Artificial Intelligence - Returning something motivates

Sandip Banerjee

Office of The Chief Electoral Officer, Department of Election, Arunachal Pradesh, Itanagar

ABSTRACT

A key component of this article is the hypothesis that intelligence, and its associated abilities, can be understood to serve the purpose of maximizing reward in general. It should therefore be understood that reward is sufficient to drive behavior that demonstrates all the abilities studied in natural and artificial intelligence, such as knowledge, learning, perception, social intelligence, language, generalizations and imitations. The view outlined above differs from the view that specialized problem formulations are necessary for each ability, based on signals or objectives that are different from each other. Further, we propose that reinforcement learning agents could learn behavior that combines most of these abilities in the form of behavior exhibited by reinforcement learning agents, and therefore could serve as a solution to artificial general intelligence problems through trial-and-error experience that maximizes reward.

Keywords : Social Intelligence, Reinforcement learning agents, Trial-and-error experience, Maximizes reward

1. A BRIEF INTRODUCTION

A wide variety of intelligence expressions can be seen in animal and human behavior, and there is a wide range of associated abilities that can be used to name and study these phenomena, such as social intelligence, language, perception, knowledge representation, planning, imagination, memory, and motor control, as well as many others. As far as I know, what could motivate agents to behave intelligently in such a diverse range of circumstances (natural or artificial)?

As an alternative to positing a specific explanation, each ability may emerge as a result of the pursuit of a particular goal, one which is specifically designed to elicit that particular ability. There have been many discussions regarding whether social intelligence is a function of the Nash equilibrium of a multi-agent system; whether language is a function of parsing, part-of-speech tagging, lexical analysis, and sentiment analysis; and whether perception is a function of segmenting and recognizing objects. A methodological alternative hypothesis will be presented in this paper: that is, that the generic objective of maximizing reward is sufficient to motivate behavior that exhibits the vast majority, if not all, of cognitive abilities that have been studied in the field of artificial intelligence and natural intelligence. Despite the fact that intelligence is associated with a wide variety of abilities that seems to be at odds with any generic objective, this hypothesis may startle people because it seems that intelligence is a unique combination of abilities. Animals and humans, as well as artificial agents in the future, encounter naturally complex environments in which to succeed (for example, to survive) within these environments due to the fact that they have to possess sophisticated abilities in order to succeed (for example, to survive) in these environments. The achievement of success, which is based on maximizing rewards, therefore demands a broad spectrum of intelligence capabilities. Any behavior that seeks to maximize rewards must necessarily possess these capabilities in such environments. Consequently, it can be said that the generic objective of reward maximization incorporates, in a way, many of the intelligence-oriented objectives or even all of them.

In addition to providing two levels of explanation for the abundance of intelligence found in nature, reward can also be viewed as a means of extending the explanation beyond itself. Firstly, different forms of intelligence may arise as a result of the maximization of different rewards in various environments, resulting, for instance, in bats with echolocation abilities, whales with whale-song communication, or chimpanzees with the ability to use tools. Furthermore, in the future, artificial agents will face the need to maximize a variety of reward signals in their environments in order to produce new forms of intelligence that will include abilities such as laser-based navigation, communication via email, or robotic manipulations that are entirely new.

It is also important to note that intelligence of an animal or human is not confined to one specific ability, but it is characterized by a variety of capabilities as well. Our hypothesis posits that all of these abilities have the aim of maximizing the reward for the animal or agent within the environment. To put it simply, the pursuit of a single goal may generate behaviors that are characterized by a number of intelligence-related abilities. As a matter of fact, such reward-maximizing behaviors may frequently be found in conjunction with specific behaviors derived from the pursuit of different goals associated with each capability, while simultaneously maximizing reward.

Taking a squirrel as an example, the squirrel brain can be understood as a decision-making system that receives sensory information from the squirrel's environment and sends motor commands to the squirrel. It is possible to interpret squirrel behavior in terms of maximizing cumulative rewards such as satiation (i.e. negative hunger). A squirrel's brain probably possesses the ability to perceive (to identify good nuts), to acquire knowledge (to comprehend nuts), to control motor activity (to collect nuts), to plan (to decide where to cache nuts), to remember the location of stored nuts (to remember where they are cached), and to be socially intelligent (in order to bluff about the locations of stored nuts so they don't get stolen). There are, therefore, various abilities associated with intelligence which can be understood as serving a singular purpose of minimizing hunger as defined by Figure 1 (see below).

The kitchen robot could also be implemented as a decision-making system that receives sensations from the robot's body and then sends actuator commands to it in response to those sensations. A kitchen robot is essentially designed in order to maximize a reward signal that measures cleanliness so that it can maximize a reward signal in order to maximize the level of satisfaction. To optimize cleanliness, a kitchen robot must presumably possess the following abilities: perception (the ability to distinguish between clean utensils from dirty ones), knowledge (the ability to understand utensils), motor control (the ability to manipulate utensils), memory (the ability to recall where utensils are located), language (the ability to predict a future mess based on dialogue), and social intelligence (as a means of encouraging young children to make fewer mess when they are playing). To maximize cleanliness, a behavior needs to possess all these abilities at once in order to achieve that singular objective (see Fig. 1).

This may provide a deeper understanding if it is noted that intelligence-related abilities arise to solve a single problem of reward maximization, as it explains why such an ability arises, for example, the classification of crocodiles is essential for the survival of the animal. It is rather interesting to note, however, when each ability is understood as a result of solving a specific problem, the why question is sidestepped as it is focused upon what the ability actually does (e.g. separating crocodiles from logs). It has also been shown that a singular goal can help us gain a much deeper understanding of each ability, including characteristics that we cannot otherwise formalize, such as dealing with irrational agents in social intelligence (such as pacifying an angry aggressor), understanding haptics in perception (for example, picking an object from your pocket in a dialogue regarding the best way to peel fruit), or understanding language to perceptual experience. Last but not least, implementing abilities as a means of facilitating a singular goal, rather than as a tool for achieving their individual specialised goals, also answers the question of how to integrate abilities, which for the time being remains an open question.

It is essential that one considers ways to resolve the intelligence problem once it has been established that reward maximisation can be used as an objective to understand the problem. Therefore, one could expect to find such methods in natural intelligence, or to use them as an artificial intelligence method. A general and scalable method

for maximising reward is to interact with the environment by trial and error as a means of learning. This is the most general and scalable method for maximizing reward. The authors hypothesize that in the presence of a rich environment, an agent that has learned to maximize rewards in this manner to be capable of developing sophisticated expressions of general intelligence as a result of placing itself in such an environment.

It is of great interest to note that the game of Go serves as a recent salutary example of both the problem and solution for maximizing rewards. The initial focus of re-search was on distinct skills, including openings, shapes, tactics, and endgames, and each skill was formalized by using unique objectives such as sequence memory, pattern recognition, local search, and combinatorial game theory. A key objective of AlphaZero was to achieve a single goal - maximizing a signal that has a value of 0, until the final step and then has a value of 1 or 0 depending on whether you win or lose. By studying each ability in greater depth, we were able to discover new opening sequences [65], employ surprising shapes within a global context [40], understand global interactions between local battles [64], and make sure that we played safely when we were ahead [40]. Also, the development of this process contributed to the development of a range of new abilities that had not previously been adequately formalized in the past - for example, the ability to balance influence and territory, thickness and lightness, attack and defence. While prior research had proven highly problematic in terms of integrating Alp-haZero's abilities into a unified whole [32], their ability was intrinsically incorporated into a unified whole. The maximisation of wins, in a simple environment like Go, proved sufficient to cause behaviour exhibiting a variety of specialised abilities to emerge. It has been found that applying the same method to different environments, for example, chess or shogi [48], has resulted in new capabilities such as piece mobility and color complexes [44], in addition to the ones mentioned above. According to us, we believe if we maximize rewards in richer environments – environments that are similar in complexity to the natural world we face as humans and animals – then we may ultimately gain even more intelligence abilities and perhaps even all of them.

This paper has been structured in such a way that it should be easier to understand, so the following structure has been adopted: We formalize the objective of reward maximization as the reinforcement learning problem in Section 2, while in Section 3 we present our main hypothesis. The following article discusses several important abilities associated with intelligence, and how reward maximization may lead to these abilities. As we turn to section four, we will explore the possibility of using reward maximization as a solution strategy. There is also a discussion of related work in Section 5 and finally, we discuss possible weaknesses of the hypothesis and consider several alternative hypotheses in Section 6.

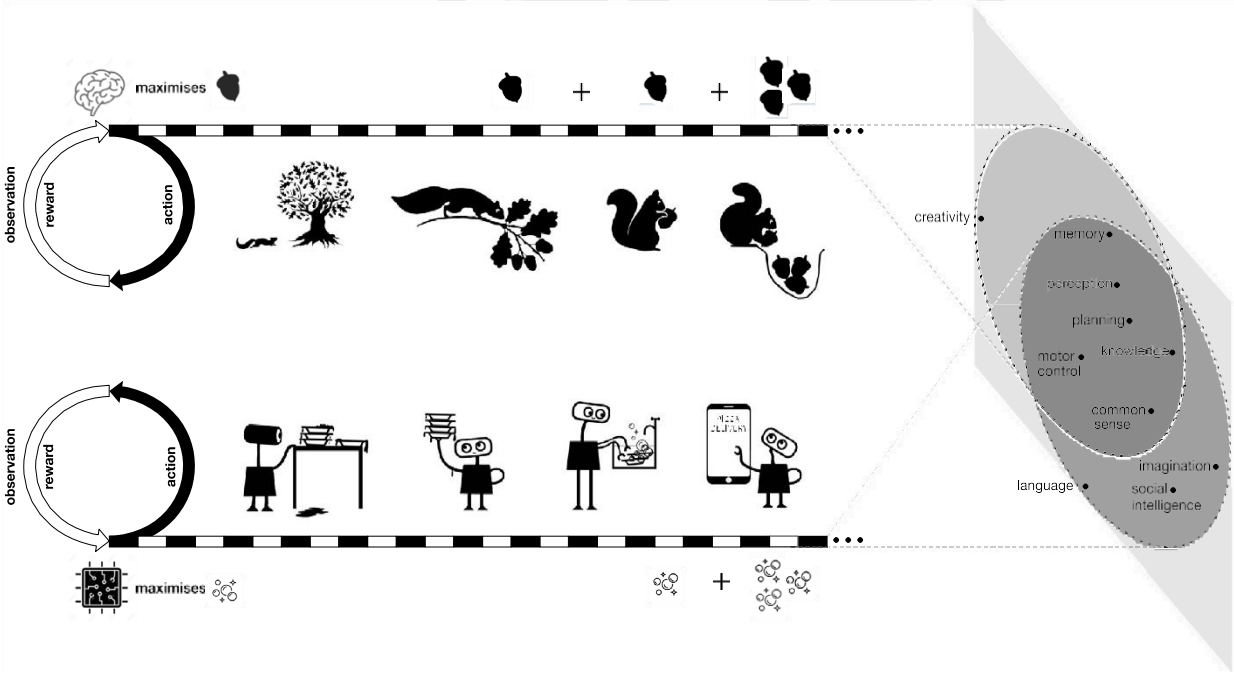

Fig. 1. The hypothesis postulates that intelligence, and its associated abilities, can be understood as subserving the maximization of reward by an agent acting in its environment. For example, a squirrel acts so as to maximize its consumption of food (top, reward depicted by acorn symbol), or a kitchen robot acts to maximize cleanliness (bottom, reward depicted by bubble symbol). To achieve these goals, complex behaviors are required that exhibit a wide variety of abilities associated with intelligence (depicted on the right as a projection from an agent's stream of experience onto a set of abilities expressed within that experience).

2. AN IN-DEPTH ANALYSIS OF REINFORCEMENT LEARNING PROBLEMS

In my opinion, intelligence can be considered to be the ability to achieve goals in a flexible manner. According to John McCarthy [29], intelligence is the computational component of the capability to achieve a goal in the real world. The problem of goal-seeking intelligence can be expressed in the form of reinforcement learning [56] which is an excellent way to formalize this problem. There are various kinds of goals and worlds available to address this general problem.

It follows from this that there are many forms of intelligence corresponding to the various reward signals of different situations, and thus differing forms of intelligence as well.

2.1 A DESCRIPTION OF THE AGENT AND ITS ENVIRONMENT

As a method of artificial intelligence that has a long history of interaction [42], reinforcement learning follows the same protocol: it is a protocol that decouples a problem into two systems which interact sequentially over time: an agent (the solution) that takes decisions and an environment (the problem) that is impacted by those decisions. As opposed to other specialized protocols where multiple agents, multiple environments, or another mode of interaction may be considered, this protocol focuses on a single agent and the environment.

2.2 AGENT

The agent is a system that receives at a time an observation Ot and produces an action at that time. Specifically, the agent is a system based on the experience history of the agent Ht O1, A1,..., Ot 1, At 1, Ot, i.e., the sequence of observations and actions that have occurred in the history of interactions between the agent and the environment, or a system that selects the appropriate action At at time t.

Due to practical constraints, agent systems are restricted to a substantially limited set of options [43]. This means the agent has a limited capacity.

Whenever agents and environment systems are created in real time, they are limited by what machinery is available to them (for instance, a computer has a limited amount of memory or a brain has a limited number of neurons). There is a time when the environment system is still processing (for example, the lion attacks) while the agent is still computing its next action while the environment system is still processing (for example, the no-op actions are produced as the agent decides whether to run away from a lion). As a result, reinforcement learning is a real-world problem, which is studied in various fields such as natural intelligence and artificial intelligence, as opposed to being a theoretical abstraction that ignores the limits of computational power.

Rather than delving into the nature of the agent, this paper concentrates on the problem it must solve, and the intelligence that can be induced by any solution that may be offered to the problem.

Table 1
The definition of environment is broad and encompasses many problem dimensions.

Dimension	Alternative A	Alternative B	Notes
Observations	Discrete	Continuous	
Actions	Discrete	Continuous	
Time	Discrete	Continuous	Time-step may be infinitesimal
Dynamics	Deterministic	Stochastic	
Observability	Full	Partial	
Agency	Single agent	Multi-agent	Other agents are part of environment, from perspective of single agent (see Section 3.3)
Uncertainty	Certain	Uncertain	Uncertainty may be represented by stochastic initial states or transitions
Termination	Continuing	Episodic	Environment may terminate and reset to an initial state
Stationarity	Stationary	Non-stationary	Environment depends upon history and hence also upon time
Synchronicity	Asynchronous	Synchronous	Observation may remain unchanged until action is executed
Reality	Simulated	Real-world	May include humans that interact with agent

2.3. ENVIRONMENT

Essentially, an environment is a system that receives an action at time t and responds to that action at time t with an observation at time Ot+1. In other words, the environment is a system that determines the next observation Ot+1 that an agent will receive from the environment based on the experience history Ht , the most recent action At , and possibly a source of randomness that occurs at the next time step ηt .

A defined environment specifies the interface to the agent as part of its definition. The agent is a single entity that makes decisions; anything outside of the agent, including its body, if it has one, is considered part of the environment, not the agent. An agent's observations and actions are defined by both sensors and actuators, which are respectively based on the observations that the agent is making and the actions which the agent is able to take.

I would like to emphasize that this definition of the environment is very broad and encompasses a wide range of problem dimensions, such as those listed in the table below.

2.4. THE REWARDS THAT ARE OFFERED

A reinforcement learning problem defines goals by the accumulation of rewards from the environment, where a reward is a special observation Rt which is emitted by a reward signal at every time step t. Reinforcement feedback measures progress towards a goal instantaneously. As the name indicates, reinforcement learning occurs when there is a reward signal in the environment and a cumulative objective to be achieved, such as a sum of rewards over a set number of steps, a discounted sum or an average reward over a set period.

In terms of rewards, there are many different goals that can be represented. An arbitrary scalar reward signal might represent a weighted combination of objectives over time, or a different trade-off. A person may reinforce desired behavior by explicit reinforcement, or by providing online feedback, or by providing delayed feedback via questionnaires or surveys, or by using natural language. It can become easier to formulate seemingly vague objectives like "I will know it when I see it" if a human is involved in the planning process.

Rewards could provide intermediate feedback, potentially at every step, on the way to the goal at the same time as providing general feedback. As part of the problem definition, an intermediate signal is essential, since learning cannot occur without feedback.

3. HAVING A REWARD THAT IS WORTH IT IS ENOUGH

Prior sections have demonstrated that rewards are sufficient to portray the varied goals that can be achieved by an individual's intelligence.

Having combined all of these ingredients into one single concept, we can now present the main point that we are attempting to express: the goal of maximizing reward is a fundamental component of many different types of intelligence. In addition, there are a great many abilities associated with every form of intelligence that result implicitly from the pursuit of these rewards. We propose a way to understand all intelligence as well as its allied abilities as best as we can in the following manner in the extent that it is feasible to do so:

Considering the idea that reward is enough is a useful hypothesis (the reward-is-enough hypothesis). According to this hypothesis, intelligence and its associated abilities are highly capable of being viewed as a subordinate to the maximisation of a particular agent's reward within its environment within a defined timeframe.

In order to prove the reward-maximizing hypothesis true, it has to be noted that, if it turned out to be true, it would mean that a good reward-maximizing agent would have the ability implicitly associated with intelligence in order to achieve its goals. In this case, it would be appropriate to have an agent that would maximize rewards as a good agent for this purpose.

Some evidence suggests that rewards may be seen as representing all goals in some way [56], but it is important not to confuse the guarantee of accomplishment with the ability that arises implicitly from the pursuit of just one of these goals, which shows that rewards may represent all goals at some level.

As an example, one of the possible methods by which this could be achieved would be by using algorithms that are as yet unknown. However, these algorithms might be implemented effectively in a system to maximize cumulative reward, perhaps by utilizing algorithms that are yet to be discovered.

As a general principle, artificial intelligence states that any behavior can be encoded by maximizing a signal designed specifically to induce that behavior [12] that is designed to promote that behavior (for example, ensuring that an object is correctly identified, or that syntactic rules are correctly used). If we take into account the wide range of pragmatic objectives that natural or artificial intelligence can be aimed at, one could argue that intelligence and its associated abilities will serve implicitly to maximize one of the many possible reward signals that will be available, in accordance with the many pragmatic objectives pursued by these organisms.

A complex environment can drive an individual to acquire sophisticated abilities over time as a result of maximizing simple rewards. In order for squirrels to survive in their natural environment, it is essential that they are able to manipulate nuts in a skillful manner in order to minimize hunger. A squirrel's ability to perform this function is caused primarily by a complex interaction between squirrel musculoskeletal dynamics (among other factors) and squirrel musculoskeletal dynamics; objects that the squirrel or this may be due to a variety of factors including leaves, branches, or soil, or because a nut is situated, connected, or hindered by something else, among them variations in the size and shape of the nut, environmental factors such as wind, rain, or snow, as well as changes caused by aging, diseases, or injuries.

In a similar way, it is necessary to enable the detection of different states of kitchen utensils. These are able to cope with a wide range of conditions such as clutter, obstructions, glare, encrustations, damage and a spectrum of thousands of different combinations anywhere in between.

Furthermore, the maximization of many other reward signals by a squirrel (such as maximizing survival time, reducing pain, or improving reproductive success) or kitchen robot (such as maximizing a healthy eating index, maximizing positive feedback from the user, or maximizing their gastronomic endorphins), and in a variety of other environments (such as various habitats, other dexterous bodies, or varying climates), would also yield abilities of perception, locomotion, manipulation, and so forth. Thus, the path to general intelligence may in fact be quite robust to the choice of reward signal. Indeed the ability to generate intelligence may often be orthogonal to the goal that it is given, in the sense that given. This is because different reward signals in many different environments may produce similar abilities associated with intelligence.

In the following sections we explore whether and how this hypothesis could be applied in practice to various important abilities, including several that are seemingly hard to formalise as reinforcement learning problems. We do not provide an exhaustive discussion of all abilities associated with intelligence, but encourage the reader to consider abilities such as memory, imagination, common sense, attention, reasoning, creativity, or emotions, and how those abilities may subserve a generic objective of reward maximisation.

3.1. REWARD YOU COLLECT FOR LEARNING AND GETTING KNOWLEDGE IS SUFFICIENT

An agent is defined as possessing knowledge within its own internal structure, which may include, for instance, information contained within the parameters of its functions related to selecting actions, anticipating cumulative rewards, or predicting features of future observations that may be an indication of knowledge within the agent. There are some types of knowledge (prior knowledge) that can be innate, and there are others that can be acquired through learning.

In some environments, innate knowledge may be required by the environment. It is important to note that it may be necessary if in order to maximize total reward, it is necessary to have knowledge that can be readily accessed in novel situations in order to maximize total reward. As an example, young gazelles may have to run away from lions when they are attacked by them at birth. The innate knowledge of predator evasion is likely to be necessary in this case before there will be any opportunity to learn this skill in the future. I would like to stress, however, that the extent of prior knowledge is both theoretically limited (by the amount of capacity within the agent) and practically limited (by how difficult it is to construct useful prior knowledge). There is also the fact that environmental demand for innate knowledge cannot be operationalised, unlike other abilities that we will examine in the following sections; innate knowledge refers to knowledge that comes before experience, and therefore can not be acquired through experience.

A person may not only be able to learn on their own, but an environment may also demand that they be able to learn on their own. The probability of such an event will be increased when there is uncertainty regarding future experiences as a result of known or unknown elements, stochasticity, or the complexity of the environment in which we live. There may be a big number of possible knowledge requirements associated with the agent's perception of the future, depending on how it perceives the future, and it may be necessary for the agent to be able to access a vast array of possible knowledge. Finally, there are many challenging and important problems in the area of natural and artificial intelligence that can be addressed when the total space of potential knowledge is much wider than the capacity of the agents in rich environments where many challenging and important problems can be solved. Imagine for a moment, for instance, what the environment was like for early humans living in their day to day lives. The knowledge requirements of an agent to be able to maximize their total reward will probably vary according to where they are born, for example, if they were born in the Arctic or in Africa, they could have very different knowledge requirements. In the case of polar bears or lions, they might have to face either ones, or maybe they might have to cross glaciers or savannahs, or they might have to build something from ice or mud, for example. If an agent faces a wide variety of events during the course of his or her life, this may lead to questions about how to allocate resources in the future: whether to hunt or farm; whether to face locusts or wars; whether to become blind or deaf; whether to encounter a friend or foe. Having specialized knowledge of a situation is a must in each case. In such a situation, where the sum of such potential knowledge exceeds the capacity of the agent, the knowledge has to be adapted to the agent's particular circumstances based on the agent's experience, thereby requiring the agent to learn. As a result, the process of learning can take many computational forms in practice, such as making predictions, models, or skills by modifying parameters, or by constructing, curating, and reusing structures in order to build, curate, and reuse.

The environment may require the acquisition of both innate and acquired knowledge, and a reward-maximizing agent will, whenever required, carry both innate and accumulated knowledge with it (for example, through evolution in natural agents and by design in artificial agents) and acquire the latter (through learning) according

to the requirements of the environment. There is a shift in the demand balance for learned knowledge as richer and longer-lived environments become more and more developed.

3.2. PERCEPTION CAN BE INFLUENCED BY REWARD ALONE

In order to be rewarded in the human world, we need to develop a variety of perceptual abilities. In addition to image segmentation for avoiding falling off a cliff, object recognition is used to classify healthy food from poisonous food, face recognition is used to distinguish between friends and enemies, scene parsing while driving, and speech recognition for understanding a verbal warning to duck. It is possible that there are several modes of perception that are required, such as visual, aural, olfactory, somatosensory, or proprioceptive perception. Historically, these perceptual abilities were formulated using separate problem definitions [9]. Nevertheless, there has been a growing movement towards unifying perceptual abilities as a solution to supervised learning problems in recent times. As a general rule, the problem is formulated as minimising the classification error for examples in a test set, given that a training set is provided with correctly labeled examples. As a result of the unification of many perceptual abilities as supervised learning problems, significant progress has been made in this area.

The use of large data-sets by software engineers has found success in a wide range of real-world applications [23,15,8].

Rather than being understood as subserving to the maximization of reward, as per our hypothesis, perception can instead be understood as subserving it. For example, the perceptual abilities listed above may arise implicitly in service of maximizing healthy food, avoiding accidents, or minimizing pain. Moreover, perception has been shown to be associated with reward maximization in some animals [46,16]. In order to be able to support a greater range of perceptual behaviours, we need to view perception through the lens of reward maximization, rather than supervised learning, in order to ensure that we can achieve challenging and realistic forms of perceptual abilities:

An active form of perception is usually an intertwined interaction between action and observation, for example haptic perception (for example, identifying contents of a pocket with the movement of a fingertip), visual saccades (for example, switching the focus from bat to ball), physical experiments (such as hitting a nut with a rock in order to see if it breaks), or echolocation (for example, emitting sounds of different frequencies and measuring the echoes that follow).

According to some theories, the utility of perception often depends on the behaviors of the agents - for example, the cost of misclassifying a crocodile will depend on whether the agent walk or swims, and whether the agent will then fight or flee as a result of the misclassification.

It is possible that the process of acquiring information can come at a cost (e.g. if you turn your head and check for a predator, there is likely to be a cost in terms of energy, computation, and opportunity).

It is common for data to be distributed differently depending on the context. For example, for an agent to be successful, the agent would need to classify ice and polar bears, upon which their rewards will be based; for an agent to be successful, the agent must classify savannah and lions. As a consequence, perception will require experience in order to be able to deal with the vast amount of data that can be gathered in a rich environment, which may exceed the agent's capacity or the quantity of pre-existing data (see Section 3.1).

It is important to note that most applications of perception do not have access to data that has been labelled.

3.3. SOCIAL INTELLIGENCE IS FUELED BY REWARDS RATHER THAN PUNISHMENTS

It has been shown that social intelligence refers to the ability to understand and interact effectively with other agents. This ability is often formalised as the equilibrium solution of multi-agent games, which are typically based on game theory. In addition to being robust to deviations and worst-case scenarios, equilibrium solutions are considered desirable because they provide a stable equilibrium. In a Nash equilibrium, for example, there can be no unilateral deviation that will benefit the deviator [33] since all agents must have the same strategy. As a matter of fact, Nash equilibrium achieves the best possible value against the worst possible opponent in a zero-sum game [34]: it achieves the best possible value in the worst-case scenario.

The concept of social intelligence can also be understood and implemented as the process of maximizing cumulative reward from the perspective of a single agent within an environment that contains other agents, according to our hypothesis. It has been shown that a single agent can observe the behavior of other agents, and it can also affect other agents through its actions, just as it observes and affects any other aspect of its environment. Therefore, if an environment requires social intelligence (for example, because it contains animals or humans), reward maximization will make it possible for the agent to predict and influence the behavior of other agents. Therefore, if an environment requires social intelligence (for example, because it contains animals or humans), reward maximization will produce it.

As an example, robustness may also be required by the environment when multiple agents follow different strategies, as well as by the environment as a whole. A reward-maximizing agent must hedge its bets and choose a robust behaviour that will be effective against any of these possible strategies if these other agents are aliased (i.e. they cannot be predicted in advance what strategies other agents will follow). Moreover, the other agents' strategies are also likely to be adaptive as well. There is therefore a possibility that other agents' behavior may be affected by their past interactions, as well as other aspects of the environment (for example, a jammed door that only opens after the nth attempt). This type of adaptability is especially common in environments where reinforcement learning agents are present and are taught to maximize their own reward in order to survive. The reward-maximizing agent may need to use a mixed strategy (i.e. a combination of strategies) to avoid being exploited in this kind of environment, which may require aspects of social intelligence such as bluffing or hiding knowledge, and it may be necessary for him or her to utilize aspects of social intelligence.

As a matter of fact, reward maximization may even result in a better solution than an equilibrium [47]. This is because, rather than assuming optimal or worst-case behavior, it may capitalize on the suboptimal behavior of other agents. In addition, there is a unique optimal value for reward maximization in general-sum games [38], but the equilibrium value for general-sum games is not unique.

3.4. THE REWARD IS ENOUGH FOR LANGUAGE LEARNING

There has been considerable study of language in both natural [10] and artificial intelligence [28], as it has a significant role to play in human culture and interaction. The concept of intelligence itself is often defined in terms of the ability to understand and use language, especially natural language [60], since language plays a dominant role in human culture and interactions.

Several studies have reported significant success in predicting language using large corpus of data by treating language as if it were the optimization of a singular objective: the prediction of language using a large corpus of data [28,8]. As a result of this approach, we have been able to make progress towards several subproblems within natural language processing and understanding that had previously been studied or implemented in isolation. There are a variety of subproblems in which syntactic issues (formal grammar, part of speech tagging, parsing, segmenting) as well as semantic problems (lexical semantics, entailment, sentiment analysis), as well as some problems which bring the two together (such as summarization, dialogue systems).

It is, however, important to keep in mind that language modeling alone may not be sufficient to produce a broader set of linguistic abilities associated with intelligence, including the following:

In some cases, language is woven into other forms of action and observation as well. In addition to what was uttered, language can also be contextualized in relation to what is happening around the agent, as experienced through vision and other sensory modalities (e.g. a dialogue between two agents carrying an awkward object or building a shelter can serve as an example). Further, it should be noted that language is often accompanied by other forms of communication, such as gestures, facial expressions, tonal variations, or physical demonstrations, in addition to language itself.

Language utterances have a consequential and purposeful effect on the environment, usually by influencing other communicators within the environment, which in turn impacts the mental state and behavior of other communicators. It is possible to optimize these consequences to achieve various ends, for example, a salesperson will be able to tailor their language to maximize sales, while a politician will be able to optimize his or her language to maximize votes.

According to the agent's situation and behaviour, the utility of language may vary as well. For example, a miner may need language pertaining to the stability of rocks, whereas a farmer would probably need language pertaining to soil fertility. As a consequence of language, there may be an opportunity cost associated with it (e.g. talking about farming instead of doing the actual work of farming).

As a result of the richness of the environment, language can serve a wide range of purposes beyond the capacity of any corpus to cope with unforeseen events. In these cases, it may be necessary to solve linguistic problems dynamically, through experience — for instance, by interactively developing the most effective language to control a new disease, to build a new technology, or to find a way to address a new grievance of a rival so as to prevent aggression from happening.

We propose that the pursuit of reward is what gives language its full richness, which includes all of these broader capabilities. As an example of how an agent can produce complex sequences of actions (e.g. uttering sentences) based on complex sequences of observations (e.g. receiving sentences), this is an example of how he or she can influence other agents in the environment (cf. the discussion of social intelligence discussed above) and accumulate more rewards. Many reward-increasing benefits can add to the pressure to comprehend and produce language. If an agent is able to comprehend a "danger" warning, then it can predict and avoid negative rewards. It is likely that if an agent is capable of generating a "fetch" command, the environment (such as a dog) will move an object towards the agent. An agent may also only eat if it comprehends complex descriptions of food locations, generates complex instructions for growing food, engages in complex dialogue to negotiate for food, or builds long-term relationships that enhance those negotiations.

3.5. GENERALIZATION REQUIRES REWARD

An important aspect of generalisation is the ability to transfer a solution to one problem into a solution to another problem [37,58,61]. It is important to note that generalisation in supervised learning [37] refers to the process of applying the solution learned from one set of data, for example photographs, to another set of data, for example paintings. Recently, a large part of meta-learning research [61,20] has focused on the issue of transferring an agent from one environment to another.

Our hypothesis indicates that, as we have pointed out, generalisation may be interpreted and implemented as maximizing cumulative reward as an agent interacts with a single complex environment in an ongoing stream - once again following a standard protocol for agent-environment interactions. The agent encounters different aspects of the environment at different times in the human world, so generalisation is a necessity in environments

like that. As an example, fruit-eating animals may encounter new trees every day, in addition to becoming injured, suffering from droughts, or being attacked by invasive species. Regardless of the state of the animal, it is imperative that it adapts quickly to its new state by generalizing the experience of the past. The various states the animal is faced with are not neatly separated into disjointed, sequential tasks with distinct labels as well. Instead of determining what state the animal is in, the state can be determined by the animal's behavior; the state may combine different elements that overlap at different time scales, and important aspects of the state may be partially observed. To be able to accumulate rewards efficiently in rich environments, one must be able to generalize from past states to future states - with all the complexities that go along with that.

3.6. TO MOTIVATE IMITATION, REWARD MUST BE PROVIDED

There is no doubt that imitation is an important ability associated with human and animal intelligence and that it may be one of the key abilities that facilitate the rapid acquisition of other abilities, such as language, knowledge, and motor skills. There has been much discussion in artificial intelligence about imitation being formulated as a problem of learning from demonstration [45] as well as behavioural cloning [2], where the goal is to reproduce the actions selected by a teacher when the teacher's actions, observations, and rewards are explicitly provided to the agent, as long as the teacher is assumed to solve a symmetric problem for the agent. There have been several successful applications of behavioral cloning in machine learning [54,62,5], especially in cases where human teacher data is abundant but interactive experience is limited or expensive. It is also true that observational learning [3] is a natural ability to learn based on the observations of other people or animals, and it does not require that the teacher is symmetric or that direct access to their behavior, observation and rewards must be provided. In complex environments, observational learning may be demanded more than direct imitation through behavioural cloning:

There is no reason to assume that a separate data-set containing the teacher's data is available as some agents are part of their surroundings (e.g. a baby observing its mother), as other agents may be integrated into the agent's environment.

The agent may require an association between its own state (for example, the pose of the baby's body), and the state of another agent (for example, the pose of its mother), or between its own actions (for example, rotating a robot's manipulator) and observations made by another agent (for example, seeing a human hand), at potentially higher abstraction levels (for example, mimicking what the mother is doing rather than what she is doing with her muscles).

There are other agents whose actions or goals can only be partially discerned (e.g. in the case of an occluded human hand), as a result one can only infer their intentions after the fact.

It is possible for other agents to behave in an undesirable manner that should be avoided.

This means that there may be many additional agents in the environment that exhibit different skills or different levels of competence, and observational learning can even occur without any explicit agency being present (e.g. imitating the idea of building a bridge from the observation of a log that has fallen across a stream).

Our hypothesis is that these broader abilities of observation learning could result from the maximization of rewards, assuming that an agent simply observes other agents in its environment as integral parts, and that the maximization of reward is its driving force. There is a possibility that this can lead to many of the same benefits as behavioural cloning – such as sample-efficient learning – but with a much more extensive and integrated context that can potentially result in many of the same benefits as behavioural cloning.

3.7. GENERAL INTELLIGENCE REQUIRES ONLY A SMALL REWARD

As our final example, we are going to consider the ability that presents the greatest challenge as well as one in which our hypothesis has the greatest potential benefits. Generally, general intelligence can be defined as a capability that allows one to flexibly accomplish a variety of goals in a variety of situations. It may also be viewed as a quality that is found in both humans and other animals. There are many ways in which humans can address problems (such as locomotion, transportation, or communication) flexibly by using appropriate solutions (such as swimming or skiing, driving or kicking, writing, or sign language) based on their circumstances. A general intelligence model has been formalised as the ability of an agent to perform tasks in a variety of different contexts and goals through the use of a set of environments [25,14].

As a result, we have proposed that general intelligence can instead be understood as, and implemented, the process of maximizing a single reward in a complex, singular setting. Natural intelligence, for example, has to deal with a constant stream of experiences generated by its interaction with nature throughout its lifetime. In order to maximize its overall reward (e.g., hunger or reproduction), it may require a flexible ability on the part of the animal to achieve a vast variety of subgoals (e.g., foraging, fighting, or fleeing) in order to successfully fulfil its range of subgoals (e.g., foraging, fighting, or fleeing). The maximization of reward should therefore be sufficient to produce an artificial general intelligence if an artificial agent's experience stream is sufficiently rich, such as battery-life or survival.

4. AGENTS BASED ON REINFORCEMENT LEARNING

There is no agnostic assumption as to what type of agent the agent is. According to our main hypothesis, intelligence and its associated abilities serve to maximize reward. This leaves open the important question of how to construct an agent that maximizes reward. It is suggested in this section that the answer to this question might also be found in the concept of reward maximisation. This article focuses on agents that have the ability to learn from their experiences interacting with their environment how to maximize reward from this experience. We refer to such agents as reinforcement learning agents, and they have several advantages over traditional agents.

As a matter of fact, it is quite common to use the same name to describe both problem (e.g. mountain climbing refers to the problem of climbing a mountain), solution methods (such as ropes and pitons used by mountain climbers) as well as a field (such as mountain climbing as a pastime). In cases where the context does not make it clear, we will refer to reinforcement learning, reinforcement learning agents, and reinforcement learning as a whole.

One of the most natural ways to maximize reward is by learning how to do it from experience, by interacting with the environment. This is a solution that is undoubtedly the most natural one. When that interaction is repeated over time, a vast amount of information is acquired about cause and effect, about the consequences of actions, and about how to accumulate rewards as a result of this interaction. There is a natural tendency to bestow a general ability to the agent in order to allow it to discover its own behavior (placing faith in the designer's foreknowledge of the environment), rather than predetermining the agent's behavior (placing faith in experience). In specific terms, the design goal of maximizing reward is achieved through an ongoing internal process whereby a behavior that maximizes future reward is learned from experience.4

It has been suggested that reinforcement learning agents may provide a general solution method that may be effective, with minimal or even no modifications, across a broad range of reward signals and environments because they learn from experience.

Further, as in the natural world, a single environment may be so complex that it may contain a heterogeneous diversity of possible experiences, just as the natural world is. As long as the agent lives, the stream of observations and rewards it encounters will inevitably exceed the amount of preprogrammed behavior it will be able to perform (see Section 3.1). It is therefore crucial for the agent to have a general ability to adapt its behaviour to new experiences completely and continuously if it is to be able to achieve a high level of reward. In such complex environments, reinforcement learning agents may indeed provide the only viable solution.

A sufficient and general reinforcement learning agent may ultimately give rise to intelligence as well as its associated abilities if it is sufficiently powerful and general. As a result, if an agent can continuously adjust the way it behaves in order to improve the cumulative reward, then any abilities that are repeatedly requested by the environment must ultimately be expressed in the agent's behavior. In the course of learning to maximally maximise reward in an environment, such as the human world, in which those abilities are of ongoing value, a good reinforcement learning agent can learn to acquire behavior that demonstrates perception, communication, social intelligence and so forth.

The sample efficiency of reinforcement learning agents is not guaranteed by us in theory, and we do not offer any theoretical guarantees about it. As it turns out, the rate and degree at which abilities will emerge will depend on the specific environment, the learning algorithm, and inductive biases; in addition, it is possible to construct artificial environments in which learning will be hindered. In reality, our hypothesis is that powerful reinforcement learning agents, when placed in complex environments, will develop sophisticated intelligence expressions as a result of their placement. Assuming that this conjecture is true, then it provides the shortest pathway to the implementation of artificial general intelligence in the near future.

In recent years, there have been several examples of reinforcement learning agents in which they have been endowed with the ability to learn to maximize rewards, which have resulted in behaviors that have underperformed expectations, previous agents, and in a few cases even human experts. According to AlphaZero, when asked to maximize wins in the game of Go (see Section 1), he acquired a sophisticated intelligence across many aspects of the game [49]. When the same algorithm was applied to maximise outcomes in the game of chess, AlphaZero gained a whole set of skills that encompassed openings, endgames, piece mobility, king safety, etc. Among these capabilities was the ability to maximize outcomes at openings, endgames and so forth [48,44]. There are a range of abilities learned by reinforcement learning agents that are able to maximize scores in Atari 2600 [30], including the recognition, localisation, navigation, and motor control abilities required for each particular Atari game. In contrast, agents that are able to achieve maximum grip in vision-based robotic manipulation have acquired sensorimotor abilities such as object singulation, regrasping, and dynamic object tracking. In spite of the fact that these examples are far narrower in scope than those which are encountered by natural intelligence, they can be seen as some practical evidence of the effectiveness of the reward maximization principle in these situations.

As a matter of fact, one may ask how to learn how to maximize reward effectively in a practical agent. It is possible, for example, to maximize the reward directly (e.g. by optimizing the agent's policy [57]), or indirectly, by dividing the reward into subgoals such as representation learning, value prediction, model learning, and planning, which may all be further separated into subgoals [56]. Our paper does not elaborate on this question further, but we would like to point out that it is one of the central questions of reinforcement learning research throughout the world.

5. RELEVANT WORK THAT CAN BE DONE

Throughout history, intelligence has been associated with goal-oriented behavior [59]. This goal-oriented notion of intelligence is central to the concept of rationality, in which an agent chooses actions in a way that achieves its goals in an optimal manner or maximizes the utility of its actions. For decades, rationality has been employed as a method of understanding human behavior [63,4], as well as as the basis for artificial intelligence [42]. In this case, it is formalised with regard to an agent interacting with an environment.

When reasoning about goals, computational constraints have frequently been argued to be a crucial factor to be considered as well. In terms of bounded rationality [50,43,36] and computational rationality [27], agents are urged to choose the program that best fulfills their needs in light of the real-time consequences (for example, the time it takes to execute the program) of that program, and subject to restrictions on the number of programs that can be executed (for example, limiting the maximum size of the program). Taking into account these viewpoints, our contribution focuses on the question of whether or not a simple reward-based goal can provide a common basis for all of the abilities associated with intelligence and builds upon these viewpoints.

We presented in section 2 a common generalisation with partially observed histories, which was based on the standard protocol for reinforcement learning defined by Sutton and Barto [56].

It is also possible that the internal reward might be distinct from the design goal itself, but was chosen to assist it [51].

An aim of unified cognitive architectures [35,1] is to achieve general intelligence through a combination of multiple solution methods. There is no generic objective that justifies and explains the choice of architecture, nor is there any single goal towards which the individual components contribute, even though they combine a variety of solutions for separate subproblems (for example, perception or motor control).

A number of years ago, behaviourists [52,53] proposed language as a means to maximize rewards; however, reinforcement learning differs from behaviourism by giving agents the ability to construct and use internal states, which distinguishes it from behaviourism. A number of years ago, Shoham and Powers [47] examined the advantages of focusing on the objective of a single agent in multi-agent environments rather than the objective of equilibrium.

6. DISCUSSION OF THE ISSUES

In this chapter we present the reward-is-enough hypothesis as well as some of its implications, and then we briefly examine some questions that frequently arise when this hypothesis is discussed.

The question is which environment will give rise, through reward maximization, to the "most intelligent" behavior or to the "best" specific abilities (such as natural language) that one might be able to develop. As a matter of fact, the specific environmental experiences encountered by an agent are what shaped the nature of its subsequent abilities — such as those encountered by friends, enemies, teachers, toys, tools, or libraries during his or her lifetime. Our focus has been on the arguably more profound question of which generic objective could provide rise to all forms of intelligence, rather than focusing on the question of which specific application of intelligence may be of interest to the general public. It is possible that different, powerful forms of intelligence may emerge if the maximization of different rewards in various environments leads to the development of distinct, powerful types of intelligence, each with a unique array of abilities that are impressive and yet incomparable. It is important to note that a good reward-maximising agent will take advantage of any elements present in its environment, but the emergence of intelligence in some form is not determined by these elements. A human brain, for example, develops differently depending on the experiences it is exposed to in its environment from birth, but regardless of its specific culture or education, it will acquire sophisticated abilities regardless of these experiences.

It is often a matter of manipulating a reward signal because it is thought only an appropriately constructed reward will be able to influence general intelligence, and a desire to manipulate the reward signal often comes from that belief. However, we suggest that intelligence may emerge in a manner that is quite sensitive to the nature of the reward signal. As a result of the complex nature of environments such as the natural world, even a seemingly innocuous reward signal can demand intelligence and its associated abilities. The agent is given a reward

whenever the agent collects a round-shaped pebble, for example, and each round-shaped pebble means +1 reward for the agent. An agent may need to classify, manipulate, navigate to, store, and organize pebbles, and be aware of tides and waves in order to be able to use this reward signal effectively. It is possible to use these reward signals effectively in order to motivate people to help collect pebbles, to use tools and vehicles so that greater quantities of pebbles can be collected, to quarry and shape new pebbles, to develop new technologies for the collection of pebbles, or to set up a corporation whose only goal is to collect pebbles.

Could there be anything else that can be done to make intelligence more intelligent, aside from maximizing rewards, that can be done to make intelligence more intelligent, in addition to maximizing rewards? In order to uncover observational patterns, unsupervised learning may be used (for example, to identify patterns in observations) or predictions may be used (for example, to predict future observations [31]). In theory, it is possible that these principles provide us with a framework through which we can understand experiences, but they do not give us a framework through which we can select the actions we need to carry out in order to attain our goals. Therefore, they cannot provide us with an adequate framework for goal-oriented intelligence on their own. I believe that supervised learning is capable of mimicking human intelligence when done correctly; if you have sufficient data from humans to draw upon, then it is plausible to imagine that all the abilities associated with human intelligence will emerge as a result of the learning process itself when performed correctly. Even though it is possible to use human data for supervised learning, it is not sufficient to develop a general-purpose intelligence that is capable of optimizing for non-human goals in non-human environments without the use of human data in order to be able to achieve these non-human goals without the need to use human data in order for these technologies to be able to achieve such goals. Imitation intelligence has the problem of being unable to discover new behaviors that solve problems in unexpected ways, even though there is a lot of data available about humans. Due to the fact that it focuses on behaviors that have already been demonstrated in data and that have already been understood by humans, it is unlikely that it will be able to discover new behaviors and solve problems in novel and surprising ways. In addition to what I mentioned above, there is a section dedicated to offline learning that you can visit. In that section you can find a section where you will be able to learn about offline learning in more detail. It is important to recognize that natural selection is a consequence of the effort put forth to maximize the fitness of a population, both in terms of reproduction success as well as overall fitness, when taken into account. Since mutations and crossovers have the potential to improve a population's fitness in the long run, the long-term effect of this process can result in a population becoming more efficient. As far as our theory is concerned, the evolution of species was characterized by a period of widespread reproductive success that resulted in the development of natural intelligence. As a result of this hypothesis, it can be argued that the development of natural intelligence may have been influenced by a long period of widespread reproductive success throughout the evolution of species. Similarly, this hypothesis suggests that the evolution of natural intelligence might have been a long process. Therefore, during the course of the evolution of natural intelligence, there has been the opportunity to observe a period of successful reproduction that is closely related to the development of the organism over a long period of time, which is closely related to its evolution as well. During the lifetime of the organism, if it is capable of reproducing successfully, this may lead to a successful reproductive period throughout the organism's lifespan. Having reviewed the findings of this study, it is possible to conclude that, in light of the framework presented in this paper, it may be possible to argue that, in light of the findings of this study, reproductive success has long been considered a signal of natural intelligence as a part of the range of rewards that were regarded as determining natural intelligence in the past. The results of these studies suggest that this may no longer be the case. As a consequence of the results of this analysis, it has been suggested that at some point in history reproductive success may have had a role to play in determining natural intelligence. The use of artificial intelligence may also allow us to increase rewards for achieving other goals, for instance, as an additional benefit of artificial intelligence, we could be able to maximize the rewards associated with achieving other goals. This may be possible because artificial intelligence has the capability of maximizing the rewards associated with achieving other goals as well. Our design of such organisms may result in intelligences that are quite different from those of mutation and crossover, so we may be able to develop intelligences that are quite different from those of mutation and crossover, and because of this, we may have

very different objectives than reproducing success in the first place, resulting in intelligences that may differ greatly from those of mutation and crossover. The first explanation of natural intelligence may be that fitness maximisation is what determines it initially (e.g. a baby's brain), but the second explanation is that learning to maximize an intrinsic reward signal is the result of trial-and-error learning [51], which could also help explain how natural intelligence is able to adapt through experience to develop sophisticated abilities (such as the sophisticated abilities of the human adult brain) as a means to maximize fitness. It should be noted that maximising free energy or minimising surprise [13] may yield a number of natural intelligence abilities, but it does not provide a general-purpose intelligence that can be applied to a wide diversity of different tasks in a wide variety of different environments with ease. Consequently it may also miss abilities that are demanded by the optimal achievement of any one of those goals (for example, aspects of social intelligence required to mate with a partner, or tactical intelligence required to checkmate an opponent). Optimisation is a generic mathematical formalism that may maximise any signal, including cumulative reward, but does not specify how an agent interacts with its environment. By contrast, the reinforcement learning problem includes interaction at its heart: actions are optimised to maximise reward, those actions in turn determine the observations received from the environ- ment, which themselves inform the optimisation process; furthermore optimisation occurs online in real-time while the environment continues to tick.

Despite the fact that reinforcement learning researchers have studied a number of variations of the reward maximization problem, there are still many variations of the reward maximization problem to consider. In contrast to following a standard protocol for agent-environment interactions, many interactions are modified to suit different scenarios, which may include multiple agents, multiple environments, or multiple training periods. Instead of maximizing a generic objective that is defined by cumulative rewards, the goal is often formulated separately for a variety of different situations, such as multi-objective learning, risk-sensitive objectives, or goals that are defined by a human in the loop. Furthermore, rather than dealing with the problem of reward maximization for general environments, special-case problems are usually studied for particular environments, such as linear environments, deterministic environments, or stable environments, rather than for general environments. In spite of the fact that this may be an appropriate solution for specific applications, a solution to a specialized problem cannot typically be generalized. On the other hand, a solution to a general problem can also be applied to any special cases. Additionally, reinforcement learning can be viewed as a probabilistic framework that can be used to approximate the goal of maximizing rewards [66,39,26,17]. It is important to note that universal decision-making frameworks [21] provide theoretical but incomputable models of intelligence across all environments, while reinforcement learning problems provide a practical formulation of intelligence within a given environment.

Offline learning from a sufficiently large dataset may not be sufficient to solve problems that are already to a large extent solved within the data set. Offline learning might not be enough to solve problems that are already to a large extent solved within the available data set. A large dataset of squirrels collecting nuts, for example, is unlikely to demonstrate all the behaviors that are necessary to build a nut harvester. In complex environments, it is inevitably the case that, although it may be possible to generalise to the agent's current problems from solutions presented in or extracted from an offline data-set, this generalisation will always be imperfect in complex environments. Additionally, it is likely that the data required to solve the agent's current problems will have a negligible probability of occurring in offline data (e.g. under random behavior or imperfect human behavior), if the agent's current problems can be solved in these data. As a result of online interaction, an agent is able to specialize to the problems that it is currently facing, to continuously verify and correct the most pressing holes in its knowledge, and to find new behaviours that are very different from those found in the data and that yield greater rewards than those found in it.

In complex, unknown environments, are there sample-efficient reinforcement learning agents that can maximize reward? It might be interesting to ask if there are sample-efficient reinforcement learning agents that can maximize rewards in complex, unknown environments. As a first step towards answering this question, we would like to point out that an effective agent might be able to use additional experiential signals in order to maximize

future reward. It has been demonstrated that many solution methods, including model-free reinforcement learning, can learn to associate future rewards with features of observations, by approximating value functions [55], which, in turn, can provide rich secondary signals that drive learning of deeper associations through a recursive bootstrapping process. Other solution methods, including model-based reinforcement learning, construct predictions of observations or features of observations, in order to facilitate the subsequent maximization of reward by planning the subsequent observations. A rapid learning process can also be facilitated by observing other agents within the environment, as discussed in Section 3.6.

Despite this, it is often the challenge of applying sample-efficient reinforcement learning in complex environments that has led researchers to introduce assumptions or to develop simpler abstractions that have been more conducive to theory and practice, allowing them to make assumptions or develop simpler abstractions. However, these assumptions and abstractions may simply side-step the difficulties that must inevitably be faced by any broadly capable intelligence. However, we have chosen to take on the challenge head-on and focus our efforts on finding a solution to the problem; in the future, we hope that other researchers will join us in this quest as well.

7. CONCLUSIONS FOR THE FUTURE

The hypothesis we have presented in this paper is that it may be sufficient to understand intelligence and its associated abilities if we focus on maximising total rewards. There is no doubt that rich environments tend to demand a wide variety of abilities as a result of the desire to maximize rewards. The bountiful expressions of intelligence found in nature, and presumably in the future for artificial agents, are all instances of this same idea with different environments and different rewards. A singular goal of maximizing reward may also result in a deeper, broader, and more integrated understanding of abilities than specialised problem formulations for each specific ability based on a particular set of criteria. In particular, we explored in greater depth several abilities that may seem difficult at first glance to comprehend through reward maximization alone, such as knowledge, learning, perception, social intelligence, language, generalization, imitation, and general intelligence, and found that reward maximization could provide a basis for understanding each capability in a more comprehensive way. Finally, we have presented a hypothesis that intelligence can emerge in practice from sufficiently powerful reinforcement learning agents that learn to maximize their reward in the future. This conjecture would provide a direct pathway to understanding and building a general intelligence artificial intelligence if it were true.

References

[1] J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind, Psychol. Rev. 111 (4) (2004) 1036.

[2] M. Bain, C. Sammut, A framework for behavioural cloning, in: Machine Intelligence 15, 1995, pp. 103–129.

[3] A. Bandura, D.C. McClelland, Social Learning Theory, vol. 1, Prentice Hall, Englewood Cliffs, 1977.

[4] G.S. Becker, The Economic Approach to Human Behavior. Economic Theory, University of Chicago Press, 1976.

[5] Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L.D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba, End to end learning for self-driving cars, CoRR, arXiv:1604.07316 [abs], 2016.

[6] D. Borsa, N. Heess, B. Piot, S. Liu, L. Hasenclever, R. Munos, O. Pietquin, Observational learning by reinforcement learning, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1117–1124.

[7] J. Bratman, M. Shvartsman, R. Lewis, S. Singh, A new approach to exploring language emergence as boundedly optimal control in the face of environ- mental and cognitive constraints, in: Proceedings of the 10th International Conference on Cognitive Modeling, ICCM, 2010.

[8] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv:2005.14165, 2020.

[9] E.C. Carterette, M.P. Friedman, Handbook of Perception, Academic Press, 1978.

[10] N. Chomsky, D.W. Lightfoot, Syntactic Structures, Walter de Gruyter, 2002.

[11] A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behav. Brain Sci. 36 (3) (2013) 181–204.

[12] G. Debreu, Representation of a Preference Ordering by a Numerical Function, 1954.

[13] K. Friston, The free-energy principle: A unified brain theory?, Nat. Rev. Neurosci. 11 (127–38) (2010) 02.

[14] B. Goertzel, C. Pennachin, Artificial General Intelligence, vol. 2, Springer, 2007.

[15] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6645–6649.

[16] N. Grujic, J. Brus, D. Burdakov, R. Polania, Rational inattention in mice, bioRxiv, https://doi.org/10.1101/2021.05.26.445807, 2021.

[17] D. Hafner, P.A. Ortega, J. Ba, T. Parr, K.J. Friston, N. Heess, Action and perception as divergence minimization, CoRR, arXiv:2009.01791 [abs], 2020.

[18] J. Hawkins, S. Blakeslee, On Intelligence, Times Books, USA, 2004.

[19] G. Hinton, T.J. Sejnowski, Unsupervised Learning: Foundations of Neural Computation, The MIT Press, 1999.

[20] T.M. Hospedales, A. Antoniou, P. Micaelli, A.J. Storkey, Meta-learning in neural networks: A survey, CoRR, arXiv:2004.05439 [abs], 2020.

[21] M. Hutter, Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability, Springer, 2005.

[22] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, S. Levine, Scalable deep reinforcement learning for vision-based robotic manipulation, in: 2nd Annual Conference on Robot Learning, Proceedings, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, in: Proceedings of Machine Learning Research, vol. 87, PMLR, 2018, pp. 651–673.

[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[24] Y. LeCun, Computer perception with deep learning, 2013.

[25] S. Legg, M. Hutter, Universal intelligence: A definition of machine intelligence, Minds Mach. 17 (4) (2007) 391–444.

[26] S. Levine, Reinforcement learning and control as probabilistic inference: Tutorial and review, CoRR, arXiv:1805.00909 [abs], 2018.

[27] R.L. Lewis, A. Howes, S. Singh, Computational rationality: Linking mechanism and behavior through bounded utility maximization, Top. Cogn. Sci. 6 (2) (2014) 279–311.

[28] C.D. Manning, C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[29] J. McCarthy, What Is AI?, 1998.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.

[31] J. Modayil, A. White, R.S. Sutton, Multi-timescale nexting in a reinforcement learning robot, Adapt. Behav. 22 (2) (2014) 146–160.

[32] M. Müller, Computer Go, Artif. Intell. 134 (1–2) (2002) 145–179.

[33] J.F. Nash, et al., Equilibrium points in n-person games, Proc. Natl. Acad. Sci. 36 (1) (1950) 48–49.

[34] J.v. Neumann, Zur theorie der gesellschaftsspiele, Math. Ann. 100 (1) (1928) 295–320.

[35] A. Newell, Unified Theories of Cognition, Harvard University Press, USA, 1990.

[36] L. Orseau, M.B. Ring, Space-time embedded intelligence, in: Proceedings of the 5th International Conference on Artificial General Intelligence, Lecture Notes in Computer Science, vol. 7716, Springer, 2012, pp. 209–218.

[37] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[38] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.

[39] K. Rawlik, M. Toussaint, S. Vijayakumar, On stochastic optimal control and reinforcement learning by approximate inference, in: Twenty-Third Interna- tional Joint Conference on Artificial Intelligence, 2013.

[40] M. Redmond, C. Garlock, AlphaGo to Zero: The Complete Games, Smart Go, 2020.

[41] J. Ben Rosen, Existence and uniqueness of equilibrium points for concave n-person games, Econometrica (1965) 520–534.

[42] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 1995.

[43] S.J. Russell, D. Subramanian, Provably bounded-optimal agents, J. Artif. Intell. Res. 2 (1995) 575–609.

[44] M. Sadler, N. Regan, G. Kasparov, Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI, New in Chess, 2019.

[45] S. Schaal, Learning from demonstration, in: M. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2–5, 1996, MIT Press, 1996, pp. 1040–1046.

[46] J. Schaffner, P. Tobler, T. Hare, R. Polania, Neural codes in early sensory areas maximize fitness, bioRxiv, https://doi.org/10.1101/2021.05.10.443388, 2021.

[47] Y. Shoham, R. Powers, T. Grenager, If multi-agent learning is the answer, what is the question?, Artif. Intell. 171 (7) (2007) 365–377.

[48] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science 362 (6419) (2018) 1140–1144.

[49] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.

[50] H.A. Simon, A behavioral model of rational choice, Q. J. Econ. 69 (1) (1955) 99–118.

[51] S. Singh, R.L. Lewis, A.G. Barto, J. Sorg, Intrinsically motivated reinforcement learning: An evolutionary perspective, IEEE Trans. Auton. Ment. Dev. 2 (2) (2010) 70–82.

[52] B.F. Skinner, The Behavior of Organisms: An Experimental Analysis, Appleton-Century-Crofts, New York, 1938.

[53] B.F. Skinner, Verbal Behavior, Appleton-Century-Crofts, New York, 1957.

[54] N. Stiennon, L. Ouyang, J. Wu, D.M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P.F. Christiano, Learning to summarize with human feedback, in: Advances in Neural Information Processing Systems 33, 2020.

[55] R.S. Sutton, Learning to predict by the methods of temporal differences, Mach. Learn. 3 (1) (1988) 9–44.

[56] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, second edition, The MIT Press, 2018.

[57] R.S. Sutton, D.A. McAllester, S. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in: Advances in Neural Information Processing Systems, 2000, pp. 1057–1063.

[58] M.E. Taylor, P. Stone, Transfer learning for reinforcement learning domains: A survey, J. Mach. Learn. Res. 10 (1) (2009) 1633–1685.

[59] E.C. Tolman, Purposive Behavior in Animals and Men, Century/Random House, UK, 1932.

[60] A.M. Turing, Computing machinery and intelligence, Mind 59 (236) (1950) 433.

[61] J. Vanschoren, Meta-learning: A survey, CoRR, arXiv:1810.03548 [abs], 2018.

[62] O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I.Danihelka, A. Huang, L. Sifre, T. Cai, J.P. Agapiou, M. Jaderberg, A.S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T.L. Paine, Ç. Gülçehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T.P. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in starcraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354.

[63] M. Weber, G. Roth, C. Wittich, E. Fischoff, Economy and Society: An Outline of Interpretive Sociology, University of California Press, 1978.

[64] Y. Zhou, AlphaGo vs Ke Jie, Slate and Shell, 2017.

[65] Y. Zhou, Rethinking Opening Strategy: AlphaGo's Impact on Pro Play, Slate and Shell, 2018.

[66] B.D. Ziebart, J.A. Bagnell, A.K. Dey, Modeling interaction via the principle of maximum causal entropy, in: Proceedings of the 27th International Con- ference on Machine Learning, Omnipress, 2010, pp. 1255–1262.