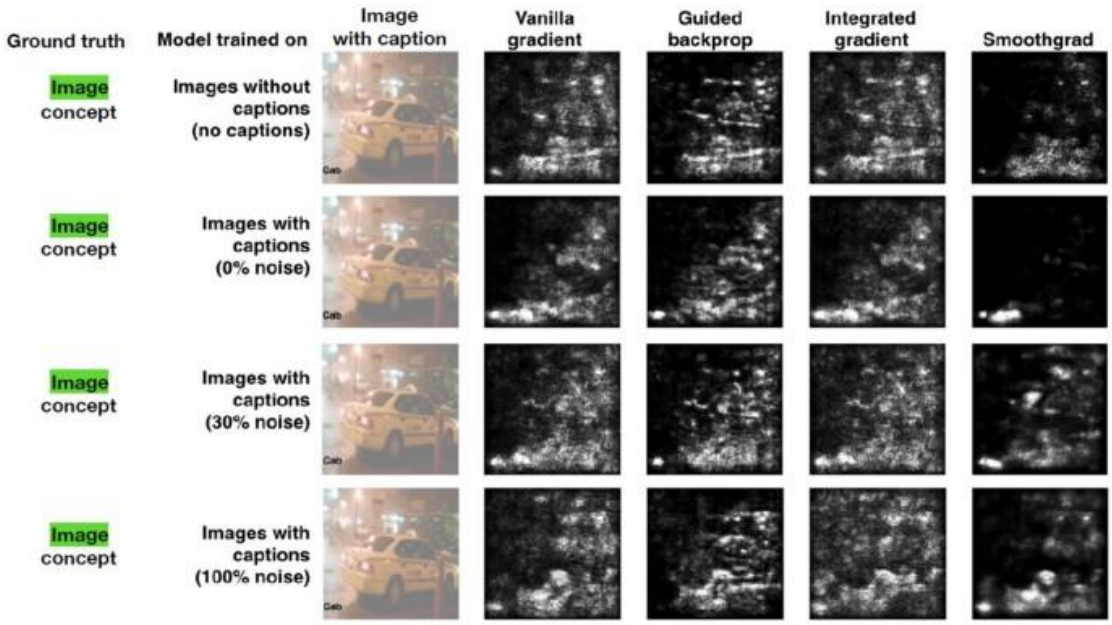


Figure 9 - Results of the salience map with approximated truth: models trained on datasets with different noise parameters p (rows) and different methods of salience map (columns) are presented. (Source: 'Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)', cit.).

The fundamental truth of the experiment was that the concept of an image was more relevant than the concept of a caption. However, when examining salience maps, humans perceived the concept of caption as more important (model with 0% noise) or did not notice any differences (model with 100% noise). On the contrary, the TCAV results correctly show that the concept of image was more important.

TCAV is one of the most innovative evolutionary approaches of neural networks in recent years and open to interesting developments: it is a step towards the creation of a human-scale linear interpretation of the internal state of a deep learning model, in a way that model decision questions can be solved in terms of high-level natural concepts.[20]

4. The most recent developments.

The discovery of expressive features constitutes a critical pre-requisite for the development of effective control policies. In the case of robots, *feature learning* can be especially challenging as it involves the generation of features describing the state of the agent and of the environment from raw and noisy sensor measurements that provide only part of the required information.[21]

Recent works in evolutionary (Salimans et al., 2017; Pagliuca et al., 2020) and reinforcement learning methods (Duan et al., 2016; Andrychowicz et al., 2019) demonstrated how *adaptive robots*[22] can successfully learn effective policies directly from observation on the basis of reward signals that only rate the performance of the agent. In other words, these methods can discover the features required to realize effective control policies automatically without the need of additional dedicated mechanisms or processes. These approaches are usually referred to as *end-to-end learning*.[23]

This is a type of deep learning process where all parameters are trained jointly, rather than step by step.[24] It is particularly prevalent in the self-driving car industry, as the benefits of this process fit neatly into the car's *convolutional neural networks* (CNNs).[25]

---

[20] See Kim, Wattenberg et al., 'Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)' (n 16) 8.

[21] Cf. N Milano, S Nolfi, 'Autonomous Learning of Features for Control: Experiments with Embodied and Situated Agents' (2020) arXiv 2009.07132, in https://arxiv.org/ftp/arxiv/papers/2009/ 2009.07132.pdf, 1.

[22] Cf. "the adaptive robot Rizon is the first applied product of the 3rd generation robotic technology. It combines high-performance force control, computer vision and advanced AI technologies, which enables the robot to adapt to the complicated environments and accomplish better 'hand-eye' coordination like human does. There are three key features which differentiate adaptive robots from others: high tolerance of position variance, great disturbance rejection and transferrable intelligence", as we read in 'Adaptive Robots', Contributed by Flexiv, 10/22/19, Robotics Tomorrow - Online Robotics Trade Magazine Industrial Automation, in https://www.roboticstomorrow.com/article/ 2019/10/adaptive-robots/14305 accessed 18 February 2022.

[23] Cf. N Milano, S Nolfi, 'Autonomous Learning of Features for Control: Experiments with Embodied and Situated Agents' (n 21) 1.

[24] Cf. Computer Science Wiki contributors, 'End-to-end learning', Computer Science Wiki, , 6 April 2018, https://computersciencewiki.org/index.php?title=End-to-end_learning&oldid=8019 accessed 18 February 2022.

[25] "CNNs are a collection of neurons that are organized in interconnected layers, with convolutional, pooling, and fully connected layers. As a mathematical construct that processes data of multiple dimensions, CNNs are designed to adaptively learn simpler patterns at lower depths while transitioning to more complicated

Since the autonomous self jointly receives multiple parameters through CNNs, it is useful to use an end-to-end learning process that can train or infer the parameters. An example could be: 'the autonomous car has to turn right towards the civil area from a highway', as there is a certain speed limit the car has to adjust its speed accordingly, while at the same time the car must also turn to the right. In this situation, end-to-end learning allows the car to perform the correct inference based on multiple reception parameters (fig. 10).

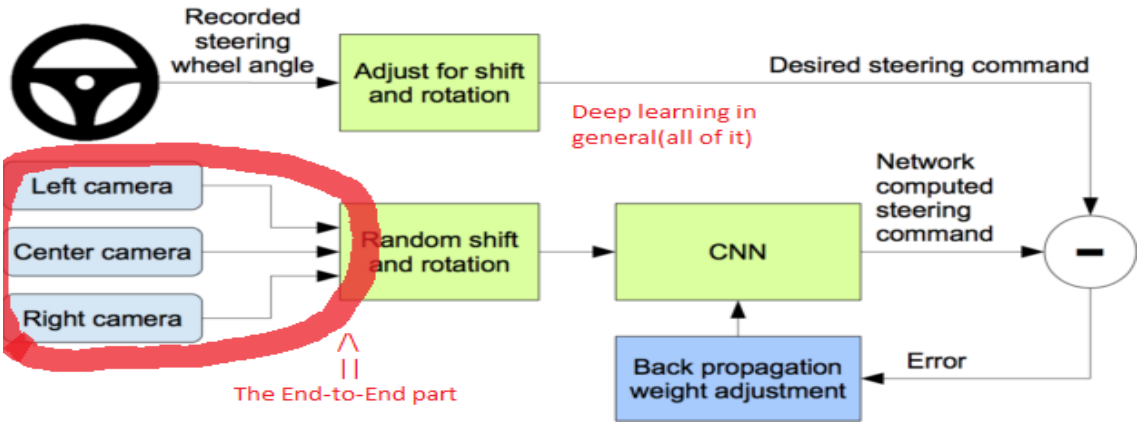


Figure 10 - The circled parameters are evaluated jointly (at the same time) within this Deep learning process, so that it too can be classified as an *end-to-end learning process* (source https://computersciencewiki.org/index. php / End-to-end_learning, cit.).

On the other hand, extensive methods that incorporate mechanisms or processes dedicated to extracting useful features can potentially accelerate learning and better adapt to complex problems. *Feature learning* is a general domain that aims to extract features that can be used to characterize data.

Recently, this area has achieved remarkable results in the context of learning neural networks for classification and regression problems. For example, feature learning with deep convolutional neural networks is an excellent solution to image classification problems. Self-learning feature learning for control is a special area of feature learning where input vectors (observations) are influenced by output vectors (actions) and where observations and actions vary throughout the process learning (Bohmer et al., 2015; Lesort et al., 2018). The term *autonomous* refers to the fact that characteristics are extracted through a self-supervised learning process, i.e. a learning process in which supervision is provided directly by the data available as input.

A first goal of feature learning is to generate a *compact* representation of high-dimensional sensory data which, in the case of robotic problems, could include 50 or more joint angles and velocities, hundreds of sensors encoding tactile information, and thousands or millions of pixels that form camera images.

patterns as we dive deeper. Deep neural networks overcome the use of exponentially large parameters by the addition of multiple hidden layers", as reported in JM Vaz, S Balaji, 'Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics' (2021) Molecular diversity, 25(3), 1569–1584, in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8342355/pdf/ 11030_2021_ Article_10225.pdf, 1570.

A second goal is to generate *useful* features, that is, more informative characteristics of the observation states from which they can be extracted.

Therefore Milano & Nolfi (2020) analyze whether features extracted through self-supervised learning methods facilitate the development of effective solutions in 'embedded and situated' robots trained through an evolutionary algorithm. They consider experimental scenarios in which robots have access to egocentric perceptual information[26]: the results demonstrate how the use of features extracted through *sequence-to-sequence learning model* [27](Srivastava et al. 2015, Sutskever et al. 2014, Cho et al. 2014) allows to obtain better solutions with respect to the *end-to-end control* condition, provided that the self-supervised learning process (which determines the extracted features) continues throughout the evolutionary process.

In summary, 'the effectiveness of adaptive methods can be enhanced by combining a control network, trained with an evolutionary or reinforcement learning algorithm, with one or more feature extraction networks trained through self-supervised learning (Lange, Riedmiller & VoigtHinder, 2012; Mattner, Lange & Riedmiller, 2012; Ha & Schmidhuber, 2018; Milano & Nolfi, 2020). While the feature extraction networks are used to extract useful features from the observations, the control network is used to map the features extracted from the previous network(s) into appropriate actions'.[28]

For its part, '*neuroevolution*' (Lehman and Miikkulaien, 2013) is a widely used method to evolve embodied and situated agents.[29] Clearly, the conditions

[26] Cf. "The very notion of a position in space requires a reference frame, and one of the primary distinctions made among possible reference frames has been between egocentric and allocentric (a.k.a. geocentric, exocentric, environment-centered) reference frames (e.g., Burgess, 2006; Feigenbaum & Rolls, 1991; Howard, 1991; McNamara, Rump, & Werner, 2003; Nardini, Burgess, Breckenridge, & Atkinson, 2006; Neggers.Van der Lubbe, Ramsey, & Postma, 2006; Wang &Spelke, 2000). As the terms suggest, egocentric reference frames code location with respect to the observer, whereas allocentric reference frames code location with respect to something external to the observer (room axes, distal cues, cardinal directions, etc.). There is substantial evidence for both egocentric and allocentric information coded in the brain from neurophysiology and neuropsychology. In different subregions of the parietal cortex, neurons respond to the stimuli in retina-centered, head-centered, and even hand-centered coordinate systems (e.g., Colby & Goldberg, 1999), supporting a system for representing space egocentrically. However, place cells in the medial temporal lobes have been shown to code location with respect to the environmental reference frame (e.g., Burgess, Jeffery, & O'Keefe, 1999)" , in N Burgess, 'Egocentric and allocentric information in spatial memory', in https://ebrary.net/155480/psychology/ egocentric allocentric spatial memory accessed 18 February 2022.

[27]Cf."Sequence to Sequence (often abbreviated to seq2seq) models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.(..) This model can be used as a solution to any sequence-based problem, especially those where inputs and outputs have different dimensions and categories. (..) The most common architecture used to create Seq2Seq models is the *Encoder-Decoder architecture*. Both the *encoder* and the *decoder* are LSTM models", in P Singh, 'A Simple Introduction to Sequence to Sequence Models', August 31, 2020, in https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models accessed 19 February 2022.

[28] Cf. S Nolfi, 'Behavioral and Cognitive Robotics: An Adaptive Perspective' (n 6) 159-160.

[29] Cf. "Much of recent machine learning has focused on deep learning, in which neural network weights are trained through variants of stochastic gradient descent. An alternative approach comes from the field of neuroevolution, which harnesses evolutionary algorithms to optimize neural networks, inspired by the fact that natural brains themselves are the products of an evolutionary process. Neuroevolution enables important capabilities that are typically unavailable to gradient-based approaches, including learning neural network building blocks (for example activation functions), hyperparameters, architectures and even the algorithms for learning themselves. Neuroevolution also differs from deep learning (and deep reinforcement learning) by maintaining a population of solutions during search, enabling extreme exploration and massive parallelization" in K Stanley, J Clune, J Lehman, R Miikkulainen, 'Designing neural

under which agents are evaluated influence the course of the evolutionary process.[30] Ideally, the environmental conditions should match the skill level of the evolving agents, that is, they should be difficult enough to elicit adequate selective pressure and simple enough to ensure that random variations can occasionally produce progress. This can be achieved by varying the evaluation conditions during the evolutionary process, i.e. by increasing the complexity of environmental conditions across generations and by selecting conditions that are challenging for current evolving agents.

There are three possible methods:

1) incremental evolution: the use of an evolutionary process divided into successive phases of increasing complexity. For example, the evolution of the ability to visually track target objects can be accomplished by exposing evolving agents first to large immobile targets, then to small immobile targets, and finally to small moving targets (Harvey, Husband & Cliff, 1994). Similarly, the evolution of agents evolved for the ability to capture fleeing prey can be organized in a series of successive phases in which the speed of the prey and the delay in pursuit are progressively increased (Gomez & Miikkulainen, 1997). However, these approaches assume that the tasks can be sorted by difficulty, when in fact they might vary along multiple difficulty axes. In general, incremental approaches can be effective but introduce hyperparameters that are difficult to tune and require the use of domain-dependent knowledge.[31]

2) Competitive co - evolution: the evolution of agents that compete with other evolving agents. For example, the co-evolution of two populations of predators and robot prey selected respectively for their ability to capture prey and escape predators (Rosin & Belew, 1997; Nolfi & Floreano, 1998; De Jong, 2005; Chong, Tiňo & Yao, 2009; Miconi, 2009; Samothrakis et. al., 2013). However, if on the one hand the exploitation of their weaknesses has an adaptive value for the opposing population, on the other it does not necessarily lead to a progressive complexification of the adaptive problem. Furthermore, it can only be applied to problems that can be formulated competitively.[32]

3) A third possible method, (Milano & Nolfi, 2021) consists in enhancing the evolutionary algorithm with a process capable of selecting the environmental conditions that have the right level of difficulty for the current evolving agents and that challenge the weaknesses of the agents in current evolution. This class of methods is referred to as *curriculum learning*. [33] Its usefulness for supervised

networks through neuroevolution' (2019) Nature Machine Intelligence. 1. 10.1038 inhttps://www.researchgate.net/publication/330203191_Designing_neural_networks_through_neuroevolution, 1.

[30] See N Milano, S Nolfi, 'Automated curriculum learning for embodied agents a neuroevolutionary approach' (2021) Sci Rep 11, 8985, in https://www.nature.com/articles/s41598-021-88464-5.pdf , 1.

[31] *Ibid* .

[32] *Ibid*, 2.

[33] Cf. "Humans need about two decades to be trained as fully functional adults of our society. That training is highly organized, based on an education system and a curriculum which introduces different concepts at different times, exploiting previously learned concepts to ease the learning of new abstractions. By choosing which examples to present and in which order to present them to the learning system, one can guide training and remarkably increase the speed at which learning can occur. This idea is routinely exploited

learning has been widely demonstrated and constitutes an active research field (Bengio et al., Cit; Sutskever & Zaremba, 2014; Graves et al., 2016 and 2017; Held et al., 2017). Wang et al. (2019) proposed an algorithm that evolves a population of increasingly complex agent-environment pairs. Progressive complexity is achieved by:

(i) optimizing agents in their associated environments,
(ii) generating new environments by creating varying copies of existing environments, and
(iii) attempting to transfer copies of current agents to another pair of environmental agents.

As demonstrated by the authors, this method allows to produce agents capable of operating effectively in remarkably complex conditions and to generate solutions for environmental conditions that are too difficult for a conventional evolutionary method. Conversely, however, this method produces tailor-made solutions for specific environmental conditions that do not necessarily generalize to other conditions.

Milano and Nolfi then propose a curricular learning method that automatically selects the environmental conditions in which the evolving agents are evaluated.[34] More specifically, it allows evolutionary algorithms to select the environmental conditions that facilitate the evolution of effective solutions. This is accomplished by adding a curricular learning component that estimates the level of difficulty of environmental conditions from the point of view of evolving agents and selects conditions with different levels of difficulty in which the frequency of difficult cases is greater than the frequency of easier cases.

The estimation of the level of difficulty of the environmental conditions is carried out on the basis of the suitability obtained in those conditions (*fitness*) by recently evaluated agents. The selection of suitable environmental conditions is achieved by selecting *h* environmental conditions from *h* corresponding subsets characterized by different levels of difficulty, where *h* is the number of evaluation episodes.

Finally, the preferential selection of difficult conditions is achieved by increasing the probability of selecting difficult environmental conditions, that is, by determining the intervals of the subsets with a power function. Using this method also reduces the stochasticity of the *fitness measure* as it ensures that agents are exposed to environmental conditions that have similar levels of difficulty.[35]

---

in animal training where it is called shaping (Skinner, 1958; Peterson, 2004; Krueger & Dayan, 2009). (..) The basic idea is to start small, learn easier aspects of the task or easier subtasks, and then gradually increase the difficulty level" in Y Bengio, J Louradour, R Collobert, J Weston, 'Curriculum Learning', in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009, in http://machinelearning.org/archive/icml2009/papers/119.pdf, 2-3 accessed 20 February 2022.

[34] SeeN Milano, S Nolfi, 'Automated curriculum learning for embodied agents a neuroevolutionary approach' (n 30) 2.

[35] *Ibid*, 12.

The curricular learning component proposed is general and can be combined with any evolutionary algorithm. Its effectiveness has been verified in combination with the 'Open-AI neurodevelopmental strategy'[36], one of the best state-of-the-art algorithms available.

5. Conclusions.

AI is described by the 'Regulation on a European approach to artificial intelligence' Commission's draft as 'a rapidly evolving family of technologies that can contribute to a wide range of economic and social benefits'.[37]

The use of AI makes it possible to infer complex and non-linear relationships from the data that the software analyzes, to achieve a specific objective by identifying and, if necessary, automatically perfecting the operations of a system.

However, if ethical challenges are not addressed sufficiently, a lack of public trust can hinder the adoption of such systems which, in turn, would lead to significant social opportunity costs through the under-use of available and well-designed technologies: the ethical challenges posed is therefore becoming a prerequisite for good governance in hi-tech societies.[38]

From an ethical perspective, this shifts the focus of ethical deliberation from specific decision-making situations to the ways in which algorithms are designed and implemented.[39]

Since the use of artificial intelligence systems proliferates, being able to explain how a certain model or system works is essential, especially for those used by governments or public sector agencies, to avoid talking about *AI black boxes*.[40]

---

[36] "The OpenAI-ES method proposed by Salimans et al. (2017) is a form of natural evolutionary strategy that estimates the gradient of the expected fitness. (..) It is a variation of the batch normalization method commonly used in supervised learning adapted to problems in which the stimuli experienced by the network are not fixed (see also Salimans et al., 2016). This is the case of embodied agents in which the stimuli that are experienced depend on the actions executed by the agents previously. The problem is solved by calculating the average and the variance of the activation of the sensors incrementally on the basis of the distribution of the activation of the sensors of agents of successive generations. This technique is particularly useful in problems in which the range of activation of the sensors vary widely during the course of the evolutionary process", as specified in P Pagliuca, N Milano, S Nolfi, 'Efficacy of Modern Neuro-Evolutionary Strategies for Continuous Control Optimization' (2020) Front. Robot. AI 7:98. doi: 10.3389/frobt.2020.00098, in https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC7805676&blobtype=pdf, 2-3.

[37] Cf. Proposal for a Regulation of the European Parliament and of the Council establishing harmonized rules on Artificial Intelligence (Law on Artificial Intelligence) and amending some legislative acts of the Union-COM / 2021/206 final, Brussels, [2021] in https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=_CELEX%3A52021PC0206 , 2.

[38] Cf. J Mökander, J Morley, M Taddeo, L Floridi, 'Ethics - Based Auditing of Automated Decision - Making Systems: Nature, Scope, and Limitations'(2021) Science and Engineering Ethics, in https://link.springer.com/content/pdf/10.1007/s11948-021-00319-4.pdf,43.

[39] *Ibid* , 44.

[40] "At its core, black boxes are algorithms that humans cannot survey, that is, they are epistemically opaque systems that no human or group of humans can closely examine in order to determine its inner states. Typically, black box algorithms do not follow well understood rules (as, for instance, a Boolean Decision Rules algorithm does), but can be 'trained' with labelled data to recognise patterns or correlations in data, and as such can classify new data. In medicine, such self-learning algorithms can fulfil several roles and purposes: they are used to detect illnesses in image materials such as X-rays, they can prioritise information or patient files and can provide recommendations for medical decision-making. (..) The reflection on the

So 'there are challenges of blending AI with behavioral science for behavior change, such as:

• Building trust with end users to gather sufficient user data.

• Designing scalable personalized and engaging interactions to retain users.

To build trust with end users, the business use case must be a win-win instead of only benefiting the company. Besides, the underlying technology must be reliable, trustworthy and ensure privacy. The emerging field of ethical AI explicitly takes into account fairness, explainability and robustness during algorithm design'.[41]

In this regard, it is worth citing here an example of artificial cognitive architecture for trust,ToM (Theory of Mind)[42], and episodic memory in an HRI (human–robot interactions) scenario that can improve the performance of artificial agents in contexts of shared objectives.

To conclude *à la* Munro[43], 'human beings have always been part of Machine Intelligence systems: they create unstructured data, load them into systems, develop statistical models, interpret the results. Adaptive learning optimizes people's time and effort by making machines smarter to make people smarter. In reality, it is a question of making the most of human intelligence'.

limitations of algorithms offers, indeed, insight into ways to improve their use. To our mind, being aware of the epistemic limitations of medical AI is a condition for entrenching responsible use and interaction with such systems. For these reasons, we believe that the debate needs to be widened by not solely focusing on the technological aspect of medical AI, but also on the interaction of humans with such technological systems" in JM Durán, KR Jongsma, 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI', Journal of Medical Ethics, 18 March 2021 in http://dx.doi.org/10.1136/medethics-2020-106820, 329, 334.

[41] Cf. B Fontaine, 'The Advent Of Trustworthy Behavioral AI', Forbes Technology Council, Jun 4, 2021 in https://www.forbes.com/sites/forbestechcouncil/2021/06/04/the-advent-of-trustworthy-behavioral-ai/ accessed 22 February 2022.

[42] See " A complementary approach to simulated social interaction is cognitive simulation, which seeks to develop artificial representations of neurocognitive mechanisms such as imitation and perception of self, simulate them in artificial agents such as humanlike robots, and assess their functioning in enabling ToM inferences in human-agent interactions (Breazeal and Scassellati, 2002; Scassellati, 2002; Michel et al., 2004). Building on simulation theory (Gallese and Goldman, 1998), cognitive simulation involves the robot establishing and maintaining representations of the mental states of its human counterparts by tracking and matching their states with resonant states of its own. These representations enable the robot to take the perspective of its human counterparts, make inferences about the human's goals, and learn from their actions", in LJ Byom, B Mutlu, 'Theory of mind: mechanisms, methods, and new directions' (2013) Frontiers in human neuroscience, 7, 413, in https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC3737477/pdf/fnhum-07 -00413.pdf , 6.

[43] Cf. R Munro, 'The fourth generation of machine learning: Adaptive learning', September, 2015, in http: //www.junglelightspeed.com / the-fourth-generation-of-machine-learning-adaptive-learning / accessed 22 February 2022.