Bulgarian Academy of Sciences
Institute of Mathematics and Informatics
Department of Algebra and Logic

Summary of the PhD Thesis

Artificial Intelligence – Definition, Realization and Consequences

What is it, how can we do it and what shall we do once we do it?

submitted by

Dimiter Dimitrov Dobrev

Professional domain: Mathematics (4.5)
Scientific discipline: Mathematical Logic

Tutor: Assoc. Prof. Lyubomir Ivanov, PhD

Sofia, 2024

This summary of PhD thesis is comprised of 34 pages, of which 3 pages of introductory sections, 27 pages of body text and 4 pages of final remarks, including a References section in which 32 titles are listed.

Contents

Introduction

The objective of this PhD thesis is to dispel two misconceptions related to Artificial Intelligence (AI) and discuss the three most important questions about AI. The two misconceptions assert that AI is a memoryless function and that AI is pseudoscience. The three questions we are going to discuss are “What is AI?”, “How can we create it?” and “What consequences will the creation of AI entail?”.

The **first misconception** describes AI as a memoryless function. This misconception is referred to as *Full Observability*, meaning that AI sees everything it has to and therefore does not need to remember anything that has happened before. There are authors who also mention the opposite, i.e. *Partial Observability*, however, this thread has remained vastly unexplored. Most researchers assume that the admission of a hidden state (memory) will make the problem highly complicated to an extent that it becomes unsolvable. Actually, the case of interest is exactly *Partial Observability* and the refusal to explore this case is a serious mistake.

Our statement that most authors regard AI as a memoryless function is not accurate enough. It would be more precise to say that what they perceive as a memoryless function is the *trained AI*. We can assume that AI is the program which searches for the trained AI (i.e. AI is a program which trains itself). The outcome of the learning process is the trained AI, so the question of interest is what a trained AI looks like. Most authors examine the set of pairs <*x*, *y*> (where *x* is the question and *y* is the answer). From this set they seek to derive a function which to each question *x* returns the corresponding answer *y*. This problem is known as *interpolation*. When the search for the interpolating function takes place within the set of neural networks, it is referred to as *neural networks*.

We are far from underestimating the Neural networks development. This approach has produced astonishing results and has thus become the mainstream method in the area of AI. Nevertheless, neural networks are limited by the lack of memory which is an important setback.

The secret of AI hides in the hidden state of the world. If we discard the hidden information, we will miss the most interesting part. The basis of AI is the “understanding” of the world, and to understand the world we must be able to describe it by means of some language for description of worlds. This is the reason why the present PhD thesis will be so much focused on discussing various languages for description of the world. The description of the hidden state is not unambiguous because there are many possible explanations of the world. We will regard the various descriptions as various hypotheses so we can conduct experiments in order to identify the correct hypothesis. We can describe the search for various descriptions (hypotheses) with the term *imagination*. Thus, AI without a hidden state is AI without imagination.

The second misconception asserts that AI is pseudoscience, therefore serious scientists should stay away from it. This is about genuine AI or Artificial General Intelligence (AGI). Most people are convinced that AGI is something impossible (science fiction). When it comes to AI, most people assume that they are talking about weak AI, which is not AI, but an imitation of AI. The very idea of weak AI was created by people who do not believe in AGI and are deeply convinced that it is something completely impossible.

This is the reason why society views those working in the field of AGI as crazy. We are lumped in with the UFO seekers and perpetuum-mobile inventors. I myself take perpetuum-mobile inventors as crazy, and I shouldn't be angry that this label is being applied to me as well.

However, in science, no idea should be rejected a priori. Many things have seemed impossible, but it turns out that their time just hasn't come yet. The main reason people fail to realize an idea is because they don't believe it's possible. The moment they believe, they are able to solve the problem very quickly. Usually they see that someone else has already realized the idea and they immediately repeat his success. Back in the time, people thought that machines would never be able to play chess. Now, a chess game program is something simple that can be explained and written in a few hours (I do it in my programming course).

The misconception that serious scientists should not deal with AI is a serious issue since it leads to lagging behind the developments in this area. A similar misconception existed many years ago in respect of cybernetics, which led to serious lagging in the area of computers (especially in the former Eastern Bloc countries).

Even among mathematical logicians there is some skepticism about AI, but logic is the mathematical discipline closest to AI and is perfectly suitable for creating AI. One can say that mathematics is the basis of all exact sciences, while logic is the math of mathematics. Each mathematical discipline builds its fundament on logic.

The logic deals with the nature of proof. What is proof? This is the human way of reasoning. But why only human? Machines can also reason and produce proof. This is why logic is the area closest to AI and logicians are best placed to conduct in-depth research in this area.

For each problem there are specialists who are best prepared and most suitable for solving this problem. If you want a painting, you will go to a painter. If you want a song, you will go to a musician. Mathematicians have proven to be the best ones in programming because their formal and abstract reasoning is very helpful in writing programs. The ones best placed to solve the AI problem are logicians because they are familiar with the abstractions of different worlds and logical formulas, and understand what is a world to be a model of a formula.

The first question is about the AI definition. Almost nobody seems to bother about this question! Although there are plethora of books and papers dedicated to AI, the question “What is AI” is very rarely raised there. However, let us note two fundamental publications where this question is central. These are Wang (1995) and Hutter (2000). We will also note a review article written in our institute (Angelova et al., 2021).

An interesting definition is provided by the Council of Europe (2022):

> *AI is actually a young discipline of about sixty years, which brings together sciences, theories and techniques (including mathematical logic, statistics, probabilities, computational neurobiology and computer science) and whose goal is to achieve the imitation by a machine of the cognitive abilities of a human being.*

The interesting aspect of this definition is the acknowledgement by the Council of Europe that mathematical logic is the foundation of AI, although many logicians would disagree with that.

The most widespread AI definition is provided by Turing and is known as the Turing Test (Turing, 1950). This is a construct where a human and a computer which imitates a human are hidden behind a curtain. If the computer imitates the human so well that we cannot tell it from the human, then this is AI. Although the Turing definition is very good, it has one important shortcoming. Turing’s definition describes a trained intellect, while we aim to define an untrained intellect. For example, a newborn baby is an intellect, but not a trained intellect yet.

In order to resolve the shortcoming of Turing’s definition, an informal AI definition was created in 2000 (Dobrev, 2000). It reads as follows: **“AI will be such a program which in an arbitrary world will cope not worse than a human.”**

In wording accessible to the general public, the above definition was published in a popular-science magazine. That publication is included in the PhD thesis (paragraph 1.1). A few years later the same definition was published in a scientific magazine (Dobrev, 2005a). That definition has received massive recognition since at present a Google search for “Definition of Artificial Intelligence” returns the abovementioned paper as the first result.

The problem with this definition is that it is not formal. The definition was formalized by introducing an intelligence quotient (IQ). Each program is assigned with a number which represents its IQ and the AI badge is awarded to the programs whose IQ is sufficiently high. That approach formed the basis of a paper published in 2019 (Dobrev, 2019b). That paper is not included in the PhD thesis because it has some shortcomings such as the assumption that the length of life and the complexity of the world are limited (i.e. fixed). Another problem is the finding that this idea had been developed in earlier publications (Hernández-Orallo et al., 1998) and (Hutter, 2000).

Another approach was adopted a couple of years later (Dobrev, 2022a). That other approach is included as part of the PhD thesis. What makes the new approach better is that the length of life and the complexity of the world are not limited. Another advantage of the new approach is the consideration of various languages for description of the world, while the first approach considered only the description derived by computable functions.

The second question: The road to creating AI goes through developing a program which is capable to “understand” the world and, on the basis of the understanding derived, plan its future behavior. “Understanding” the world means describing it by some language for description of worlds. That is why the creation of an appropriate language for description of the world is crucial to the creation of AI.

The second part of the PhD thesis is dedicated to the description of such a language, which claims to be an appropriate language for description of worlds (Dobrev, 2022b and 2023). An important part of this language are algorithms. The following is an example of an algorithm: “I will wait until the bus comes”. In (Dobrev, 2023) the author has included a definition of an algorithm which generalizes the standard definition of a computable function. The new definition describes the algorithm as a sequence of actions which can be executed in an arbitrary world. In the special case of a world which consists of an infinite tape and a head that reads from and writes on the tape, the algorithm is a Turing machine.

The third and the most important question is about the consequences which the creation of AI will entail. This does not seem to be a concern for a very large majority of people. In this regard, people are like curious children who play with a box of matches without reckoning that this can cause a fire. Few as they are, there are still scientific papers which discuss the future that will come on us after the creation of AI. An example of such a paper is Alfonseca et al. (2021). Another example is Ivanova et al. (2020) discussing the impact of AI on the labor market and how education should change in this regard.

The fundamental question is: “Is the creation of AI a must-do task for us or can we do without it?” The answer is that we do not have the luxury to choose because if we do not create AI, someone else will do it instead of us. We had better reckon and get ready for the consequences, but with due caution because we do not want to let the spirit run away from the bottle.

We will need the answer to the third question only when we have created AI. This does not mean that we should wait until AI comes by and only then start wondering “And now what?”. Waiting until the time when the question becomes urgent will be a grave mistake. Once AI is here, it will be too late to ponder “And where we go now?”. We should have raised this question well before the actual appearance of this machine (i.e. of this program because, as you will see below, AI is a program rather than a machine).

The contributions of this PhD thesis:

1. The PhD thesis proposes an informal definition of AI which is very widespread today (this is the first result returned by Google in response to searches for “Definition of Artificial Intelligence”).

2. The PhD thesis provides a rigorous mathematical definition of AI. We do not claim full credit for the proposed rigorous mathematical definition since it is an improvement of the AI definition initially created by Hernández-Orallo in 1998 (Hernández-Orallo et al., 1998) and substantially improved by Marcus Hutter in 2000 (Hutter, 2000).

3. The PhD thesis introduces a language for description of worlds such that the description can be searched automatically without human assistance. While other languages for description of worlds have been created before our language, the advantage of our language is the automated search for the description of the world, while the other languages are based on the premise that some human being has already understood the world and is now going to describe it in the respective language.

Abstract of the PhD thesis

The computable functions class is one of the key focus areas of research in mathematical logic. Although computable functions are very important for logicians, they have always felt there is not enough space for them in this set and have always tried to break loose from its bounds.

For example, in Turing and enumeration degrees of unsolvability logicians add an oracle and this is how they expand the set of computable functions. In this PhD thesis we will also use oracles, however, these oracles will be different in two aspects: On one hand, there will be less oracles because we are only interested in those which help us predict the next few steps. On the other hand, we will use unusual oracles which are not considered in enumeration degrees.

The problem we solve in this PhD thesis is a practical one. We have some function *f* and want to describe it. We will name this *f* function *world* and the language by which we will describe it will be termed *language for description of worlds*. The cardinality of these languages will be countable, meaning that our languages cannot describe all functions out there. In our terminology, the set of functions described by one language will be the *interpretation of that language*. I.e. the problem is to find a description such that the corresponding function in the interpretation of the language is proximal to the *f* function we are looking for.

We know too little about the *f* function. All we know is a finite part of this function (lived experience). Therefore, it is not impossible to obtain, through the interpretation of some language, a function which is proximal to *f*. The function we will find in this way will be even identical with *f* over lived experience.

We are not aiming to describe the world just for the sake of getting some description. We are looking for a description which will enable us look into the future and see what is going to happen after several steps. On the basis of this prediction, we will select the best move. This is what makes our oracles unusual.

For example, one of these unusual oracles is the Enemy. This is an agent who plays against us and always chooses the move which will hurt us most. Indeed, the Enemy is a very bizarre oracle since it depends on *f* – the function in which we have embedded it. The Enemy will behave differently in different worlds. Actions which are disruptive in a given world may not be disruptive or can even be helpful in another world.

The good thing about Oracle Enemy is that it enables us predict the future. We can use the *Min-Max* algorithm to see a few steps ahead. In this algorithm *Min* stands for our expectation that

the Enemy will select the move which will hurt us most, while *Max* reflects the assumption that we will play the moves which are most advantageous to us.

Another example is the Oracle of Randomness. We can imagine this oracle as a dice which returns several possibilities. Each of these possibilities has its own probability. This oracle can also be used to predict the future, but we will use the *Max-Average* algorithm instead of *Min-Max*. Here we cannot predict what exactly the oracle will do, but can foresee several potentialities and calculate their average.

A third example is the Oracle of Noise. This agent is a constant, however in his constancy there is some noise. Moreover, the probability of that noise is unknown. In this case we will predict the future by considering one and only one possible action of agent Noise with the assumption that the prediction may be biased by some noise.

“What is a language for description of worlds?” This is a question the answer to which you will not find in the PhD thesis. What you will find here are only examples of such languages. The first chapter of the PhD thesis provides two definitions of Artificial Intelligence (AI). The first definition says that AI is a program whose performance will not be worse than that of a human being. The second definition is more interesting. It is a formal definition and uses a language for description of worlds.

As a first step, the formal AI definition determines the best performing policy and that definition is made on the basis of some language for description of worlds. In the second step AI is defined as a computable policy which is sufficiently proximal to the best performing policy.

The PhD thesis puts forth the proposition that the formal AI definition does not depend on the language for description of worlds used for arriving at the definition. That proposition has not been proven and probably cannot be proven at all.

The selected language for description of worlds can be used as a basis for creating a program which satisfied the AI definition. Although in theory this program will halt after a finite number of steps, in practice it is so inefficient that only an infinitely fast computer would be able to run it.

Before we can create a program which satisfies the AI definition, we need a language for description of worlds which enables us predict the future. This is where we bump into the finding that not every extension of the set of computable functions is suitable and not every such extension will work for us.

Furthermore, the PhD thesis puts forth the proposition that although the AI definition does not depend on the language for description of worlds, the efficiency of a definition-compliant program created on the basis of a certain language strongly depends on the language we would select and use for describing worlds.

The second part of the PhD thesis deals with a particular world and a particular language for description of worlds by which that world been described.

That particular world is the world of the chess game, although it has one interesting peculiarity: one step in our chess world is *not* one move of a chess piece. Moving a chess piece takes many steps. The agent “focuses” it’s gaze on only one square of the chessboard at a time. Before moving a piece, the agent must shift his focus to that piece, lift the piece, then refocus on the new square and finally drop the lifted piece there. Thus, moving a chess piece is the execution of an algorithm which consists of multiple steps. This is where the main advantage of the new language for description of worlds comes from. With the new language, the prediction of the future happens not by looking a few steps forward, but by looking a few *large* steps forward. A large step is the execution of an algorithm.

One of the most important contributions of this PhD thesis is the definition of the term *algorithm*. In the existing literature an algorithm is not dissociated from a computable function. Typically, each program is assigned with the computable function which the program computes.

In this PhD thesis, an algorithm is a sequence of actions executed in some world. The result from the execution of these actions depends on the world in which we are going to execute them. For example, the Turing machine is typically understood as an inseparable system consisting of a head and an infinite tape. In this PhD thesis, the algorithm is the head of Turing's machine, while the infinite tape is the world in which the algorithm is executed. If we change the world, the result from the execution of the algorithm will also change. For example, if we replace the infinite tape with a finite one, the algorithm will produce a finite function. A sorting algorithm can sort different objects in different worlds. For instance, in the pharmacy world the algorithm can sort medicine bottles.

The important thing about predicting the future is to ensure that the small steps are replaced with large ones (i.e. the execution of algorithms). This how we can look into the more distant future, because thinking in small steps will trigger a combinatorial explosion and we will not be able to see further than the tips of our noses.

The third part of the PhD thesis addresses certain philosophical issues related to the implications arising from the creation of AI. That part is not mathematical and does not include a single theorem or even definition. Nevertheless, this is an important part of the PhD thesis insofar as mathematics is not only about the creation of formal constructs or the provision of formal proof. A mathematician should also consider the consequences of his formal reasoning.

One of the questions we raise in the final part of the PhD thesis is "Should AI technology be accessible by everyone?" Our answer is that AI is dangerous technology and serious papers in the AI domain should be classified. This answer will put the future reviewer of this PhD thesis in an awkward situation. If the reviewer endorses this statement and believes that the PhD thesis is a serious contribution to AI, they should try to keep it away from the public eye, and write a negative review. In the opposite case, if the reviewer concludes that the PhD thesis does not say anything serious about AI, they should again issue a negative review.

The reader of the PhD thesis will not come across of any substantial mathematical statement that has been supported by proof. While there are five substantial mathematical statements in the first part, none of them has been proven and we even believe that some of them cannot be proven at all. The PhD thesis provides two AI definitions, the first of which is an informal one. The second definition is formal, but with one exception: we want the policy to be sufficiently proximal to the best performing policy. While "sufficiently proximal" is not a formal concept, we consider a set of algorithms which are infinitely proximal, and this is already a formal concept. ("Infinitely proximal" means that for each ε the parameter h of an algorithm has a value at which the distance between that algorithm and the best performing strategy is lesser than ε.)

The reason for the many informal statements in the PhD thesis is that AI is a relatively new area where the main challenge is not how to solve known formal problems, but how to formalize the terms and problems which emerge in this area. While AI research dates back to Alan Turing (meaning that the area is not that new), AI still needs some serious formalization and clarification of its key concepts.

1 What is Artificial Intelligence?

1.1 The informal definition

In this PhD thesis we are going to discuss the following questions: “Do we have to know what is AI?” and “What is intelligence?”. After that we are going to give a definition of Artificial Intelligence. Finally, from this definition we are going to get an algorithm which after a final number of steps will discover AI.

1.1.1 AI – What is this?

Do we have to know what is AI? This question can be easily answered: Yes, if we want to find it then our task will be a lot easier if we know what is the thing we are looking for. Failing to define AI, our position will not differ from that of the Alchemists who sought for the Philosopher's stone but almost had no idea what they were searching for.

The most widely spread definition of AI is the so called Turing's test. Alan Turing was a British mathematician famous for the invention of the theoretical Turing machine and for the deciphering of the German codes during World War II.

The Turing's test is quite simple. We place something behind a curtain and it speaks with us. If we can't make difference between it and a human being then it will be AI. However, this definition exists from more than fifty years, so we are going to create a newer and a more up-to-date one.

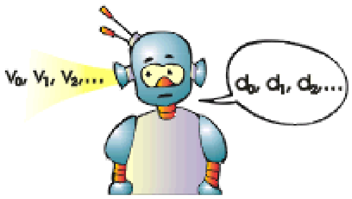
Turing's definition suggests that, an Intellect is a person with knowledge gained through the years. If this is so, then what about a newly born baby? Is it an Intellect? Our answer will be "yes". Our definition of an intellect will be: a thing that knows nothing but it can learn. At this point we differ from most people who imagine a university professor when they hear the word Intellect.

Before giving a formal definition of AI we will make it clear that we accept the thesis of Church, stating that every calculating device can be modelled by a program. This means that we are going to look for AI in the set of programs. We will suppose that AI is a step device living in a kind of world. At each step it receives information (from the world) and influences (at the world) by the information it works out. Also, we will assume that the information received and worked out at each step will be a finite amount. Let's say it gets **n** bits and works out **m** bits.

After this clarification we can state informally our definition. **AI will be such a program which in an arbitrary world will cope not worse than a human.**

The next task will be to formalise this definition in order to use it and to search for AI with it. First, what is a world for us? These will be two functions **World(s, d)** and **View(s)**.

The first will take as arguments the state of the world and the influence that our device has on the world at this step. As a result, this function will return the new state of the world (which it will obtain on the next step). The second function will inform us what does our device see. An argument of this function will be the world's state and the returned value will be the information that the device will receive (at a given step). Also, we have to add one $\mathbf{s}_0$. It will be the world's state when our device was born. During its life the world will go through the states $\mathbf{s}_0$, $\mathbf{s}_1$, $\mathbf{s}_2$, ... The device will influence the world with the information it works out at each step $\mathbf{d}_0$, $\mathbf{d}_1$, $\mathbf{d}_2$, ... Also, AI will receive information from the world $\mathbf{v}_0$, $\mathbf{v}_1$, $\mathbf{v}_2$, ... It is clear that $\mathbf{s}_{i+1} = \mathbf{World}(\mathbf{s}_i\ , \mathbf{d}_i)$ and $\mathbf{v}_i = \mathbf{View}(\mathbf{s}_i)$.



We have everything up to this moment. We have a world and a device that lives in it. However, there is one thing missing - the meaning of life. What is life without pain and joy, a philosopher would say. That is why we will introduce meaning of life. This will be an evaluation to tell us whether one row $\mathbf{v}_0$, $\mathbf{v}_1$, $\mathbf{v}_2$, ... is better than another.

Most people think that they have spent their life better if they have seen more Swiss resorts and less coal-mines. More or less our definition of the meaning of life will be the same. We will pick out two bits from $\mathbf{v}_i$ and call them victory and loss. The aim will be to get more victories and fewer losses.

1.2 The formal definition

1.2.1 The AI Definition and a Program Which Satisfies this Definition

We will consider all policies of the agent and will prove that one of them is the best performing policy. While that policy is not computable, computable policies do exist in its proximity. We will define AI as a computable policy which is sufficiently proximal to the best performing policy. Before we can define the agent's best performing policy, we need a language for description of the world. We will also use this language to develop a program which satisfies the AI definition. The program will first understand the world by describing it in the selected language. The program will then use the description in order to predict the future and select the best possible move. While this program is extremely inefficient and practically unusable, it can be improved by refining both the language for description of the world and the algorithm used to predict the future. This can yield a program which is both efficient and consistent with the AI definition.

1.2.2 Introduction

Once, I was talking to a colleague and he told me: *'Although we may create AI someday, it will be a grossly inefficient program as we will need an infinitely fast computer to run it'*. My answer was: *'You just give me this inefficient program which is AI, and I will improve it so that it becomes a true AI which can run on a real-world computer'*.

Today, in this PhD thesis I will deliver the kind of program I asked my colleague to give me at that time. I will set out an inefficient program which satisfies the AI definition. I will go further and suggest some ideas and guidance on how this inefficient program can be improved to become a real program which runs in real time. My hope is that some readers of this PhD thesis will succeed to do this and deliver the AI we are looking for.

How inefficient is the program described here? In theory, there are only two types of programs – ones which halt and ones which run forever. In practice however, some programs will halt somewhere in the future, but they are so inefficient that we can consider them as programs which run forever. This is the case with the program described here — formally it halts, but its inefficiency makes it unusable (unless the computer is infinitely fast or the world is extremely simple).

What is the definition of AI? We will define AI as a policy. An agent who follows this policy will cope sufficiently well. This is true for any world, provided however that there are not any fatal errors in that world. If a fatal error is possible in a given world, the agent may not perform well in that particular world, but his average performance over all possible worlds will still be sufficiently good.

Which worlds we will consider as possible? The world's policies are continuum many. If we do not have any clues as to what the world should be, then we cannot have a clue about what the expected success of the agent should look like. We will assume that the world can be described and such description is as simple as possible (this assumption is known as *Occam's* *razor*). In other words, we will choose a language for description of worlds and will limit our efforts only to the worlds described by that language. The worlds whose description is simpler (shorter) will be preferred (will carry more weight).

This PhD thesis will consider several languages for description of the world. The first language will describe deterministic worlds. This language will describe the world by means of a computable function, which will take the state of the world and the action of the agent as input and return the new state of the world and the next observation as output. If we know the initial state of the world and agent's actions, this function will give us the life of the agent in that world.

The second language will describe non-deterministic worlds – again by a computable function, but with one additional argument. This argument will be randomness. In this case, we will need to know one more thing in order to obtain the agent's life in that world. We will need to know what that randomness has been.

We will define AI by these two languages and will make the assumption that these two definitions are identical. We will make even the assumption that the AI definition does not depend on our choice of language for description of worlds, and all languages produce the same definition of AI.

On the basis of these two languages we will make two programs which satisfy the AI definition. These two programs will calculate approximately the same policy, but their efficiency would be dramatically different. Therefore, the choice of language for description of the world will not affect the AI definition, but will have a strong impact on the efficiency of the AI obtained through the chosen language.

Contributions

This PhD thesis improves the AI definition initially provided by Hernández-Orallo et al. in 1998 and then substantially improved by Marcus Hutter in 2000. More precisely, this PhD thesis introduces two improvements:

1. An AI definition which does not depend on the length of life. Papers (Orallo 1998 and Hutter 2000) do provide an AI definition, however, the assumption there is that the length of life is limited by a constant and this constant is a parameter of the definition.

2. An AI definition which does not depend on the language for description of the world. The language in Hutter (2000) is fixed. Thus, paper Hutter (2000) implies that there is only one possible way to describe the world.

1.2.3 Related work

1.2.3.1 General Intelligence

Let us first note that the meaning which we imply in *artificial intelligence* in this paper is *artificial general intelligence*. Other authors have discussed two types of AI which they describe as *narrow* and *general* (sometimes as *weak* and *strong*). I believe that a more appropriate pair of terms for the two types of AI is *fake* and *genuine AI*.

Let us illustrate this statement using the example of diamonds. Both intelligence and diamonds are classified in two categories – natural and artificial. Artificial diamonds are further divided in two subcategories – *genuine* (consisting of carbon) and fake (made of glass). Today, when we say artificial diamonds we mean ones made of carbon. Now let us image that we are living in the 19th century when nobody was yet able to make artificial diamonds from carbon. What people in the 19th century meant by artificial diamonds were diamonds made of glass – shiny pieces that look like diamonds but in fact are not. Today we call these glass pieces fake diamonds.

A genuine artificial diamond is every bit as good as a natural diamond. In terms of hardness and transparency these two diamonds are equal. However, they differ in price because an artificial diamond is much cheaper than a natural one although it may be superior in terms of size and purity.

The same applies to artificial intelligence. Artificial general intelligence is by all measures as good as natural intelligence, and can even be better in terms of speed, memory and “smartness”. Certainly, the price of artificial intelligence will be much lower than that of natural intelligence. Today, in the 21st century, natural intelligence is even priceless because you cannot buy it.

Regarding narrow artificial intelligence, it looks like intelligence, but it is not. When we come to have artificial general intelligence one day, narrow AI programs will be called *fake artificial intelligence* or *intelligence-mimicking programs*.

Nowadays most papers dedicated to AI actually mean some narrow or fake AI. In this paper by AI we will mean general or genuine AI.

1.2.3.2 The Intuitive Definition

Now let us proceed with an overview of the papers dedicated to the definition of artificial intelligence. This definition is very important and actually drills down to the most important question about AI. Nonetheless, these papers a very few because most researchers never bother themselves with the question “What is AI?” – there are just a few researchers who do. The reason is that our colleagues simply do not believe in AI. If you do not believe in ghosts you do not ask yourself “What is actually a ghost?”. Recently I attended a lecture given by one of the leading experts in the area of AI (Solar-Lezama, 2023). He said “No matter how smart AI is, there will always be some human who is smarter than it”. Evidently, this colleague of ours does not believe in AI and cannot imagine that one day AI will be smarter than any human.

Although the papers dedicated to the AI definition are not so many, there are still some of them. Very good overviews of these papers can be seen in Wang 2019 and in the works of Hernández-Orallo (2012, 2014a, 2014b, 2014c, 2017). Here we will offer a shorter overview in which we will try to say things that have not been said in the mentioned overview papers.

The first intuitive (informal) definition of AI was provided by Alan Turing and is known as the Turing Test (Turing, 1950). That definition is perfect in its simplicity. Nonetheless, there is a significant problem with it. What the Turing Test defines is trained intellect (i.e. intelligence plus education). We would like have a definition of untrained intellect (i.e. pure intelligence without education). The first definition of pure intelligence was provided by Pei Wang in 1995 (Wang, 1995). It reads as follows:

Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources.

Subsequently, Pei Wang’s definition was improved in 2000. That improvement was published in Dobrev (2000). Today, it is the first result listed by *Google* on the topic of AI Definition. The first result returned by *Google* in response to a query for *Definition of Artificial Intelligence* is the paper of Dobrev (2005a), which is an improved version of Dobrev (2000). Here is the improved version of Pei Wang’s definition:

AI will be such a program which in an arbitrary world will cope not worse than a human.

What is the gist of the improvement? First, what Pei Wang has defined is intelligence, while the improved version defines artificial intelligence. That improvement is not significant, because the real question is “What is intelligence?”. The fact that AI is a program is a direct corollary from Church thesis (Church, 1941) which says that any information system can be emulated by a computer program.

Here is the significant aspect of the improvement of Pei Wang’s definition: While Wang wants the intelligence to be able to cope in a concrete world (in its environment), according to the improved version the intelligence must be able to cope in an arbitrary world. What makes this improvement significant? In the end of the day, for us it is important that AI is able to cope well in its own environment, because this is the important environment we are interested in. However, AI should not be dependent on the environment because we wish to be able to deploy it in various environments (worlds) such that each deployment is successful regardless of the environment. Although we can perfectly say that the real world is what matters to us, this world is not unique. The place and time of birth make a big difference. If either of these parameters were to change, we would find ourselves in a very different world. Obviously, Pei Wang was clearly mindful that there is not just one world, which is why he added to his definition the phrase *while operating with insufficient knowledge and resources*. I.e. Pei Wang wants AI to be able to cope in difficult circumstances as well, implying that if it succeeds when it is difficult it will also succeed when it is easy. Of course, things are difficult for those who are uneducated and poor. It would be much easier when one is equipped with knowledge and resources.

Another improvement of Pei Wang's definition relates to the fact that his definition does not say how well AI should cope. Wang implies that AI will either cope or fail, but we know that some cope better than others. That is, how well AI can cope, and therefore its level of intelligence, is important. The improved version of the definition says that AI should cope not worse than a human. Although benchmarking to a human makes the definition informal, it is still important because we should identify the level of intelligence which is sufficient for us to accept that a given program covers the necessary level of intelligence to be recognized as AI.

1.2.4 Terms of the problem

Let the agent have *n* possible actions and *m* possible observations. Let Σ and Ω be the sets of actions and respectively observations. In the observations set there will be two special observations. These will be the observations *good* and *bad*, and they will provide rewards 1 and -1. All other observations in Ω will provide reward 0.

We will add another special observation – *finish*. The agent will never see that observation ($finish \notin \Omega$), but we will need it when we come to define the model of the world. The model will predict *finish* when it breaks down and becomes unable to predict anything more. For us the *finish* observation will not be the end of life, but rather a leap in the unknown. We expect our AI to avoid such leaps in the unknown and for this reason the reward given by the *finish* observation will be -1.

Definition 1: The tree of all possibilities is an infinite tree. All vertices which sit at an even-number depth level and are not leafs will be referred to as action vertices and those at odd-number depth levels will be observation vertices. From each action vertex there will depart *n* arrows which correspond to the *n* possible actions of the agent. From each observation vertex there will depart *m+1* arrows which correspond to the *m* possible observations of the agent and the observation *finish*. The arrow which corresponds to *finish* will lead to a leaf. All other arrows lead to vertices which are not leafs.

Definition 2: In our terms the world will be a 3-tuple $<S, s_0, f>$, where:

1. S is a finite or countable set of internal states of the world;

2. $s_0 \in S$ is the initial state of the world; and

3. $f: S\times\Sigma \rightarrow \Omega\times S$ is a function which takes a state and an action as input and returns an observation and a new state of the world.

The *f* function cannot return observation *finish* (it is predicted only when *f* is not defined and there is not any next state of the world). What kind of function is *f* – computable, deterministic or total? The answer to each of these three questions can be *Yes*, but it can also be *No*.

Definition 3: A deterministic policy of the agent is a function which assigns a certain action to each action vertex.

Definition 4: A non-deterministic policy of the agent is a function which assigns one or more possible actions to each action vertex.

When the policy assigns all possible actions at a certain vertex (moment) we will say that at that moment the policy does not know what to do. We will not make a distinction between an agent and the policy of that agent. A union of two policies will be the policy which we get when choose one of these two policies and execute it without changing that policy. Allowing a change of the chosen policy will lead to something else.

Definition 5: Life in our terms will be a path in the tree of all possibilities which starts from the root.

Each life can be presented by a sequence of actions and observations:

$$a_1, o_1, \dots, a_t, o_t, \dots$$

We will not make a distinction between a finite life and a vertex in the tree of all possibilities because there is a one-to-one correspondence between these two things.

Definition 6: The length of life will be *t* (the number of observations). Therefore, the length of life will be equal to the length of the path divided by two.

Definition 7: A completed life is one which cannot be extended. In other words, it will be an infinite life or a life ending with the observation *finish*.

When we let an agent in a certain world, the result will be a completed life. If the agent is non-deterministic then the result will not be unique. The same applies when the world is non-deterministic.

1.2.5 The grade

Our aim is to define the agent's best performing policy. For this purpose we need to assign some grade to each life. This grading will give us a linear order by which we will be able to determine the better life in any pair of lives.

Let us first determine how to measure the success of each life *L*. For a finite life, we will count the number of times we have had the observation *good*, and will designate this number with $L_{good}(L)$. Similar designations will be assigned to the observations *bad* and *finish*. Thus, the success of a finite life will be:

$$Success(L) = \frac{L_{good}(L) - L_{bad}(L) - L_{finish}(L)}{|L|}$$

Let us put L_i for the beginning of life *L* with a length of *i*. The *Success(L)* for infinite life *L* will be defined as the limit of $Success(L_i)$ when *i* tends to infinity. If this sequence is not convergent, we will take the arithmetic mean between the limit inferior and limit superior.

$$Success(L) = \frac{1}{2}.\left(\liminf_{i\to\infty}\big(Success(L_i)\big) + \limsup_{i\to\infty}\big(Success(L_i)\big)\right)$$

By doing this we have related each life to a number which belongs to the interval *[-1, 1]* and represents the success of this life. Why not use the success of life for the grade we are trying to find? This is not a good idea because if a world is free from fatal errors then the best performing policy will not bother about the kind of moves it makes. There would be one and only one maximum success and that success would always be achievable regardless of the number of errors made in the beginning. If there are two options which yield the same success in some indefinite time, we would like the best performing policy to choose the option that will yield success faster than the other one. Accordingly, we will define the grade of a completed life as follows:

Definition 8: The grade of infinite life *L* will be a sequence which starts with the success of that life and continues with the rewards obtained at step *i*:

$$Success(L), reward(o_1), reward(o_2), reward(o_3), \ldots$$

Definition 9: The grade of finite and completed life *L* will be the same sequence, but in this sequence for $i>t$ the members $reward(o_i)$ will be replaced with *Success(L)*:

$$Success(L), reward(o_1), \ldots , reward(o_t), Success(L), Success(L), \ldots$$

(In other words, the observations that come after the end of that finite life will receive some expectation for a reward and that expectation will be equal to the success of that finite life.)

In order to compare two grades, we will take the first difference. This means that the first objective of the best performing policy will be the success of entire life, but its second objective will be to achieve a better reward as quickly as possible.

1.2.6 The expected grade

Definition 10: For each deterministic policy *P* we will determine *grade(P)*: the grade we expect for the life if policy *P* is executed.

We will determine the expected grade at each vertex *v* assuming that we have somehow reached *v* and will from that moment on execute policy *P*. The expected grade of *P* will be the one which we have related to the root.

We will provide a rough description of how we relate vertices to expected grades. Then we will provide a detailed description of the special case in which we look for the best grade, i.e. the expected grade of the best performing policy.

Rough description:

1. Let *v* be an action vertex.

Then the grade of *v* will be the grade of its direct successor which corresponds to action *P(v)*.

2. Let *v* be an observation vertex.

2.1. Let there be one possible world which is a model of *v*.

If we execute *P* in this world we will get one possible life. Then the grade of *v* will be the grade of that life.

2.2. Let there be many possible worlds.

Then each world will give us one possible life and the grade *v* will be the mean value of the grades of the possible lives.

The next section provides a detailed description of the best performing policy. The main difference is that when *v* is an action vertex, the best performing policy always chooses the highest expected grade among the expected grades of all direct successors.

1.2.7 The best performing policy

As mentioned above, we should have some clue about what the world looks like before can have some expectation about the success of the agent. We will assume that the world can be described by some language for description of worlds.

Let us take the standard language for description of worlds. In this language the world is described by a computable function (this is the case in Hutter, 2000). We will describe the computable function *f* by using a Turing machine. We will describe the initial state of the world as a finite word over the machine alphabet. What we get is a computable and deterministic world which in the general case is not a total one.

Definition 11: A world of complexity k will be a world in which:

1. The f function is described by a Turing machine with k states.

2. The alphabet of that machine contains $k+1$ symbols ($\lambda_0, \dots, \lambda_k$).

3. The initial state of the world is a word made of not more than k letters. The alphabet is $\{\lambda_1, \dots, \lambda_k\}$, i.e. the alphabet of the machine without the blank symbol λ_0.

Here we use the same k for three different things as we do not need to have different constants.

We will identify the best performing policy for the worlds of complexity k (importantly, these worlds are finitely many). For this purpose we will assign to each observation vertex its best grade (or the expected grade if the best performing policy is executed from that vertex onwards).

Let us have life $a_1, o_1, \dots , a_t, o_t, a_{t+1}$.

Let this life run through the vertices $v_0, w_1, v_1, \dots , w_t, v_t, w_{t+1}$,

where v_0 is the root, v_i are the action vertices and w_i are the observation vertices.

Now we have to find out how many models of complexity k are there for vertex v_t.

Definition 12: A deterministic world is a model of v_t when in that world the agent would arrive at v_t if he executes the corresponding actions ($a_1, \dots , a_t$). The models of each action vertex are identical with the models of its direct successors.

Definition 13: The best performing policy for the worlds of complexity k will be the one which always chooses the best grade (among the best grades of the direct successors).

Definition 14: The best grade of vertex w_{t+1} (for worlds of complexity k) is determined as follows:

Case 1. Vertices v_t and w_{t+1} do not have any model of complexity k.

In this case the best grade for w_{t+1} will be *undef*. At this vertex the policy will not know what to do (across the entire subtree of v_t) because the best grade for all successor vertices will be *undef*.

If we do not want to introduce an *undef* grade, we can use the lowest possible grade – the sequence of countably many -1s. The maximal grade will be chosen among the vertices which are different from *undef*. Replacing *undef* with the lowest possible grade will give us the same result.

Case 2. Vertices v_t and w_{t+1} have one model of complexity k.

Let this model be D. In this case there are continuum many paths through w_{t+1} such that D is model of all those paths. From these paths (completed lives) we will select the set of the best paths. The grade we are looking for is the grade of these best paths. Each of these paths is related to a deterministic policy of the agent. We will call them the best performing policies which pass through vertex w_{t+1}.

This is the procedure by which we will construct the set of best deterministic policies: Let P_0 be the set of all policies which lead to w_{t+1}. We take the success of each of these policies in the world D. We create the subset P_1 of the policies which achieve the maximum success. Then we reduce P_1 by selecting only the policies which achieve the maximum for $reward(o_{t+2})$ and obtain subset P_2. Then we repeat the procedure for each $i>2$. In this way we obtain the set of the best deterministic policies P. (The best ones of those which pass through vertex w_{t+1} as well as the best ones for the paths which pass through vertex w_{t+1}. As regards the other paths, it does not matter how the policy behaves there.)

$$P = \bigcap_{i=0}^{\infty} P_i$$

We can think of P as one non-deterministic policy. Let us take some $p \in P$. This will give us the best grade:

$$Success(p), reward(o_{t+1}), reward(o_{p,t+2}), reward(o_{p,t+3}), \ldots$$

Here we drop out the members *reward(o*$_i$*)* at $i \leq t$ because they are uniquely defined by v_t. The next member depends on w_{t+1} and D, but does not depend on p. The remaining members depend on p.

Another way to express the above formula is:

$$\max_{p \in P_0} Success(p), reward(o_{t+1}), \max_{p \in P_1} reward(o_{p,t+2}), \max_{p \in P_2} reward(o_{p,t+3}), \ldots$$

Case 3. Vertices v_t and w_{t+1} have a finite number of models of complexity k.

Let the set of these models be M. Again, there are continuum many paths through w_{t+1} such that each of these paths has a model in M. These paths again form a tree, but while in case 2 the branches occurred only due to a different policy of the agent, in this case some branches may occur due to a different model of the world. Again, we have continuum many deterministic policies, but now they will correspond to subtrees (not to paths) because there can be branches because of the model. Again we will try to find the set of best performing deterministic policies and the target grade will be mean grade of those policies (the mean grade in M).

We will again construct the set of policies P_i. Here P_1 will be the set of policies for which the mean success reaches its maximum. Accordingly, P_2 will be the set of policies for which the mean *reward(o*$_{t+2}$*)* reaches its maximum and so on. This is how the resultant grade will look like:

$$\max_{p \in P_0} \sum_{m \in M} q_m . Success(m,p), \sum_{m \in M} q_m . reward(o_{m,t+1}), \max_{p \in P_1} \sum_{m \in M} q_m . reward(o_{m,p,t+2}), \ldots$$

If we take some $p \in P$, the resultant grade will look like this:

$$\sum_{m \in M} q_m . Success(m,p), \sum_{m \in M} q_m . reward(o_{m,t+1}), \sum_{m \in M} q_m . reward(o_{m,p,t+2}), \ldots$$

Here q_i are the weights of the worlds which have been normalized in order to become probabilities. In this case we assume that the worlds have equal weights, i.e.:

$$q_i = \frac{1}{|M|}$$

■

What we have described so far looks like an algorithm, however, rather than an algorithm, it is a definition because it contains uncomputable steps. The so described policy is well defined, even though it is uncomputable. Now, from the best grade for complexity k, how can we obtain the best grade for any complexity?

Definition 15: The best grade at vertex v will be the limit of the best grades at vertex v for the worlds of complexity k when k tends to infinity.

How shall we define the limit of a sequence of grades? The number at position *i* will be the limit of the numbers at position *i*. When the sequence is divergent, we will take the arithmetic mean between the limit inferior and limit superior.

Definition 16: The best performing policy will be the one which always chooses an action which leads to the highest grade among the best grades of the direct successors.

What makes the best performing policy better than the best performing policy for worlds of complexity *k*? The first policy knows what to do at every vertex, while the latter does not have a clue at the majority of vertices because they do not have any model of complexity *k*. The first policy can offer a better solution than the latter policy even for the vertices at which the latter policy knows what to do because the first policy also considers models of complexity higher than *k*. Although at a first glance we do not use Occam's razor (because all models have equal weights), in earnest we do use Occam's razor because the simpler worlds are calculated by a greater number of Turing machines, meaning that they have a greater weight.

1.2.8 The AI definition

Definition 17: AI will be a computable policy which is sufficiently proximal to the best performing policy.

At this point we must explain what makes a policy proximal to another policy and how proximal is proximal enough. We will say that two policies are proximal when the expected grades of these two policies are proximal.

Definition 18: Let *A* and *B* be two policies and $\{a_n\}$ and $\{b_n\}$ are their expected grades. Then the difference between *A* and *B* will be $\{\varepsilon_n\}$, where:

$$\varepsilon_n = \sum_{i=0}^{n} \gamma^i (a_i - b_i) = \varepsilon_{n-1} + \gamma^n (a_n - b_n)$$

Here γ is a discount factor. Let $\gamma=0.5$. We have included a discount factor because we want the two policies to be proximal when they behave in the same way for a long time. The later the difference occurs in time, the less impact it will have.

When *n* goes up, $|\varepsilon_n|$ may go up or down. We have made the definition in this way because we want the difference to be small when the expected grade of policy *A* hovers around the expected grade of policy *B*. I.e., if for *n-1* the higher expected grade is that of *A* and for *n* the higher expected grade is that of *B*, then in ε_n the increase will offset the decrease and vice versa.

Definition 19: We will say that $|A\text{-}B|<\varepsilon$ if $\forall n\ |\varepsilon_n|<\varepsilon$.

1.2.9 A program which satisfies the definition

We will describe an algorithm which represents a computable policy. Each action vertex relates to an uncompleted life and the algorithm will give us some action by which this life can continue. This algorithm will be composed of two steps:

1. The algorithm will answer the question 'What is going on?' It will answer this question by finding the first *k* for which the uncompleted life has a model. The algorithm will also find the set *M* (the set of all models of the uncompleted life, the complexity of which is *k*). Unfortunately, this is uncomputable. To make it computable we will try to find efficient models with complexity *k*.

Definition 20: An efficient model with complexity *k* will be a world of complexity *k* (definition 11), where the Turing machine uses not more than *1000.k* steps in order to make one step of the life (i.e. to calculate the next observation and the next internal state of the world). When the machine makes more than *1000.k* steps, the model will return the observation *finish*.

The number *1000* is some parameter of the algorithm, but we assume this parameter is not very important. If a vertex has a model with complexity k, but does not have an efficient model with complexity k, then $\exists n\ (n>k)$ such that the vertex has an efficient model with complexity n.

2. The algorithm will answer the question 'What should I do?'. For this purpose we will run h steps in the future over all models in M and over all possible actions of the agent. In other words, we will walk over one finite subtree and will calculate *best* for each vertex of the subtree (this is the best expected grade up to a leaf). Then we will choose an action which leads to the maximum by *best* (this is the best partial policy).

1.2.10 Is this AI?

Does the algorithm described above satisfy our AI definition? Before that we must say that the algorithm depends on the parameters h and ε. In order to reduce the number of parameters, we will assume that ε is a function of h. For example, this function can be $\varepsilon=h^{-0.5}$.

Statement 1: When the value of h is sufficiently high, the described algorithm is sufficiently proximal to the best performing policy.

Let the best performing policy be P_{best}, and the policy calculated by the above algorithm with parameter h be P_h. Then statement 1 can be expressed as follows:

$$\forall\varepsilon>0\ \exists n\ \ \forall h>n\ (\ |P_{best} - P_h|<\varepsilon\ )$$

Although we cannot prove this statement, we can assume that when h tends to infinity then P_h tends to the best performing policy for the worlds the complexity of which is k. When t tends to infinity, k will reach the complexity of the world or tend to infinity. These reflections make us believe that the above statement is true.

1.2.11 Conclusion

We examined three languages for description of the world. On the basis of each language, we developed an AI definition and assumed that all three definitions are the same. Now we will make an even stronger assertion:

Statement 5: The AI definition does not depend on the language for description of the world on the basis of which the definition has been developed.

We cannot prove this statement although we suppose that it is true. We also suppose that the statement cannot be proven (similar to the thesis of Church).

Although we assumed that the AI definition does not depend on the language for description of the world, we kept assuming that the program which satisfies this definition strongly depends on the choice of language. The comparison between the first two languages clearly demonstrated that the second language is far more expressive and produces a far more efficient AI.

Let us look at one more language for description of worlds – the language described in Dobrev (2022b, 2023). That language describes the world in a far more efficient way by defining the term 'algorithm'. The term 'algorithm' enables us plan the future. For example, let us take the following: 'I will wait for the bus until it comes. Then I will go to work and will stay there until the end of the working hours.' These two sentences describe the future through the execution of algorithms. If we are to predict the future only by running h possible steps, then h will necessarily become unacceptably large.

The language described in Dobrev (2023) is far more expressive and lets us hope that it can be used to produce a program which satisfies the AI definition and which is efficient enough to work in real time.

2 How can we create AI?

2.1 Language for Description of Worlds

We will reduce the task of creating AI to the task of finding an appropriate language for description of the world. This will not be a programing language because programing languages describe only computable functions, while our language will describe a somewhat broader class of functions. Another specificity of this language will be that the description will consist of separate modules. This will enable us look for the description of the world automatically such that we discover it module after module. Our approach to the creation of this new language will be to start with a particular world and write the description of that particular world. The point is that the language which can describe this particular world will be appropriate for describing any world.

2.2 Introduction

This PhD thesis presents a new approach to the exploration of AI. This is the Event-Driven (ED) approach. The underlying idea of the ED approach is that instead of absorbing all input/output information, the model should reflect only the events which matter ("important events").

Every action is an event. Every observation is an event, too. If the model were to reflect each and every action and observation, it would end up overloaded with an enormous amount of information. The overloaded situation can however be avoided when the model is limited only to certain important events. This leads us to the idea of Event-Driven models.

A disadvantage (or perhaps an advantage) of the ED model is that it does not describe the world completely, but only partially. More precisely the ED model describes a certain class of worlds (the worlds which comply with a certain pattern).

What makes this PhD thesis different? The mainstream approach to dealing with multi-agent systems is based on the assumption that the world is given (known) and what we need to find is a policy. In other words, the world is part of the known terms of the problem while the policy is the unknown part. This PhD thesis is different because we will assume that instead of being given, the world is unknown and has to be found.

The assumption that the world is given implies that we have a relation which provides a full description of the world. Conversely, in this PhD thesis we will try to provide partial descriptions of the world and will do so by employing ED models.

Structure of the language: The description of the world will not be similar to a homogenous system consisting of one single layer. Our description of the world would rather be structured as a multilayer system. In Figure 1 we have presented our multilayer structure as a pyramid where the first layer is the base of the pyramid:

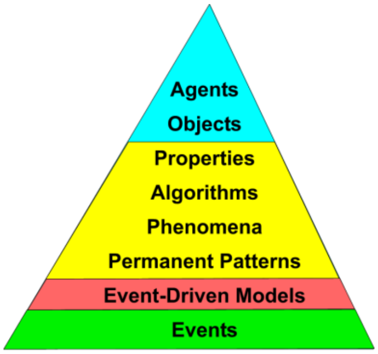


Figure 1

Thus, events will form the first layer of our system. We will use these events in order to describe the Event-Driven (ED) models. Importantly, the states of the ED models must not be the same (some states must be different). For this condition to be fulfilled, there must be at least one state with a special occurrence in it so that the state can be distinguished from other states. We will name this special occurrence *distinction*. The set of the distinctions we will designate as *trace*.

Our first idea about the trace is that it is fixed. (For example, let us have a state in which is always cold.) Then our notion of a trace will evolve and we will have a *moving* trace which can appear and then disappear or move from one state to another. (In our example the coldness could move to the room next door.)

The next layers of the pyramid will consist of various patterns (we will present all of them by ED models). We will begin with the permanent patterns, i.e. those which are observed all the time. While most authors assume that all patterns are permanent, in this PhD thesis we reckon that in addition to the permanent ones there other, impermanent patterns which are observed only from time to time. We will name these impermanent patterns *phenomena*. In other words, just as traces can be *fixed* or *moving*, patterns can also be *permanent* or *phenomena*.

Algorithms are also impermanent patterns (observed only when an algorithm is being executed). We will present algorithms as sequences of events. Typically, algorithms are understood as sequences of *actions* on the basis of the assumption that the protagonist is always the one who executes the algorithm. In this PhD thesis the algorithm will be any pattern and if the actions of an agent are aimed at maintaining the pattern, then we will say that the agent is the one who executes the algorithm.

When a phenomenon is associated with the observation of an object, we will call that phenomenon a *property*. This takes us to an abstraction of a higher order, namely the abstractive concept of *object*. Objects are not directly observable, but are still identifiable through their properties.

The next abstraction we get to will be *agents*. Similar to objects, we cannot observe agents directly, but can still gauge them on the basis of their actions. In order to describe the world, we have to describe the agents which live that world and explain what we know about these agents. The most important thing to describe about agents is whether they are our friends or foes and accordingly what will they try to do to us by their actions – help us or disrupt us?

The descriptions above relate to computable worlds (ones that can be emulated by a computer program). While any presence of a non-computable agent in the world would make the world itself non-computable, there is another way to make non-computable worlds. We may add a rule which depends on the existence of some algorithm (more precisely, on the existence of an execution of that algorithm). The question “Does an execution of the algorithm exist?” is non-computable (halting problem, Turing (1937)).

Contributions

1. Event-Driven model. (The concept has already been introduced by Dobrev (2018), but that paper did not provide an interpretation of the ED model. This is important, because interpretations are what makes models meaningful and distinguish between adequate and inadequate models.)

2. This PhD thesis proves that Markov Decision Process (MDP) is a special case of an ED model and that an ED model is the natural generalization of MDP.

3. Simple MDP. We will simplify the MDP to obtain a more straightforward model which can describe more worlds.

4. Extended model. This is the model in which the state knows everything. We will use this model in order to introduce an interpretation of events and ED models. (Although the Extended model was introduced in Dobrev (2019a), in that version of the model the state knows only what has happened and what is going to happen, but does not know what is possible to happen. Therefore, the state of the Extended model in Dobrev (2019a) knows nothing about the missed opportunities. In Dobrev (2019a) the Extended model is referred to as “maximal”.)

5. A definition of the concept *algorithm*. We have presented the algorithm as a sequence of events in arbitrary world. Further on, we present the Turing Machine as an ED model found in a special world where an infinite tape exists. Thus we prove that the new definition generalizes the *Turing Machine* concept and expands the *algorithm* concept.

6. A language for description of worlds such that the description can be searched automatically without human intervention.

2.3 The chess game

Which concrete world are we going to use in order to create the new language for description of worlds? This will be the world of chess.

Let us first note that we will want the world to be partially observable because if the agent can see everything the world will not be interesting. If the agent sees everything, she will not need any imagination. The most important trait of the agent is the ability to imagine the part of the world she does not see at the current moment.

For the world to be partially observable we will assume that the agent sees just one square of the chessboard rather than the entire board (Figure 2). The agent’s eye will be positioned in the square she can see at the moment, and the agent will be able to move that eye from one square to another so as to monitor the whole board. Formally speaking, there is not any difference between seeing the whole board at once and exploring it by checking one square at a time – in either case one gets the full picture. There will not be any difference only if you know that by moving your sight from one square to another you will monitor the whole chessboard. In practice the agent does not know anything, so she will need to conjure up the whole board, which however will not be an easy process and will require some degree of imagination.

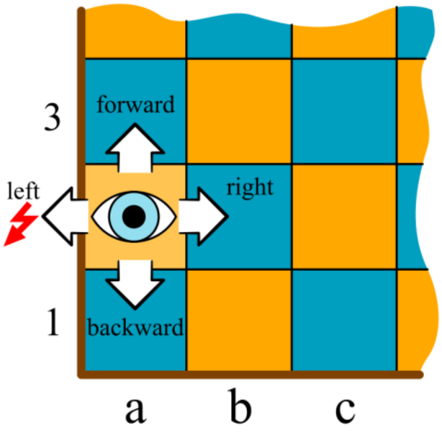


Figure 2

In Figure 2, the agent has her eye in square **a2** and can move it in all the four directions. (Right now she cannot move it left because this is the edge of the board and a move to the left would be incorrect.) In addition to moving the eye in the four directions, the agent can perform two other actions: *“Lift the piece you see right now”* and *“Put the piece you lifted in the square you see right now”*. We will designate the two additional actions as *Up* and *Down*. These six actions will enable the agent monitor the chessboard and move the chess pieces, and that’s everything one needs to play chess.

2.3.1 A chess game with a single player

We will examine two versions of a chess game – a game with a single player and a game with two players.

How do you play chess with a single player? You first move some white piece, then turn the board around, play some black piece and so forth.

We will start by describing the more simple version in which the agent plays against herself. This version is simpler because in that world there is only one agent and that agent is the protagonist. Next we will examine the more complicated version wherein there is a second agent in the world and that second agent is an opponent of the protagonist.

The question is what can be the goal when we play against ourselves?

2.3.2 The goal

While the authors of most papers dedicated to AI choose a goal, in this PhD thesis we will not set a particular goal. All we want is to describe the world, and when we figure out how the world works we will be able to set various goals. In chess for example our goal can be to win the game or lose it. When we play against ourselves, the goal can vary. When we play from the side of the white pieces our goal can be *“white to win”* and vice versa.

Understanding the world does not hinge on the setting of a particular goal. The Natural Intelligence (the human being) usually does not have a clearly defined goal, but that does not prevent him from living.

There are two questions: “What’s going on?” and “What should I do?” Most authors of AI papers rush to answer the second question before they have answered the first one. In other words, they are looking for some policy, and a policy can only exist when there is a goal to be pursued by that policy. We will try to answer only the first question and will not deal with the second one at all. Hence, for the purposes of the present PhD thesis we will not need a goal.

In most papers the goal is defined through *rewards* (the goal is to collect as many rewards as possible). When referring to Markov decision process (MDP) we will assume that rewards have been deleted from the definition because we need them only when we intend to look for a policy, while this PhD thesis is not about finding policies.

2.4 Conclusion

Our task is to understand the world. This means we have to describe it but before we can do so we need to develop a specific language for description of worlds.

We have reduced the task of creating AI to a purely logical problem. Now we have to create a language for description of worlds, which will be a logical language because it would enable the description of non-computable functions. If a language enables only the description of computable functions, it is a programing and not a logical language.

The main building blocks of our new language are Event-Driven models. These are the simple modules which we are going to discover one by one. With these modules we will present patterns, algorithms and phenomena.

We introduced some abstractions. Our first abstraction were objects. We cannot observe the objects directly and instead gauge them by observing their properties. A property is a special phenomenon which transpires when we observe an object which possesses that property. Thus the property is also presented through an ED model.

Then we introduced another abstraction – agents. Similar to objects, we cannot observe agents directly and can only gauge them through their actions.

We created a language for description of worlds. This is not the ultimate language, but only a first version which needs further development. We did not provide a formal description of our language and instead exemplified it by three use cases. That is, instead of describing the language we found the descriptions of three concrete worlds – two versions of the chess game (with one and two agents, respectively) and a world which presents the functioning of the Turing Machine.

Note: It is not much of a problem to provide a formal description of a language which covers all the three worlds, but we are aiming elsewhere. The aim is to create a language which can describe any world, and provide a formal description of that language. This is a more difficult problem which we are yet to solve.

We demonstrated that through its simple constituent modules, the language for description of worlds can describe quite complicated worlds with multiple agents and complex relationships among the agents. The superstructure we build on these modules cannot hover in thin air and should rest on some steady fundament. Event-Driven models are exactly the fundament of the language for description of worlds and the base on which we will develop all abstractions of higher order.

3 What shall we do once we are done with creating AI?

3.1 AI should not be an Open Source Project

Who should own the Artificial Intelligence technology? It should belong to everyone, properly said not the technology *per se*, but the fruits that can be reaped from it. Obviously, we should not let AI end up in the hands of irresponsible persons. Likewise, nuclear technology should benefit all, however it should be kept secret and inaccessible by the public at large.

3.1.1 Introduction

Many advocate the idea that AI technology should be disseminated in an unrestricted manner and even that it should be an Open Source Project. These proponents well include responsible and earnest figures such as President Macron (Macron, 2018). Here we will try argue a little bit with these people and highlight to them how inappropriate and even devastating such a scenario would be.

When President Macron (Macron, 2018) refers to open algorithms, probably he tends to mean the ownership of these algorithms. While it is not a bad idea to let everyone own them, it does not mean that the code of these algorithms should be available to all. Similarly, a nuclear power plant may be owned by the State, i.e. by all of its citizens, which is not to say that the technologies used to run the plant are publicly available and anyone can pick up the drawings and assemble a power plant in their backyard.

What we see now is a grossly irresponsible attitude to the technology of Artificial Intelligence. We are still in a very nascent phase of AI and can hardly image the kind of mighty power and unsuspected opportunities lurking in there. What happened in 1896? In that year Henri Becquerel (Becquerel, 1896) found that if one placed a lump of uranium ore on a photographic plate and put the two in a drawer, after some time the plate gets bleached. If one inserted an object such as a key between the two, a rendition of the key will appear on the plate. Although Becquerel had thus discovered the phenomenon of radioactivity, at that time he was unable to imagine the potential hidden in this technology. Becquerel’s experiment was more of an amusement and a magician’s trick. This is exactly what happens with AI now. Many interesting and entertaining experiments are being made, but people have not the slightest idea where this technology can take us to.

Back in 1896, were people able to foretell how mighty and ominous nuclear technology is? They had no clue. A clue did not emerge before it was found how much energy is released from the fission of atomic nuclei.

Can we now figure out the dangerousness of AI technology? Yes, and many reasonable people are aware although they do not fully understand the actual width and depth of this discovery.

Every reasonable and responsible person should consider whether they should take part to the development of this new technology or leave that to harebrained and irresponsible individuals.

This article deals with tech disasters which are avoidable rather than with the inevitable and unavoidable consequences from AI. For example, if we gave a chain saw to a harebrained person then he can fell the whole forest and that is unavoidable consequence. If however the fool cut their leg, that would be a tech disaster which could have been avoided in case the fool was less stupid and more cautions.

We say that one extremely powerful intellect is something dangerous. This is not a new idea. Adorno and Horkheimer (2002) already said in their time that reason can be another form of barbarity. They said that intellect can help people alter nature in an indiscriminate and barbaric way. In Adorno et al. (2002) they did not refer to artificial intelligence, but to the bureaucratic machine. Since the similarities between AI and a bureaucratic machine are more than the differences between the two, what they said in Adorno et al. (2002) may well be applied to our topic. In Adorno et al. (2002) the authors explored a scenario where a group of people use the bureaucratic machine as a weapon for oppressing the others (a totalitarian State). However, Adorno et al. (2002) does not deal with a scenario where the bureaucratic machine spins out of control and turns its workings against the will of humans, because this is not possible. Society as a whole can always change the laws and the rules which govern the bureaucratic machine. This means that society as such cannot lose control of the bureaucratic machine, but the individual member of society lacks any control whatever. From the individual’s perspective the bureaucratic machine is an existing reality which he or she can nowise change. The situation with AI is similar. Society as a whole will retain control on AI, provided we are not stupid enough to let it run away, but for the single individual AI will be something carved in stone which cannot be altered. The latter belongs to the unavoidable consequences from the emergence of AI and falls therefore outside the scope of this article.

3.1.2 What may go wrong?

A disaster, caused intentionally or inadvertently, may occur. In the history of nuclear technology, the names of two such disasters are Hiroshima and Chernobyl. The first was caused willfully, while the latter was the result of stupidity and negligence.

An intended disaster occurs when someone decides to use a technology maliciously, i.e. as a weapon.The notion of maliciousness as used here is relative, because all creators of weaponry believe they are creating something useful and, while killing people, they save the lives of other people. The usual explanation is that we kill less in order to save more or at least we kill members of alien nations to save members of our nation.

Can AI kill? What about the laws of robotics (Asimov, 1950) postulated by Isaac Asimov? Do they work or not? Essentially, these laws are no more than good wishes without any binding effect on AI developers. At present, the technologies which claim to have something in common with AI are mainly used in weapon systems. One example are the so called “smart bombs”. The commonest assertion is that a stupid bomb kills indiscriminately, while a smart bomb kills only those who we have told it kill. Thus, a smart bomb is described as less bloodthirsty and more humane. The latter is rather an excuse for those who develop smart bombs. These bombs are more powerful than stupid ones and enable us kill more people than we could possibly do with the old stupid bombs.

Let us imagine that a tech disaster has occurred, somebody has let the spirit out of the bottle and AI has spun out of control. If we consider AI as a weapon, let us imagine that somebody uses it against us. Thus, our attitude to AI as a weapon will flip to highly negative because the attitude to a weapon strongly depends on who is using it: us or somebody else against us.

Do we stand any chance of surviving such a tech disaster? The answer is that we do not have any chance at all. Science-fiction movies such as *The Terminator* (Cameron, 1984) depict humans waging wars against robots, however these are very stupid robots, much more stupid than humans. The truth about AI is that it will outsmart by far any human being. Therefore, the very notion of fair rivalry between humans and robots is meaningless. It is equally meaningless to run a fair rally between humans and motor vehicles. Humans have no chance because vehicles are much faster.

This is to say that we cannot afford such a tech disaster for the mere reason that we will not survive it. From a historical perspective, mankind has been through many natural calamities and tech disasters, however, the impacts of these have always been confined at some local level. The Halifax accident, for example, caused an explosion which destroyed the city. Nevertheless, the devastations were local and limited to one city. A nuclear war is cited as the most dreadful catastrophic scenario, but the impacts of even this scenario would be local. A nuclear wars may destroy all cities, but at least a couple of villages will survive. Therefore, even an eventual nuclear war does not pose a risk of the magnitude we may face if control on AI is lost.

3.1.3 Conclusion

We should take AI technology very seriously and restrict the access to all programs that have something to do with that technology. We should also classify AI-related articles and go as far as locking all computers to prevent random people make experiments with AI technology. Such experiments should only be allowed for persons who are sufficiently intelligent and responsible. At present, medical doctors need to demonstrate compliance with a range of requirements before they are allowed to practice. Conversely, anyone can undertake AI research at will. This needs to be changed and AI researchers should become subject to certain requirements.

Letting AI spin out of control is a tech disaster which may occur from either stupid or irresponsible behavior. People with inferiority complexes, who may seek to acquire absolute power and become “Masters of Universe”, should not be allowed to deal with AI.

On the other hand, independent researchers should be allowed to publish (in classified magazines) and thus earn recognition for their work. It goes without saying that when independent researchers publish their works they should be given a date stamp and a guarantee that nobody would be able to challenge or steal their merit.

What we mean by a tech disaster are things which are avoidable in principle and can be avoided if we are sufficiently smart and responsible. These things are: the control on AI to be lost, AI to be used as a weapon or a group of people to use AI as a means to oppress everyone else.

AI technology has many other implications which are unavoidable. One consequence we are unable to avoid is people losing their jobs. But, besides being unable, we are also unwilling to avoid it because nobody likes to be forced to work. We would be happy do work for pleasure, but hate to work because we have to.

Many people worry that secret services tamper with their computers. Secret services today see and know everything. God also sees and knows everything, but nobody seems to worry about that. Of course, God is discrete and good-minded. He will not tell your wife that you are cheating, neither will he use the information in your computer to make some private gains. Secret services are also discrete, but not always good-minded. Not coincidentally, the ugliest bandits are usually former or even acting officers of secret services.

We do not need to worry that secret services see everything. This is unavoidable. It is silly to worry about unavoidable things which we cannot change at all. We said that the persons working for these services are responsible ones but it does not mean they are responsible enough. The way to avoid the problem is not to play hide-and-seek with secret services, but control them and make sure only the right people work for these services.

My assertion is that we should let secret services control our computers officially and *not* under cover. If we did so we would forget about problems such as hacks, viruses or SPAM. Our computers will be safe and reliable. We may even use secret services as a backstop and ask them to recover the information lost when our hard disk goes bust. They store that information anyway.

While who works for secret services is important, an even more important question is who will be allowed to do AI research. These should be smart, reasonable and responsible persons, free of inferiority complexes or criminal intents. If we made this kind of research and experimentation open to anyone who wishes so, the tech disasters to follow will dwarf Hiroshima and Chernobyl to near-miss incidents.

3.2 What will our life look like once AI is here ?

AI is about man creating a being which is incomparably smarter than man himself. Although that being will be good-minded and will serve us faithfully, it can still be a problem for us because now we pride ourselves for being the smartest beings – smarter than all other animals and even machines. Our intelligence gives us the self-confidence to claim that we are the most sophisticated product of evolution. Now we manage planet Earth and decide which animal or plant deserves to live, which should be allowed to reproduce, proliferate and occupy more space, and which must be constrained only in natural reserves.

The question is what will our life look like once AI is here. When that time comes, our life will appear deceitfully easy. We will not have to bother about our food or livelihood, will not have to work and even will not have to entertain each other because AI will deliver entertainment much better than any human entertainer could do. Despite the deceitful simplicity of life, natural selection will continue and some of us will be on a trajectory to survival, while others will be on a trajectory to extinction.

What we mean by natural selection is reproduction instead of death. Almost no one will have to die after AI comes by. Organ repair and cloning techniques will make the human body practically perpetual, but we may still decide to set some cutoff point and say that nobody will be allowed to live longer than 120 years. How many people shall we let live on the Earth? We may keep them at 7 billion or increase the number to 70 or even 700 billion, but any case we will need to set some limit because in a congested world people will be mutually disruptive to their affairs, moreover there will not be any space left for other species.

What are we going to do after AI is here? Since we will not have to work in order to make our living, the only meaningful mission would be to engage in reproduction. Although reproduction has been and continues to be key today, in today’s world we need to work before we can reproduce. When working will not be important anymore, money will become irrelevant because it will lose its basic function as a measure of how good people are at their jobs. Then, which criterion will drive natural selection? The criteria now are: intellect, beauty, health, education, strength, bravery, swiftness, honesty, religion and worldview.

Strength and swiftness were very important in the past, but now, when machines are much stronger and swifter than us humans, these two traits are not the most important ones. When machines become smarter than us, intellect will not be the most important criterion, either. Bravery is a complicated criterion. On one side, brave people win, but on the other side the bravest ones tend to brake their necks. Honesty is similar. The most successful businesspeople and politicians are those who rarely shine with honesty, but the most dishonest ones end up in jail. From an evolution advantage in the past, education now tends to be a disadvantage. My teacher of fine arts used to say that the percentage of spinster women among university graduates is much higher than the average, and when a lady makes it to a doctoral degree the situation becomes nothing short of desperate. In other words, should education continue to be a virtue and should the more educated people be given more chances?

As regards health, good health is certainly valuable, but if everyone is healthy it cannot be a criterion. Beauty is another very important criterion, but it is very subjective. Who will sort out the beautiful from the ugly? Given that deep cosmetic corrections are available even now, with AI we will able to craft our appearance whichever way we like.

Who will determine the new natural selection criteria? Perhaps we should leave this to the smarter one, i.e. AI? The gardener is the one to decide which flower is nice, worth to propagate and have more space in the garden. If we, humans, want to be the criteria setters, then we should consider what will happen. In this case religion and worldview will be the most important criteria. The dominant worldview will have the upper hand and provide better opportunities to those who share the religion of the majority.

Conclusion

This PhD thesis leads to certain conclusions. The main conclusion is that Artificial Intelligence is around the corner and will soon be here to change fundamentally the way we live.

The main change will be that the price of labor will be zero. We are accustomed to the fact that due to technical progress labor becomes increasingly cheaper. For example, digging a pit was costly when people used to dig manually, but now when we have excavators this cost has dropped dramatically. We are accustomed to the fact that goods become increasingly cheaper in the market because the labor component of their cost continually decreases over time. For the sake of clarity we will say that cheaper goods do not cost less money, but less of something which remains invariable, such as gold (although gold also gets cheaper since machines excavate a lot more gold than people were ever able to mine manually). We are accustomed to the fact that labor gets cheaper, but are not ready for the time when the price labor will be zero.

A sequel of this PhD thesis and of its main conclusion is Dobrev's patent (2021a). The patent describes a traffic management scheme for automated metro systems (automated means that trains travel autonomously without train drivers). An automated metro system is not AI just as an automated coffee maker is not AI, either. Certainly, automated metro is something more complex than an automated coffee maker, but still it is not AI. We already know what AI is because in this PhD thesis there is a definition of Artificial Intelligence. Thus, we already know or at least the author thinks he knows what AI is.

Dobrev's patent (2021a) is not a direct outcome of this PhD thesis, but the reason for its creation is the main conclusion made here. When labor will cost nothing it will not make sense to invest in labor. Most inventions aim to save some human labor. All these inventions will become meaningless. Dobrev's patent (2021a) aims to save time and energy – these are resources which will remain valuable even when AI is up and running. Certainly, energy will also become free of charge when nuclear fusion reactors come online, but time will remain a valuable resource and we will continue our efforts to make things more time-efficient.

In 1988 the Bulgarian city of Varna hosted a logic conference dedicated to the 90th anniversary of Arend Heyting. Among the attendees there was a former assistant of Turing. His name was Gandhi and he was the doyen of the conference. He told us a story: Before the war Turing decided to convert all his savings in silver and bury his silver somewhere in the ground, hoping to offset the anticipated devaluation of money. He was reluctant to invest in other assets such as buildings because no one knew which buildings would be intact when the war ends.

Thus, Turing's decision to buy silver is not a direct sequel of the war, but came as an indirect consequence of his anticipations related to the war. Similarly, Dobrev's patent (2021a) is not a direct sequel of this PhD thesis, but its creation and the investment in obtaining a patent were motivated by the main conclusion made in this PhD thesis.

Then Gandhi continued his story and told us that Turing meanwhile forgot the place where he had hidden his silver. After the end of the war he made numerous attempts to find it, but, unfortunately, without success. Gandhi himself went on some of Turing's silver searching expeditions.

The conclusion is that we may try to prepare for major future events such as war or the advent of AI, but these efforts would hardly be successful. The sheer magnitude of such events makes it very difficult for people to figure out which policy is the best to follow.

Publications

The author has four publications which have become part of the content of this PhD thesis:

Chapter 1.

Dobrev, D. (2000). AI – What is this. *PC Magazine – Bulgaria, 11/2000, pp. 12–13 (for the English version see https://dobrev.com/AI/definition.html).*

Dobrev, D. (2022a). Definition of AI and a program that satisfies this definition, *viXra:2210.0120.*

Chapter 2.

Dobrev D. (2022b). Language for Description of Worlds. Part 1: Theoretical Foundation. *Serdica Journal of Computing 16(2), 2022, pp. 101-150.*

Dobrev D. (2023). Language for Description of Worlds. Part 2: The Sample World. *Serdica Journal of Computing 17(1), 2023, pp. 17-54.*

Chapter 3.

Dobrev, D. (2019c). AI Should Not Be an Open Source Project. *International Journal "Information Content and Processing", Volume 6, Number 1, 2019, pp. 34-48.*

The author has 19 other publications and one patent, which are related to the subject of the PhD thesis but are not part of its content:

Dobrev, D. (1993). First and oldest application. *1993. http://dobrev.com/AI/first.html.*

Dobrev, D. (2001). AI - How does it cope in an arbitrary world. *In: PC Magazine - Bulgaria, February'2001, pp.12-13 (on http://dobrev.com/AI/world.html in English).*

Dobrev, D. (2005a). A Definition of Artificial Intelligence. *In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-74.*

Dobrev, D. (2005b). Formal Definition of Artificial Intelligence. *International Journal "Information Theories & Applications", vol.12, Number 3, 2005, pp.277-285.*

Dobrev, D. (2005c). Testing AI in one Artificial World. *Proceedings of XI International Conference "Knowledge-Dialogue-Solution", June 2005, Varna, Bulgaria, Vol.2, pp.461-464.*

Dobrev, D. (2005d). AI in Arbitrary World. *Proceedings of the 5th Panhellenic Logic Symposium, July 2005, University of Athens, Athens, Greece, pp.62-67.*

Dobrev, D. (2007a). Parallel between definition of chess playing program and definition of AI. *International Journal "Information Technologies & Knowledge ", vol.1, Number 2, 2007, pp.196-199.*

Dobrev, D. (2007b). Two fundamental problems connected with AI. *Proceedings of Knowledge - Dialogue - Solution 2007, June 18 - 25, Varna, Bulgaria, Volume 2, p.667.*

Dobrev, D. (2008a). Second Attempt to Build a Model of the Tic-Tac-Toe Game. *June'2008 (represented at KDS 08), published in IBS ISC, Book 2, p.146.*

Dobrev, D. (2008b). The Definition of AI in Terms of Multi Agent Systems. *December, 2008, arXiv:1210.0887 [cs.AI].*

Dobrev, D. (2013a) Comparison between the two definitions of AI. *arXiv:1302.0216 [cs.AI]*

Dobrev, D. (2013b). Giving the AI definition a form suitable for the engineer. *arXiv:1312.5713 [cs.AI].*

Dobrev, D. (2014). Comparison between the two definitions of AI. *International Conference "Mathematics Days in Sofia", July 2014, Sofia, Bulgaria, pp. 28-29.*

Dobrev, D. (2017a). Incorrect Moves and Testable States. *International Journal "Information Theories and Applications", Vol. 24, Number 1, 2017, pp.85-90.*

Dobrev, D. (2017b). How does the AI understand what’s going on. *International Journal “Information Theories and Applications”, Vol. 24, Number 4, 2017, pp.345-369.*

Dobrev, D. (2018). Event-Driven Models. *International Journal "Information Models and Analyses",* Volume 8, Number 1, 2019, pp. 23-58.

Dobrev, D. (2019a). Minimal and Maximal Models in Reinforcement Learning. *International Journal “Information Theories and Applications”,* Vol. 26, Number 3, 2019, pp. 268-284.

Dobrev, D. (2019b). The IQ of Artificial Intelligence. *Serdica Journal of Computing, Vol. 13, Number 1-2, 2019, pp.41-70.*

Dobrev, D. (2021a). A metro management method where trains travel without stopping at each and every metro station. *BG patent No 67273 B1/ 15.03.2021, patent application No 112419 of 01.12.2016 (https://dobrev.com/patent.pdf).*

Dobrev, D. (2021b). Before We Can Find a Model, We Must Forget about Perfection. *Serdica Journal of Computing, Vol. 15, Number 2, 2021, pp. 85-128.*

Declaration

The author confirms that this PhD thesis is the result of his genuine scientific work. The use of prior results is acknowledged by appropriate references.

References

Adorno, T. & Horkheimer, M. (2002) Dialectic of Enlightenment. Stanford University Press, 2002, ISBN: 9780804736336.

Alfonseca, M., Cebrian, M., Anta, A., Coviello, L., Abeliuk, A. & Rahwan I. (2021) Superintelligence Cannot be Contained: Lessons from Computability Theory. *Journal of Artificial Intelligence Research,* Vol. 70, 2021, pp. 65-76.

Angelova G., M. Nisheva-Pavlova, A. Eskenazi, Kr. Ivanova (2021) Role of Education and Research for Artificial Intelligence Development in Bulgaria until 2030. *Mathematics and Education in Mathematics, Proceedings of the Fiftieth Spring Conference of the Union of Bulgarian Mathematicians 2021, ISSN:1313-3330, pp. 71-82.*

Asimov, I. (1950). I, Robot (The Isaac Asimov Collection ed.). New York City: Doubleday. ISBN 0-385-42304-7.

Cameron, J. (1984). The Terminator. *American science-fiction action film.*

Church, A. (1941) The Calculi of Lambda-Conversion. *Princeton: Princeton University Press.*

Council of Europe (2022) What's AI? https://www.coe.int/en/web/artificial-intelligence/what-is-ai

Becquerel, H. (1896). "Sur les radiations émises par phosphorescence". Comptes Rendus. 122: 420–421.

Dobrev, D. (2000). AI - What is this. *PC Magazine - Bulgaria, 11/2000, pp.12-13 (on https://dobrev.com/AI/definition.html in English).*

Dobrev, D. (2005a). A Definition of Artificial Intelligence. *In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-74.*

Dobrev, D. (2018). Event-Driven Models. *International Journal "Information Models and Analyses",* Volume 8, Number 1, 2019, pp. 23-58.

Dobrev, D. (2019a). Minimal and Maximal Models in Reinforcement Learning. *International Journal "Information Theories and Applications",* Vol. 26, Number 3, 2019, pp. 268-284.

Dobrev, D. (2019b). The IQ of Artificial Intelligence. *Serdica Journal of Computing, Vol. 13, Number 1-2, 2019, pp.41-70.*

Dobrev, D. (2019c). AI Should Not Be an Open Source Project. *International Journal "Information Content and Processing", Volume 6, Number 1, 2019, pp. 34-48.*

Dobrev, D. (2021a). Метод за управление на метрото, при който влаковете се движат без да спират на всички спирки. *BG патент № 67273 B1/ 15.03.2021 г., заявка № 112419 от 01.12.2016 (https://dobrev.com/patent.pdf).*

Dobrev, D. (2022a). The AI Definition and a Program Which Satisfies this Definition, *arXiv:2212.03184 [cs.AI].*

Dobrev D. (2022b). Language for Description of Worlds. Part 1: Theoretical Foundation. *Serdica Journal of Computing 16(2), 2022, pp. 101-150.*

Dobrev D. (2023). Language for Description of Worlds. Part 2: The Sample World. *Serdica Journal of Computing 17(1), 2023, pp. 17-54.*

Hernández-Orallo, J., & Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. Proc. intl symposium of engineering of intelligent systems (EIS'98), February 1998, La Laguna, Spain (pp. 146–163). : ICSC Press.

Hernández-Orallo, Dowe, D.L. (2012). IQ tests are not for machines, yet. *Intelligence (2012), doi:10.1016/j.intell.2011.12.001*

Hernández-Orallo, Dowe, DL. (2014a). How universal can an intelligence test be?. *Adaptive Behavior. 22(1):51-69. doi:10.1177/1059712313500502.*

Hernández-Orallo, J.; Dowe, DL.; Hernández Lloreda, MV. (2014b). Universal psychometrics: measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research. 27:50-74, ISSN:1389-0417. doi:10.1016/j.cogsys.2013.06.001.*

Hernández-Orallo, Javier Insa-Cabrera. (2014c). Definition and properties to assess multi-agent environments as social intelligence tests. *arXiv:1408.6350 [cs.MA].*

Hernández-Orallo, J. (2017) Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review 48, 397–447, ISSN:0269-2821. https://doi.org/10.1007/s10462-016-9505-7.*

Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv:cs.AI/0004001 [cs.AI]*

Ivanova Kr., M. Nisheva, E. Eskenazi, G. Angelova, N. Maneva (2020) Artificial Intelligence in and for Education in Bulgaria – Measures for Achievement Reliable Intelligent Growth. *Proc. of the 13th Nat. Conf. "Education and Research in the Information Society", 2020, pp. 7-20.*

Macron, E. (2018). Emmanuel Macron Talks to WIRED About France's AI Strategy. *31 of March, 2018, www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy*

Solar-Lezama, A. (2023) AI will program itself: synthesis, learning and beyond. *Lecture from "INSAIT Series on Trends in AI & Computing", April 3, 2023, Sofia University.*

Turing, A. (1937). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*. Wiley. s2-42 (1): pp. 230–265.

Turing, A. (1950) Computing machinery and intelligence. *Mind, 1950.*

Wang, P. (1995) Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. *Ph.D. Dissertation, Indiana University.*

Wang, P. (2019) On Defining Artificial Intelligence. *Journal of Artificial General Intelligence 10(2) 1-37, 2019.*