# On The Verge Of A New Relationship Between Man And Artificial Learning Machines?

Eugénio Oliveira
Faculty of Engineering, University of Porto
and LIACC- Artificial Intelligence and Computer Science Lab
Porto, 4200-465, Portugal
eco@fe.up.pt

## Abstract

We are now living an exciting time in which astonishing results coming from the A.I. field trigger the usual expectations and fears about intelligent man-machine relationship. Some believe that the basic ingredients for the artificial general intelligence are already here and, in their opinion, it is just a matter of putting it all together. It seems that we do not have to program computers anymore; they will program themselves. Should this statement frighten us?

Our position is both of recognizing the big potentialities of new A.I. developments and, simultaneously, to warn against yet another overselling and possibly damaging stage in the A.I. field.

We advocate that, intelligent systems, relying in evolving machine learning algorithms, fully autonomous software agents or robots, should always follow the "human in the loop" principle, ensuring that the responsibility for all future intelligent entity activities can be traced back to some recognized and accountable individuals or organizations.

**keywords:** machine learning, master algorithm.

## 1 Introduction

*"Our goal is to figure out the simplest program we can write such that it will continue to write itself by reading data, without limit, until it knows everything there is to know."* [4][1]
"The Master Algorithm" is indeed a remarkable book that makes us thinking and exercising our critical opinion without denying both the beauty and dangers of its main message.

"Machine learning is remaking science, technology, business, politics, and war ..." [4]. Although this claim may be accepted as partially true, it also reveals a well-known tendency for overselling a specific research topic, trying to ignore that often, Machine Learning (ML) algorithms work together with a multitude of other different algorithms in order to get the things done. For example, when saying: Googles self-driving car taught itself how to stay on the road; no engineer wrote an algorithm instructing it, step-by-step, how to get from A to B, it seems that it is all about Machine learning. But no. There is also a need for, at least, competences on advanced computer vision and systems control, trajectory planning, sensing and perception algorithms. Moreover you also need computer systems' distributed architectures and modules (or even software agents) coherent interaction and coordination.

It is true that ML algorithms look like artifacts that produce new artifacts. In some way, a "Master Algorithm" would be a powerful and absolute General-purpose learner, a kind of "Holy Grail" which, in reality, will be very difficult to find.

Science is mostly based on observations, gathering data, and inferring models in which data fit. And thus, it would seem perfectly reasonable to argue that ML over Big Data will enlarge the scope of science and will give us unlimited knowledge.

However, it definitively seems to me that, up to now, those algorithms work over data that, although gathered in large amounts, have a quite simple structure. You do not need extra knowledge to build up a theory that explains those extracted patterns.

On the other hand, there are situations for which this approach will not be enough. Since you have huge amounts of data available about climate all over the world for many decades, why is it climate for the next months still so hard to preview? It is may be because there is a need for more sophisticated human reasoning over that big data that goes beyond those patterns that can be directly extracted from that data.

It is true that, using ML, past-based patterns can be made available and may be useful for many different kind of situations. Even so, being guided in our

[1] All text between quotes with no reference attached, is taken from this same reference [4].

so-called preferences for future activities, literature reading, political voting or wine drinking, may be indeed a bad idea since it looks like you are being held by your hand as a child. And, who knows, whenever you decide to act differently from what was previewed, when you are upset with your past choices and decide to do it differently, it may happen that you will become suspicious, seen as a disruptive person, half a way to become a potential terrorist...

Fortunately, I really believe that the claim that the "Master Algorithm" would derive all knowledge in the world - past, present, and future - from data is, as seen by today's perspective on ML possibilities, a clear exaggeration.

To present the "Master Algorithm" as a unified theory that "will make sense of everything we know to date" seems to me like overselling the real capabilities of a useful tool to infer some specific knowledge as it is the case of ML algorithms.

Yes, "at its core, machine learning is about prediction", but nobody can prove that this really means "predicting what we want, the results of our actions, how to achieve our goals, how the world will change" except for limited and stereotyped situations.

Saying that machine learning algorithms are finding out the solution for a problem by "just adding data" is misleading. First because it is not "just". Data is, per se, a big problem to be dealt with. We have to solve many problems for appropriately acquire it, select it, prepare it, represent it. Then "adding" may also be not so simple. Combining pieces of data or subsets of data also poses problems that, if not solved, may jeopardize the all operation. Finally, "data" is also something that needs to be better defined.

Are we ready to derive useful knowledge from any kind of data sets? Of course not. You may supply hundreds of thousands of medical cases about, let us say, different cancer types, but if you miss a few tenths of cases regarding specific situations, they will always remain invisible to the inferred algorithms.

Sundar Pichai, chief executive of Google, is an A.I. enthusiast and he assures that " Google is going to be AI first". Although he is confident that A.I. will make available a general tool designed for general purposes in general contexts, he also adds, and I fully agree, that "for the moment, at least, the greatest danger is that the information we're feeding them [A.I.-enhanced assistants] is biased in the first place" [6].

Next section will be about the Master Algorithm and needed data, and the following sections, 3 and 4, about two possible ingredients to help on making the so-called "Master" more acceptable to us.

## 2 "Master Algorithm" Claims

As an experimented researcher on the A.I. field, I was challenged by the main hypothesis intended to be proved in the book [4]:

> *"All knowledge - past, present, and future - can be derived from data by a single, universal learning algorithm".*

The point is that it is not only true that ML and A.I. are, once again being oversold, as it is true that we cannot yet derive an ultimate paradigm to build up a single and definitive universal learning algorithm.

ML was an A.I. important research topic that steadily grew since the seventies mainly driven by research on symbolic learning. Other emergent different approaches to the same goal, automatic learning, were always more or less rejected as not belonging to the same ML-A.I. tribe.

Connectionists and evolutionary-based algorithms have often been seen by the former ML researchers, as proposing research directions waiting for their respective dead-end. ML researchers even saw themselves like the elite of A.I., working upon the only topic that, in fact, deserved to be considered as doing real A.I.

It is amazing that now, not only ML is claiming to be "the" A.I. but also it is willing to encompass all the other approaches to automatic learning.

In the referred book, the "Master Algorithm" is foreseen as the result of combining precisely the five already existent machine learning paradigms and schools: the symbolists, connectionists, evolutionaries, Bayesians, and "analogizers".

Despite the fact that two classes of the existent learning algorithms, althoug brillant, are, in fact, extreme simplifications of brain machinery and evolution laws, the author really believes that the current state of the art is enough for the definitive paradigm shift leading to, if not yet the "Master Algorithm", at least closest to it anyone has come.

In our opinion, nothing proves that, exactly now, we came to the situation in which we recognize to have all the needed bricks to build up a solid staircase leading to the universal learning capability.

Chaining and mixing those existent different machine learning principles may not be enough to solve the overall problem. Even if we accept the power of data, it may take more than collected observations to directly induce natural selection "as Darwin did".

Was it just only a matter of observing data? I do not believe it was only that. Notice that many people, many

brains, and all along many, many years were not (and are not) getting everybody to the same conclusions even in the presence of the same available data. And this may be because we need still more than simple data. Which points out to what we are missing here that could, who knows, be the most important: Some kind of ability that some brains have developed, and others did not, to extract, in some context, more sophisticated knowledge from the same data. And, perhaps, there are "hundreds" of needed capabilities to be developed in the future that, even the most gifted brains cannot yet imagine.

I recall that the Theory of Unification is needed because quantum physics only deals with the very small, Einstein's general relativity theory deals with the very big and we need a theory that works everywhere.

However, physicists do not think that the Unification theory will come out of a kind of combination of the previous two theories mentioned before. They are still looking for something radically new. The same will happen, in my humble opinion, with the so-called "Master Algorithm" and it is an over simplification to believe that it will precisely come out of the ML algorithms that we already know nowadays.

I am not as radical as those who state that "big data is not the new oil; it's the new snake oil". But, nevertheless, I would be more cautious in targeting the possible goals of current ML algorithms working over big data as the "ultimate learning machine".

The author claims that through the "Master Algorithm", by mining a "vast amount of patient data and drug data combined with knowledge mined from biomedical literature is how we will cure the cancer".

It would be wonderful. Who wants to deny such a great possibility? I am rather optimistic on the possibilities of interpreting extracted patterns from appropriated mined data. Although nothing tells me to corroborate that possible medical achievement, I also see nothing making me to definitively deny that possibility. And the goal is so appealing that I would prefer to be optimistic. I thus concede this possibility and agree with the author of [4].

However, the real problem is when the goals to achieve are not as universally desired as the cure of cancer but, instead, they are much more controversial.

"*Ours*" against the "*Others*" in which politic opinions is concerned, for example; companies versus consumers in the market; etc. For those purposes, data collection and data mining outcomes become not at all crystal clear and may lead to artificially justified dominance in many different aspects.

Stating that "[computers] they'll even guess what we want before we express it." is the first step to impose to you what you should want. Following this trend, I will not be so sure that "Just because computers can learn does not mean they magically acquire a will of their own".

Does not everything come out of data? Saying that "they don't get to change the goals" seems to me misleading, even at the present moment, where some kind of BDI type of cognitive agents are already able to be pro-active and autonomously set some of their own possible (intermediate) goals leading to some other ultimate, predefined, intentions (the "I" in the "BDI" agents' architecture) in mind. "The Master Algorithm" is not just a passive consumer of data; it can interact with its environment and actively seek the data it wants" Does this not mean that the "Master Algorithm" has a kind of will of itself? How can it decide on what will be relevant and what is not? Or does it rely, or depend, on something that masters it, selecting and supplying what is considered to be the needed and relevant (or could it also be the misleading) data?

Moreover, to the author of [4], the well known John Holland, pioneer of genetic algorithms, decided to turn Darwin's natural selection theory into an algorithm. But he recognizes that although Genetic Algorithms (GAs) central piece of knowledge is a previously known fitness function, evolution as a whole has no known purpose. He then exemplifies with diagnosing capability in order to justify how easy it would be the design of such a fitness function. I believe that the issue is not as simple as that. What it implies, in my humble opinion, is that the evolutionary machinery of the GAs only drives you to solutions whose characteristics you have already selected beforehand. That is why you specify and apply a fitness function. You then will be able to capitalize the final "evolutionary" outcome of an a priori biased intention.

As the author said this is more like selective breading, through a pre-selected specific fitness function, than natural selection. Yes, GAs breed programs like they may one day breed robots for the sake of ... well I do not know to serving what purposes.

## 3 The Human In The Loop

A few years ago I and two former PhD students of mine, we had designed a so-called autonomous software system that was able to propose solutions for unexpected plan disruptions in the context of Airlines Operations Control [1].

Each different software expert in dealing with, and trying to solve, a specific aspect of the problem, from

aircraft landing delays or aircraft unavailability due to malfunctions, to crew members absence, start to look for the best solution for the problem in hands. Several different autonomous agents of the overall multi-agent system worked together and collaborate in finding a good solution to the problem minimizing the effects of the unexpected situation.

However, we soon got to the conclusion that the way we built the so-called autonomous system could lead to a too much biased, although seen as "optimum" like solution. And of course, according to that solution, the airline company itself would always be the winner.

What about the legitimate interests of the crew members or the real individual interests of the passengers? It was no problem to try to find out new weighted solutions taking also into account all different perspectives. They were not as much appealing to the company as the first one but, nevertheless, they could be accommodated together with marginal impact.

The real, fair and final solution, calculating a combined utility taking the different perspectives into account, was only achieved when we included the human in the loop principle, giving some authorized officer the responsibility, without escape, of explicitly weighting those several different perspectives in order to find out what could be, according to his decision, and under his responsibility, the best compromise for the ultimate policy justifying that same solution. In some contexts passengers, or even crew members, may be more important than the immediate and direct company interests. Someone has to be responsible for the choice. Moreover, software agents behind the scene, also included a learning capability, trying to improve, each time, the way they negotiate with others to make their proposal better accepted by the other ones. Curiously enough, the learning algorithm I was applying does not perfectly fit in the five tribes learning algorithms classification proposed in [4]. Since the system needed to learn with very few examples, we were using a reinforcement "Q-learning" algorithm.

Later on, the "human in the loop" component became again a corner stone of the final semi-decentralized multi-agent system we have designed. This time the objective was managing ship damages when they are under severe conditions, either weather conditions or external attacks. In these scenarios all the monitoring capabilities and solutions finding (plan of actions and resources allocation) came out of the automatic software agents capabilities, including the very same learning algorithm using reinforcement learning. However, it was mandatory that at least some part of the command chain was replicated for interfering with the decision system at different levels in the all decision process. No way of forgetting the intrinsic responsibilities assigned to humans (officials and commanders) in charge regarding the acceptance of the proposed solutions at different moments in time. This was indeed a relevant factor in the possible acceptance of the semi-automatic solution for delicate and sensitive problems like the one of managing a ship in harsh situations.

Is it an answer to A.I. and ML potential dangers just to include the human in the loop"? It might be. However we should not forget that "Drones can fly autonomously with the help of learning algorithms; although they are still partly controlled by human pilots". And, despite being monitored by humans, I am not sure of the drones goodness in many different situations ...

# 4 Emotion-Like States

Back in 1997, I published a short paper about "Robots as responsible Agents" [7]. My naive approach, twenty years ago, was that the then novel cognitive software agents architecture based on "mentalistic" concepts like "Beliefs", "Desires" and "Intentions" could bring a positive influence to the designing of more self-aware robots controlled by the software agents.

I was proposing a two-layer architecture, using symbolic representation for dealing with knowledge and goals at the deliberative level and sub-symbolic neural networks for implementing specific behaviors at the reactive level.

One of the main problems we were addressing was how to make these two levels to communicate, to interact and to cooperate without being completely depending from each other which could lead to deadlocks.

Regarding intelligent robots, I doubt that with the current hardware limitations and capabilities we may make them evolve for a much more intelligent-like kind of entity. However at least we could combine the two different levels of decision making, one relying on some kind of instinctive, reactive capability and the other displaying a more cognitive intelligent behavior. While the former level should be implemented in a sub-symbolic way through neural networks, the latter was based on the already mentioned BDI architecture. The main issue became then how to make those layers to work together.

Planning, learning, classification, intentions-guided decision making may in certain situations (I would say in most of them) take control of the intelligent robot. In other specific scenarios we may expect that reactive behavior is the best decision for the sake of survivability or efficiency. The problem is that a vast gray area exists where both capabilities may overlap and even compete

for the robot's control.

We also proposed the use of a modal logic (intentional logic) to correctly define what could be the persistent goals for an agent (controlling a robot) to pursue and the conditions for giving them up. My real implemented mobile robot never solved this kind of schizoid behavior in some particular situations in which reasoning and reaction were both of paramount importance.

It is not here the place to go into details on this problem. It was only about 5 years later that I realize that one important and decisive component of human-like reasoning is deeply related with emotions. Contrary to what many past scientists and philosophers advocate, human emotional states are essential for human reasoning and decision making.

Neuroscientists of the last decades proved that reason and emotion are intrinsically intermingled [2] and, thus, we, computer scientists in the quest for real artificial intelligent entities, should take this relationship into account.

I have tried to give a contribution to logically define an emotions-based BDI agent architecture making it possible to make decisions while also taking some primitive emotions into account [3]. Individual perception of the personal risks any situation involves, as well as knowing individual capabilities it might be available for dealing with it, are crucial factors influencing an individual (agent, robot) final decision.

Past experiences, in different scenarios and with different meanings can be translated into kind of primitive emotions (fear, anxiety, ...) through accumulator kind of variables. Accumulators gain some kind of energy through specific stimulus sometime in the past and they discharge their energy following specific decay curves.

Including these "emotion-like" states in the reasoning loop will make more difficult to take decisions that possibly leads to bad results in terms of causing pain. This implies artificial and, let us say, intelligent decision-making may benefit in taking into consideration this more human-like factors, like emotion states, in order to become more human friendly.

# 5 Conclusion

The really interesting and challenging book, "The Master Algorithm", poses a number of questions to those who do not get too much excited by the periodic promises about "ultimate" solutions for replicating general artificial intelligence, all-purpose learning capabilities and intelligence. Despite the serious proposals the book include, regarding the quest for super algorithms for machine learning through the combination of the existing ones, they also ring a bell to those who would like to give humans always the possibility to monitor artificial systems in any situation. Do not let them "leave" and "learn" without tight supervision.

Therefore, there are, by now, at least two main, somewhat simple, mechanisms to help in preventing harmful and irresponsible autonomic decisions of an artificial entity: First and most effective, always include the human in the loop of the artificial systems decision-making capabilities. This should happen at different levels, if needed, and making it clear and explicit the human responsibility at any relevant and decisive choice point. Second and more subtle, continue to develop more environment and emotion dependent mechanisms, providing that they can be transparent and explicable. Make this more human-like emotion states also a relevant ingredient of the final decision making process.

We are not yet capable of fully define what intelligence really is. However we expect that intelligent behavior will be highly performant in solving complex problems in large unknown and possibly unstructured environments. To achieve that, first much more accurate perception, coming from different sensors need to be connected to the "decision-maker" (kind of a brain). Second what makes intelligence more evident is the recognition that decisions were made taking into account some kind of both individual and social common sense that, unfortunately, still is ill-defined.

Many facets of intelligence can be formulated "as goal driven or, more generally, as maximizing some utility function" [5] in the sense that also "the (biological) goal of animals and humans is to survive and spread. The goal of AI systems should be to be useful to humans". Period!

It may also be the case that representing kind of emotion states, may contribute for more intelligent and human-friendly decision-making.

Finally we are not doing research like running in an Olympic competition where the rules are well known and the targets to reach are unambiguous. We are wandering around in a vast territory, exploring and discovering more and more different lands. A large territory where large portions still remain uncharted. And this seems to me that does not point out to a solution that resembles more like to be a matter of assembling and combining several known pieces to build up a kind of Frankenstein.

# References

[1] António Castro, Ana Paula Rocha, and Eugénio Oliveira. *A New Approach for Disruption Management in Airline Operations Control*. Studies in Computational Intelligence, V. 562. Springer, 2014.

[2] António Damasio. *Descartes' error: emotion, reason and the human brain.* Avon Books, 1994.

[3] Eugénio Oliveira David Pereira and Nelma Moreira. Formal Modelling of Emotions in BDI Agents. In *Computational Logic in Multi-Agent Systems*, Lecture Notes of Computer Science, V.5056, pages 62–81. Springer, 2008.

[4] Pedro Domingos. *The master algorithm: how the quest for the ultimate learning machine will remake our world.* Basic Books, 2015.

[5] Marcus Hutter. *Universal Artificial Intelligence, Sequential Decisions based on Algorithmic Probability.* Springer, 2005.

[6] Guideon Lewis-Kraus. The great a.i. awakening. In *The New York Times Magazine*, Dec. 14, 2016.

[7] Eugénio Oliveira. Robots as responsible Agents. In *Proceedings of The IEEE International Conference on Systems, Man and Cybernetics. Computational Cybernetics and Simulation, V.3*, pages 2275 – 2279, 1997.