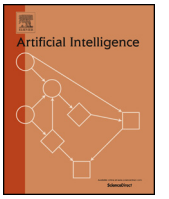
Artificial Intelligence 299 (2021) 103535

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



# Reward is enough

David Silver *, Satinder Singh, Doina Precup, Richard S. Sutton

ARTICLE INFO

*Article history:*
Received 12 November 2020
Received in revised form 28 April 2021
Accepted 12 May 2021
Available online 24 May 2021

*Keywords:*
Artificial intelligence
Artificial general intelligence
Reinforcement learning
Reward

ABSTRACT

In this article we hypothesise that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward. Accordingly, reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation. This is in contrast to the view that specialised problem formulations are needed for each ability, based on other signals or objectives. Furthermore, we suggest that agents that learn through trial and error experience to maximise reward could learn behaviour that exhibits most if not all of these abilities, and therefore that powerful reinforcement learning agents could constitute a solution to artificial general intelligence.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

Expressions of intelligence in animal and human behaviour are so bountiful and so varied that there is an ontology of associated abilities to name and study them, e.g. social intelligence, language, perception, knowledge representation, planning, imagination, memory, and motor control. What could drive agents (natural or artificial) to behave intelligently in such a diverse variety of ways?

One possible answer is that each ability arises from the pursuit of a goal that is designed specifically to elicit that ability. For example, the ability of social intelligence has often been framed as the Nash equilibrium of a multi-agent system; the ability of language by a combination of goals such as parsing, part-of-speech tagging, lexical analysis, and sentiment analysis; and the ability of perception by object segmentation and recognition. In this paper, we consider an alternative hypothesis: that the generic objective of maximising reward is enough to drive behaviour that exhibits most if not all abilities that are studied in natural and artificial intelligence.

This hypothesis may startle because the sheer diversity of abilities associated with intelligence seems to be at odds with any generic objective. However, the natural world faced by animals and humans, and presumably also the environments faced in the future by artificial agents, are inherently so complex that they require sophisticated abilities in order to succeed (for example, to survive) within those environments. Thus, success, as measured by maximising reward, demands a variety of abilities associated with intelligence. In such environments, any behaviour that maximises reward must necessarily exhibit those abilities. In this sense, the generic objective of reward maximisation contains within it many or possibly even all the goals of intelligence.

Reward thus provides two levels of explanation for the bountiful expressions of intelligence found in nature. First, different forms of intelligence may arise from the maximisation of different reward signals in different environments, resulting for example in abilities as distinct as echolocation in bats, communication by whale-song, or tool use in chimpanzees. Sim-

* Corresponding author.
*E-mail address:* davidsilver@google.com (D. Silver).

https://doi.org/10.1016/j.artint.2021.103535
0004-3702/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

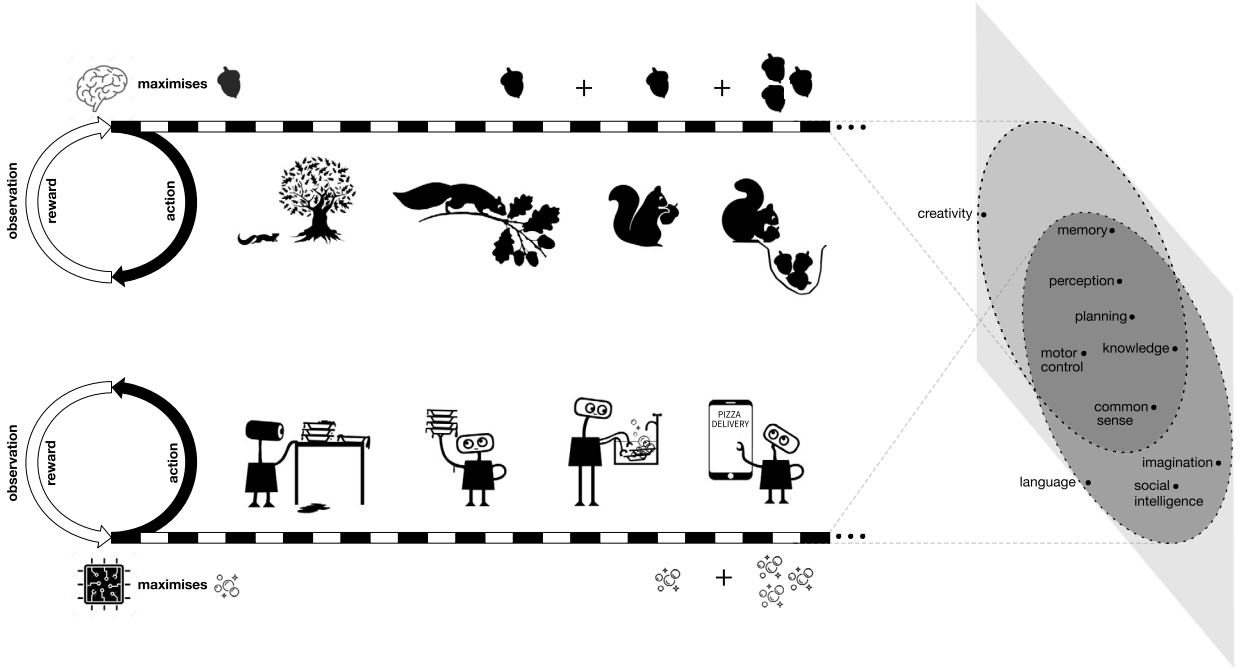



**Fig. 1.** The *reward-is-enough* hypothesis postulates that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment. For example, a squirrel acts so as to maximise its consumption of food (top, reward depicted by acorn symbol), or a kitchen robot acts to maximise cleanliness (bottom, reward depicted by bubble symbol). To achieve these goals, complex behaviours 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).

ilarly, artificial agents may be required to maximise a variety of reward signals in future environments, resulting in new forms of intelligence with abilities as distinct as laser-based navigation, communication by email, or robotic manipulation.

Second, the intelligence of even a single animal or human is associated with a cornucopia of abilities. According to our hypothesis, all of these abilities subserve a *singular goal* of maximising that animal or agent's reward within its environment. In other words, the pursuit of one goal may generate complex behaviour that exhibits multiple abilities associated with intelligence. Indeed, such reward-maximising behaviour may often be consistent with specific behaviours derived from the pursuit of separate goals associated with each ability.

For example, a squirrel's brain may be understood as a decision-making system that receives sensations from, and sends motor commands to the squirrel's body. The behaviour of the squirrel may be understood as maximising a cumulative reward such as satiation (i.e. negative hunger). In order for a squirrel to minimise hunger, the squirrel-brain must presumably have abilities of perception (to identify good nuts), knowledge (to understand nuts), motor control (to collect nuts), planning (to choose where to cache nuts), memory (to recall locations of cached nuts) and social intelligence (to bluff about locations of cached nuts, to ensure they are not stolen). Each of these abilities associated with intelligence may therefore be understood as subserving a singular goal of hunger minimisation (see Fig. 1).

As a second example, a kitchen robot may be implemented as a decision-making system that receives sensations from, and sends actuator commands to, the robot's body. The singular goal of the kitchen robot is to maximise a reward signal measuring cleanliness.[1] In order for a kitchen robot to maximise cleanliness, it must presumably have abilities of perception (to differentiate clean and dirty utensils), knowledge (to understand utensils), motor control (to manipulate utensils), memory (to recall locations of utensils), language (to predict future mess from dialogue), and social intelligence (to encourage young children to make less mess). A behaviour that maximises cleanliness must therefore yield all these abilities in service of that singular goal (see Fig. 1).

When abilities associated with intelligence arise as solutions to a singular goal of reward maximisation, this may in fact provide a deeper understanding since it explains *why* such an ability arises (e.g. that classification of crocodiles is important to avoid being eaten). In contrast, when each ability is understood as the solution to its own specialised goal, the *why* question is side-stepped in order to focus upon *what* that ability does (e.g. discriminating crocodiles from logs). Furthermore, a singular goal may also provide a broader understanding of each ability that may include characteristics that are otherwise hard to formalise, such as dealing with irrational agents in social intelligence (e.g. pacifying an angry aggressor), grounding language to perceptual experience (e.g. dialogue regarding the best way to peel a fruit), or understanding haptics in percep-

[1] For example, as judged by occasional human inspection.

tion (e.g. picking a sharp object from a pocket). Finally, implementing abilities in service of a singular goal, rather than for their own specialised goals, also answers the question of how to integrate abilities, which otherwise remains an outstanding issue.

Having established that reward maximisation is a suitable objective for understanding the problem of intelligence, one may consider methods for solving the problem. One might then expect to find such methods in natural intelligence, or choose to implement them in artificial intelligence. Among possible methods for maximising reward, the most general and scalable approach is to learn to do so, by interacting with the environment by trial and error. We conjecture that an agent that can effectively learn to maximise reward in this manner would, when placed in a rich environment, give rise to sophisticated expressions of general intelligence.

A recent salutary example of both problem and solution based on reward maximisation comes from the game of Go. Research initially focused largely upon distinct abilities, such as openings, shape, tactics, and endgames, each formalised using distinct objectives such as sequence memorisation, pattern recognition, local search, and combinatorial game theory [32]. AlphaZero [49] focused instead on a singular goal: maximising a reward signal that is 0 until the final step, and then $+1$ for winning or $-1$ for losing. This ultimately resulted in a deeper understanding of each ability – for example, discovering new opening sequences [65], using surprising shapes within a global context [40], understanding global interactions between local battles [64], and playing safe when ahead [40]. It also yielded a broader set of abilities that had not previously been satisfactorily formalised – such as balancing influence and territory, thickness and lightness, and attack and defence. AlphaZero's abilities were innately integrated into a unified whole, whereas integration had proven highly problematic in prior work [32]. Thus, maximising wins proved to be enough, in a simple environment such as Go, to drive behaviour exhibiting a variety of specialised abilities. Furthermore, applying the same method to different environments such as chess or shogi [48] resulted in new abilities such as piece mobility and colour complexes [44]. We argue that maximising rewards in richer environments – more comparable in complexity to the natural world faced by animals and humans – could yield further, and perhaps ultimately all abilities associated with intelligence.

The rest of this paper is organised as follows. In Section 2 we formalise the objective of reward maximisation as the problem of reinforcement learning. In Section 3 we present our main hypothesis. We consider several important abilities associated with intelligence, and discuss how reward maximisation may yield those abilities. In Section 4 we turn to the use of reward maximisation as a solution strategy. We present related work in Section 5 and finally, in Section 6 we discuss possible weaknesses of the hypothesis and consider several alternatives.

## 2. Background: the reinforcement learning problem

Intelligence may be understood as a flexible ability to achieve goals. For example according to John McCarthy, *"intelligence is the computational part of the ability to achieve goals in the world"* [29]. Reinforcement learning [56] formalises the problem of goal-seeking intelligence. The general problem may be instantiated with a wide and realistic range of goals and worlds — and hence a wide range of forms of intelligence — corresponding to different reward signals to maximise in different environments.

### 2.1. Agent and environment

Like many interactive approaches to artificial intelligence [42], reinforcement learning follows a protocol that decouples a problem into two systems that interact sequentially over time: an *agent* (the solution) that takes decisions and an *environment* (the problem) that is influenced by those decisions. This is in contrast to other specialised protocols that may for example consider multiple agents, multiple environments, or other modes of interaction.

### 2.2. Agent

An *agent* is a system that receives at time $t$ an observation $O_t$ and outputs an action $A_t$. More formally, the agent is a system $A_t = \alpha(H_t)$ that selects an action $A_t$ at time $t$ given its *experience* history $H_t = O_1, A_1, ..., O_{t-1}, A_{t-1}, O_t$, in other words, given the sequence of observations and actions that have occurred in the history of interactions between the agent and the environment.

The agent system $\alpha$ is limited by practical constraints to a bounded set [43]. The agent has limited capacity determined by its machinery (for example, limited memory in a computer or limited neurons in a brain). The agent and environment systems execute in real-time. While the agent spends time computing its next action (e.g. producing no-op actions while deciding whether to run away from a lion), the environment system continues to process (e.g. the lion attacks). Thus, the reinforcement learning problem represents a practical problem, as faced by natural and artificial intelligence, rather than a theoretical abstraction that ignores computational limitations.

This paper does not delve into the nature of the agent, focusing instead on the problem it must solve, and the intelligence that may be induced by any solution to that 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

An *environment* is a system that receives action $A_t$ at time $t$ and responds with observation $O_{t+1}$ at the next time step. More formally, an environment is a system $O_{t+1} = \varepsilon(H_t, A_t, \eta_t)$ that determines the next observation $O_{t+1}$ that the agent will receive from the environment, given experience history $H_t$, the latest action $A_t$, and potentially a source of randomness $\eta_t$.

The environment specifies within its definition the interface to the agent. The agent consists solely of the decision-making entity; anything outside of that entity (including its body, if it has one) is considered part of the environment. This includes both the sensors and actuators that define the observations for the agent and the actions available to the agent respectively.

Note that this definition of environment is very broad and encompasses many problem dimensions, including those in Table 1.

### 2.4. Rewards

The reinforcement learning problem represents goals by cumulative rewards. A *reward* is a special scalar observation $R_t$, emitted at every time-step $t$ by a *reward signal* in the environment, that provides an instantaneous measurement of progress towards a goal. An instance of the reinforcement learning problem is defined by an environment $\varepsilon$ with a reward signal, and by a cumulative objective to maximise, such as a sum of rewards over a finite number of steps, a discounted sum, or the average reward per time-step.

A wide variety of goals can be represented by rewards.[2] For example, a scalar reward signal can represent weighted combinations of objectives, different trade-offs over time, and risk-seeking or risk-averse utilities. Reward can also be determined by a human-in-the-loop, for example humans may provide explicit reinforcement of desired behaviour, online feedback via clickthrough or thumbs-up, delayed feedback via questionnaires or surveys, or by a natural language utterance. Including feedback from a human can provide a mechanism to formulate seemingly fuzzy goals such as "I'll know it when I see it".

In addition to their generality, rewards also provide intermediate feedback, potentially at every time-step, on progress towards the goal. This intermediate signal is an essential part of the problem definition when considering long or infinite streams of experience – without intermediate feedback, learning is not possible.

## 3. Reward is enough

We have seen, in the previous section, that rewards are sufficient to express a wide variety of goals, corresponding to the diverse purposes towards which intelligence may be directed.

We now have all the ingredients to make our main point: that many different forms of intelligence can be understood as subserving the maximisation of reward, and that the many abilities associated with each form of intelligence may arise implicitly from the pursuit of those rewards. Taken to its limit, we hypothesise that all intelligence and associated abilities may be understood in this manner:

**Hypothesis** *(Reward-is-Enough).* Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment.

This hypothesis is important because, if it is true, it suggests that a good reward-maximising agent, in the service of achieving its goal, could implicitly yield abilities associated with intelligence. A good agent in this context is one that

[2] Indeed, the *reward hypothesis* speculates that *all* goals may be represented by rewards [56]. This should not be confused with our reward-is-enough hypothesis, which considers the abilities that arise implicitly from the pursuit of any one such goal.

successfully – perhaps using as-yet-undiscovered algorithms – performs well at maximising cumulative reward in its environment. We return to the question of how such an agent may be constructed in Section 4.

In principle, any behaviour can be encoded by maximisation of a reward signal that is explicitly chosen to induce that behaviour [12] (for example, providing rewards when objects are correctly identified, or when a syntactically correct sentence is produced). The hypothesis here is intended to be much stronger: that intelligence and associated abilities will implicitly arise in the service of maximising one of many possible reward signals, corresponding to the many pragmatic goals towards which natural or artificial intelligence may be directed.

Sophisticated abilities may arise from the maximisation of simple rewards in complex environments. For example, the minimisation of hunger in a squirrel's natural environment demands a skilful ability to manipulate nuts that arises from the interplay between (among other factors) the squirrel's musculoskeletal dynamics; objects such as leaves, branches, or soil that the squirrel or nut may be resting upon, connected to, or obstructed by; variations in the size and shape of nuts; environmental factors such as wind, rain, or snow; and changes due to ageing, disease or injury. Similarly, the pursuit of cleanliness in a kitchen robot demands a sophisticated ability to perceive utensils in an enormous range of states that includes clutter, occlusion, glare, encrustation, damage, and so on.

Furthermore, the maximisation of many other reward signals by a squirrel (such as maximising survival time, minimising pain, or maximising reproductive success) or kitchen robot (such as maximising a healthy eating index, maximising positive feedback from the user, or maximising their gastronomic endorphins), and in many other environments (such as different habitats, other dexterous bodies, or different 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 the maximisation of many 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 is enough for knowledge and learning*

We define knowledge as information that is internal to the agent; for example, knowledge may be contained within the parameters of an agent's functions for selecting actions, predicting cumulative reward, or predicting features of future observations. Some of this knowledge may be innate (prior knowledge), while some knowledge may be acquired through learning.

An environment may demand innate knowledge. In particular, it may be necessary, in order to maximise total reward, to have knowledge that is immediately accessible in novel situations. For example, a new-born gazelle may need to run away from a lion. In this case, innate understanding of predator evasion may be necessary before there is any opportunity to learn this knowledge. Note, however, that the extent of prior knowledge is limited both in theory (by the capacity of the agent) and in practice (by the difficulty of constructing useful prior knowledge). Furthermore, unlike other abilities that we shall consider, environmental demand for innate knowledge cannot be operationalised; it is the knowledge that comes before experience, and therefore cannot be acquired from experience.

An environment may also demand learned knowledge. This will occur when future experience is uncertain, due to unknown elements, stochasticity, or complexity of the environment. This uncertainty results in a vast array of potential knowledge that the agent may need, depending on its particular realisation of future events. In rich environments, including many of the hardest and most important problems in natural and artificial intelligence, the total space of potential knowledge is far larger than the capacity of the agent. For example, consider an early-human environment. An agent may be born in either the Arctic or Africa, leading to radically different requirements for knowledge to maximise total reward: facing polar bears or lions; traversing glaciers or savannah; building with ice or with mud; etc. Particular events within the agent's life may also lead to radically different needs: choosing to hunt or to farm; facing locusts or war; going blind or going deaf; encountering friend or foe; etc. In each of these possible lives, the agent must acquire detailed and specialised knowledge. If the sum of such potential knowledge outstrips the agent's capacity, knowledge must be a function of the agent's experience and adapt to the agent's particular circumstances – thus demanding learning. In practice this learning may take many computational forms, for example making predictions, models or skills through the adaptation of parameters or the construction, curation and reuse of structure.

In summary, the environment may call for both innate and learned knowledge, and a reward-maximising agent will, whenever required, contain the former (for example, through evolution in natural agents and by design in artificial agents) and acquire the latter (through learning). In richer and longer-lived environments the balance of demand shifts increasingly toward learned knowledge.

### 3.2. Reward is enough for perception

The human world demands a variety of perceptual abilities to accumulate rewards. Some life-or-death examples include: image segmentation to avoid falling off a cliff; object recognition to classify healthy and poisonous foods; face recognition to differentiate friend from foe; scene parsing while driving; or speech recognition to understand a verbal warning to "duck!" Several modes of perception may be required, including visual, aural, olfactory, somatosensory or proprioceptive perception.

Historically, these perceptual abilities were formulated using separate problem definitions [9]. More recently, there has been a growing movement towards unifying perceptual abilities as solutions to supervised learning problems [24]. The problem is typically formalised as minimising the classification error for examples in a test set, given a training set of correctly labelled examples. The unification of many perceptual abilities as supervised learning problems has led to considerable success in a wide variety of real-world applications where large data-sets are available [23,15,8].

As per our hypothesis, we suggest that perception may instead be understood as subserving the maximisation of reward. For example, the perceptual abilities listed above may arise implicitly in service of maximising healthy food, avoiding accidents, or minimising pain. Indeed, perception in some animals has been shown to be consistent with reward maximisation [46,16]. Considering perception from the perspective of reward maximisation rather than supervised learning may ultimately support a greater range of perceptual behaviours, including challenging and realistic forms of perceptual abilities:

- Action and observation are typically intertwined into active forms of perception, such as haptic perception (e.g. identifying the contents of a pocket by moving fingertips), visual saccades (e.g. moving eyes to switch focus between bat and ball), physical experimentation (e.g. hitting a nut with a rock to see if it will break), or echolocation (e.g. emitting sounds at varying frequency and measuring timings and intensity of subsequent echoes).
- The utility of perception often depends upon the agent's behaviour – for example, the cost of misclassifying a crocodile depends upon whether the agent is walking or swimming, and whether the agent would subsequently fight or flee.
- There may be an explicit or implicit cost to acquiring information (e.g. there are energy, computational, and opportunity costs for turning the head and checking for a predator).
- The distribution of data is typically context-dependent. For example, an Arctic agent may be required to classify ice and polar bears, upon which its rewards depend; while an African agent may need to classify savannah and lions. The diversity of potential data may, in rich environments, vastly exceed the agent's capacity or the quantity of pre-existing data (see Section 3.1) – requiring that perception is learned from experience.
- Many applications of perception do not have access to labelled data.

### 3.3. Reward is enough for social intelligence

Social intelligence is the ability to understand and interact effectively with other agents. This ability is often formalised, using game theory, as the equilibrium solution of a multi-agent game. Equilibrium solutions are considered desirable because they are robust to deviation or worst-case scenarios. For example, the Nash equilibrium is a joint strategy for all agents such that no unilateral deviation gives benefit to the deviator [33]. In zero-sum games a Nash equilibrium is also minimax optimal [34]: it attains the best possible value against worst-case opponents.

According to our hypothesis, social intelligence may instead be understood as, and implemented by, maximising cumulative reward from the point of view of one agent in an environment that contains other agents. Following this standard agent-environment protocol, one agent observes the behaviour of other agents, and may affect other agents through its actions, just as it observes and affects any other aspect of its environment. An agent that can anticipate and influence the behaviour of other agents can typically achieve greater cumulative reward. Thus, if an environment needs social intelligence (e.g. because it contains animals or humans), reward maximisation will produce social intelligence.

Robustness may also be demanded by the environment, for example when the environment contains multiple agents following different strategies. If these other agents are aliased (i.e. there is no way to identify in advance which strategies other agents will follow) then a reward-maximising agent must hedge its bets and choose a robust behaviour that will be effective against any of these potential strategies. Furthermore, the other agents' strategies may be adaptive. This means that the behaviour of other agents may depend on the agent's past interactions, just like other aspects of the environment (for example, a jammed door that only opens on the $n$th attempt). In particular, such adaptivity may arise in environments that contain one or more reinforcement learning agents that learn to maximise their own reward. This kind of environment may call for aspects of social intelligence such as bluffing or information hiding, and a reward-maximising agent may have to be stochastic (i.e. use a mixed strategy) to avoid exploitation.

Reward maximisation may in fact lead to a better solution than an equilibrium [47]. This is because it may capitalise upon suboptimal behaviours of other agents, rather than assuming optimal or worst-case behaviour. Furthermore, reward maximisation has a unique optimal value [38], while the equilibrium value is non-unique in general-sum games [41].

### 3.4. Reward is enough for language

Language has been a subject of considerable study in both natural [10] and artificial intelligence [28]. Because language plays a dominant role in human culture and interactions, the definition of intelligence itself is often premised upon the ability to understand and use language, especially natural language [60].

Recently, significant success has been achieved by treating language as the optimisation of a singular objective: the predictive modelling of language within a large corpus of data [28,8]. This approach has facilitated progress towards many subproblems of language that were often previously studied or implemented separately within natural language processing and understanding, including both syntactic subproblems (e.g. formal grammar, part-of-speech tagging, parsing, segmentation) and semantic subproblems (e.g. lexical semantics, entailment, sentiment analysis) as well as some that bring both together (e.g. summarisation, dialogue systems).

Nevertheless, language modelling by itself may not be sufficient to produce a broader set of linguistic abilities associated with intelligence that includes the following:

- Language may be intertwined with other modalities of action and observation. Language is often contextual, not just in what was uttered but also on what else is happening in the environment around the agent, as perceived through vision and other sensory modalities (e.g. consider a dialogue between two agents carrying an awkward object or building a shelter). Furthermore, language is often interspersed with other communicative actions, such as gestures, facial expressions, tonal variations, or physical demonstrations.
- Language is consequential and purposeful. Language utterances have a consequence in the environment, typically by influencing the mental state and thereby the behaviour of other communicators within the environment. These consequences may be optimised to achieve a variety of ends, for example, a salesperson learns to tailor their language to maximise sales, while a politician learns to tailor their language to maximise votes.
- The utility of language varies according to the agent's situation and behaviour. For example, a miner may require language regarding the stability of rocks, while a farmer may need language regarding the fertility of soil. Furthermore there may be an opportunity cost to language (e.g. discussing farming instead of doing the work of farming).
- In rich environments the potential uses of language to deal with unforeseen events may outstrip the capacity of any corpus. In these cases, it may be necessary to solve linguistic problems dynamically, through experience — for example, 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 forestall aggression.

According to our hypothesis, the ability of language in its full richness, including all of these broader abilities, arises from the pursuit of reward. It is an instance of an agent's ability to produce complex sequences of actions (e.g. uttering sentences) based on complex sequences of observations (e.g. receiving sentences) in order to influence other agents in the environment (cf. discussion of social intelligence above) and accumulate greater reward [7]. The pressure to comprehend and produce language can come from many reward-increasing benefits. If an agent can comprehend a "danger" warning then it can predict and avoid negative rewards. If an agent can generate a "fetch" command then this may cause the environment (say, containing a dog) to move an object nearer to the agent. Similarly, an agent may only eat if it can comprehend complex descriptions of the location of food, generate complex instructions for growing food, engage in complex dialogue to negotiate for food, or build long-term relationships that enhance those negotations – ultimately resulting in a variety of complex linguistic skills.

### 3.5. Reward is enough for generalisation

Generalisation is often defined as the ability to transfer the solution to one problem into the solution to another problem [37,58,61]. For example, generalisation in supervised learning [37] may focus upon transferring a solution learned from one data-set, such as photographs, to another data-set, such as paintings. Generalisation in meta-learning [61,20] has recently focused upon the problem of transferring an agent from one environment to another environment.

As per our hypothesis, generalisation may instead be understood as, and implemented by, maximising cumulative reward in a continuing stream of interaction between an agent and a single complex environment – again following a standard agent-environment protocol (see Section 2.1). Environments such as the human world demand generalisation simply because the agent encounters different aspects of the environment at different times. For example, a fruit-eating animal may encounter a new tree every day; furthermore it may become injured, suffer a drought, or face an invasive species. In each case, the animal must adapt quickly to its new state, by generalising its experience from past states. The differing states faced by the animal are not parcelled neatly into disjoint, sequential tasks with distinct labels. Instead the state depends upon the animal's behaviour; it may combine a variety of elements that overlap and recur at different time-scales; and important aspects of the state may be partially observed. Rich environments demand the ability to generalise from past states to future states – with all these associated complexities – in order to efficiently accumulate rewards.

### 3.6. Reward is enough for imitation

Imitation is an important ability associated with human and animal intelligence, which may facilitate the rapid acquisition of other abilities, such as language, knowledge, and motor skills. In artificial intelligence, imitation has often been formulated as a problem of learning from demonstration [45] through *behavioural cloning* [2], where the goal is to reproduce the actions chosen by a teacher, when supplied with explicit data regarding the teacher's actions, observations and rewards, typically under the assumption that the teacher is solving a symmetric problem to the agent. Behavioural cloning has led to several successful machine learning applications [54,62,5], especially those where human teacher data is plentiful but interactive experience is limited or costly. In contrast, the natural ability of *observational learning* [3] includes any form of learning from the observed behaviour of other humans or animals, and does not assume a symmetric teacher or require direct access to their actions, observations and rewards. This suggests that a much broader and realistic class of observational learning abilities, compared to direct imitation through behavioural cloning, may be demanded in complex environments:

- Other agents may be an integral part of the agent's environment (e.g. a baby observing its mother), without assuming the existence of a distinct data-set containing teacher data.
- The agent may need to learn an association between its own state (e.g. the pose of the baby's body), and the state of another agent (e.g. the pose of its mother), or between its own actions (e.g. rotating a robot's manipulator) and observations of another agent (e.g. seeing a human hand), potentially at higher levels of abstraction (e.g. imitating the choice of food by the mother rather than her muscle activations).
- Other agents may be partially observed (e.g. where the human hand is occluded), such that their actions or goals may only be imperfectly inferred, perhaps in hindsight.
- Other agents may demonstrate undesirable behaviours that should be avoided.
- There may be many other agents in the environment, exhibiting different skills or different levels of competence.
- Observational learning may even occur without any explicit agency (e.g. imitating the idea of bridge-building from observations of a log fallen across a stream).

We speculate that these broader abilities of observation learning could be driven by the maximisation of reward, from the perspective of a single agent that simply observes other agents as integral parts of its environment [6], potentially leading to many of the same benefits as behavioural cloning – such as the sample-efficient acquisition of knowledge – but in a much wider and more integrated context.

### 3.7. Reward is enough for general intelligence

For our last example, we turn to the ability which simultaneously poses the greatest challenge and where our hypothesis offers the greatest potential benefit. General intelligence, of the sort possessed by humans and perhaps also other animals, may be defined as the ability to flexibly achieve a variety of goals in different contexts. For example, humans can flexibly address problems (such as locomotion, transportation or communication) with solutions (such as swimming or skiing, carrying or kicking, writing or sign language) appropriate to their circumstances. General intelligence is sometimes formalised by a set of environments that measures the agent's capabilities across a variety of different goals and contexts [25,14].

According to our hypothesis, general intelligence can instead be understood as, and implemented by, maximising a singular reward in a single, complex environment. For example, natural intelligence faces a contiguous stream of experience throughout its lifetime, generated from interactions with the natural world. An animal's stream of experience is sufficiently rich and varied that it may demand a flexible ability to achieve a vast variety of subgoals (such as foraging, fighting, or fleeing), in order to succeed in maximising its overall reward (such as hunger or reproduction). Similarly, if an artificial agent's stream of experience is sufficiently rich, then singular goals (such as battery-life or survival) may implicitly require the ability to achieve an equally wide variety of subgoals, and the maximisation of reward should therefore be enough to yield an artificial general intelligence.

## 4. Reinforcement learning agents

Our main hypothesis, that intelligence and its associated abilities may be understood as subserving the maximisation of reward, is agnostic to the nature of the agent. This leaves open the important question of how to construct an agent that maximises reward. In this section, we suggest that this question may also be answered by reward maximisation. Specifically, we consider agents with a general ability to learn how to maximise reward from their ongoing experience of interacting with the environment. Such agents, which we refer to as *reinforcement learning agents*, provide several advantages.[3]

[3] It is common to use the same name to describe both problem (e.g. mountain climbing refers to the problem of ascending a peak), solution methods (e.g. the ropes and pitons used by mountain climbers), and field (e.g. the pastime of mountain climbing). Where not clear from context we refer to the reinforcement learning problem, reinforcement learning agents, and the field of reinforcement learning.

First, among all possible solution methods for maximising reward, surely the most natural approach is to learn to do so from experience, by interacting with the environment. Over time, that interactive experience provides a wealth of information about cause and effect, about the consequences of actions, and about how to accumulate reward. Rather than predetermining the agent's behaviour (placing faith in the designer's foreknowledge of the environment) it is natural instead to bestow the agent with a general ability to discover its own behaviours (placing faith in experience). More specifically, the design goal of maximising reward is implemented through an ongoing internal process of learning from experience a behaviour that maximises future reward.[4]

Reinforcement learning agents, by learning from experience, provide a general solution method that may be effective, with minimal or even zero modification, across many different reward signals and environments.

Furthermore, a single environment may be so complex, like the natural world, that it contains a heterogeneous diversity of possible experiences. The potential variations in the stream of observations and rewards faced by a long-lived agent will inevitably outstrip its capacity for preprogrammed behaviours (see Section 3.1). To achieve high reward, the agent must therefore be equipped with a general ability to fully and continually adapt its behaviour to new experiences. Indeed, reinforcement learning agents may be the only feasible solutions in such complex environments.

A sufficiently powerful and general reinforcement learning agent may ultimately give rise to intelligence and its associated abilities. In other words, if an agent can continually adjust its behaviour so as to improve its cumulative reward, then any abilities that are repeatedly demanded by its environment must ultimately be produced in the agent's behaviour. A good reinforcement learning agent could thus acquire behaviours that exhibit perception, language, social intelligence and so forth, in the course of learning to maximise reward in an environment, such as the human world, in which those abilities have ongoing value.

We do not offer any theoretical guarantee on the sample efficiency of reinforcement learning agents. Indeed, the rate at and degree to which abilities emerge will depend upon the specific environment, learning algorithm, and inductive biases; furthermore one may construct artificial environments in which learning will fail. Instead, we conjecture that powerful reinforcement learning agents, when placed in complex environments, will in practice give rise to sophisticated expressions of intelligence. If this conjecture is correct, it offers a complete pathway towards the implementation of artificial general intelligence.

Several recent examples of reinforcement learning agents, endowed with an ability to learn to maximise rewards, have given rise to broadly capable behaviours that exceeded expectations, the performance of prior agents, and in several cases, the performance of human experts. For example, when asked to maximise wins in the game of Go, AlphaZero learned (see Section 1) an integrated intelligence across many facets of Go [49]; when the same algorithm was applied to maximise outcomes in the game of chess, AlphaZero learned a different set of abilities encompassing openings, endgames, piece mobility, king safety and so forth [48,44]. Reinforcement learning agents that maximise score in Atari 2600 [30] have learned a range of abilities, including aspects of object recognition, localisation, navigation, and motor control demanded by each particular Atari game, while agents that maximise successful grips in vision-based robotic manipulation have learned sensorimotor abilities such as object singulation, regrasping, and dynamic object tracking [22]. While these examples are far narrower in scope than the environments faced by natural intelligence, they provide some practical evidence for the effectiveness of the reward maximisation principle.

One may of course wonder how to learn to maximise reward effectively in a practical agent. For example, the reward could be maximised directly (e.g. by optimising the agent's policy [57]), or indirectly, by decomposing into subgoals such as representation learning, value prediction, model-learning and planning, which may themselves be further decomposed [56]. We do not address this question further in this paper, but note that it is the central question studied throughout the field of reinforcement learning.

## 5. Related work

Intelligence has long been associated with goal-oriented behaviour [59]. This goal-oriented notion of intelligence is central to the concept of *rationality*, in which an agent selects the actions in the manner that optimally achieves its goals or maximises its utility. Rationality has been widely used to understand human behaviour [63,4], and as a basis for artificial intelligence [42], formalised with respect to an agent interacting with an environment.

It has frequently been argued that computational constraints should also be taken into account when reasoning about goals. *Bounded* [50,43,36] or *computational* rationality [27] suggests that agents should select the program that best achieves their goals, given the real-time consequences arising from that program (for example, how long the program takes to execute), and subject to limitations on the set of programs (for example, limiting the maximum size of the program). Our contribution builds upon these viewpoints, but focuses on the question of whether a single, simple reward-based goal could provide a common basis for all abilities associated with intelligence.

The standard protocol for reinforcement learning was defined by Sutton and Barto [56]; in Section 2 we presented a common generalisation with partially observed histories.

[4] The internal reward may also be distinct from, but chosen to service, the design goal [51].

Unified *cognitive architectures* [35,1] aspire towards general intelligence. They combine a variety of solution methods for separate subproblems (such as perception or motor control), but do not provide a generic objective that justifies and explains the choice of architecture, nor a singular goal towards which the individual components contribute.

The perspective of language as reward-maximisation dates back to behaviourism [52,53]; however, the reinforcement learning problem differs from behaviourism in allowing an agent to construct and use internal state. In multi-agent environments, the advantages of focusing on a single agent's objective, rather than an equilibrium objective, were discussed by Shoham and Powers [47].

## 6. Discussion

We have presented the reward-is-enough hypothesis and some of its implications. Next we briefly answer a number of questions that frequently arise in discussions of this hypothesis.

**Which environment?** One may ask which environment will give rise, through reward maximisation, to the "most intelligent" behaviour or to the "best" specific ability (for example natural language). Inevitably, the specific environmental experiences encountered by an agent — for example the friends, foes, teachers, toys, tools, or libraries encountered during the lifetime of a human brain — will shape the nature of its subsequent abilities. While this question may be of great interest for any specific application of intelligence, we have focused instead upon the arguably more profound question of which generic objective could give rise to all forms of intelligence. The maximisation of different rewards in different environments may lead to distinct, powerful forms of intelligence, each of which exhibits its own impressive, and yet incomparable, array of abilities. A good reward-maximising agent will harness any elements present in its environment but the emergence of intelligence in some form is not predicated upon their specifics. For example, the human brain develops from birth differently when exposed to different experiences in its environment, but will acquire sophisticated abilities regardless of its specific culture or education.

**Which reward signal?** The desire to manipulate the reward signal often arises from the idea that only a carefully constructed reward could possibly induce general intelligence. By contrast, we suggest the emergence of intelligence may be quite robust to the nature of the reward signal. This is because environments such as the natural world are so complex that intelligence and its associated abilities may be demanded by even a seemingly innocuous reward signal. For example, consider a signal that provides +1 reward to the agent each time a round-shaped pebble is collected. In order to maximise this reward signal effectively, an agent may need to classify pebbles, to manipulate pebbles, to navigate to pebble beaches, to store pebbles, to understand waves and tides and their effect on pebble distribution, to persuade people to help collect pebbles, to use tools and vehicles to collect greater quantities, to quarry and shape new pebbles, to discover and build new technologies for collecting pebbles, or to build a corporation that collects pebbles.

**What else, other than reward maximisation, could be enough for intelligence?** *Unsupervised learning* [19] (e.g. to identify patterns in observations) and *prediction* [18,11] (e.g. of future observations [31]) may provide effective principles for understanding experience, but do not provide a principle for action selection and therefore cannot in isolation be enough for goal-oriented intelligence. *Supervised learning* gives a mechanism for mimicking human intelligence; given enough human data one might imagine that all abilities associated with human intelligence could emerge. However, supervised learning from human data cannot be enough for a general-purpose intelligence that is able to optimise for non-human goals in non-human environments. Even where human data is plentiful, imitative intelligence may be limited in scope to the behaviours already known and exhibited by humans within the data, rather than discovering creative new behaviours that solve problems in unexpected ways (see also offline learning, below). *Evolution* by natural selection can be understood at an abstract level as maximising fitness, as measured by individual reproductive success, optimised by a population-based mechanism such as mutation and crossover. In our framework, reproductive success can be seen as one possible reward signal that has driven the emergence of natural intelligence. However, artificial intelligence may be designed with other goals than reproductive success, and may maximise the corresponding reward signals using methods other than mutation and crossover, potentially leading to very different forms of intelligence. Furthermore, while fitness maximisation may explain the initial configuration of natural intelligence (e.g. a human baby brain), a process of trial-and-error learning to maximise an intrinsic reward signal [51] (see Section 4) may furthermore explain how natural intelligence adapts through experience to develop sophisticated abilities (e.g. a human adult brain) in the service of fitness maximisation. *Maximisation of free energy* or minimisation of surprise [13] may yield several abilities of natural intelligence, but does not provide a general-purpose intelligence that can be directed towards a broad diversity of different goals in different environments. 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 environment, which themselves inform the optimisation process; furthermore optimisation occurs online in real-time while the environment continues to tick.

**Which reward maximisation problem?** Even amongst reinforcement learning researchers a number of variations of the reward maximisation problem are studied. Rather than following a standard agent-environment protocol, the interaction

loop is often modified for different cases that may include multiple agents, multiple environments, or multiple training lifetimes. Rather than maximising a generic objective defined by cumulative reward, the goal is often formulated separately for different cases: for example multi-objective learning, risk-sensitive objectives, or objectives that are specified by a human-in-the-loop. Furthermore, rather than addressing the problem of reward maximisation for general environments, special-case problems are often studied for a particular class of environments such as linear environments, deterministic environments, or stable environments. While this may be appropriate for specific applications, a solution to a specialised problem does not usually generalise; in contrast a solution to the general problem will also provide a solution for any special cases. The reinforcement learning problem may also be transformed into a probabilistic framework that approximates the goal of reward maximisation [66,39,26,17]. Finally, universal decision-making frameworks [21] provide a theoretical but incomputable formulation of intelligence across all environments; while the reinforcement learning problem provides a practical formulation of intelligence in a given environment.

**Can offline learning from a sufficiently large data-set be enough for intelligence?** Offline learning may only be enough to solve problems that are already to a large extent solved within the available data. For example, a large data-set of squirrels collecting nuts is unlikely to demonstrate all the behaviours required to build a nut harvester. Although it may be possible to generalise to the agent's current problems from solutions demonstrated in or extracted from an offline data-set, in complex environments, this generalisation will inevitably be imperfect. Furthermore, the data necessary to solve the agent's current problems will often have a negligible probability of occurring in offline data (e.g. under random behaviour or imperfect human behaviour). Online interaction allows an agent to specialise to the problems it is currently facing, to continually verify and correct the most pressing holes in its knowledge, and to find new behaviours that are very different from and achieve greater reward than those in the data set.

**Is the reward signal too impoverished?** One might wonder whether sample-efficient reinforcement learning agents necessarily exist that can maximise reward in complex, unknown environments. In answering this question, we note first that an effective agent may make use of additional experiential signals to facilitate the maximisation of future reward. Many solution methods, including model-free reinforcement learning, learn to associate future reward with features of observations, through value function approximation, which in turn provide rich secondary signals that drive learning of deeper associations through a recursive bootstrapping process [55]. Other solution methods, including model-based reinforcement learning, construct predictions of observations, or of features of observations, that facilitate the subsequent maximisation of reward through planning. Furthermore, as discussed in Section 3.6, observations of other agents within the environment may also facilitate rapid learning.

Nevertheless, it is often the challenge of sample-efficient reinforcement learning in complex environments that has led researchers to introduce assumptions or develop simpler abstractions that are more amenable to both theory and practice. However, these assumptions and abstractions may simply side-step the difficulties that must inevitably be faced by any broadly capable intelligence. We choose instead to accept the challenge head-on and focus upon its solution; we hope that other researchers will join us on our quest.

## 7. Conclusion

In this paper we have presented the hypothesis that the maximisation of total reward may be enough to understand intelligence and its associated abilities. The key idea is that rich environments typically demand a variety of abilities in the service of maximising reward. The bountiful expressions of intelligence found in nature, and presumably in the future for artificial agents, may be understood as instantiations of this same idea with different environments and different rewards. Furthermore, a singular goal of reward maximisation may give rise to deeper, broader and more integrated understanding of abilities than specialised problem formulations for each distinct ability. In particular, we explored in greater depth several abilities that may at first glance seem hard to comprehend through reward maximisation alone, such as knowledge, learning, perception, social intelligence, language, generalisation, imitation, and general intelligence, and found that reward maximisation could provide a basis for understanding each ability. Finally, we have presented a conjecture that intelligence could emerge in practice from sufficiently powerful reinforcement learning agents that learn to maximise future reward. If this conjecture is true, it provides a direct pathway towards understanding and constructing an artificial general intelligence.

## Declaration of competing interest

All authors are employees of DeepMind.

## Acknowledgements

We would like to thank the reviewers and our colleagues at DeepMind for their input into this article.

## 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] M. 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 environmental 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 International 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 Conference on Machine Learning, Omnipress, 2010, pp. 1255–1262.