Journal of Future Robot Life 1 (2020) 23–33
DOI 10.3233/FRL-190103
IOS Press

On the probabilistic mind of a robot

Marco Roccetti [*,**], Luca Casini and Giovanni Delnevo
University of Bologna, Italy

Abstract. In this article, we discuss what differentiates an *artificial* mind from a human one, during the process of *making a choice*. We do this without any intention to debunk famous arguments according to which there are aspects of human consciousness and expertise that cannot be simulated by artificial humanlike *cognitive* entities (like a robot, for example). We will first put in evidence that artificial minds, built on top of Deep Neural Network (DNN) technologies, are probabilistic in nature, and follow a very clear line of reasoning. Simply told, their reasoning style can be assimilated to a process that starts from a bunch of example data and learns to point to the most *likely output*, where the meaning of *likely*, here, is neither vague or fuzzy, but it obeys well-known probability theories. Nonetheless, as such, choices could be made by those intelligent entities that fail to pass human *plausibility* criteria, even if those chosen are those with high probability values. We will provide an (obvious) explanation for this apparent paradox, and we will demonstrate that an artificial mind based on a DNN can be driven to translate probabilities into choices that humans judge as plausible. Finally, we will show how an artificial probabilistic mind can be made to learn from its errors, till the point where it exhibits a cognitive behavior comparable to that of a human being.

1. INTRODUCTION

Whether one is passionate for robotics or for Artificial General Intelligence (AGI), one cannot disregard the fact that the scene is now mostly dominated by the cognitive aspects surrounding what is called the *mind* of an artificial entity, such as a robot. While fully autonomous robots and AGI are not yet here, and may not be for a while, artificial *minds*, built on the top of technologies such as Deep Neural Networks (DNN), have already been coping with various forms of intellectual challenges. Emerging as powerful tools for the continuous interaction they have with the real external world (i.e. with big data), they have begun to take high-stakes decisions that were never before imagined. But despite their successes we humans often have a hard time trusting that such kinds of artificial *minds* are doing their jobs well.

It is beyond doubt that decision making is a complex process, consisting of making a choice and then taking responsibility for the consequences of the resulting events. If what we call the artificial *mind* of a robot is just an algorithm, or a mathematical function powered by a DNN, without any resemblance to a self-aware entity, can it really be considered responsible for anything? And should the designers of such an artificial *mind* be held responsible for its functioning, or rather should it be the end-users who decide how and when to employ it? While we are aware that many such questions go beyond our expertise as computer scientists, and are best addressed by philosophers (Alaieri & Vellino, 2016; Floridi & Cowls, 2019; Hew, 2014), our modest opinion is that the situation can be described with simpler and alternative arguments.

What we discuss here is a reflection on the difference of what it means to *make a choice* for a human and for an artificial *mind*, respectively. Ignoring, just for a while, all the buzzwords and the hype

*Corresponding author. E-mail: marco.roccetti@unibo.it.
**Department of Computer Science and Engineering, Alma Mater Studiorum – University of Bologna, Italy.

behind AGI and robotics, we would like to simplify the issue and posit that modern artificial *minds*, based on DNNs, are essentially *probabilistic models*. Oversimplifying a complex mechanism, this can be reduced to a process that starts from a bunch of example data and learns to point to the most *likely* output. Following this simplistic line of reasoning, who can spot the differences between a human and an artificial *mind*? In fact, what if we replace the (big) data with human experience, and the learning algorithm with the common sense of a human being who interacts with the consequences? Aren't people really making choices based on what they *expect to happen*, based on their past experiences?

Even if the reader believes that we are being too frugal and embracing a minimalist approach to tackle a big problem, a fundamental difference between a human and an artificial *mind* can be recognized just at this precise point of our reasoning. It amounts to the fact that, despite the ubiquitous presence of *fortuity* in our lives, we everyday dispute its role and value, and how to translate it into the form of probable/non probable events in order to make a decision. Far from discussing here the adventures of *Chevalier De Mere*, *Blaise Pascal* and *Pierre de Fermat* that kick-started the study of probability, transforming it into a sound mathematical theory, disagreements on how to reason about the likelihood of a given event are conducted among human beings every day. These are disagreements that our cognitive biases often exacerbate, especially when we humans ignore evidence that runs counter to the specific hypotheses we favor (Climehaga, 2018).

And here comes the *good news* about how an artificial *mind* behaves. Please follow our argumentation carefully: given our difficulty as humans in applying *pure* probabilistic reasoning, and the fact that most of the time we are asked to make choices in the presence of imperfect knowledge and information. We could say that, at least from a probabilistic point of view, we as humans almost never develop reasoning mechanisms that lead to decisions whose motivations can be justified under the form of a *sound probabilistic model*. That is not true with artificial *minds*. And even if DNN technologies are often described as *black boxes*; in the context of a DNN-based artificial *mind*, those probabilistic factors that have led to a given decision can be quantified to a very precise extent. In other words, what we are trying to say is that the decisions taken by an artificial *mind* always have valid motivations, so long as we are given the probabilistic values that that artificial mind has computed for the different alternatives in play (Roccetti et al., 2019; Villani et al., 2018). Simply told, the meaning of *likely*, for artificial minds, like those we have discussed so far, is never vague or fuzzy, but it obeys well-known probability theories.

Nonetheless, choices could still be made by those intelligent entities that fail to pass human *plausibility* criteria, even if they are those with higher probability values. For this reason, in the remainder of this paper, after a careful review on how probabilistic models have been employed over the years to implement artificial entities with a humanlike reasoning style (Section 2), we will first illustrate what lies behind a probabilistic reasoning model based on a DNN technology, and then we will discuss how an artificial mind can be driven to translate probabilities into choices that humans judge as *plausible* (Section 3). Finally, we will show how an artificial probabilistic mind, based on a DNN, can learn from its errors, up to the point where it exhibits a cognitive behavior comparable to those of a human being (Section 4). Conclusions terminate this paper in Section 5.

2. RELATED WORK

Here we survey the main models and systems that have made use of probabilistic reasoning to emulate humanlike cognitive functions.

Quite soon after computers emerged, scientists began to imagine whether there was some way to make machine *think*, allowing them to operate autonomously, without being explicitly programmed to perform each action at any given time. Those were the glorious early days of Artificial Intelligence (Jefferson, 1949). Trying to make computers *think* was a process that started from looking at how humans do think, and then trying to transfer this process into something that a machine could compute (Pomerol, 1997). Thus, reasoning was framed as a problem of *logic*, where, given the context of the problem, the machine would decide the right action to take, or the right answer to give, following a set of rules defined as sound by the designers.

Those methods worked reasonably well, but had severe limitations in the things that could be modeled, both conceptually and technically: for example, think of the context of a problem that cannot be completely known, or that can change over time (like a rover exploring a faraway planet or a natural language to be understood where a certain word was never seen before), or even think of the situation when a problem does not admit an exact solution, being too vast (as with chess, for example) to be represented in memory and codified by means of a formal theory. (Even the computation itself could take too long to terminate or might never terminate at all). While some of these limitations could somehow be surpassed from a technical viewpoint, the conceptual limitations remain and open the path to an important question: how can computers treat uncertainty? Hence a paradigm shift was needed in order to allow computers to act in a way that resembled human-like reasoning, when uncertainty has arrived on the scene. This is the reason why we now address the topic of probabilistic reasoning.

Making short a very long history, at the beginning just two systems were designed that followed a probabilistic reasoning style, to make humanlike decisions: MYCIN (Shortliffe & Buchanan, 1975) and PROSPECTOR (Hart et al., 1978). MYCIN was in essence a rule-based system that, given some data about a patient, was able to suggest what antibiotic, and in what dosage, said patient should take. Similarly, PROSPECTOR was designed to mimic the reasoning of a geologist for estimating the likelihood of the presence of a mineral ore deposit in a given location, on the basis of certain measurements and selected observations of the ground. This kind of approach (called rule based) is what came to be known as an *expert system*.

Moving quickly along this path, after the use of rules to derive conclusion came the above idea to implement some kind of *probabilistic inference*, to derive conclusions as soon as new evidence was collected. Along this path, Bayesian statistics has been the privileged mathematical theory, with which the relationship between a hypothesis and some evidence was modeled over these latest thirty years. This use of Bayes was motivated by the reason that it allows us to compute how the posterior probability of a certain outcome changes in response to a newly acquired knowledge. Many of the probabilistic systems that were proposed employed graphical structures, as pioneered by the *Turing Prize* winner Judea Pearl, who first designed a so-called Bayesian Network (Pearl, 2014); that is, a connected graphical representation of many *events* (as nodes), linked with each other through arcs to represent the important concept of *causality*, or even simply mutual influence among those events.

Without even mentioning computational complexity issues, an obvious and unfortunate downside to that kind of system amounted to the need for a human expert to transfer his/her knowledge into a structure, representable as a Bayesian network. Sometimes this can easily be achieved, as most of the knowledge is readily available, but in many cases, with so many variables at play, this approach is almost unmanageable. We could say that this shortcoming was the main reason for the partial demise of research on such systems, and to a general decrease in interest and confidence towards similar systems and models.

Modern Deep Neural Networks (DNN) are neural networks with many layers of stacked neurons. They were already on the AI scene in the 1950s, yet the lack of: (i) efficient algorithms for training them, (ii) enough data on which to train them, and (iii) powerful machines to run them, has hindered their development until the 21st century. Modern day DNNs are finally able to overcome many of the obstacles that previous approaches could not. Essentially, they do not need to incorporate explicitly defined knowledge to reason over an argument, as they are able to learn from data in a computationally efficient way. However, one of the fiercest criticisms raised against them is that, even though they are based on probability theory, like other previous models that managed uncertainty they are considered as *opaque* in their learning process, in the sense that, while it is still possible to see what neurons led to a certain conclusion, it is difficult to provide formal mathematical descriptions that link the learning activity of those neurons to that specific conclusion. This is a very problematic issue, and motivates the abundance of recent studies along this line of research. Nonetheless, it is a fact that, without DNNs, those tasks that need the understanding and interpretation of complex structures, such as image recognition and natural language processing for example, would not be capable of automation in the same way that they are today. Examples of less opaque DNN-based processes (referred to as "Explainable AI") are now starting to appear, and there are also many efforts to reconcile traditional Bayesian thinking and neural networks so as to employ the best of both worlds (Yoon et al., 2018; Liu et al., 2016).

In the end, beyond the problem of the interpretability of a DNN, it is important to highlight the real benefit that neural networks have brought in terms of knowledge elicitation. As we mentioned above, it is often the case when there is no way to define a clear set of rules that are able to describe a real phenomenon, even if we have plenty of data generated by that phenomenon. A DNN can easily make sense of that data, while discovering features and relationships that would otherwise be drowned under the weight of that data itself. With a well-trained DNN we can derive the most *likely* decision along with its corresponding probability value. As we will elaborate further in the following sections, those probability values are what best characterizes the nature of those artificial entities.

3. A PROBABILISTIC MIND

Let us consider how an artificial mind works that follows a probabilistic method of reasoning, based on a DNN technology. We do that with a simple example, where, to simplify matters, making a choice is just binary: one can choose either to do something or not to do it.

Consider, first, the left graph (a) of Fig. 1. Here are represented the predictions (i.e., the probabilistic values) that a given human being is either ill (blue curve) or not (red one), after (s)he has taken some radiological test for example, as computed by a DNN-powered artificial *mind*. In the (fortunate) case of Fig. 1(a), there is a clear separation between *good* and *bad*, as the artificial *mind* makes no mistakes, in the sense that a line could be drawn separating those subjects to whom the algorithm has assigned a non-zero probabilistic value of being ill from those to whom a non-zero probabilistic value of being healthy was assigned. Under these simplistic conditions, we can imagine that human beings have no problem to trust the decisions taken by such an artificial *mind* (that is, all reds are healthy, all blues are ill, without any doubt).

Unfortunately, this is a rare case, occurring in very simple situations where perfect and complete contextual information is made available. Instead, in the real world, crystal-clear circumstances are rarely given, and perfect information is rarely provided, to assist in decision making. Realistic situations, based on real-world data, look more like the case depicted in the right hand graph (b) of Fig. 1. Bad

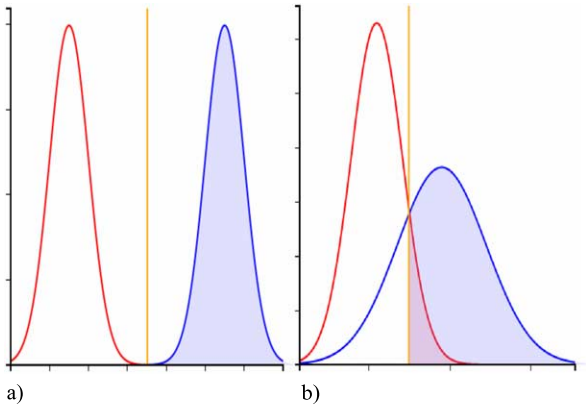


Fig. 1. Probabilistic distributions with two classes (blue: positive, red: negative). Yellow line: decision threshold.

initial data, noise, and even the role of fortuity may result in imbalanced as well as overlapping probabilistic distributions, as computed by an artificial *mind*. And the bad news is that in these cases it is almost impossible to avoid some kind of mistake in decision making.

And, if mistakes cannot be avoided, one good idea to contrast the consequent negative effects would be that of minimizing just those errors whose implications can be considered more harmful than others. In the simple case of Fig. 1, graph (b), for example, all this translates into choosing a decision threshold (the vertical yellow line in the Figure). The different areas drawn at the intersection of the two probabilistic curves and the threshold indicate, respectively: the True Negatives (TN, correctly recognized as healthy human beings), the True Positives (TP, actually ill and recognized as such), the False Negative (FN, predicted as healthy, yet actually unwell) and the False Positives (FP, predicted as unwell yet in reality healthy). Moving the threshold towards the right would lower the rate of the FPs, but would also entail missing many TPs (many are unwell but we are not able to predict them as such). Moving the threshold towards the left would entail missing fewer TPs, yet at the cost of more people being judged as unwell, when in fact they are actually healthy (FP). We could also set the threshold at the intersection of the two curves, minimizing both errors, but this is equivalent to saying that both types of error are equivalent, and that might not always be true in certain circumstances.

What then is the best threshold? No one knows. The answer depends on the task at hand. Sometimes it would be better to take certain actions only when we experience high probabilistic values which suggest that choice, ignoring dubious options (for example, we would not decide that a patient has to undergo surgery if we are not sure (s)he suffers from a given disease, as confirmed by a test whose positivity to that disease should show a very high probability value). Conversely, on other occasions it is better to take action with a minimal suspicion that some conditions can be verified, because treating a false positive can cause less damage than missing a true one. Alternatively, looking at the opposite side of the coin, we might even want to eliminate the cases that are definitely negative (as they can be considered "not problematic"), instead putting more focus and investigation on those cases that are dubious. Beyond *the* intricacy of these arguments lies the most important point that we would like to emphasize, and with which we would like to finalize the first part of this discussion: While it is true that a process that leads to a choice in an artificial *mind* is probabilistic in its nature, nonetheless it translates from the realm of possible alternatives to a set of real actions *following a very clear line of reasoning*, yet depending on a probabilistic decision threshold. Further, we have already anticipated that we humans could set that probabilistic threshold to drive the final output.

Now comes the consequent and final crucial point: while it is well recognized that human beings, leveraging on their past experiences and current feelings, employ a similar type of threshold (some-

times *unconsciously*) to take their decisions, is it essential that artificial minds have such a threshold expressly programmed? Who will set this threshold? Will it still be we humans in the future, as it is now, or will we allow artificial *minds* to define it by themselves? In the next Subsection, we will provide an example of how to translate probabilities into choices that humans judge as both *plausible* and *useful*, adjusting the probabilistic decision threshold we have previously introduced.

3.1. Adjusting the decision threshold

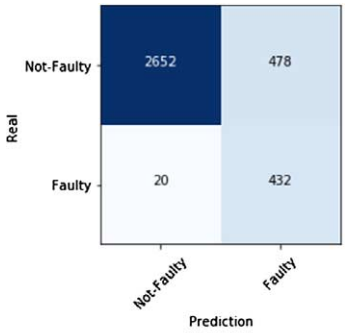
With reference to the above discussion, and to give a pragmatic example of the importance of the role that the decision threshold can play, we briefly touch upon a case we recently studied (Roccetti et al., 2019). We worked with an Italian company to spot malfunctioning mechanical water meters which are used to measure how much water is consumed over a given period of time. Those meters are periodically read by humans, who then report a number of so-called water consumption *readings*. The problem is to forecast if/when a meter is going to break down, based on a set of previous readings. This is a typical unbalanced problem, as the number of meters that are going to be faulty is typically fewer than those that work well. Hence, we trained a DNN able to make such predictions, that based its work on a GRU (Gated Recurrent Unit, with a few hundreds of neurons) to manage those long series of water consumption *readings*. This DNN was able to make predictions on failures with an overall accuracy of almost 90%.

Yet 90% is not 100%. Said simply, we can still have: False Positives (meters predicted as faulty but which will not break down), False Negatives (meters predicted as not faulty but which will break down), True Positives (for which the prediction of being faulty is confirmed in the reality), and True Negatives (for which the prediction of not being faulty is not contradicted by subsequent events).

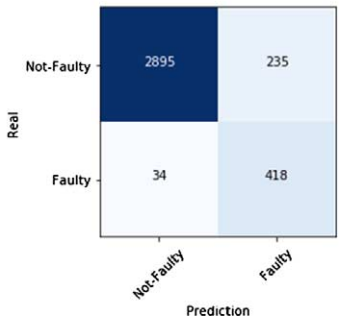
The reader will appreciate that the prediction of (not) being faulty comes from our DNN with an associated probability value. If such probability value of (not) being faulty is greater than our decision threshold, that meter is predicted as (not) faulty. Otherwise, the prediction is reversed. In Fig. 2, we show by how much the proportion of water meters correctly predicted as (not) faulty may change, depending on the value we choose for the decision threshold.

The case represented in Fig. 2(b) is where we attempt to find the point that lies at the intersection of the two curves of Fig. 1(b), yielding a threshold of 0.5. Mathematically speaking, this could be the *optimal* choice if the number of faulty meters would be equal to the number of those which work well. The fact is that, with a threshold of 0.5, if we consider the total number of correct predictions over the total number of predictions, we find a general prediction precision of almost 92% (i.e. we get a relative error of just 8%). Nonetheless, if we focus our attention on the set of just the faulty meters (a smaller number), this precision of the DNN predictions falls to 64%. And this is only half the story. When we misclassify those 235 negatives (not faulty) meters as positives, we are not only making a generic error, but rather this implies that more than one third of the meters we suggested for replacement will actually be replaced for no good reason. Framed this way, that threshold does not seem to be optimal after all.

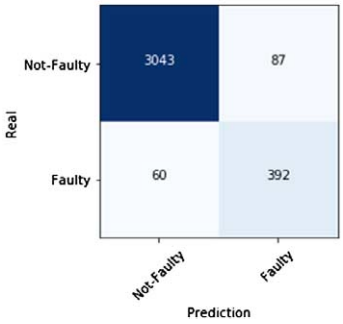
The next reasonable approach seems to be to move the threshold to a higher value or, in layman's words, to only select as positives those meters for which we are very confident that they are faulty. Figure 2(c) illustrates this situation (with a threshold of 0.8). As can be observed, this time the situation changes radically. There is a higher prediction precision in general (96%), and the precision for the faulty meters has increased to 82%. With this setup, we would therefore suggest replacing only 1 in 5 of the meters previously predicted to be faulty, which is significantly more acceptable. But here we



a) Threshold = 0.2



b) Threshold = 0.5



c) Threshold = 0.8

Fig. 2. A probabilistic mind: moving the decision threshold.

have the problem that the number of meters that are faulty and that we are not able to recognize as such is increased from 34 to 60.

At this point of the discussion it is interesting to attempt to understand what happens if we reduce the decision threshold, for example to 0.2 (Fig. 2(a)). As expected, the performance deteriorates in terms of both the general prediction precision and the prediction precision for just the faulty meters, to the values of 86% and 47% respectively. Nonetheless, there is the benefit that we have minimized the number of false negatives – the number of faulty meters that are not recognized as such falls to just 20 meters. From the perspective of a possible collaboration between an artificial, truly probabilistic mind, and a human intelligence, the selection of such a low threshold should not be neglected. With this low threshold, we could have an artificial mind that *eliminates from consideration* almost all of the meters

that are definitely working well, (exactly 2,672 meters, with an error of only 20), leaving to human decision makers the further analysis of the remaining 910 meters, with the aim of determining, using alternative methods, how many meters are (not) faulty. In some sense, this hybrid method resembles a typical pattern, often employed for introducing a technological innovation in a given setting: An intelligent machine eliminates (a large amount of) the simple cases, leaving to human experts the decisions on the most difficult cases.

Thus, a combination of the two methods would appear to be reasonable. A low threshold for minimizing the false negatives and a higher threshold for minimizing the false positives, with the other cases left for human consideration. In our case this would amount to a decision-making setting that automatically: (i) replaces 392 meters, (ii) considers 2,652 meters to be good; and (iii) leaves to a human decision maker the question of what to do with the remaining 538 *dubious* situations. In conclusion, there is not any *a priori* optimal threshold, but rather a series of choices that need to be made when implementing and deploying these models in the real world.

Apart from the crucial problem of choosing the decision threshold, we have clearly explained that a probabilistic artificial mind (e.g. a robot) can make errors of a different type. We devote the next section to a discussion on the interesting question of whether a probabilistic artificial mind can learn from its own errors.

4. RECOVERING FROM ERRORS

The ability to learn from one's own errors is a subject for a problematic discussion, even if we refer to humans. We do not intend to go deeply into this topic because it is outside our professional reach, but It is sufficient for us to remind our readers that optimistic points of view have always existed to oppose more pessimistic ones. Just to cite an example of this ongoing dispute, Syed supports the pessimistic vision, up to the point where he cites the *God complex* in order to provide an explanation for the human inability to recover from one's own errors, even for humans who have great expertise in their field of competence (e.g., senior medical doctors) (Syed, 2015). On the contrary, Firestein and others maintain the opposite position, taking *science* as the most prominent example for their thesis. They refuse to accept the vision of *science* as a path of incremental successes, while claiming, instead, that each step of progress only surfaces out of failure (Firestein, 2015; Coelho and McClure, 2005).

Ignoring that philosophical discussion, we next discuss artificial minds from the perspective that errors must considered as being inevitable for robots. It is easy for many people to overlook the fact that artificial minds, in the form of DNNs, are often trained using examples of those choices that humans have previously made in similar circumstances. What often is missing in the discussion is the awareness that, in general, we have no guarantee that those decisions taken by humans are the best ones. Hence, in those cases, errors made by robots might be simply the consequences of wrong decisions made by humans. But this digression moves us too far from the central point of the thesis of this section, which is simply that a probabilistic artificial mind can be trained to learn from its own errors, provided that it is based on DNN technology.

An example of this is another recent case that we have studied (Delnevo et al., 2019). We were given an underwater scenario, where the problem is to find the *optimal* path between two different points. The complexity is created by a situation where much of the relevant information is unknown, or simply incomplete. Nonetheless, when human experts (i.e. geologists) approach this problem, they are driven towards the solution by two main factors. The first is the length of the route they are looking for. They usually look for the shortest path. The second is the need to avoid underwater obstacles –

an issue that depends on the morphology of the seabed, which is, in turn, largely unknown. Even though, at first sight, this could be seen as a typical pathfinding problem, and hence modelled using traditional path optimization techniques, this is not in fact the case. In order to work well, traditional approaches require precise knowledge of a set of data that we do not possess in this specific situation. It is worth mentioning that many methods employed by geologists to compute an optimal underwater route leverage on the partial information provided by a vessel that navigates and scours the seabed for data that, nonetheless, are returned with a high degree of imprecision, depending on several different causes. Even human experts, under these circumstances, often make decisions that appear as optimal only because many relevant details are unknown.

For our experiment, we used a DNN-based artificial mind to automatically derive decisions comparable to those taken by geologists. In particular, we trained a multi-layer perceptron, composed of some hidden layers with several dozen neurons (our DNN), with the aim of identifying the optimal route from a given starting point to a final destination within a given seabed corridor. We trained our DNN with several routes linking two different points, as computed by the team of geologists with whom we collaborated, and then we asked our DNN to find a route between two other different points, within that same seabed corridor. Generally speaking, we got a good result, yielding a precision of 86% when we measured the ability of our DNN to find the next good step along the optimal path (to be clear: 86% is the ratio between the number of the right choices over the totality of choices made by our DNN).

Nonetheless, 86% is not 100%, in the sense that many times our DNN took the wrong decision to get out of the seabed corridor, due to motivations we cannot even understand given this scenario of incomplete knowledge. At that point, in order to to respond to the crucial question whether our DNN-based artificial mind could learn from its own errors, we developed the following procedure: With every run of our DNN, as soon as it erroneously got out of the corridor we redirected it towards a good point within the corridor, and then we stored this error. Essentially, we collected all the points along the path where our DNN had made an error, along with the appropriate correction. Gathering all the errors and the corresponding corrections, we re-trained our DNN, showing it the entire list of errors/corrections. Then, we asked our re-trained DNN to try to find again an optimal route linking the initial two points.

Table 1 reports on twelve different experiments (routes) conducted with our DNN. The second column shows how many errors were made with the first run, while the third column illustrates the number of errors our DNN makes for the same route after our re-training procedure. We can easily observe that in many cases (namely, experiments # 2, 3, 5, 6, 7, 10, and 11) our re-trained DNN made no more mistakes at the second run.

In some other cases, (namely, experiments # 1, 4, 8, 9, and 12), it still made some errors at the second run, yet they are fewer than before. To solve this problem in these cases we again re-trained our DNN, by showing it again all the mistakes made at the second run, and then asked it to find an optimal route with a third run. The final result has been that, with the third run, our DNN makes no more errors. To conclude this section, we have to admit that, while we still do not still have evidence that our procedure is general enough to be successfully applied to any kind of problem, it reveals nonetheless that there is a trend for a DNN to learn from its own errors, following a very simple and efficient method, like the one we have adopted in our study.

Table 1
Recovering from errors

Experiment	Initial errors	Errors after re-training
1	20	14
2	285	0
3	10	0
4	30	8
5	1	0
6	8	0
7	493	0
8	41	18
9	457	12
10	37	0
11	4	0
12	115	39

5. CONCLUSIONS

Starting from the reasonable point that human brains and cognition are not yet well understood by neuroscientists, our opinion is that the predictions of realizing *humanlike artificial* minds in the near future, regardless of the implementation technology, are nonsensical rather than optimistic (Parry et al., 2016; Bassett & Gazzaniga, 2011). What we have discussed in this paper is different. We have provided some reflections and examples concerning the point that DNN technologies can be integrated into probabilistic models suitable for implementing *cognitive entities*. Such cognitive entities can exhibit semantics of behavior that can be useful for humans and, to some extent, are also clear and predictable, while depending on the data on which such models are trained.

Finally, we are confident that machines (e.g. *robots*) can be created that will take intelligent decisions and actions, even if they do not strictly imitate human nature (Ramamoorthy & Yampolskiy, 2018; Laird et al., 2017).

REFERENCES

Alaieri, F. & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In *Social Robotics*. Lecture Notes in Computer Science (pp. 159–168). doi:10.1007/978-3-319-47437-3_16.

Bassett, D.S. & Gazzaniga, M.S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, *15*(5), 200–209. doi:10.1016/j.tics.2011.03.006.

Climehaga, N. (2018). Infinite value and the best of all possible worlds. *Philosophy and Phenomenological Research*, *97*(2), 367–392. doi:10.1111/phpr.12383.

Coelho, P.R.P. & McClure, J.E. (2005). Learning from failure. *American Journal of Business*, *20*(1). doi:10.1108/19355181200500001.

Delnevo, G., Roccetti, M. & Mirri, S. (2019). Intelligent and good Machines? The Role of domain and context codification. *Mobile Networks and Applications*. To appear.

Firestein, S. (2015). *Failure: Why Science Is so Successful*. Oxford University Press.

Floridi, L. & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*.

Hart, P.E., Duda, R.O. & Einaudi, M.T. (1978). PROSPECTOR: A computer-based consultation system for mineral exploration. *Mathematical Geology*, *10*(5), 589–610. doi:10.1007/BF02461988.

Hew, P.C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, *16*(3), 197–206. doi:10.1007/s10676-014-9345-6.

Jefferson, G. (1949). The mind of a mechanical man. *British Medical Journal*, *1*(4616), 1105–1110. doi:10.1136/bmj.1.4616.1105.

Laird, J.E., Lebiere, C. & Rosenbloom, P.S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*(4), 13. doi:10.1609/aimag.v38i4.2744.

Liu, Q., et al. (2016). Probabilistic reasoning via deep learning: Neural association models. Preprint. Available at: arXiv:1603.07704.

Parry, K., Cohen, M. & Bhattacharya, S. (2016). Rise of the machines: A critical consideration of automated leadership decision making in organizations. *Group & Organization Management*, *41*(5), 571–594. doi:10.1177/1059601116643442.

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pomerol, J.C. (1997). Artificial intelligence and human decision making. *European Journal of Operational Research*, *99*(1), 3–25. doi:10.1016/S0377-2217(96)00378-5.

Ramamoorthy, A. & Yampolskiy, R. (2018). Beyond mad? The race for artificial general intelligence. *ITU J*, *1*, 1–8.

Roccetti, M., et al. (2019). Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *Journal of Big Data*, *6*(1), 1–23.

Shortliffe, E.H. & Buchanan, B.G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*(3–4), 351–379. doi:10.1016/0025-5564(75)90047-4.

Syed, M. (2015). *Black Box Thinking: Why Most People Never Learn From Their Mistakes – But Some Do*. Penguin.

Villani, C., Bonnet, Y. & Rondepierre, B. (2018). *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. France: Conseil National du Numérique.

Yoon, Ki., et al. (2018). Inference in probabilistic graphical models by graph neural networks. Preprint. Available at: arXiv:1803.07710.