A High Level Theory on the Nature of Intelligence and Consciousness

Arnau Garriga-Casanovas

***Abstract*—Research into artificial intelligence has increased significantly in recent years. However, the fundamental question of what intelligence is and how it works remains open to some extent. Traditional definitions of intelligence are broad and lack clarity regarding its nature and mechanisms. The nature of consciousness is another matter that has been widely explored with multiple theories but for which we do not have a final agreed theory, especially in terms of its relation to intelligence. In this work, we present a preliminary theory of the nature of intelligence and its working mechanisms. We contrast it against observations to show that our theory is consistent with observed reality. We also use our theory of intelligence to offer a theory on the nature of consciousness, as well as its relation to human understanding and intent. We then show that our theories can be consistent with the theory of evolution. Similarly, we explore examples of how our theory applies to specific cases to show its consistency across applications. Lastly, we outline how our theory can pave the way towards artificial general intelligence. Our theory is unproven.**

[1]

I. Introduction

Research on artificial intelligence (AI) has increased significantly in recent years, but it there is still no general agreement about what intelligence is and how it works. A conventional definition of intelligence is that it is the ability of an agent to achieve its objectives in an environment [1]. This broad definition, however, does not clarify much about the nature of intelligence or how it works, and makes it difficult to create an artificial general intelligence (AGI) based on it.

There are multiple paradigms about intelligence. A classical one identifies intelligence with rationality, and defines intelligent agents as those that act in a perfectly rational manner, maximising a utility function in a bounded optimality problem, as pioneered by [2] and elaborated in [3]. Logical AI has seen significant interest and development [4], [5], with formalisms and techniques reaching a high level of maturity e.g. see [6]. Another paradigm focuses on non-logicist intelligence, which includes brain-based approaches to AI [7], and broader approaches using neurocomptuational techniques such as neural networks [8]. Lastly, other fields such as psychology consider that there exist between multiple kinds of intelligence, with different theories offering different breakdowns of intelligence [9]–[11]. All these paradigms, however, do not lead to a clear agreement on a complete theory of intelligence, and generally do not offer a clear blueprint for the implementation of intelligence. In addition, these formalisms offer limited insight into the nature of human thought and how to create an AGI that resembles or surpasses human intelligence.

More recently, in [12], the authors hypothesise that intelligence can be understood as the maximisation of a reward by an agent acting in its environment. This "reward-is-enough" theory indicates that AI, and potentially AGI, can be created using reinforcement learning (RL) for an agent with a singular reward in a given complex environment. Elements of this "reward-is-enough" theory are used as part of the intelligence theory presented in this paper.

Another fundamental question intrinsically related to intelligence is whether humans have free will and where it stems from. The existence of a human intelligence capable of achieving objectives typically implies some form of free will. However, the source of that free will in a human brain, which can be viewed as a complex electrochemical system, the behaviour of which is determined by a set of laws of physics, is unclear.

In this paper, we present a preliminary theory offering a possible explanation of what intelligence is and how it works. Our theory links intelligence with the ability to generate thoughts. We also provide various examples that are consistent with the proposed theory, and we derive consequences from the theory, which match practical observations. In addition, we use the formulated theory to explain the sensation of free will and consciousness that humans experience.

The theory presented here elucidates the way intelligence works, and thus can be used as a basis to create an AGI. As such, we also suggest initial, high level guidelines about the way to create an AGI based on our theory, and we outline a potential high level path to achieve it.

It should be noted that definitions of intelligence can be to some extent arbitrary, and vary depending on the objective of the analysis. In this work, we propose a definition focused on the intelligence used to tackle complex problems and in deciding strategic actions, which in the literature is sometimes referred to as "system 2" or "slow thinking" [13]. As such, our work does not focus on primal forms of intelligence such as those used for object manipulation or self locomotion, nor in basic forms of intelligence consisting of a direct relation between an input and an immediately corresponding output, but rather in advanced forms of intelligence such as that used in science, engineering or business. This focus on complex intelligence is aimed at understanding intelligence in general and paving the way towards AGI. The word intelligence from this point onward thus refers to complex intelligence typically involving thought.

[1]The author is affiliated with the Mechatronics in Medicine Laboratory, Hamlyn Centre, Imperial College London. This work was conducted outside grants received. Email: a.garriga-casanovas14@imperial.ac.uk

This work is based on the analysis and observations of the author, both in terms of observed behaviour in other people, reports by other people, and in terms of introspective observation of the mental process used by the author when trying to use intelligence. This work should be understood as a personal collection of ideas and views forming a personal, preliminary theory of intelligence rather than a formal research paper. This work is an unproven theory that needs to be validated experimentally. We also provide guidelines to achieve this.

The nature of this paper is primarily philosophical, although it also intersects with AI and brain science. It should be noted that the aim is to provide a more clear understanding of what intelligence is at a high level, and not to explain the specific electrochemical processes in the human brain that are linked to intelligence.

II. Intelligence

A. Human Thought

We first present a theory of what human thought is in order to then explain what intelligence is and how it uses thoughts.

Humans receive information regarding the environment through their sensors, typically visual, auditory, haptic, olfactory, and gustatory. This information is perceived by the brain, and then the information is memorised with a certain degree of loss, which can be viewed as a degree of abstraction. In general, a full video of an experience is not recorded; only a high level overview of events with specific elements to which attention may be focused. It is a compression with loss of information.

We conjecture that the way the human brain thinks is by generating simulated sensory inputs and playing out simulations based on experience to see where they lead. The process of generating simulations based on each situation to see where they lead is what we can define as human thought. The setting at the beginning of each simulation can be based on the physical inputs in each situation. The brain looks for similar experiences in the memory, and generates combinations of those with various degrees of variation to see where each simulation leads.

Human thought therefore is experienced in a similar way as perceiving information from the environment, with the main difference that instead of receiving the information from the sensors, it comes from simulations that are based on combinations of learned memories and patterns based on those.

One simulation (thought) played out by the brain can then lead to a new simulation. This is what we refer to as chain of thought. For example, a person ordering coffee in the evening may be asked whether they prefer regular or decaffeinated. This may trigger the brain to generate a first simulation of the rest of their day after drinking regular coffee, which may involve attending the gym later in the day for which the person needs caffeine. This may lead the person to generate another simulation regarding the equipment needed for the gym, which includes trainers. This may lead the person to generate a simulation contrasting the need for trainers with the fact that the person did not bring the trainers that day, triggering a simulation that the person cannot attend the gym session. This leads to a final simulation where the person is not attending the gym and therefore does not need coffee. These thoughts are simulations generated based on elements of the previous simulation which set the scenario for the new simulation.

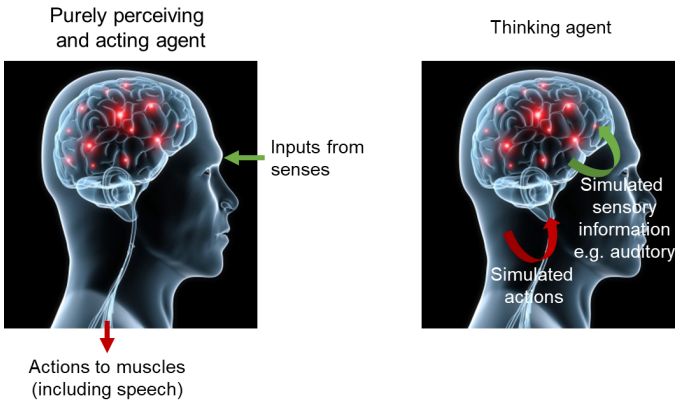

Fig. 1. Conceptual illustration of (left) a person purely receiving sensory inputs from the environment and using a low level mechanism to relate them directly to an action to the muscles, which may include speech; and (right) a person thinking, which involves generating simulations of sensory information together with simulations of actions and evaluating their outcome.

The simulations (thoughts) can be based on images, audio (typically words forming an inner speech), and in less frequent cases can also be generated using other senses such as olfactory, gustatory, or haptic (including internal sensations, e.g. in the stomach or heart). Multimodal simulations are also possible, and in most cases combine visual and auditory content.

There are cases of people who lack the ability to generate mental imagery, a condition known as aphantasia [14], [15]. In those cases, thoughts can be generated using simulations involving other modes, such as auditory, haptic, or simulations based on smell and taste. This is consistent with the result of experiments with aphantasic people who are given an image, and are later asked to draw the image from memory. They commonly report that they rely on verbal lists of items in the image to later produce a drawing of it [16].

There are also cases of people who lack the metacognition to be aware of their thoughts, but still rely on visual, verbal, or other modes of mental simulation to think. We conjecture that in those cases, the simulations (thoughts) take place behind the scenes in their mind while they do not notice it.

B. Intelligence Theory

Our proposed theory is that intelligence is the capability that, at its core, involves first generating a set of simulations (thoughts) based on each scenario that are relevant to it, then evaluating the outcome of the simulation against a reward function, and lastly selecting the simulation that leads to the highest reward to implement it as the course of action. The implementation involves following an action pattern similar to that of the successful simulation adapted to the scenario at hand.

There are therefore four key ingredients to achieve intelligence in an individual or more general agent:

1. The ability to identify patterns in the information perceived in the past to find structure and relations in it (the more interesting are consequential relations). In some fields, this is referred to as the ability to create a model of observed reality.

2. The ability to relate the current situation to other situations in the past that are stored in memory, akin to a similitude function, to be able to react to the current situation.

3. The ability to generate realistic simulations of future developments that are relevant to each situation using the patterns identified in the past. This is the part that is missing to some extent in current large language models (LLMs) and similar generative AI developments.

4. The ability to evaluate the outcome of each simulation against a reward function, to then select the most suitable to implement.

These ingredients deserve clarification. Beginning with the fourth, to achieve intelligence it is necessary to have a reward function that maximises things that we consider useful according to a predefined set of primal objectives.

Regarding the third, an intelligent entity needs to be capable of generating simulations that are useful, which means that they are broad enough to consider all relevant cases in each scenario but not excessively broad to waste time on pointless simulations. These simulations need to be generated based on the present inputs in each situation. This means that given a situation in which an individual finds itself, they need to be capable of generating simulations based on that scenario that are accurate based patterns identified in the past and that are useful to determine actions. These simulations in simple cases are possible ways in which the future may play out depending on possible actions. In more complex cases of thought, the simulations are abstractions that in one way or another contribute to the determination of the eventual actions of the individual in the future.

It should be noted that the simulations generated, or thoughts, are not strictly limited to future ways in which things may play out in a direct manner. The simulations may involve more abstract simulations, such as, generating mental diagrams to analyse the current scenario, and the subsequent application of a set of procedures previously trained, to then reach a conclusion. These more abstract simulations are ultimately also ways to simulate future ways in which things may play out in an indirect manner. Indeed, for example when studying a structural engineering problem, a person may perform a mathematical analysis to decide how to dimension a structure. The mathematical analysis is a prediction of how the structure will behave for a given design according to a set of laws of physics that the person has learned either directly or indirectly through books. The laws of physics in this context can be viewed as predictions of how the reality will play out given a set of inputs. Thus, by performing a mathematical analysis based on laws of physics, the person is ultimately predicting the future behaviour of the structure for different cases of structural

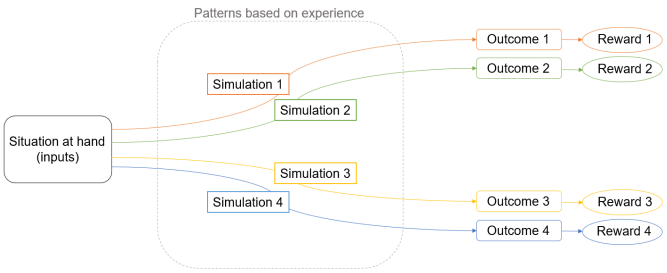

Fig. 2. Conceptual illustration of intellectual process used to make decisions. It begins with a given situation or problem, then a set of simulations are generated (thoughts), which in simple thought processes are ways in which the situation may play out depending on potential actions and in more complex thought processes are abstractions of strategies as described in subsequent sections of this paper, then these simulations lead to a set of outcomes, and these are finally evaluated against a reward function to select the best course of action.

dimensions, evaluating the result, and selecting the one that maximises its reward function. The procedure to perform the future prediction however is abstract and complex, and it has been learned by the person from previous experience. It is the best way to generate simulations given that particular type of problem to maximise its reward. These abstract thoughts, or simulations, are further elucidated in the next subsection.

Regarding the second point at the beginning of this section, intelligence requires the capability of finding similarity between situations. Specifically, intelligence involves the capability of checking the current situation against the information stored about past situations to find relations to similar situations. Finding these relevant simulations may be seen as a matter of assigning weights to different relations between the situation at hand and memorised scenarios depending on their relevance. The level of intelligence will depend on the quality of those weights as discussed in the next subsection. Attention mechanisms and transformer architectures currently used in AI are a possible way to find similar situations between that at hand and the ones stored in the training memory.

Lastly, the first element necessary to achieve intelligence is likely the most determining in intelligence and the most complex. It involves identifying relations in memorised experiences to generate simulations that are relevant and match reality. This can be seen as creating a model of the perceived world which is a set of relations between variables abstracted from the perceived world and therefore defined. The most interesting relations in the perceived world to create simulations that are realistic are consequential relations with a certain degree of abstraction, since these allow an agent to generate simulations that flow forward for a given scenario and a defined strategy in terms of a set of actions.

It should be noted that intelligence in this work is understood as complex intelligence, as described in section I. It should also be noted that the ability to generate some kind of simulations (thinking), is likely present in the majority of humans and potentially in some other life forms. The speed

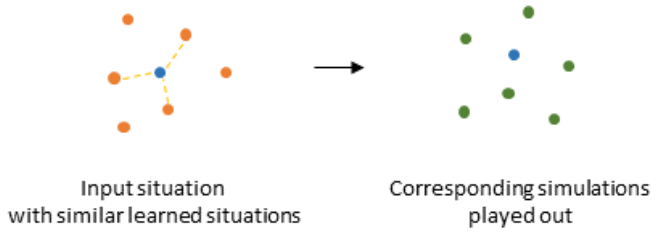

Fig. 3. Conceptual illustration of the process of using a similitude function to generate simulations. The situation at hand is abstracted by its features, illustrated as a blue point in the left point cloud, and compared against the situations in the training data. Relevant situations in the training data (illustrated as three dashed lines) are then used to generate simulations by inputting the features of the situation at hand and playing out a simulation that combines the main elements of the relevant situations in the training data. The other simulations played out shown in green are based on the training data inputs shown in orange.

at which the simulations are executed, and the potential for parallelisation, may vary between people. In this work, we do not consider whether speed and parallelisation varies or not, and we leave it for future work. Computers also possess the capability to generate simulations at various degrees of speed and parallelisation.

The definition of intelligence considered in this work is thus akin to what is also known as ”system 2” thinking, ”slow thinking”, or ”conscious thinking” in some literature, e.g. [13].

C. Intelligence Implementation

The way an intelligent agent makes decisions and defines a course of action is by generating simulations of situations similar to the one at hand, and then selecting the one that maximises the reward function. This selected simulation is the decision to be acted out. The agent then produces actions by comparing the abstract summary of the selected simulation and the abstract summary of the situation at hand, and applying the actions in the simulation to the situation at hand. It does so by taking the situation at hand, abstracting it, comparing it to the abstraction of the selected simulation, applying the actions in the abstracted simulation to the abstract of the situation, and then going back to the detailed version of the situation at hand but with the actions substituted by those of the simulation.

The generation of the simulations is performed using a similitude function, as conceptually illustrated in Figure 3. The features of the situation at hand are extracted and compared against the training data to identify relevant situations and strategies in the training. A first simulation is then played out by combining the relevant simulations and performing an abstract form of interpolation between them.

Giving advice

A relevant case to consider at this point to illustrate this theory is the one that arises when a person A gives advice to another person B. Typically, the person asking for advice describes a situation briefly with only a small set of elements. The short and abstract description usually serves person A to already give advice. In some cases, person A may ask for additional information where it sees a bifurcation in the way simulations play out depending on a given factor, but this is typically in the form of a small number of additional factors, and then person A is ready to give advice. This indicates that person A can generate simulations relevant to the situation that person B faces, and can develop these simulations based only on a small set of elements describing the situation. Person A can then choose among the possible simulations to select the one that maximizes a reward function and give that advice to person B.

D. Intelligence Levels and Illustration

Humans, other life forms, and artificial agents may have different levels of intelligence, which in some fields are quantified as intellectual quotient (IQ). According to the previous sections there are four main ingredients needed for intelligence. In this subsection, we relate those ingredients to more practical elements in the creation of intelligence to help explain how those ingredients map to intelligence in agents and how they can lead to different IQ levels.

E. Reality matching

A central factor in achieving intelligence is the accuracy of the simulations that are generated given a situation, which means whether the thoughts or simulations that a person executes given an input are a good representation of reality. According to the previous section, this relates to ingredient 1, i.e. having a good model of reality. The development of this ingredient depends on the training that a person underwent during its life and the relations formed from that, i.e. the experiences that the person has received over their life and the patterns that it formed in their brain.

For example, a less intelligent person may have incorrectly identified a relation that if they are associated with a certain zodiac sign, they will have bad luck on certain actions. This person will then avoid those actions, reducing their choices, and thus leading to worse outcomes. The more intelligent person will have identified the fact that there is no relation between the zodiac sign and their luck, thus they will generate simulations and take actions based on a more accurate set of relations that represent reality, which will lead to better outcomes.

Focus

Another factor in achieving intelligence is the usefulness of the simulations that are generated given a situation, which means whether the thoughts or simulations that a person executes given an input are relevant or not. This relates to ingredient 2, i.e. having a good similarity function that evokes relations to past experiences that are useful. An intelligent person, when faced with a given situation, will have a brain (or simulation unit) that evokes a small number simulations that are relevant to the situation at hand and that match the reality of how things play out.

Distractions in thought, which can be seen as simulations that interject but do not contribute to maximising the reward, can be considered a (negative) part of the similarity function, given that distractions may form part of the thoughts that a person executes, even though these are not relevant.

Efficiency of thought

Efficiency of thought relates predominantly to ingredient 3. It can be loosely defined as the level of detail that a person considers in each simulation, and how long the simulation runs. An intelligent person will have a brain that only considers essential elements in the abstraction of each situation and only runs each simulation for the minimum required time, which allows for shorter thinking times for each simulation, and thus more potential simulations considered. The brain of an intelligent person will need to be trained to evaluate the outcome of a simulation once it reaches certain milestones or a certain period of time has elapsed.

Reward

Reward levels determine the scores that each simulation receives, and thus the course of action that a person chooses. An intelligent person will have rewards that lead that person to choose the simulation that makes it achieve its goals, instead of choosing one that is less beneficial. This relates to ingredient 4. It should be noted that the correct selection of rewards depends on the arbitrary goals set for the person, and thus any set of rewards could be considered acceptable. However, in the conventional definitions of intelligence, which is to achieve one's goals as defined by each exercise, the rewards need to be defined such that the person feels happy when the goals for each exercise are achieved.

People with unconventional reward functions would typically not be considered intelligent even if they always achieved their goal in their reward function because the objective in conventional intelligence definitions is to achieve the goal of exercises that can range from a wide variety of fields. As such, the reward levels to achieve intelligence in more conventional definitions need to be such that the person always achieves the goal of an exercise for all exercises typically considered in intelligence definitions.

Simple and complex intelligence

The comparison between simple and complex thought processes is useful to illustrate our theory of intelligence in general. When we think simply, we just simulate potential strategies that we may follow and we look at how the situation is likely to unfold for each of the potential strategy according to our knowledge of the world. Then we select the strategy that leads to the highest reward. By contrast, when we think deeply, which applies when addressing complex problems, such as those in mathematics or physics, we generate simulations that do not directly involve us following strategies in terms of actions. Instead, the simulations generated include items such as diagrams and words that begin potential paths to address a problem. These potential paths are based on our previously learned strategies. We then advance each simulation, following each of the paths, and see where it leads. If we get stuck or reach a point that we consider a milestone based on our previous training, we evaluate that situation. In the evaluation, we check whether our reward function considers that it is a desirable situation or not. It should be noted that an important part of this evaluation is to check for contradiction. We do that by checking if the situation that we reached in the exploration has features that contradict our previously learned information about the world.

Example of problem solving and reality matching

For example, when solving a simple mathematical problem, we may find that after manipulating an equation working strictly in the real numbers domain we reach a point where we have the square root of a negative number. Our previous training may include information that said that, when faced with this situation, it is not possible to determine the square root of a negative number, which will indicate us that the simulation we have followed is not a desirable course of action. We then cancel that course of action and follow a different path.

An interesting case occurs in this same example if we do not know that the square root of a negative number is not desirable (we may have never learned that). We may then apply the same rules of square root to the number as if it was positive and add a negative sign in the end, guessing that this is the correct approach. If a teacher tells us this is wrong, we will learn it after.

Alternatively, if we are the first person ever to encounter this problem, we will need to check what we mean by it. We will need to evaluate what we are trying to achieve, and how that relates to the observed world. We will need to see how every operation we are following matches the observed world. We will then be generating a parallel simulation in the real world, to which the mathematical analysis corresponds. The simulation in the real world based on our observations will tell us what is possible and not, and what the result needs to be based on observation. We will then define the mathematical rules and processes accordingly. Thus, our rational processes, no matter how abstract or complex, are rooted in the real world, and are simulations of how things will play out based on our observations. We then create abstractions and operate based on rules extracted from the real world.

F. Abstraction, category fit and conducting research

Abstraction and category fit

An important step when generating simulations, and particularly when tackling complex problems such as the mathematical problems described previously, is using abstract concepts that encompass a set of elements and actions identified in the observed world. In an intelligent person, it is important to have a clear relation between abstract concepts and the more practical elements to which they correspond. This is part of the world model described in point 1 of the intelligence theory. The process of abstraction can be seen as a process of knowledge distillation used in machine learning.

When we think about complex questions, and therefore generate simulations that are distant from any previous ones learned, it is common and necessary for us to check whether

each element fits each abstract concept category used in the thought process. This is particularly important when conducting research or trying to advance our knowledge by exploring thought processes, or simulations, which are not directly related to previous ones that we have seen, as in the previous mathematical problem example.

We conjecture that we check whether a new element fits in a category by defining features of that category, and checking the degree to which the new element is related to those features.

Specification of new, complex problem solving

The following is a high level illustration of the process we hypothesise we follow to solve a new complex problem by applying abstraction and category fitting together with heuristics. In general, an intelligent individual has learned a set of abstract processes or strategies to solve problems encountered in the past; these can be seen as potential approaches to solve problems. In these processes, there are a set of abstract categories, and the person fits the inputs in each situation to the abstract categories, and proceeds with the learned process to solve the problem.

When faced with a new problem, the person tries a set of previously learned approaches, starting with those approaches that are more similar to the problem at hand. The person inserts the inputs from the problem at hand to the initial categories of the given approach, and then proceeds with the abstract steps of the approach previously learned using the given inputs. The person then looks at the outcome reached to see if it has contradictions or not, and more generally what the reward function of the outcome is. Here, contradictions can be seen as penalty terms that lower the reward.

In general, the way we apply the abstract processes is by generating simulations with the given inputs of the problem and seeing how they evolve. The abstract processes may involve concatenating a set of shorter processes, where the resulting outputs of a process are fitted as the category inputs to a following process.

Example of new heuristic problem solving

As an example, we can consider a (not very good) student presented with the following problem: find x given the equation $x^3+x=130$. The student in this example has never seen such a problem or any similar third order equations. To solve the problem, the student will begin generating simulations, trying different approaches that they have learned in the past. They may start by trying to isolate x by factoring it out, as $x(x^2+1)=130$.

Here the student has followed a previously learned abstract approach of isolating the variable in an equation. They have substituted the inputs in the problem to the learned approach, and reached the equation above. The student now evaluates the situation reached, and realises that they do not know how to proceed forward. The student has also learned in past situations that this is not a desirable situation, and thus it has a low reward. As such, this approach is abandoned.

The student then tries another approach of simply plugging in random numbers in the first equation. They plug in 10, but 10^3+10 does not equal 130. They plug in 2, but it does not work either. After two failed attempts the student abandons the approach.

The student then tries a third abstract approach which is to factorize the numbers in the equation, and reaches the result of 130 factorised as 2, 5, 13. The student considers this a step forward in the reward function and keeps it.

The student then does not know any other approaches and since it has time, tries again the previous approach of substituting numbers. This time, the student tries the factorised values because in the past they have seen a case where they were useful. Thus, the latest approach is to substitute values in the equation which correspond to the factors of the independent terms in the equation. The student tries 5 and finds that it solves the equation.

The student then cements the learning of the new abstract approach which involves factorising the independent terms of an equation and trying those factors in the equation to see if they solve it.

We know that this approach is not suitable and reaching a solution was luck. However, until exploring further, this student may have learned this approach and may use it, either in mathematics, or in other problems that seem similar and where the given inputs can be inserted into this abstract approach. This heuristic process, with significantly more complexity and exploration in thought, is how we create knowledge. In the future, the student will find that in many other problems the approach does not work, and thus may realise that it was luck but not an appropriate approach. As such, their knowledge will change after exploring further. The knowledge will keep evolving until it is consistent with the world and the abstract approaches always work for a given set of problems. At that point the knowledge will stabilise, since it will be a good set of strategies and a good representation of the world.

General knowledge generation

A similar heuristic process applies when we advance knowledge in general. A person first identifies a simple pattern that can be abstracted. When faced with new questions about the world, the person applies those patterns, and checks whether the results are desirable or have contradictions. If the results are desirable and without contradiction, the pattern is added to the knowledge.

Analysis and research thought process

The process of checking approaches and categories for a given problem or situation is a common process in research and in deep thoughts. When conducting research or solving complex problems that are previously unseen, we generate simulations based on the inputs and previous approaches, and see if the current inputs fit the abstract concepts in the approaches. We then see where they lead, and if the current elements in the inputs and their evolution through the approach still fit the abstract categories in which they have been inserted. If so, we may reach a satisfactory result, and we have advanced our knowledge or solved a problem.

Repeated exploration

The heuristic process of trying different abstract approaches to a given situation is continuous and typically with a low

efficacy since the selection of possible approaches to a new problem tends to be arbitrary. When faced with a problem, we keep trying approaches repeatedly in our minds, dedicating a few seconds to each approach, potentially for hours or days until we find a suitable approach. It should be noted that often, the different approaches that we try are branches of a given approach, or variations of a given approach.

It appears to be part of the human nature to try repeatedly many approaches in our minds that do not lead to satisfactory outcomes, until we find a suitable one to act out.

G. Firmware

We conjecture that the reward function and some fundamental patterns and low level routines in the simulations explored by the brain are akin to a firmware in the brain that is innate. These elucidated in this subsection.

The fundamental reward function in the brain is hypothesised to be the result of an evolutionary process, which encourages traits that improve the survival of a species. This reward function may involve a combination of factors such as surviving, which translates as lack of physical pain or hunger in the reward function, and reproducing to pass on genetic material, which translates as rewards for social behaviours that maximise the chances for reproduction. This is widely discussed in reinforcement learning and evolutionary literature. The reward function may also include low level factors such as lack of contradiction when checking the outcome of simulations, that are needed to select appropriate courses of action and relations. As such, if a simulation leads to contradiction, it creates a penalty. The fundamental factors in the reward function are a central philosophical question which is widely discussed in the literature, and are not further elaborated here.

Fundamental patterns may also be part of human firmware, which in practice are low level routines executed by the brain at each moment. In this work we do not delve into the discussion of which routines are part of the firmware as opposed to learned, but we speculate that the firmware routines revolve around three main types:

i) Recurrent short routines such as evaluating whether to make a decision based on the simulations run up to a given time or to run more simulations delaying the moment to make a decision. We speculate these to be part of the firmware, and to be fine tuned by learning from previous similar scenarios. This basic routine is one of the first executed constantly when new inputs appear, and consists on checking whether in the past, with those inputs and scenario, it was better to think or to act based on a first reaction.

ii) Checking for contradiction. We also speculate this to be fine tuned using learning from experience.

iii) Segmenting the observed world to define objects and more general entities.

The reason for speculating these routines to be part of the firmware is that they are common in most people and to some extent in animals.

III. Consciousness and understanding

A. Free will

The theory presented here implies that humans have no free will. Humans are born with an initial hardware and firmware, they then receive sensory input, they create memories and patterns from that, and they execute simulations (thoughts) based on the previous stored memories and patterns by combining them. These simulations then are evaluated by the reward function to determine a course of action that is then implemented. The firmware and hardware determine the combination and execution of simulations, and the evaluation against a given reward function. There is no free will to make decisions.

B. Consciousness

Consciousness and the feeling of being phenomenally conscious are therefore an illusion. The perception of consciousness may arise when a person experiences an internal speech and video when it is generating simulations, i.e. it is thinking. This speech constantly playing in the mind may lead the person to form the sense of consciousness and may give the person the impression that it can control the simulations being generated, i.e. thinking freely, and make decisions. However, the simulations being generated, i.e. thoughts, which include the voice and video in the head, are generated automatically based on every given situation and established relations in the brain based on past experiences, according to the theory presented in this paper. Thus, there is no one controlling the fundamental direction in which the internal speech and video play. It plays automatically based on what is has been learned, its firmware, and the constant inputs. This theory of consciousness agrees with an existing theory termed illusionism [17].

The appearance of the sensation of self also deserves attention. A person initially will perceive the world and identify other entities, such as objects, animals, and other persons. This is likely done via a low level pattern recognition structure similar to deep learning that identifies groups of matter that are structurally together as separate entities. The person will then perceive its own body, mostly through vision and also other senses, and identify the elements of the body as matching those of other persons around. Thus, the person will identify that there is an entity corresponding to its own body, and that the thoughts of action that it experiences, as well as the sensations it experiences, match the actions performed and sensations encountered by that one body. The person will thus establish a relation between that one body and its thoughts and perception. The person then establishes that there is one individual that is directly related to its perception and thoughts. This leads to the identification of that individual as self. The simulations generated by the brain and played in the brain are then associated as the consciousness of that self.

C. Empiricism

This theory agrees with the traditional philosophical theory spearheaded by David Hume that everything we know about the world is through senses. We perceive the world through

our senses and we identify relations that we store in our memory. Our thoughts are simulations generated by combining memories, originally formed from sensory inputs.

D. Studying and understanding

An important question is what it means to study a matter, for example to prepare for an exam, which relates to the concept of understanding. To study is to generate as many simulations relating to a matter as possible and see if the outcomes are desirable or if they lead to the absurd (contradiction). Once a person has considered all angles pertaining to a matter, which means running all simulations they can conceive related to the matter, the person will say that it understands the matter, and will be able to quickly answer questions about it given that they will already have identified the best simulations. The best simulations will gain weight in their brain, and thus will be run faster and with higher reward. They will not need to run again simulations that lead to the absurd that have already been explored. Every time the person faces a similar scenario when dealing with a matter that has been understood, the person will directly evoke the most suitable simulations already identified.

In some instances, particularly when studying, we may consider a matter understood once we can enter a small set of inputs of a problem at hand and generate simulations that lead to satisfactory outputs without contradictions. This applies particularly when we can satisfactorily solve a simple problem that relates to a matter, leading to a feeling that we have understood it. However, it can occur that later on we face a more difficult problem where entering the inputs and executing an analogous simulations does not lead to a satisfactory output, and instead leads to contradiction. At that point, we realise that we did not have a good understanding of the matter. This illustrates the fact that our feeling of understanding of a matter can be subjective and is affected by the complexity of the problems considered.

E. Knowledge and intent

In general, knowledge arises when a person or agent is able to generate simulations about a matter that lead to outcomes without contradiction that match reality. This differs from reshuffling information, which does not necessarily require knowledge about the matter.

Intent similarly arises when the agent is aware of the potential outcomes given a situation and possible actions. An algorithm that generates a single output without having generated a simulation of the outcome does not have intent. Conversely, intelligence does involve intent since it involves generating those simulations of possible outcomes.

F. Sensory simulations for consciousness

There is a philosophical theory that holds that any phenomenally conscious thought is reducible to sensory experience [15], [18]. The theory presented in this paper extends that theory to include all thought to be rooted on previous sensory experience.

IV. Evolution

The theory presented here indicates that intelligence is based on a small set of basic principles. These are the ability to identify patterns in memorised information, to relate the current situation to similar ones in the memory, to generate simulations based on it, and to evaluate these against a reward to select the most suitable one for action. It is possible for these capabilities to have appeared in an evolutionary process, and this would be consistent with established evolutionary theory.

To illustrate this, we can consider an early unicellular organism without intelligence, such as that in Figure 4 (left). The organism may resemble Euglenophyta but its mechanisms should be considered entirely fictional to illustrate this example. This organism may have a flagellum that it uses to propel itself and thereby intercept nutrients. The organism initially may regularly move the flagellum in a predefined pattern and randomly intercept nutrients. Through mutation, a light sensor may appear in the organism. This sensor and a set of chemical reactions also obtained through mutation can lead to a mechanism by which the flagellum tends to propel the organism towards areas with more nutrients (for example areas with more light), and thus the organism has an advantage. The mechanism may initially be a set of chemical reactions without intelligence in it.

Through an evolutionary process of random mutations and survival of the fittest, the organism may develop more sensors and dexterity to the flagellum. Through the same process, it may also develop a mechanism to store information regarding situations where it perceived a set of inputs from the sensors, it executed a set of actions, and it obtained a certain amount of nutrients, the reward. The mechanism can simply be a set of chemical reactions that activate given a set of inputs, and instead of sending a signal to the flagellum to perform actions, it generates signals equivalent to those inputs perceived in similar past situations when the flagellum was activated in a certain pattern, and it then triggers a reward at the end based on the final sensory inputs generated by the mechanism. Thus, when faced with a new scenario in terms of sensor inputs, it can run the sensor inputs through its memorised information using this mechanism to evaluate the different outcomes based on different patterns of flagellum activation, and then execute the most suitable one.

This overall mechanism and process are still a set of reactions like in the basic flagellum with a single sensor, but in this case the reactions will involve generating simulations before choosing one action. At this stage, we propose to say that a fundamental form of intelligence appeared.

Once this fundamental intelligence appeared, it can evolve in complexity. Relations can be formed in the network of reactions generated for each sensor input to match the reality, and thus generate various simulations.

V. Examples and Corollaries

A. Deaf people

An interesting case to study is that of people that are congenitally deaf. They cannot use voices (auditory speech) to

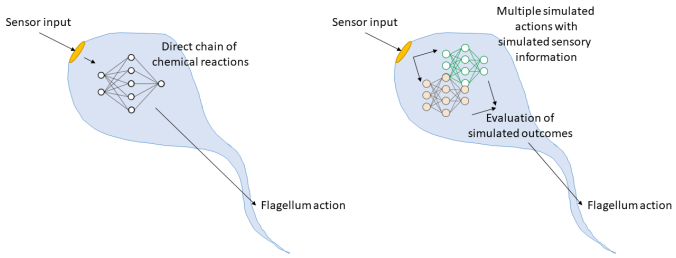

Fig. 4. Conceptual illustration of (left) generic organism (resembling Euglenophyta but not necessarily related) in primal case where it receives sensory input and through a direct chain of chemical reactions creates an output; and (right) generic organism in more advanced case where it receives sensory input, it then simulates possible actions and their outcomes based on its memorised information, and selects the action with highest reward to implement. The simulations are based on previous experience, and are neural networks where the input is the sensory input plus previously executed actions over the course of the actions, and the output is the situation reached at the completion of one of the actions over an arbitrary action time that is also learned.

think because they have never heard voices. We hypothesise that the mapping in their brains relates images, as well as haptic, gustatory and olfactory sensations, and uses combinations of these to think. Deaf people when thinking may also employ words that they have seen written, which in turn are images but can be considered simpler and thus at a higher level of abstraction, making the thought more efficient. The intelligence in that case involves having identified relations between those images, and then generating simulations of images based on previously the memorised information.

There are instances of people born deaf who report hearing voices when talking to other people. We hypothesise that these experiences to be combinations of haptic perceptions of vibration together with visual simulations associated to the conversation, which may be partly supported by tangential evidence [19]. However, these differ from the auditory simulations generated by individuals who can hear.

B. Blind people

Another interesting case study is that of congenitally blind people. In an analogous manner as in the previous subsection, blind people cannot use images to think. Thus, the mapping in their brain is likely to identify relations and patterns in words and more general audio, in addition to haptic, olfactory and gustatory sensations. Blind people are capable of reaching equivalent IQ levels as people who can see, which indicates that it is possible to achieve intelligence similar to that of humans without relying on images. The thoughts, or simulations, generated by blind people are likely to be based on words and general audio when facing complex situations and reasoning on complex tasks. Words are computationally less expensive than images, and the case of blind people suggests that AI can be built relying on relations between words.

C. Brain disconnect and people mumbling

The intelligence theory presented here is consistent with an evolutionary process to develop intelligence, as previously described. When a person thinks, the brain appears to switch mode from continuously receiving input and acting outputs, to generating simulations involving ways in which the inputs will develop for certain possible actions taken. When activating this switch, the brain in thinking mode in a way disconnects the inputs and outputs to stop receiving information or acting out strategies.

The disconnect when thinking is likely to be imperfect. As a result, we sometimes see people who are thinking deeply that mumble or whisper what they are thinking about without realising it, sometimes acting as if they were talking alone. They also sometimes perform actions that correspond to their thoughts. This evidence is consistent with the theory presented here.

D. Complete aphantasia

There are cases of people who report thinking without generating any mental simulations of any kind. In this theory, we argue that those cases are either due to the fact that i) they lack the metacognition to realise their thoughts, or ii) there is no thought process in the relation between an input and an output, meaning that no simulations are generated. In the latter case, the definition of intelligence presented here implies that there is no thought and thus no intelligence involved, since it is purely a complex process of a mechanism relating an input to an output. In general, we speculate that people with complete aphantasia are more likely to lack metacognition to realise their thoughts rather than lack intelligence as defined here.

E. Science

Science at its core involves finding relations between magnitudes that we can observe and describe. This can be understood according to the theory presented in this paper as follows. In science, we identify a relation, and then we try to see if it always applies in the reality that we can observe. We do that by entering every potential input that matches the relation that we identified and checking if indeed the output prescribed by the hypothesised relation occurs. If it is confirmed for all conceivable inputs then we consider the relation established and the theory confirmed.

F. Lack of intelligence

The study of the lack of intelligence is useful to understand intelligence as its antithesis. We consider a person unintelligent when their actions and decisions do not achieve the goals.

From the perspective of this theory, an unintelligent person is someone who either:

i) Does not run sufficient simulations given a situation

ii) Runs simulations that do not match reality because the relations it formed in its brain either do not match the reality or overlook elements of it

iii) Does not complete the simulations in sufficient detail and breadth to make a good decision about a situation.

Thus, the person then selects a decision based on a either a small number of simulations (thoughts) or some simulations

that do not match reality, which leads to actions that do not achieve the most desirable goals. When observing a stupid person act, we can say that they are short-sighted or foolish because their actions lead to outcomes that are far from their goals. Instead, in that situation, us as observers will have run many more simulations more extensively, and we can see that other simulations lead to outcomes closer to the goals than the course of action selected by the unintelligent person.

Lack of knowledge is generally not perceived as stupidity unless inadvertently. If we know that a person is not aware of some information necessary to create simulations that relate to a given exercise, we will see that the person does not select the best simulation and thus course of action because they cannot generate the correct simulation. But we will not call the person stupid because we are aware that even if they run all infinitely possible simulations, they would not be able to generate the correct simulation because they lack information. However, if we are not aware that the person is lacking some information, we may accidentally call the person stupid because we assume that they have the information needed to generate relevant simulations, but they have not generated the necessary simulations to select the best one and thus decide the best course of action.

It is then common for us to check with a person whether they had the information needed and whether they considered the simulation we find most suitable. If they have, and instead chose a different simulation and course of action, we typically define them as unintelligent.

G. Optimal Intelligence

A maximum level of intelligence exists for a given set of resources, e.g. a given human brain and body. The maximum level of intelligence consists on simulating all scenarios that are relevant to a given situation and then selecting the best one to act. The key is to consider the relevant scenarios in each case.

An AI can usually outperform a human one by simulating a larger amount of scenarios in parallel. Even if this AI does not always consider the most relevant scenarios first, it will include them given the large number of simulations it can perform in parallel.

H. Personalities

The rewards and connections can be tuned to generate different types of persons that we observe in humanity. For example, a person with short attention span may have a fundamental process in the brain that gives low rewards to any long thought, so the person will quickly select a simulation without exploring all relevant options.

I. Intelligence definition boundaries

The definition of intelligence presented here specifies a key feature of intelligence as the ability to generate simulations in the mind, given an input situation. It therefore distinguishes intelligent agents as separate from those that simply possess a chain of mechanisms that relate an input to an output action. This boundary in the definition of intelligence also means that intelligent agents are those that have knowledge of a situation, intent, and an illusion of consciousness, all of which stem from the ability to generate simulations given an input scenario.

It should be noted, however, that even in intelligent agents, given an input situation, the generation of simulations and evaluation of them to select an output is in fact a long chain of mechanisms that relate an input to an output. As such, intelligence can be argued to resemble a complex form of the mechanisms found in non-intelligent organisms that relate inputs to outputs, and the boundary between intelligence and the absence of it can be somewhat arbitrary. The theory presented in in this paper places the boundary of intelligence at the ability to generate simulations and evaluate them, but it should be noted that this boundary is indeed somewhat arbitrary. This boundary is selected in relation to both the nature of the tasks that we can consider to require intelligence, and the implications over knowledge, intent, and consciousness, both of which relate to the ability to generate simulations according to this theory.

VI. Artificial General Intelligence

The theory presented here provides a blueprint to replicate intelligence artificially. AGI involves, for a given scenario with a set of inputs, generating a set of relevant simulations based on training information, evaluating the possible outcomes, and then selecting the best one to act. This requires having identified relations and patterns in the training information that match reality, so that the simulations agree with reality. The firmware in AGI needs to specify the reward function and a set of low level routines, as previously described in the firmware section.

A. Comparison between LLMs and natural intelligence

There are various differences between the pseudointelligence achieved by large language models (LLMs) and the theory of intelligence presented in this work.

Currently the most advanced LLM rely on a transformer architecture. In this, the machine takes a sequence of words as input, and outputs another sequence of words that maximise a specified reward function. The machine does not understand the words, based on the definition of understanding presented in this paper. The machine uses an embedding to encode the words, or rather parts of words, processes them, then produces an encoded output, and finally reverses the embedding to generate the final outputted words. Thus, it learns the relation between inputs in the form of sequences of numbers and corresponding outputs, also in the form of sequences of numbers.

This LLM approach to intelligence in a human would be the equivalent of writing answers in a completely foreign language without understanding the relation between words and the meaning they represent in the world, based on a large set of examples of sequences of words in and out. When given a new input prompt, the person would write a collection of words that are similar to those in the examples

in the training data set. This can lead to satisfactory results in some cases, but it involves no understanding of the world according to the definition of understanding proposed here and no intent beyond writing sequences of words to satisfy a reward function. This can be referred to as artificial clueless intelligence (ACI).

This form of ACI, or pseudointelligence, from LLMs can achieve impressive and useful results, but it differs from the theory of human intelligence as presented in this work. The key difference is that, given an input, human intelligence generates simulations of the possible ways in which the situation can play out (or possible abstract approaches in the case of complex situations), depending on possible actions, which in this case are the outputs created by the agent. This capability of generating simulations of possible outcomes to then select the most appropriate action is currently not present in LLMs.

As such, intent is a key difference between LLMs and human intelligence as presented here. LLMs do not have an intent since they do not generate simulations of the way in which things can play out depending on their actions. Instead, they produce an output that is a sequence of words to maximise a reward function. Human intelligence, on the other hand, involves intent according to our theory. Humans will generate simulations of possible future scenarios, and select an action that leads to a desired outcome.

This also applies to conversations. Humans participate in conversations with an intent, saying collections of words that are intended to lead to an outcome that benefits them. Conversational AIs, instead, typically participate in conversations without an idea of how things may play out or the final outcome. They generate words to satisfy a reward function with no longer term plan.

B. High level AGI architecture

The theory in this paper can serve as a preliminary guideline that may be used to help in the development of AGI. Building such AGI, however, requires an architecture capable of generating simulations akin to our thoughts. This architecture is not directly available in the technology available today. The present challenge is therefore to marry the theory in this work to the technologies available, such as transformers and more general LLMs, to create AGI. In this subsection, we offer a potential high level outline for doing that relying on a multimodal transformer architecture that combines words and images. Future work will explore the details.

The first ingredient for intelligence, a model of reality, can be considered to be available in the form of current LLMs. The second ingredient, the ability to relate the current situation to the training data, can also be obtained from transformers. The situation used as input is akin to the prompts together with the context used as input for transformers.

The third ingredient to build AGI is to generate simulations based on the input that are relevant to it. For this, we propose the development of a transformer that generates multiple such simulations in parallel that are stored in a cache. For simple problems, these represent potential ways to act together with their outcome. These can be directly evaluated, which is the last ingredient to build AGI, to select the most suitable one. This represents the final step for the AGI.

For complex problems, these simulations generated by the transformer typically serve as intermediate simulations that add context and also narrow down the attention to the matter at hand. Multiple iterations can then be performed until eventually a simulation is produced that collapses to a solution that is outputted to be acted out. As in the previous simpler problems case, the final step is to evaluate the simulations against the reward function, which can be performed using ML classifiers, potentially built on a transformer architecture, given the sequential and multimodal nature of the data in the simulations.

As an example of simple problem solving by this AGI architecture, we can consider a game of chess. The AGI is a player, and in this case it would generate verbal simulations of the type: ”if I move the bishop to this position, then the opponent can move their queen to that position, which would leave my rook unprotected; otherwise if I move the bishop to this other position, then I check their king which forces the opponent to move their queen to protect it, and thus my rook is safe”. The AGI would accompany these verbal simulations with visual simulations of diagrams of the chess board that match the movements. In this case, the AGI directly simulates the possible outcomes based on potential moves to act, typically using a horizon of a few steps, and then evaluates them against the reward function. The AGI finally selects the move that, for the horizon considered, leads to the highest reward. It should be noted that the choice of the horizon length to consider depends on the time pressure to make a decision. As previously discussed in section II-G, the horizon choice can be part of a firmware that can be fine tuned from training data using machine learning architectures.

As an example of more complex problem solving by this AGI using intermediate simulations, we can consider part of a business case that involves finding the expected revenue of a business for one of various potential scenarios to select the highest one. The AGI needs to initially generate a set of intermediate simulations, here separated by semicolons. These can be: ”we are comparing various scenarios and we now need to compute the revenue for one of them; our goal is to find the revenue after 5 years; the data provided for this scenario is giving us the current revenue of $10M$ and the projected annual growth of 8 per cent; we can compute the new values for each subsequent year; the equation to calculate revenue after n years is $R_5 = R_0 G^n$ where R_i is revenue at year i and G is the annual growth rate”. At this point, the AGI needs to substitute values into the equation. Given the intermediate simulations that provide context, it can use them to narrow down the attention and perform the calculations by assigning high attention weights to the specific data values. To perform maths, it can generate subsequent simulations of the type: ”the annual growth rate of 8 per cent needs to be expressed in the form to be used in the equation; this is 1.08;

the total growth after $n = 5$ years is 1.08^5; we need to multiply it times the current revenue of $10M$". Now it can compute the mathematical operations because, from all the intermediate simulations, it has narrowed down the attention to the point that the next simulation is an immediate operation that can be performed directly from training data (if simple) or using a simple application programming interface (API) to a tool such as a calculator. To do so, it can simulate: "the total growth rate is $1.08^5 = 1.47$; to multiply the total growth rate times the current revenue of $10M$ we can use the fact that the revenue is 10 in a decimal system and therefore we just need to move the point one digit to the right; the resulting predicted revenue is $14.7M$". It is interesting that in the first operation in this last example, the AGI relies on a calculator API, whereas in the second operation, the AGI can use a different strategy that does not require an API nor having seen the specific product $10 * 1.47$. Instead, it can use a strategy of intermediate simulations to compute the product. Humans use a similar process. Lastly, from all these intermediate simulations, together with those for the costs, the final solution collapses to the result "the predicted revenue in this scenario is $14.7M$". The same applies to the other scenarios.

More generally, given a situation, each of the intermediate simulations is generated based on the training data, the prompt, and the other intermediate simulations. These are intended to either advance towards the solution or to add context that helps narrow down the attention to the matter at hand.

The architecture using intermediate simulations that then collapse to a final outcome that is acted out is similar to that in chain of thought [20]. The intermediate simulations, however, are more flexible and they do not necessarily contribute to solving the problem. They are generated thoughts that stay in the cache and can help the model reach the outcome but not necessarily do so.

Two scores could be used to evaluate the simulations in this potential implementation of AGI with intermediate steps. The first is the reward function, which measures the usefulness of the simulation relative to the final outcome. The second is the usefulness of the simulation as context towards the final outcome. Each simulation can be evaluated against these two scores. Simulations that score low on both are discarded. Simulations that score high on context but low on final outcome are kept as context, with a given weight that can be understood as the probability that it will be relevant. This weight gradually decreases, fading away with time as new simulations are added. Thus, the cache becomes akin to a bag of many simulations, with each of them typically consisting of either one sentence, a few key words, an important data value extracted from the prompts, an image, or a video, and each of those having different weights, which can be interpreted as the strength of each thought. These simulations in the bag provide context and focus the attention on the key elements of the problem, gradually advancing it towards the solution using heuristics. Lastly, simulations that score high on outcome are those that either collapse to become the selected output to be acted out or directly lead to the final simulation that is the output.

The intermediate simulations and simulations of potential outcomes based on possible actions by the AGI represent thoughts and define intent.

C. Hallucinations

A current question for AGI is how to mitigate hallucinations. Hallucinations are common, especially in mathematical questions when using current LLMs. The reason for hallucinations is that the LLM relies excessively on training data to produce the sequence of words in the output. One strategy to mitigate them is to give significant weight to the mathematical data in the prompt so that the output must contain or be produced based on the specific data in the prompt. However, this can be difficult to perform in general. Instead, by generating multiple intermediate simulations given a prompt, these intermediate simulations can break down the problem and extract the specific mathematical data to which high attention may be given, to then assemble into the final output.

Humans also suffer from hallucinations, where they misread or misremember data. When creating an AGI based on human intelligence, one can aim to mitigate hallucinations, but it may not be necessary to fully eliminate them.

D. Future Developments

The intelligence theory presented in this work can provide guidelines for a foundation to create an AGI that mimics human intelligence, as briefly outlined in the previous subsections. Such AGI would have intent and would be capable of conducting research and solve complex problems. The development of such AGI according to the theory presented here can borrow from LLMs, using them to generate the simulations needed for each situation. LLMs can thus serve as a fundamental element of AGI on which the ingredients presented in this work can be added.

Another possibility for the future is that AI develops in a way that differs from human intelligence as described in this work. Such AI could have a different architecture from human intelligence but achieve similar results. For example, LLMs may become significantly larger, may add capabilities such as checking its outputs against factual data to minimise hallucinations, and may be then capable of producing credible results that are similar to those of human intelligence. This technology may become very valuable, but according to this work it will differ from human intelligence since it will lack the simulations generated by humans and thus the intent from human intelligence. Such AI may also lack the ability to conduct research or solve new complex problems, and may become akin to a machine that reshuffles existing information and outputs it.

E. Proofs

The theory presented in this work is consistent with observations of reality, but it is unproven. There are two main avenues to attempt to prove the theory, both of them challenging.

The first one involves a large amount of experimental observation of the human brain, including e.g. advanced functional MRI, electrodes, dissections, and other physical measurements, together with extensive interviews and observations of human behaviour to prove the theory.

The second one is to create an AGI based on this theory and experimentally compare it with humans. This second avenue is the most promising, given that the creation of an AGI would have a value in the order of at least a few percent of the global gross domestic product, and therefore proving the theory would be well aligned with the significant financial incentives needed to fund the research.

VII. Concluding remarks

We presented a preliminary theory of the nature of intelligence and its working mechanisms. Our theory considers that intelligence fundamentally involves, for each given situation, generating a set of simulations about ways in which the situation may play out, or in more complex problems about potential approaches (which can be viewed as abstract simulations about the way a situation may play out), and evaluating them against a reward function to select the most suitable one to act out. These simulations need to be relevant and match reality. We conjectured that this is achieved by identifying relations in the information perceived through sensors, which are then used to generate simulations that are relevant to each situation. We also showed our theory to be consistent with various observations of human behavior and reality. We outlined how knowledge might develop through a heuristic process. We described human understanding and intent through the lens of our theory, and showed these to be intimately related to our definition of intelligence. Furthermore, we outlined how intelligence might develop through an evolutionary process that could be consistent with our theory.

Our work also provides a potential explanation for what human consciousness is. Given our theory, humans with intelligence can be viewed as machines that identify relations in the information they perceive and that they use to create useful simulations. Human consciousness is the illusion that emerges as each person generates simulations, which commonly involve audio and video, as they think. We conjecture that the fundamental processes that govern which scenarios are simulated is not controlled by any free will, and instead is determined by the genetics and inputs that each person received throughout life, which determine the thoughts of a person at each given time. Thus, we conclude that free will is an illusion.

References

[1] P. J. Werbos, "Intelligence in the brain: A theory of how it works and how to build it," *Neural Networks*, vol. 22, no. 3, pp. 200–212, 2009.

[2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach 3rd edition*. Saddle River, NJ: Prentice Hall, 2009.

[3] M. Hutter, *Universal Artificial Intelligence*. Berlin: Springer., 2005.

[4] V. Lifschitz, *Formalizing Common Sense: Papers by John McCarthy*. Norwood, New Jersey: Ablex Publishing Corporation, 2006.

[5] J. McCarthy and P. J. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," in *Readings in artificial intelligence*. Elsevier, 1981, pp. 431–450.

[6] M. Hutter, *Commonsense Reasoning*. San Francisco, CA: Morgan Kaufmann, 2006.

[7] A. Rodriguez, J. Whitson, and R. Granger, "Derivation and analysis of basic computational operations of thalamocortical circuits," *Journal of cognitive neuroscience*, vol. 16, no. 5, pp. 856–877, 2004.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] H. Gardner, *Multiple intelligences: The theory in practice.* Basic books, 1993.

[10] D. Goleman, *Emotional intelligence*. Bloomsbury Publishing, 2020.

[11] R. J. Sternberg *et al.*, *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive, 1985.

[12] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artificial Intelligence*, vol. 299, p. 103535, 2021.

[13] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[14] R. Keogh and J. Pearson, "The blind mind: No sensory visual imagery in aphantasia," *Cortex*, vol. 105, pp. 53–60, 2018.

[15] P. Lennon, "Aphantasia and conscious thought," *Oxford Studies in Philosophy of Mind Volume 3*, p. 131, 2023.

[16] W. A. Bainbridge, Z. Pounder, A. F. Eardley, and C. I. Baker, "Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory," *Cortex*, vol. 135, pp. 159–172, 2021.

[17] K. Frankish, "Illusionism as a theory of consciousness," *Journal of Consciousness Studies*, vol. 23, no. 11-12, pp. 11–39, 2016.

[18] J. Prinz, "The sensory basis of cognitive phenomenology," *Cognitive phenomenology*, vol. 174, 2011.

[19] J. R. Atkinson, K. Gleeson, J. Cromwell, and S. O'Rourke, "Exploring the perceptual characteristics of voice-hallucinations in deaf people," *Cognitive neuropsychiatry*, vol. 12, no. 4, pp. 339–361, 2007.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.