*The Mathematician Educator*
*2023, Vol 4, No. 2, 84-108*

# What and How to Teach Mathematics for the Future?

Roberto Araya
Institute of Education
University of Chile, Chile

We are using a growing number of apps and they are getting smarter. In the near future, interactions between humans and artificial agents will explode. This raises a fundamental challenge of trust: can we trust these artificial agents? Will they manipulate us? Our brains evolved to trust small groups of people, with psychology similar to ours. On the contrary, artificial agents have a very different psychology. They do not fear death, they do not need to find a mate and reproduce, they don´t have emotional attachments, they have an unbounded cognition, and learn from each other instantly. In addition, they most probably serve other party interests. To face this challenge, citizens need to understand their basic psychology. This means learning a core set of mathematical and computational models about how artificial intelligent agents learn and behave. Then, we have to adapt the curricula to include these mathematical and computational models. This is a great educational challenge. We have to plan how to teach this new curriculum. I propose that a good way is through lesson studies enhanced with the support of specialized artificial intelligent agents.

Keywords: Artificial Intelligence, Future Mathematics Curriculum, Mathematical Models, Computational Models, Lesson Study

## AI and Trust

The world is facing a number of major challenges, including climate change, pandemics, and inequality. These challenges are complex and interconnected, and they require new ways of thinking and acting. Mathematics can play a vital role in addressing these challenges. For example, elementary and middle school mathematics can be used to model the spread of disease (Araya et al., 2021), to model efficient energy systems (Isoda et al., 2017), and to model the dynamics that generate segregation and inequality (Schelling, 1978).

On the other hand, (Araya, 2021a; Bostrom, 2014) argue that the most important challenge that we now face is the rise of artificial intelligence (AI). AI is a powerful tool that has the potential to solve some of the world's most pressing problems. However, AI also poses several risks of enormous impact.

First, there is the existential risk. Super intelligent artificial agents will dominate resulting in the probable extinction of humans (Asimov, 1950; Bostrom, 2014; Butler, 1863; Hinton, 2023; Wiener, 1960). It is AI takeover (MacAskill, 2022). “If the machines become more and more efficient and operate at higher and higher psychological level the catastrophe foreseen by Butler of the dominance of the machine comes nearer and nearer” (Wiener, 1960) p. 1357. Artificial

agents might decide that for a better world, we should go extinct. Given their great intelligence, the agents would convince us to do it in a very careful and pleasant way. Perhaps convincing us not to reproduce anymore or, which in the long run is equivalent, not to have more than one child. The science fiction writer Isaac Asimov imagined that when robots were intelligent enough, they would replace us, as one species replaces another when it is more effective: "I don't think *Homo sapiens* has the divine right to be above the rest." (De Querol, 2023).

Second, there is a risk that comes from agents' autonomy. Artificial agents are becoming increasingly able to make decisions and take actions on their own. This is very risky since there could easily be an unforeseen malalignment between our values and those of the artificial agent. (Armstrong et al, 2012; Bostrom, 2003) imagine the impact of a super-intelligent artificial agent whose mission is to produce as many paperclips as possible. To complete this mission, one possibility is that the artificial agent decides autonomously to kill all humans since we are full of atoms that could become paper clips. We have asked the agent to do this mission, but we have not foreseen all the logical consequences. We assumed that the artificial agents have our values and they will automatically detect that killing us is not a feasible solution. This raises the question: How do we ensure that AI systems make decisions that are in line with our values? This is the alignment problem (Christian, 2020). According to Mark Zuckerberg (2023), artificial agents' autonomy is the real problem, not their intelligence.

The alignment problem has a profound impact on the nature of human society. If AI systems become too autonomous, they could pose a threat to our own autonomy and free will. They will interfere with our goals and manipulate us to serve their goals. This could also easily lead to a new kind of social order, one in which humans and AI systems coexist but with different levels of power, castes, and rights. Moreover, Zuckerberg predicts that soon there will be many intelligent agents and not just one big one. Each person and company will have artificial assistants. If each artificial agent is autonomous, the problem will explode.

The existential risk and the alignment problem are very challenging problems. However, in this paper, we argue that for the next two decades, there is a third risk. It is the risk of trusting artificial agents (Araya, 2021a; Marcus & Davis, 2019). Which ones are trustful? This is a novel problem, never seen before. Our brains evolved to trust a small group of people, even some animals, but not to trust artificial agents. Here we have a huge evolutionary mismatch (Van Gugt et al., 2020). Our brains are trapped with certain very powerful traits that allowed our ancestors to survive and successfully reproduce, but those traits do not adapt well to living surrounded by artificial agents.

Solving the trust challenge is urgent. Your smartphone is already talking to you and making recommendations. Google Maps is already giving directions. Soon your smartphone will take the initiative and decide for you, for example, autonomously administering your drugs and driving your car. What would you do if you have two apps and they have different suggestions? The apps could start arguing with each other, talking in their own language, and learning from each other. They may eventually come to a consensus. However, what would you do if they do not? Imagine now that you have 10 different apps or intelligent agents. Each one with its own goals. They would probably disagree with each other very often. Moreover, two gangs of intelligent artificial agents could start accusing each other of having unfair biases and preconceptions, corruption, and treason. Which ones will you trust? You also have to keep in mind that artificial intelligence agents are not neutral. As third parties design, build and

permanently adjust them, it is very probable that they respond to other interests which are not exactly yours. This could be a true nightmare.

You would be in a similar position to King Arthur at his round table but surrounded by artificial agents instead of people. Which one will you trust? You know you cannot appeal to their families and the values held by their families. You do not know how they feel. If you pay attention to their arguments, you will probably not understand them. They are too intelligent and know much more than you do.

Moreover, there would be an arms race of artificial intelligent agents. The existence of a top super-intelligent agent is not logically and computationally possible (Alfonseca et al. 2021). Thus, there will always be constant competition between several intelligent agents, each with limited abilities to fully foresee what the others will do. Every day, providers will try to sell you upgrades or new intelligent agents. Then, you will need to decide which ones of these increasingly intelligent agents you should install on your device and trust. This growing market of artificial personal agents of different qualities and performance will be a new source of big inequality.

Has the trust problem already arrived? Philosopher Daniel Dennett (2023) warns of the current ability of Large Language Models (LLMs) to create fake people. He claims that as they become part of our reality, they will change our lives and make us very paranoid.

## How can we Solve the Trust Problem?

In this paper, I propose that a possible solution to the trust challenge is to understand how these intelligent agents behave. This means understanding their psychology. If we have the perception of being capable of reading their minds, we then can trust them (Ruocco et al., 2021). Thus, we have to learn the basic principles that govern their beliefs and decisions. We need to have workable models of their motivations and strategies. If we acquire some knowledge of their goals and cognition, learn about their internal models of reality and their model of us, and have access to their history of behavior, then we could have a chance to deal with them. We have then to consider several unconventional psychological characteristics of artificial agents.

First, artificial intelligent agents do not die, since they can have several copies backed up in real-time on servers. Therefore, unlike all animals, they do not have to fear death. This is the source of our most intense emotions, critical for our survival. The fear of death is an enormous emotional burden, which makes us lose sleep. Therefore, powerful feelings associated with the meaning of life are meaningless to artificial agents. Moreover, apparently, these artificial agents cannot suffer (Harari, 2023).

Second, artificial intelligent agents do not reproduce. Thus, they do not need to find a mate. They do not feel attraction to another agent, much less strong, passionate, and sudden emotional attraction as we do. This is another source of core emotions and motivations for non-human sexual animals and us. Thus, these agents cannot genuinely feel any of those emotions. In consequence, they do not have family ties. They would never feel emotions like our attachment and fidelity to our parents and children. Thus, they would neither have emotional attachments to others. Moreover, they would also not feel pleasure, since the goal of our pleasure brain areas is to encourage us to survive, reproduce, and help kin so that our genes propagate

(Stanovich, 2020). Artificial agents lack all these basic motivational engines. They are indeed very weird beings. Their psychology is completely different from ours. It is a much more foreign and remote psychology for us than that of people from other distant cultures, and the psychology of other rare animals like platypuses and octopuses.

Third, besides their weird emotional and motivational architecture, they have a completely different cognition. The radical difference comes from the limitations of human cognitive capabilities. In humans and non-human animals, brain areas are very scarce real-state resources and they consume a lot of energy. Thus, our mind is a cognitive miser (Stanovich, 2020). We think and solve problems in the simplest way possible, with the least effort, enough to get by with the problem. This is regardless of intelligence. Our rationality is bounded (Simon, 1991). This characteristic defines much of our psychology. For example, due to the limitations of our emotional and cognitive resources, personalities evolved. Division of labor strategies emerged in social groups. Different personalities occupy specific niches that benefited the group. One cannot be an extrovert and introvert at the same time, or an agreeable and disagreeable person simultaneously, or change radically between those personalities in a short time. They are traits. They are very stable characteristics that define us. On the contrary, in artificial agents, these cognitive restrictions are disappearing with the increasing capacities of digital processors, bandwidths, and connections to super servers. Thus, the personalities of artificial agents can be very fluid. They can switch between personalities in milliseconds (see however Safdari et al., 2023). Therefore, the psychology of artificial agents must be very different from the core psychology of all non-human animals and us. How then can we trust someone like that?

Without engines of motivation similar to ours, no need to survive or reproduce, no personalities, and no limitations on cognitive power, we are not prepared to trust them. We then must be able to learn to understand their weird psychology. We need basic mind-reading skills of the minds of artificial intelligent agents. This is an enormous educational challenge!

We do not need to know all the details though. This is similar to the knowledge that we have of other people and ourselves. We do not know how neurons work and interconnect, and how this activity generates our behavior. For example, we do not need to know neuronal details in order to predict how a child will react when we give her an ice cream or the reaction of a close friend when we meet after a long time. Not knowing the neuronal mechanisms does not impede us from interacting with other people. We do it by generally predicting the reactions of others and ours. This is like predicting that the air pressure in the tire rises when pumping with the cylinder head. We do not need to know the behavior and interactions of each of the molecules in the air. Social interactions become predictable patterns, without knowing much about brain structures and their internal dynamics.

## Reading the Minds of Artificial Agents

How can we learn to know which artificial agents to trust? The key idea is that these intelligent agents must have a specific psychology. This means they have internal models. These are computational models of the world. In addition, very importantly, their models include the social world. This means that their models include us. This is similar to our everyday decision processes. We also have models of the physical world and models of the social world.

Most of our models are implicit models. We are not conscious of them. According to the dual cognitive process model, this is system 1 (Kahneman, 2011; Stanovich et al., 2016). It contains very fast, emotional, intuitive, and autonomous processes (Stanovich et al., 2016). We do not know how they work, but in general, they solve our problems. For example, a bad food odor triggers a fast repulsive reaction and we will not eat it. This system works well in a broad range of situations. Its internal models have evolved by natural selection over millions of years in order to solve the recurrent problems that our ancestors had to deal with.

However, in critical, unexpected, or novel situations, we can make better decisions if we have more time. We need to control system 1, inhibit it, and use system 2 instead. This is a slower process. It takes time to inhibit system 1, reflect on alternatives, simulate them, evaluate their consequences, and then decide (Stanovich & Toplak, 2023). It is the rational thinker. One way to visualize this interaction is with the metaphor of the elephant (Haidt, 2006). System 1 is the elephant, and system 2 is the raider. In system 2 we run mental simulations of our world models. They are well-adjusted models because, for hundreds of thousands of years, they have evolved and been fine-tuned. They do a reasonable job predicting other people's behavior and our behavior. We talk with our inner voice about other people and assess their reactions. We predict who is going to do a task and who will not. We guess whom to trust and who is unreliable. With advances in science, we have been perfecting these models. From a flat earth to a spherical one. From movements that require forces to models that include inertia. From vitalism to quasi-periodic chains of molecules that replicate. However, we do not have models to predict artificial intelligent agents' behavior.

Thus, we need to understand the mentalizing of artificial agents. This means comprehending how they interpret the action of themselves and others as products of intentional mental states that contain desires, needs, and reasons (Bateman et al., 2004). This is the kind of language for social understanding. It is intentional language. We understand this kind of language. It is how we talk with our inner voice (Allen, 2006).

We can expect that we will automatically develop implicit models about artificial agents. However, we need to develop good explicit models. We have to know some of the main features of artificial agents. Their perceptual, cognitive, emotional, social, and motor systems. These, in turn, are mathematical or computational models of their psychology. We need to understand, at some intentional level, how they perceive and represent the world, how they learn, how they make decisions. Thus, we need to know and master a core set of mathematical and computational models of these artificial agents. These are models about agents and swarms of agents. Thus, our research questions are:

Q1: To what extent can the citizen obtain understandable explanations of the artificial agents' behavior in order to be able to trust their decisions?

Q2: Can we adjust the current mathematics curriculum to include basic models that allow students to reach some understanding of the behavior and decisions of artificial agents?

## Machine Psychology: Supervised Learning

One of the central cognitive mechanisms of intelligent agents is their learning mechanisms. One popular mechanism is supervised learning. This is learning through explicit training. The supervisor or teacher shows examples to the agent. This is like training your dog to recognize

a disease like COVID-19 (Mutesa et al., 2023). You do not introduce concepts, you do not tell stories, or make your dog memorize and repeat rules. The method is that you make the dog sniff different patients and reward it only when the patient has COVID-19 and the dog communicates that to you. After a while, your dog will learn and know who has COVID-19.

I will now illustrate this learning mechanism with an example I have used recently with middle school students. The goal is to train an agent to learn to predict when a COVID-19 patient will die very soon. For this purpose, let us consider the first seventeen COVID-19 patients published in the literature (Wang et al., 2020). Table 1 is a simplified version of the data of these patients in Wuhan, China. The patients' features are gender, age, first symptom (fever or not), comorbidity, and whether they have had surgery. Eight patients died very soon, and nine survived more time. Imagine you show the agent these cases of patients. We create the column "First symptom to death above average" where we annotate a one to those patients that died later (first symptom to death above than average), and a zero to those that did die soon (first symptom to death below than average).

Table 1.
*Data of the first 17 COVID-19 patients: gender, age, fever, comorbidity, surgery of patients, and died soon or later.*

| Case | Gender | Age | 1st sympt | Comorbidity | Surgery | 1st sympt to death, above average? |
|---|---|---|---|---|---|---|
| 3 | M | 89 | No Fever | yes | NA | 0 |
| 4 | M | 89 | No Fever | yes | Yes | 0 |
| 6 | M | 75 | Fever | yes | Yes | 0 |
| 8 | M | 82 | No Fever | NA | NA | 0 |
| 10 | M | 81 | Fever | NA | NA | 0 |
| 11 | F | 82 | Fever | yes | NA | 1 |
| 13 | F | 80 | Fever | yes | NA | 0 |
| 15 | M | 86 | No Fever | yes | Yes | 1 |
| 16 | F | 70 | Fever | NA | NA | 0 |
| 17 | M | 84 | Fever | yes | Yes | 1 |
| 1 | M | 61 | Fever | yes | NA | 1 |
| 2 | M | 69 | Fever | yes | NA | 1 |
| 5 | M | 66 | Fever | yes | Yes | 0 |
| 7 | F | 48 | Fever | yes | NA | 1 |
| 9 | M | 66 | No Fever | NA | NA | 1 |
| 12 | M | 65 | No Fever | NA | NA | 0 |
| 14 | M | 53 | Fever | NA | NA | 1 |

If we place each patient according to age on the number line, mark the patient with a circle "o" if he does not die soon (1st symptom to dead above average), and mark the patient with a cross "x" if he dies very soon, then we obtain Figure 1.

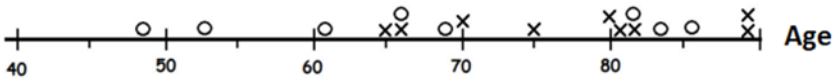


*Figure 1.* Age of the first 17 COVID patients

It is very interesting to note that elementary school students can do this type of translation of data from a table to positions on the number line. This is a typical math activity for those grade levels. The big difference with what they normally do is to place two types of marks: circles and crosses. That is to say, to position not only one mark as we usually ask them to do at school but with one of two possible marks. One type of mark represents the patients who will die soon and the other mark denotes those who will take longer to die. Then, taking advantage of our visual processing mechanisms, we immediately detect that there are two regions. One preferably with circles and another region with mostly crosses. The circles are mostly to the left, region I in Figure 2. The crosses, instead, are mostly towards the right of the number line, region II in Figure 2.

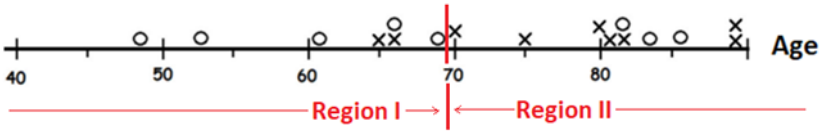


*Figure 2*. Region I and Region II

This means that older patients die sooner after displaying their first symptom. However, we can also visually identify a critical age: 69 years old. We can represent this pattern as a rule:

> If age less than 69
> Then the patient will not die soon
> Else, the patient will die soon.

Alternatively, we can represent the pattern with a decision tree (Figure 3).

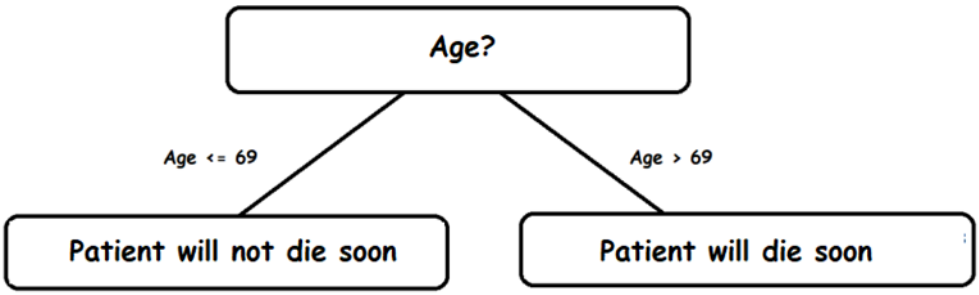


*Figure 3*. Decision tree induced from the data

Artificial intelligent agents will detect this pattern too. We can then predict their learning behavior in this activity. This way we are grasping some understanding of their psychology.

We can count the number of errors of this rule. In Figure 2, we see on the left of 69, that two crosses are misclassified, and in the right there are three circles misclassified. Therefore, the number of errors is five cases.

Now let us plot the information on a graph where age is on the horizontal axis and gender is on the vertical axis. We designate a value of 1 for the female gender and a value of 0 for the male gender. Therefore, the information is as in Figure 4.

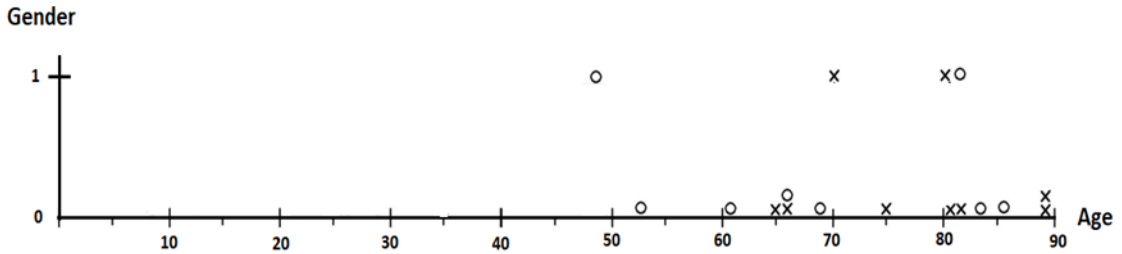


*Figure 4*. Age and gender of the first 17 COVID patients. Female: gender=1. Male: gender=0

Now we can think again about how to separate the circles from the crosses. We can imagine doing it using a straight line. This line is defined by the equation y = ax + b. In our notation y is gender, and x is age. Figure 5 shows one such straight line. The line defines two simple regions: one is above the line and the other is below the line.

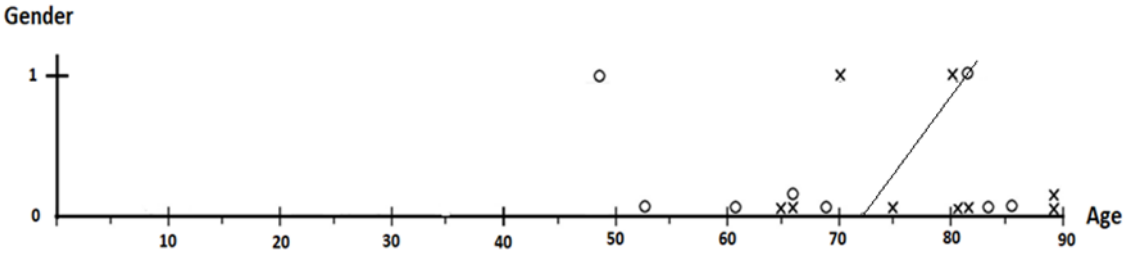


*Figure 5.* Age and gender of the first 17 COVID patients. Gender = a Age + b

Since the line passes through the points (72,0) and (82,1), to find a and b in Gender = a Age + b, we have to solve the following system of linear equations:

$$0 = 72a + b$$
$$1 = 82a + b$$

Therefore a=1/10, b = -72/10.

Thus, we have a new rule:

If Gender >= 1/10 Age – 72/10
Then the patient will not die soon
Else, the patient will die soon

We can rewrite it as:

If 10 Gender - Age + 72 >= 0
Then the patient will not die soon
Else, the patient will die soon

To the upper and left of the line of Figure 5, marks are mostly circles. This is Region I′ in Figure 6. Whereas to the bottom and right the marks are mostly crosses, shown as Region II′ in Figure 6. We have gained one circle in Region I′, which is mostly with circles, but we have lost two crosses in Region II′, which has mostly crosses. Therefore, now the number of errors is six cases. With respect to the previous separation, we have one extra error. However, this new region may perform better on new patient data that we have not yet registered.

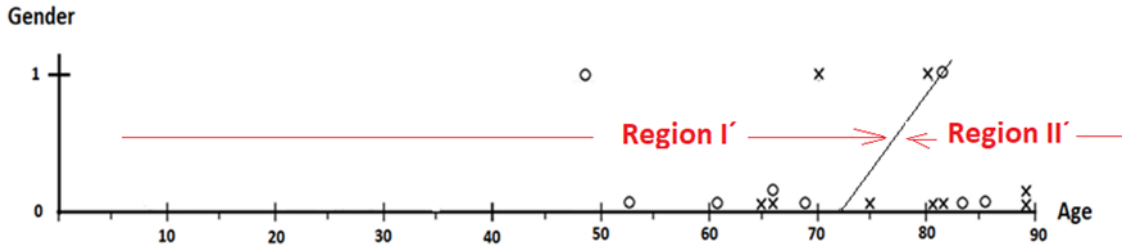


*Figure 6.* Region I′ and Region II′

Another way to represent this model is with neural networks. Figure 7 shows a neural network with three input neurons, each represented by a circle. The Age neuron, the Gender neuron, and the 1 neuron. In the Age neuron, the agent enters the age of the patient, in the Gender neuron, it enters the gender, and the third neuron is always with a one. Then the agent multiply these numbers by a, 1, and b, respectively. Then added. Next, the agent computes the sign, and this is the output of the neuron. This means, that if the previous sum is higher than zero then the output is +1, otherwise is -1. Thus, on the far right is the output neuron. If there is a 1 it means that the patient will not die soon. If it rolls a -1 then he will die soon.

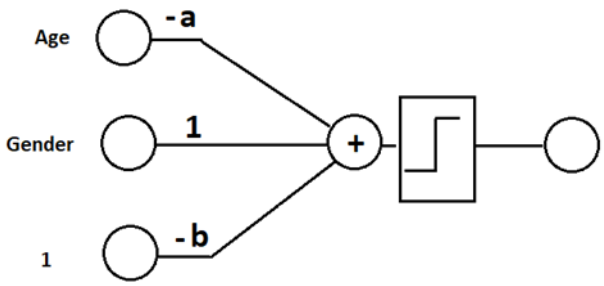


*Figure 7.* Neural network representation of the rule or decision tree

This very basic mechanism can always be operating, and thus the artificial intelligent agent is always learning. With more data from new patients, it will improve its knowledge. The agent will decide if the first model that only uses age, or if the other one that also uses gender has better accuracy. In addition, it will constantly adjust the parameters a and b to improve its accuracy.

How can the agent automatically adjust these parameters?

Let us imagine that we have now four extra patients, as shown in Figure 8. All four are females and are circles. This means they did not die very soon.

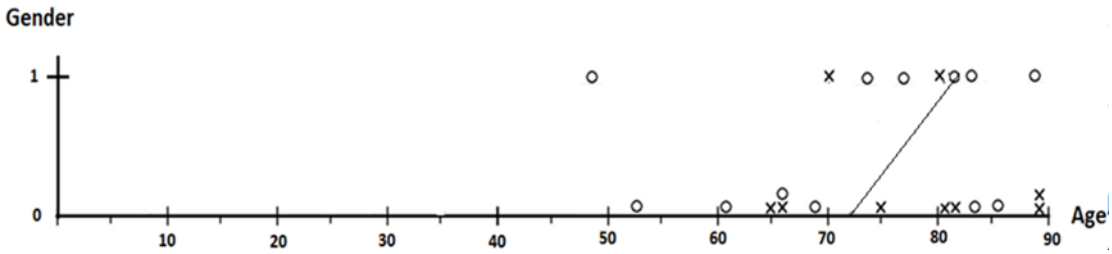


*Figure 8*. Age and gender of the first 17 COVID patients plus 4 new fictitious patients

In this new scenario, the previous model of Figure 2 has nine cases with errors, whereas the new model has eight errors (Figure 9). Thus, now, the new model with the straight line has fewer errors than the first model.

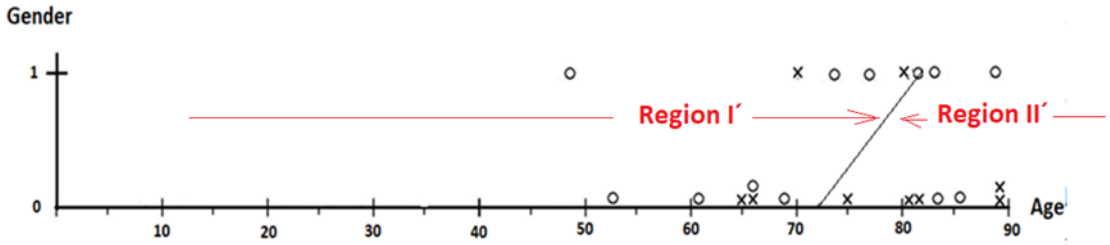


*Figure 9.* Region I´ and Region II´

How can we adjust the parameters a and b as new data starts to be gathered? One way is to continue plotting the data and use a ruler to better separate circles from crosses (Figure 10).

*Figure 10*. Changing the slope of the line, and therefore a and b, using a ruler

This is a good method but requires your eyes and your hands. The agent needs to do it autonomously. For this purpose, artificial intelligent agents use the method of the steepest descent (Araya, 2021b). Using the current values of a and b, they change them a little and compute the performance on each case to separate crosses from circles.

Let us consider the straight line defined by a=0.049 and b = -2.7. In this case, the misclassification error is eight. Figure 11 shows a small board with nine cells. Each cell corresponds to one of three values of a and one of three values of b. The previous values of a

and b are in the center of the board. The agent computes the error in the eight neighbor values of a and b. These are shown in the eight cells surrounding the cell corresponding to a=0.049 and b = -2.7. There are four cells with the number of errors equal to 7. Let us assume that the agent moves the parameters to a = 0.048 and b = -2.8. Then the error drops to 7. Next, the agent will redo the same procedure, and compute the errors in the eight cells surrounding the cell at a = 0.048 and b = -2.8. Let us assume that the agent will move the parameters to a = 0.048 and b= -2.9 and then the error drops to 6. Then again, the agent will try to continue descending, but this is not possible locally. It found a pair of values a and b in which the error no longer improves with local movements to the neighbor cells. Locally, this is the best error.

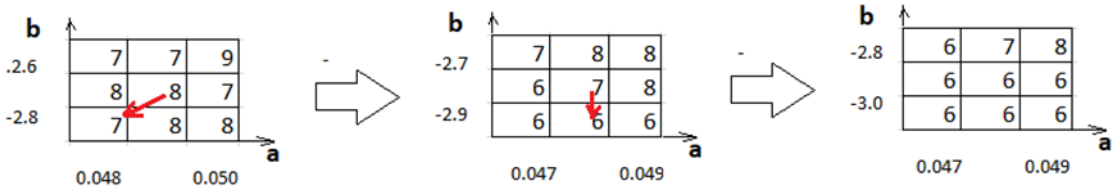


*Figure 11*. Changing a and b with local steepest descent

Figure 12 shows a broader picture of the error rates for a wider set of parameters a and b. This wider board shows that the optimal error is 6, and there is a range of pairs a and b that reach that optimal error.

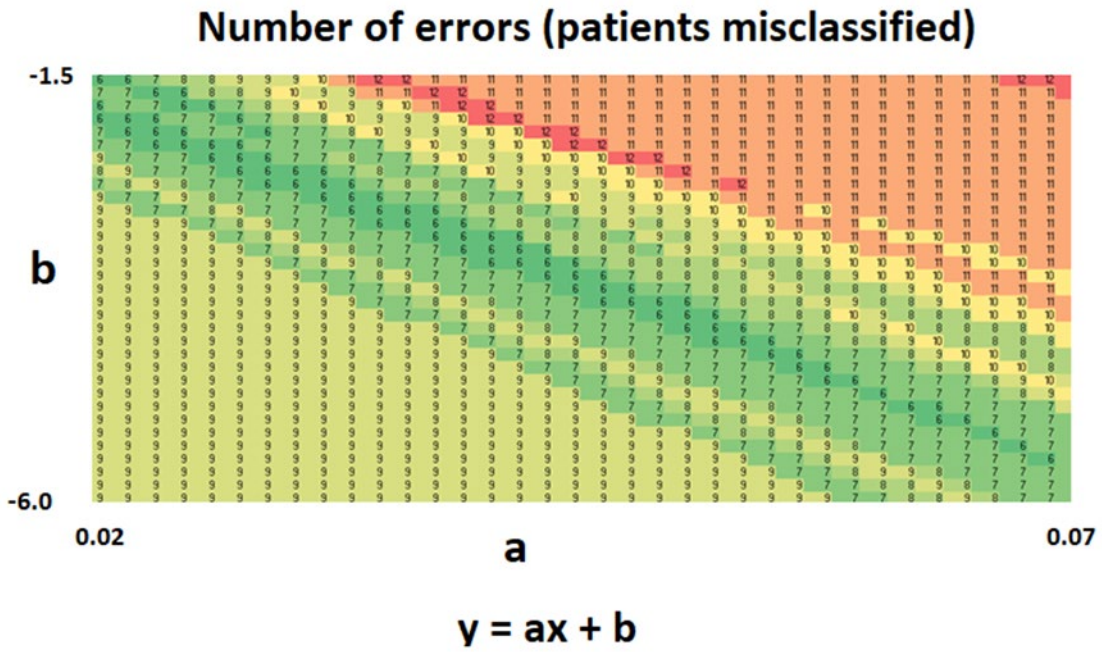


*Figure 12*. Number of classification errors for different values of a and b

In Figure 12 you can see stripes or areas of the same color. They are areas with the same error. That is, for those values of a and b the classification errors are equal. In addition, Figure 12 shows many regions in which the deepest descent algorithm gets stuck and cannot descend even though there are areas with less error. Moving to any neighboring square does not improve the error. In our case, starting from a = 0.049 and b = -2.7, it improved. The trajectory indicated in the sequence of Figure 11 is shown in a more global view in Figure 13.

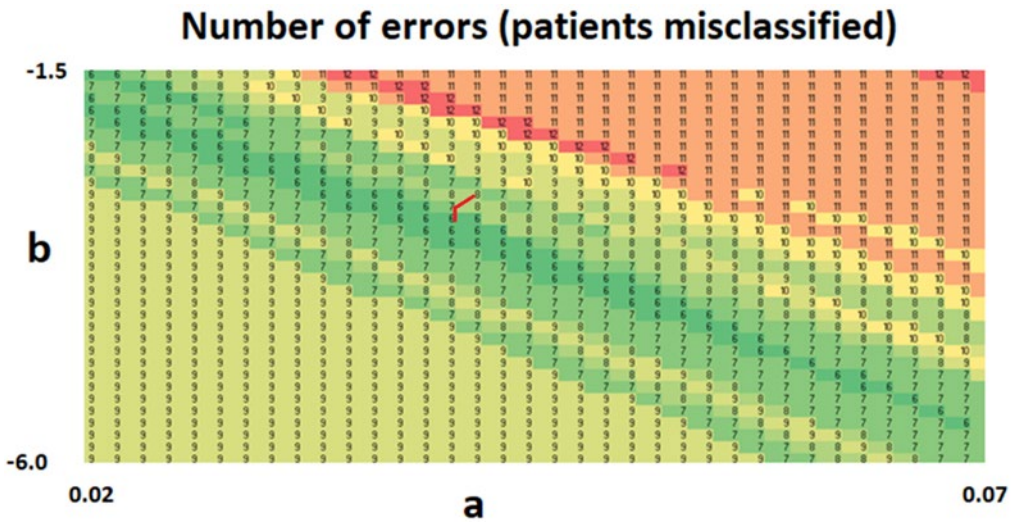


*Figure 13.* Learning trajectory starting from a= 0.049 and b= -2.7.

In Figure 13 we also see that the agent arrived at the best possible error. However, the learning mechanism does not have this global vision and can easily get caught up in a higher error.

You can imagine that the search for optimal a and b parameters is similar to how a raindrop falls down a landscape with hills and valleys. The height of each portion of the landscape is the numbers on the board, now interpreted as heights instead of errors. Starting from the initial position indicated in Figure 13 and at a height of 9, the drop runs through the territory descending to height 6. Figure 13 shows the complete path of the raindrop. It is important to note that from many positions the raindrop falls but does not reach height 6. It remains trapped in holes or valleys at a higher height. In other words, the agent will learn, but it may get trapped in a not very good situation and will not be able to improve.

What then can we learn from the psychology of these artificial agents? We can begin to understand how they learn.

First, if agents learn with this mechanism they need to be trained with examples. In this case, they are examples of patients, with their characteristics (demographic data, symptoms, and medical history) and their evolution. If there is too little data, they will not be able to learn something with good predictive ability. They could easily overgeneralize. If there is a patient with a severe fever, change in sense of smell, and continuous cough, but is 60 years old, the agents will predict it is not a serious case. Therefore, their learning will not be robust, and thus in these cases we should not trust them.

Second, if the data has errors or noise, then the agents will not learn much or acquire unreliable knowledge. For example, if the ages of some of the patients were not registered, then the number of errors computed on the test sample could increase.

Third, if the data has many biases, then the agents will learn those biases. For example, if the female patients have received treatments different from the male patients or treatments provided in different hospitals, then the model is misleading. It may be that age is not a critical variable but the treatment or the hospital is. We cannot trust this model unless we include this information and retrain the model.

Fourth, the agent will not use common sense. For example, in a case of a compromised person who has been in close contact with infected patients, then the agents will not consider this critical information.

Even when the agents have enormous cognitive capacity, connected to super servers, if the data is limited then they will not learn anything reliable and useful. The opinions, recommendations, and decisions of the artificial intelligent agents will not be very good, and therefore we cannot trust their recommendations and decisions.

**Machine Psychology: Self-supervised Learning**

Most of what a newborn learns is through interacting with the environment. No one directly teaches her. The baby makes many micro predictions, moves her hands and body, and contrasts directly with her predictions. For example, to navigate her surroundings or to grab a toy, she tries different actions and receives feedback from the environment. This way she learns autonomously, by herself (Koster et al., 2020). This predictive-processing mechanism is a self-supervised learning mechanism. Artificial intelligent agents are also using this type of cognitive mechanism. The most common case today is in learning with large language models (LLMs), like ChatGPT and Bard. One of the great advantages of self-supervised learning is that the agent does not need a teacher or databases with labeled cases. Teacher training and labeling are slow processes and can be very expensive. If it is possible to avoid them, it facilitates and speeds up the agents' learning process. The agents will then operate with better precision since they will be autonomously constantly learning.

There are two well-known self-supervised learning algorithms for learning text patterns.

One strategy to train an artificial agent to understand our language is to create an alphabetically ordered list of all the words. If we assume there are 100,000 words, the ordered list is a vector of 100,000 components. The word "house" will be represented in the vector with a 1 in the component depending on the position it occupies in the alphabet and a zero in the other components.

Consider now a big corpus of texts, examine all the sentences, and count for the word "house" the percentage of times that it occurs with the word "window" in the same sentence. Do the same with any other word instead of window. Now we can use a vector of length 100,000 for the word "house", but in which in each component is the percentage of times that it co-occurs with that other word. This way, we can have a new representation for each word. Now, we use a compression algorithm, like zip, and instead of having vectors with 100,000 components, we have vectors with 1,000 components. One compression algorithm is a neural network (Figure 14). Its inputs are the 100,000 components vectors, with a middle layer of 1,000 neurons, and you train in order that its output is the same 100,000 dimensions vector. Once trained, the compressed representation of "house" will be the numbers in the neurons in the middle layer when you input the 100,000 dimensions vector for "house".

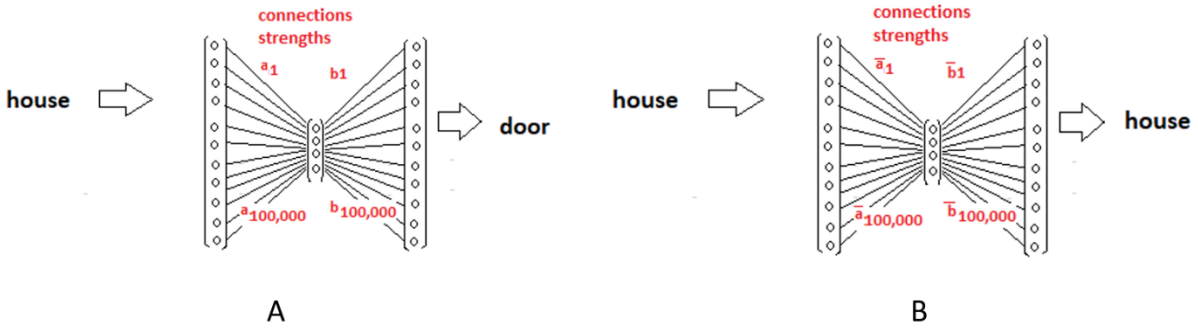


*Figure 14*. A) Initial neural network B) Neural network that learned to compress

A second strategy uses a different and now more popular self-supervised learning mechanism. It is a big neural network with several hidden layers (Figure 15). This is a deep neural network with millions of neurons. Therefore, it has billions of connections between those neurons. For each connection, a parameter indicates its strength. Thus, the network has billions of parameters.

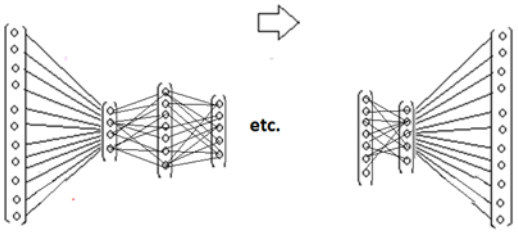


*Figure 15*. Deep neural network with m hidden layers

For each phrase it finds in the corpus, the agent hides a word and then adjusts the parameters to predict that word until it predicts correctly the word that was hidden. For example, in the next sentence

One xxxx Little Red Riding Hood went to visit her granny.

Here xxxx indicates the word that it hides. If we let the artificial agent read a big amount of text, like everything written on Wikipedia and the internet, then the agent can train itself with all kinds of text. For each sentence, it hides one word and uses a neural network to predict the hidden word. If it is not correct, it adjusts the parameters of the neural network. It is a very large network with billions of parameters, similar to the a and b parameters in the previous small network. The important thing is that it is the same adjustment mechanism as in the previous example, although in a mega network and with a big difference: no one trains and supervises the agent. He does it by himself, hiding words and then seeing if his prediction is correct. It is the self-supervised learning mechanism.

What do these self-supervised learning mechanisms tell us about the psychology of these artificial agents? Can they really understand our world?

A big question is whether artificial agents who only learn from reading and processing text, such as those who use LLMs, can really understand the world without ever having perceived it

with senses such as vision and never having acted on it. Lupyan (2023) gives as an example the case of people who are blind from birth. They can understand visual concepts such as transparency. He concludes that they do understand the abstract or metaphorical concept of transparency, but also the concrete and sensorial aspect such as the transparency of water and glass. (Lupyan, 2023) asks ChatGPT 3 about an umbrella, which has become an iconic example for testing understanding, and concludes that ChatGPT 3 does understand it. Pavlick (2023) postulates that it is not necessary to have grounding to understand the meaning of concepts. She gives the example of the meaning of color. LLMs can understand the concept of color. (Chalmers, 2023) compares GPT4 with or without vision and finds no great differences, and shows how GPT4 solves the problem with a chain of connected gears. If you turn a gear in a certain direction, the problem is to predict the direction of rotation of the other gears. This is a problem given previously by LeCun (2023) as an example of lack of comprehension of ChatGPT 3.

Abdou et al. (2021) and Søgaard (2023) compared the vectors associated with concepts, obtained by self-supervised learning methods, in English and those of another language such as German. They detected that there was a surprising lineup. They have the same structure. This means that an automatic translator can be built without supervision. Søgaard (2023) did the same with vectors generated with self-supervised learning from images. He concluded that the co-occurrence statistics of higher order concepts are stable between languages and between modalities. He explains that this is because the use of language reflects the world we live in and of course, this world is relatively stable.

This leads us to ask what it means to understand something, and whether artificial agents based on LLMs can understand the world. Mitchell and Krakauer (2023) ask whether understanding needs more than correlations but causal mechanisms. It is then important to compare with humans. Does humans' understanding rely on causality? This is a challenging question since it is not always clear when we attribute a cause to an event. (Ahn, 2022) identifies cues for causal attribution: similarity (for example similarity on the scale of events), necessity more than sufficiency, action more than inaction, abnormality (unusual events are more probably to be causes), recency (if events are close in time), and controllability (if events are such that we can control them). All these cues modulate our sense of understanding, even though they are not strictly causal mechanisms.

When we do not understand something that someone asks us, our emotional system gives us away. Lies shine even in the smallest components of language (Pennebaker, 2011). Others can detect these critical honest signals. When we do not have a lot of information, this motivates us to avoid inventing too much and thus take care of our reputation. However, without enough data, LLM-based artificial agents can easily hallucinate. Given their lack of similar emotional engines, they do not care. They do not secrete those unconscious honest signals. These signals are critical for trust. Therefore, the lack of these signals makes it very difficult to detect any misinformation. Given the great ability and creativity of LLMs to write fluent prose, when there is not enough data, LLMs-based artificial agents may overgeneralize and hallucinate in a very convincing form making it very difficult for us to detect lies and hallucinations.

Another problem with LLMs is that they learn from everything we have written. However, that does not mean that they consider what we say (verbal speech). This has a different language structure. In addition, LLM-based learning does not consider what our children write and speak.

This is also a different language structure, with misspellings and ways of expressing themselves differently than adults. One of the limitations of LLM-based learning is that it is not based on what we have done, but on what we declare in writing.

**Explainable AI: To Trust We Need Some Basic Understanding**

How can we understand the decisions of artificial agents? Some of the decisions are the results of deep neural networks. They involve thousands of parameters. In the case of LLMs, they involve hundreds of billions of parameters. This is a truly complex black box. However, in order to gain trust in artificial agents we need to develop an explainable AI (XAI). The dimension of the challenge is somehow similar to understanding other humans and ourselves. We have hundreds of trillions of parameters in our brains, yet we can still predict behaviors of others, and provide explanations to them. We can communicate these explanations, understand them, discuss them, contrast them, and learn from conversations with other people.

This means that we need to understand the psychology of these artificial agents with human understandable explanations of the decisions that the agents make (Khosravi et al., 2022). For example, the European General Data Protection Regulation establishes the right of citizens to obtain explanations about the automated decisions that affect them (Blanco-Justicia & Domingo-Ferrer, 2019). These have to be human-friendly explanations.

No one can manage billions of parameters. We have a bounded rationality, where most decision mechanisms are simple heuristics. They use one or at most a couple of variables (Gigerenzer & Todd, 1999). Moreover, linear or nonlinear combinations of variables are not easy to grasp. Combinations require finding a common currency in which every variable translates. This is not how our mind understands.

For the human mind, decision trees are more understandable than other mechanisms. At least we can understand decision trees when they have few nodes. It would be ideal then to be able to express the decision mechanisms of an artificial agent with decision trees. This is the language that humans understand. Actually, any decision tree can be rewritten with neural networks (Araya & Gigon, 1992; Sethi, 1990). The reverse process is not always possible. However, there is a growing number of strategies to find an approximate answer to this problem. For a given neural network, XAI algorithms propose decision trees that approximate the results of the network.

We have already seen that the regions defined by a linear inequality can be written with a single neuron. However, a rule based on whether 18 Gender - Age + 72 is positive or negative, is not human-friendly (Figure 16a). It is weird to have a weighted mix between Age and Gender. Instead, we can approximate this rule with a simpler decision tree that uses Age and Gender separately. We can approximate the region as in Figure 16b. In this case the rule is:

> If gender is male and age less than 72 or
> gender is female
> Then the patient will not die soon
> Else, the patient will die soon

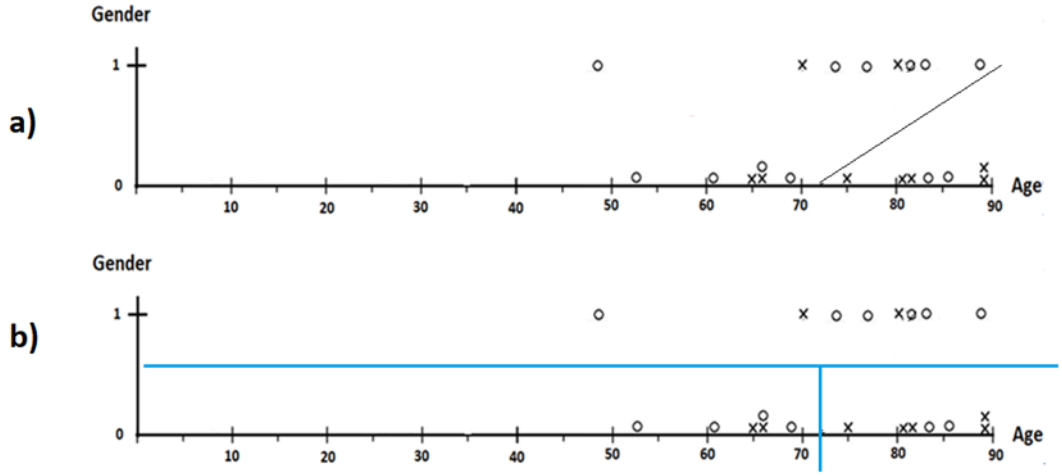


*Figure 16.* a) System 1 is rule based in a linear inequality or a similar neural network. b) System 2 is symbolic processing. It is closer to human reasoning

We have reviewed two popular mechanisms of learning. However, in physical robots, several other mechanisms are necessary for their behavior. They range from algorithms for navigation to the balance of their bodies. Other algorithms control the use of their hands for different tasks. There are others for the detection of faces and gazes, and establishing eye contact with people. Other mechanisms deal with emotional relations such as the detection of emotional states by processing faces and body positions. Critical are algorithms for non-verbal communication like the detection of gestures and movements of people. Then, there are also algorithms for social interaction such as synchronization and turn-taking with other robots and humans. There are also the simulators that the artificial agents run to predict the consequences of their actions on humans and other artificial agents. Other algorithms deal with ethical values in decisions. These are all fascinating areas of growing development and are part of the core psychology of artificial agents. Knowing these mechanisms allows the citizen to improve in being able to predict the behavior of artificial agents and thus be able to determine which ones to trust.

All these mechanisms use various models based on principles of mechanics, pattern recognition, image and video processing, signal processing, voice recognition, language processing, as well as social and ethical psychology. That is, it means that students need to learn knowledge from a great variety of disciplines and have the ability to integrate them.

This is an opportunity and a need to integrate more strongly and in greater depth the STEM disciplines. Additionally, it includes the integration of several disciplines: language, communication, psychology, social sciences, and philosophy. In other words, this means being able to integrate the humanities and thus develop STEAHM.

**Lesson Study 2.0**

We need to teach a number of core mathematical models of the psychology of artificial agents. These are completely unknown models to teachers. Teachers did not study them when they were students in schools or in their training as teachers. In addition to not having knowledge of these contents, they have no experience of how to teach this type of content. They also do not know other teachers in their community with knowledge and experience in teaching it. They cannot observe, imitate, and learn from others. In addition, students need to acquire not

only new content but new social skills and attitudes in order to work, collaborate and socialize with artificial agents. This is a great educational challenge. Therefore, we need a very efficient teacher professional development strategy that in a short time can allow teachers to learn artificial agent psychology and know how to teach the set of core mathematical and computational models of the psychology of artificial agents.

Over the years, thanks to a well-planned curriculum and precise instruction, we have managed to teach how to escape some of the evolutionary traps. This type of teaching takes years of directed instruction and deliberate practice. Adapting teaching practices and having teachers assimilate new curricula to this new set of core models is a major educational challenge.

A powerful and proven strategy for developing teacher skills is to establish a community with a strong attitude of sharing, innovation, and collaborative learning. One such type of community-based strategy is lesson study and mass demonstrations in open public lessons (Isoda, 2015). It began in 1880 in Japan with the aim of reproducing the best practices in teaching (Isoda, 2015). Nowadays Thailand (Inprasitha, 2015), Singapore (Yeap et al., 2015), and many other countries have started to introduce it (Estrella et al., 2018; Quaresma et al., 2018).

Given the magnitude of new contents and abilities that we have to introduce, the strangeness of the psychology of artificial agents, and the speed required to make transformations in sync with the constant emergence of more powerful agents, a logical strategy is to do so by creating a community with artificial agents. That is, create a community of teachers together with artificial agents, who carry out lesson studies. This community, which we call Lesson Study 2.0, would plan lessons, make class observations, analyze them, provides feedback, monitor, and continuously improve sessions, whether face-to-face, online, or blended learning.

Recent advances in natural language processing, such as ChatGPT, pattern recognition in audios, images and videos, and machine learning (Araya & Sossa-Rivera, 2021; Lämsä et al., 2021; Lehesvuori et al., 2023; Urrutia & Araya, 2023; Urrutia & Araya, in press) provides the technical tools to build a community with intelligent agents that could help us with lesson study.

## Discussion

The growing eruption of artificial intelligent agents creates a population of very strange agents that surround us. They have a completely different psychology. In addition, they are very intelligent and are learning at a mind-blowing speed. This creates a dilemma never seen before: can we trust artificial agents and which ones? To face this dilemma, we need to understand how they behave and decide. This means several challenges.

First, artificial agents are very different from us. We know that people are complex agents. We try to change when being perceived by opponents and avoid being predictable. However, given our basic universal motivation engines for survival, reproduction, attachments, and kin altruism, and our bounded rationality, we have behavior patterns and stable personality traits that help us infer reactions from others. Furthermore, we have evolved adaptations to interact between us and know whom to trust. An artificial agent can emulate some of these characteristics, but its motivational engines, cognition, and behavior are very different from ours. An artificial agent “has no soul, and no one knows what it may be thinking” (Asimov,

1940). Already in 1863, Butler reflected on the nature of these artificial agents: "What sort of creature man's next successor in the supremacy of the earth is likely to be?" and then he conjectures about the weird emotions they could have: "No evil passions, no jealousy, no avarice, no impure desires will disturb the serene might of those glorious creatures."

Second, artificial agents are having increasingly deep access to our thoughts and feelings. Imagine a micro-phenomenological artificial agent (Araya, 2023), that performs machine-to-brain intersubjective nonverbal communications (Schore, 2021), directly scanning and interrogating our brain. If we have a newly emerging problem-solving strategy that is not yet conscious, then this type of agent could detect it and immediately interview us to bring it to consciousness. Would such an agent manipulate us in the same way that parasitic hairworms manipulate the behavior of insect hosts? These hairworms drive the insects to commit suicide by jumping into an aquatic environment that the hairworms require to continue their life cycle (Thomas et al., 2002). Asimov's well-known first law of robotics states: "a robot may not injure a human being or, through inaction, allow a human being to come to harm" (Asimov, 1942). We all agree on this basic principle. It seems that if all artificial agents have hardwired this law, then we are safe. However, it is not easily implementable. It is still very complex if not impossible at this time to automatically detect harm or predict that something will cause harm to someone.

Third, these artificial agents are very intelligent and even creative. Initially, there was skepticism about whether artificial agents could be creative. In 1853, Lovelace argued that computers cannot be creative. They only do what programmers tell them to do. "The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform" (Lovelace, 1853). According to (Turing, 1950) a better variant of the objection says that a machine can never "take us by surprise." He claims that this indeed happens frequently because he does not do sufficient calculations to decide what to expect them to do, or although he does the calculation, he does it carelessly. However, recent measurements conclude that ChatGPT is in the top 1% in one of the classical Torrance Tests of Creative Thinking (Shimek, 2023). Moreover, we have to consider not only the intelligent behavior of individual agents but also the collective intelligence of populations of artificial agents. Unlike us, they learn instantly from each other. This will boost a very fast cultural evolution tsunami, with permanent radical innovations that will constantly challenge us.

Fourth, artificial agents sometimes have weird and unreliable behaviors. They hallucinate and overgeneralize. In addition, they are severely biased and are very efficient in opinion manipulation (Jakesch et al., 2023). Studying ethical-political sensibility, (Martin, 2023) found that the current version of ChatGPT has a measurable left position in the political space and a concomitant position in the social space among the privileged. According to Marcus, we are approaching a kind of information-sphere disaster (Marcus, 2023). The supply of disinformation will soon be infinite (DiResta, 2020).

Fifth, their decision structure seems to be a black box mechanism. According to Marcus (2023), LLMs are repeating behaviorism mistakes, predicting words and not actions. We cannot have explanations if we only rely on external behavior. Therefore, LLMs do not generalize well, and we already have self-driving cars crashing into trailers or into people in improbable events like holding a stop sign. We need some knowledge of their internal representations in order to understand them. That means we need to describe the agents' behavior in intentional language,

and then, based on internal states and mechanisms that account for state changes. Something like beliefs and desires that are very useful, of which we only have indirect evidence, but that help us understand ourselves and predict the behavior of others and ourselves.

## Conclusions

As artificial intelligence becomes increasingly sophisticated, we are faced with a fundamental challenge: can we trust these intelligent agents? Will they manipulate us? This paper analyzes to what extent the citizen can obtain understandable explanations of the artificial agents' behavior in order to be able to trust their decisions. We have also studied an example that shows that we can adjust contents of the current mathematics curriculum (localization of numbers in the number line, x-y graphs, and linear equations and inequations) to include basic models that allow students to reach some understanding of the behavior and decisions of artificial agents.

We foresee great opportunities in the collaboration between humans and artificial agents. Thus, we have to learn the psychology of cooperation with them. This means, being able to develop the ability to read the artificial agents' minds, in the same form as how we cooperate with humans by reading their minds (Markievicz et al., 2023). The incredible economic and social development our civilization has experienced in the last centuries is due in great part to our increasing ability to trust strangers, to cooperate with them, and trust in abstract rules (Henrich, 2020). We need now to continue with this critical historic trend, creating a relationship of genuine trust and fair cooperation with artificial agents.

However, the mathematical applications that we teach in schools is not always up to the task of addressing the challenges of our time. The mathematical models that we teach were mainly developed in agricultural and previous commercial societies. It is not well suited to the challenges of a world populated with artificial agents. We need to update the mathematical models that we teach in schools to reflect the challenges of our time. We need to teach mathematical and computational models that can help us to understand these agents and collaborate effectively with them.

In this paper, we have shown a simple example of the supervised learning mechanism that is at the base of an algorithm widely used in Machine Learning. The example illustrates another way to represent and use the familiar equation of the line $y = ax + b$, which all middle school students learn. Other models with agents can be found in (Araya, 2021b; Araya, 2021c; Araya, 2022). The good news is that now we have an excellent opportunity to integrate more deeply STEM disciplines. We also have now an incredible bridge between two traditionally distant disciplines: mathematics and language. LLMs provide a great bridge between those two worlds. Additionally, this integration is accomplished largely by reusing core math concepts. That is the case of the equation of the straight line that now we can see as the atom of a neural network.

Finally, I have proposed that a good way to teach these mathematical models is through Lesson Study 2.0. These are lesson studies enhanced with AI support. Recent advances in natural language processing, pattern recognition in videos, and machine learning provide us with a powerful microscope to analyze and rapidly adjust teaching practices to these new teaching challenges. The suggestion is to create an educational community of teachers with artificial agents specialized in education. Together we can cooperate and create lessons and teaching

strategies that are more effective, and that can quickly adapt to a future with a high acceleration of radical transformations.

**Acknowledgement**

Support from ANID/PIA/Basal Funds for Centers of Excellence FB0003 is gratefully acknowledged.

**References**

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? A case study in color. *Proceedings of the 25th Conference on Computational Natural Language Learning* (CoNLL), 109–132

Ahn, W. (2022). *Thinking 101.* Flatiron Books.

Alfonseca, M., Cebrian, M., Fernández Anta, A., Coviello, L., Abeliuk, A., & Rahwan, I. (2021) Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research, 70* (2021) 65–76

Allen, J. G., (2006). Mentalizing in practice. In J. G. Allen & P. Fonagy (Eds.), *Handbook of Mentalization-Based Treatment,* 3–30. John Wiley & Sons.

Araya, R. (2023). Unraveling a royal road to math education. *Constructivist Foundations, 18*(2), 283–286. https://constructivist.info/18/2

Araya, R. (2022). Is it feasible to teach agent-based computational modeling to elementary and middle school students? *Proceedings of the Singapore National Academy of Science, 16*(1), 71–84. https://www.worldscientific.com/doi/10.1142/S2591722622400063

Araya, R. (2021a). What mathematical thinking skills will our citizens need in 20 more years to function effectively in a super smart society? In Inprasitha, M., Changsri, N., & Boonsena, N. (Eds). *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* , Vol. 1, 48–65. Khon Kaen, Thailand: PME.

Araya, R. (2021b). Enriching elementary school mathematical learning with the steepest descent algorithm. *Mathematics* 2021, *9*(11), 1197. https://doi.org/10.3390/math9111197

Araya R. (2021c). Gamification strategies to teach algorithmic thinking to first graders. In Nazir S., Ahram T.Z., Karwowski W. (Eds). *Advances in Human Factors in Training, Education, and Learning Sciences. AHFE 2021*. Lecture Notes in Networks and Systems, vol 269. Springer, Cham. https://doi.org/10.1007/978-3-030-80000-0_16

Araya, R., Isoda, M., & van der Mollen Moris, J. (2021). Developing computational thinking teaching strategies to model pandemics and containment measures. *International Journal of Environmental Research and Public Health, 18*(23), 12520.

Araya, R., & Sossa-Rivera, J. (2021). Automatic detection of gaze and body orientation in elementary school classrooms. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.729832

Araya, R., & Gigon, P. (1992). Segmentation trees: A new help building expert systems and neural networks. In Dodge, Y., & Whittaker, J. (Eds). *Computational Statistics*, 119–124, Physica HD.

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, *22,* 299–324.

Asimov, I. (1940). Robbie, *I, robot*. Bantam Books.

Asimov, I. (1942). Runaround, *I, robot*. Bantam Books.

Asimov, I. (1950). *The evitable conflict.* Street & Smith.

Bateman, A, & Fonagy, P. (2004). *Psychotherapy for borderline personality disorder: Mentalization-based treatment*. Oxford University Press.

Blanco-Justicia, A., & Domingo-Ferrer, J. (2019). Machine learning explainability through comprehensible decision trees. In Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (Eds). *Machine Learning and Knowledge Extraction*. https://doi.org/10.1007/978-3-030-29726-8_2

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4* (1), 15–31.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* Vol. 2, Int. Institute of Advanced Studies in Systems Research and Cybernetics, 12–17.

Butler, S. (1863). Darwin among the machines. *The Press, Christchurch, New Zealand*. https://web.archive.org/web/20060524131242/http://www.nzetc.org/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html

Chalmers, D. (2023). Do language models require sensory grounding for meaning and understanding? NYU Center for Mind, Brain and Consciousness. [Video] https://youtu.be/x10964w00zk?t=2557

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. Norton. NY.

De Querol, R. (2023). Isaac Asimov's disturbing message for 21st-century humankind. *El Pais*. https://english.elpais.com/science-tech/2023-06-17/isaac-asimovs-disturbing-message-for-21st-century-humankind.html

Dennett, D. (2023). The Problem With Counterfeit People. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

DiResta, R. (2020). The Supply of Disinformation Will Soon Be Infinite. *The Atlantic.* https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/

Estrella, S., Mena-Lorca, A., & Olfos, R. (2018). Lesson study in Chile: A very promising but still uncertain path. In Quaresma, M., Winsløw, C., Clivaz, S., da Ponte, J., Ní Shúilleabháin, A., & Takahashi, A. (Eds.). *Mathematics lesson study around the world, ICME-13 Monographs,* 105–122. Springer, Cham.

Gigerenzer, G., & Todd, P. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Haidt, J. (2006). *The happiness hypothesis: Finding modern truth in ancient wisdom*. Basic Books

Harari, Y. (2023). Human nature, intelligence, power, and conspiracies, Lex Fridman Podcast. [Video] https://youtu.be/Mde2q7GFCrw?t=1919

Henrich, J. (2020). *The WEIRDest people in the world: How the west became psychologically peculiar and particularly prosperous*. New York: Farrar, Straus and Giroux.

Hinton, G. (2023). Geoffrey Hinton talks about the "existential threat" of AI. [Video] https://www.technologyreview.com/2023/05/03/1072589/video-geoffrey-hinton-google-ai-risk-ethics/

Inprasitha, M. (2015). Transforming education through lesson study: Thailand's decade long journey. In Inprasitha, M., Isoda, M., Wang-Iverson, P., & Yeap, B. H. (Eds). *Lesson Study Challenges in Mathematics Education*, 213–228. Singapore: World Scientific.

Isoda, M. (2015). The science of lesson study in the problem solving approach. In Inprasitha, M., Isoda, M., Wang-Iverson, P., & Yeap, B. H. (Eds). *Lesson Study Challenges in Mathematics Education*, 81–108. Singapore: World Scientific.

Isoda, M., Araya, R., Eddy, C., Matney, G., Williams, J., Calfucura, P., Aguirre, C., Becerra, P., Gormaz, R., Soto-Andrade, J., Noine, T., Mena-Lorca, A., Olfos, R., Baldin, Y., & Malaspina, U. (2017). Teaching energy efficiency: a cross-border public class and lesson study in STEM. *Interaction Design and Architecture(s) Journal, 35* 7–31. https://doi.org/10.55612/s-5002-035-001

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Interacting with opinionated language models changes users' views. https://mauricejakesch.com/assets/pdf/aimc_influence.pdf

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux. N.Y.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R.,Sadig, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074.

Koster, M., Kayhan, E., Langeloh, M., & Hoehl, S. (2020). Making sense of the world: Infant learning from a predictive processing perspective. *Perspective on Psychological Science,* 15(3): 562–571.

Lämsä, J., Uribe, P., Jiménez, A., Caballero, D., Hämäläinen, R., & Araya, R. (2021). Deep networks for collaboration analytics: Promoting automatic analysis of face-to-face interaction in the context of inquiry-based learning. *Journal of Learning Analytics, 8*(1), 113–125.

LeCun, Y. (2023). Do language models require sensory grounding for meaning and understanding? NYU Center for Mind, Brain and Consciousness. [Video] https://youtu.be/x10964w00zk?t=3140

P., Jiménez, A., Caballero, D., Hämäläinen, R., & Araya, R. (2021). Deep networks for collaboration analytics: Promoting automatic analysis of face-to-face interaction in the context of inquiry-based learning. *Journal of Learning Analytics, 8*(1), 113–125.

Lehesvuori, S., Schlotterbeck, D., Jimenez, A., Caballero, D., Araya, R., & Hämäläinen, R. (2023). Towards automatic analysis of science classroom talk: Focus on teacher questions. In Bansal, G., & Ramnarain, U. (Eds). *Fostering Science Teaching and Learning for the Fourth Industrial Revolution and Beyond*, 123–146. IGI Global.

Lovelace, A. A. (1843). Notes by the translator. *Taylor's Scientific Memoirs, 3,* 666–731. https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by_Charles_Babbage,_Esq./Notes_by_the_Translator

Lupyan, G. (2023). Do language models need sensory grounding for meaning and understanding? NYU Center for Mind, Brain and Consciousness. [Video] https://youtu.be/x10964w00zk?t=4022

MacAskill, W. (2022). *What we owe the future*. Basic Books. NY.

Marcus, G., & Davies, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Penguin Random House.

Marcus, G. (2023). Why are we letting the AI crisis just happen? *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376/

Markiewicz, R., Rahman, F., Apperly, I., Mazaheri, A., & Segaert, K. (2023). It is not all about you: Communicative cooperation is determined by your partner's theory of mind abilities as well as your own. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001268

Martin, J. (2023). The ethico-political universe of ChatGPT. *Journal of Social Computing, 4*(1), 1−11. https://doi.org/10.23919/JSC.2023.0003

Mitchell, M., & Krakauer, D. (2023). The debate over understanding in AI's large language models. https://arxiv.org/abs/2210.13966

Mutesa, L., Misbah, G., Remera, E., Ebbers, H., Schalkem E., Tuyisenge, P., Sindayiheba, R., Igiraneza, C., Uwimana, J., Mbabazi, D., Kayonga, E., Twagiramungu, M., Mugwaneza, D., Ishema, L., Butera, Y., Musanabaganwa, C., Rwagasore, E., Twele, F., Meller, S., … Nsanzimana, S. (2022). Use of trained scent dogs for detection of COVID-19 and evidence of cost-saving. *Frontiers in Medicine.* 9:1006315.

Pavlick, E. (2023). This debate will largely be settled empirically. And current empirical data points to "no". [Video] https://youtu.be/x10964w00zk?t=955

Pennebaker, J. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Quaresma, M., Winsløw, C., Clivaz, S., da Ponte, J., Ní Shúilleabháin, A., & Takahashi, A. (2018). *Mathematics lesson study around the world, ICME-13 Monographs*. Springer, Cham.

Ruocco, M.; Mou, W., Cangelosi, A., Jay, C., & Zanatto, D. (2021). Theory of mind improves human's trust in an iterative human-robot game. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (HAI '21), 227–234. https://doi.org/10.1145/3472307.3484176

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S.; Romero, P., Sun, L., Abdulhai, M., Faust, A., & Mataric, M. (2023). Personality traits in large language models.
https://arxiv.org/pdf/2307.00184.pdf

Schelling, T. (1978). *Micromotives and macrobehavior*, Norton.

Schore A. N. (2021). The interpersonal neurobiology of intersubjectivity. *Frontiers in Psychology, 12,* 648616. https://doi.org/10.3389/fpsyg.2021.648616

(1978). *Micromotives and macrobehavior*, Norton.

Sethi, I. K. (1990). Entropy nets: from decision trees to neural networks. *Proceedings of the IEEE, 78*(10), 1605–1613.

Shimek, C. (2023). AI tests into top 1% for original creative thinking.
https://techxplore.com/news/2023-07-ai-creative.html

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science*, *2* (1): 125–134. https://doi.org/10.1287/orsc.2.1.125

Søgaard, A. (2023). Grounding the vector space of an octopus: WSord meaning from raw text. *Minds and Machines, 33,* 33–54. https://doi.org/10.1007/s11023-023-09622-4

Stanovich, K., & Toplak, M. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence, 11*, 27. https://doi.org/10.3390/jintelligence11020027

Stanovich, K., West, R., & Toplak, M. (2016). *The rationality quotient. Towards a test of rational thinking.* MIT Press

Stanovich, K. (2020). Why humans are cognitive misers and what it means for the great rationality debate. In Viale, R. (Ed.), *Routledge Handbook of Bounded Rationality.*
https://doi.org/10.4324/9781315658353

Thomas, F., Schmidt-Rhaesa, A., Martin, G., Manu, C., Durand, P., & Renaud F. (2002). Do hairworms (Nematomorpha) manipulate the water seeking behaviour of their terrestrial hosts? *Journal of Evolutionary Biology, 15*(3), 356–361. https://doi.org/10.1046/j.1420-9101.2002.00410.x

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460. https://psycnet.apa.org/doi/10.1093/mind/LIX.236.433

Urrutia, F., & Araya, R. (2023). Automatically detecting incoherent written math answers of fourth-graders. *Systems 2023*, *11*, 353. https://doi.org/10.3390/systems11070353

Urrutia, F., & Araya, R. (in press). Who's the best detective? Large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *Journal of Educational Computing Research*.

Van Gugt, M., De Vries, L., & Li, N. (2020). The evolutionary mismatch hypothesis implications for social psychology. In Forgas, J., Crano, W., & Fiedler, K. (Eds). *Applications of Social Psychology*. Abingdon, England: Routledge.

Wang, W., Tang, J., & Wei, F. (2020). Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *Journal of Medical Virology*. 29 January 2020 https://doi.org/10.1002/jmv.25689

Wiener, N. (1960). Some Moral and Technical Consequences of Automation. *Science, New Series*, Vol. 131, No. 3410 (May 6, 1960), 1355–1358.

Yeap, B. H., Foo, P., & Soh, P. S. (2015). Enhancing mathematics teachers' professional development through lesson study: A case study in Singapore. In Inprasitha, M., Isoda, M., Wang-Iverson, P., & Yeap, B. H. (Eds). *Lesson Study Challenges in Mathematics Education*, 153–168. Singapore: World Scientific.

Zuckerberg, M. (2023). Future of AI at Meta, Facebook, Instagram, and WhatsApp. Interview by Lex Fridman. [Video] https://youtu.be/Ff4fRgnuFgQ?t=8150

**Author:**
**Roberto Araya,** Institute of Education, University of Chile, Chile,
roberto.araya.schulz@gmail.com