Autonomous Agents and Multi-Agent Systems (2022) 36:41
https://doi.org/10.1007/s10458-022-09575-5

# Scalar reward is not enough: a response to Silver, Singh, Precup and Sutton (2021)

Peter Vamplew[1] · Benjamin J. Smith[2] · Johan Källström[3] · Gabriel Ramos[4] · Roxana Rădulescu[5] · Diederik M. Roijers[6,7] · Conor F. Hayes[8] · Fredrik Heintz[3] · Patrick Mannion[8] · Pieter J. K. Libin[6,9,10] · Richard Dazeley[11] · Cameron Foale[1]

Accepted: 2 July 2022 / Published online: 16 July 2022

## Abstract

The recent paper "Reward is Enough" by Silver, Singh, Precup and Sutton posits that the concept of reward maximisation is sufficient to underpin all intelligence, both natural and artificial, and provides a suitable basis for the creation of artificial general intelligence. We contest the underlying assumption of Silver et al. that such reward can be scalar-valued. In this paper we explain why scalar rewards are insufficient to account for some aspects of both biological and computational intelligence, and argue in favour of explicitly multi-objective models of reward maximisation. Furthermore, we contend that even if scalar reward functions can trigger intelligent behaviour in specific cases, this type of reward is insufficient for the development of human-aligned artificial general intelligence due to unacceptable risks of unsafe or unethical behaviour.

**Keywords** Scalar rewards · Vector rewards · Artificial general intelligence · Reinforcement learning · Multi-objective decision making · Multi-objective reinforcement learning · Safe and ethical AI

## 1 Introduction

Recently, Silver et al. [70] posited that the concept of reward maximisation is sufficient to underpin all intelligence. Specifically they present the *reward-is-enough* hypothesis that "Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment", and argue in favour of reward maximisation as a pathway to the creation of artificial general intelligence (AGI). While others have criticised this hypothesis and the subsequent claims [47, 58, 64, 69], here we make the argument that Silver et al. have erred in focusing on the maximisation of scalar rewards. The ability to consider multiple conflicting objectives is a critical aspect of both natural and artificial intelligence, and one which will not necessarily arise or be adequately addressed by maximising a scalar reward. In addition, even if the maximisation of a scalar

✉ Peter Vamplew
p.vamplew@federation.edu.au

Extended author information available on the last page of the article

reward is sufficient to support the emergence of AGI, we contend that this approach is undesirable as it greatly increases the likelihood of adverse outcomes resulting from the deployment of that AGI. Therefore, we advocate that a more appropriate model of intelligence should explicitly consider multiple objectives via the use of vector-valued rewards.

Our paper starts by confirming that the reward-is-enough hypothesis is indeed referring specifically to scalar rather than vector rewards (Sect. 2). In Sect. 3 we then consider limitations of scalar rewards compared to vector rewards, and review the list of intelligent abilities proposed by Silver et al. to determine which of these exhibit multi-objective characteristics. Section 4 identifies multi-objective aspects of natural intelligence (animal and human). Section 5 considers the possibility of vector rewards being internally derived by an agent in response to a global scalar reward. Section 6 reviews the relationship between scalar rewards, artificial general intelligence (AGI), and AI safety and ethics, before providing our proposal for a multi-objective approach to the development and deployment of AGI. Finally Sect. 7 summarises our arguments and provides concluding thoughts.

## 2 Does the reward-is-enough hypothesis refer to scalar rewards?

Before we argue against the use of scalar rewards, we first establish that this is in fact what Silver et al. are advocating. While the wording of the *reward-is-enough* hypothesis as quoted above does not explicitly state that the reward is scalar, this is specified in Sect. 2.4 (“A reward is a special scalar observation $R_t$”) where the authors also state that a scalar reward is suitable for representing a variety of goals or considerations which an intelligent agent may display:

> A wide variety of goals can be represented by rewards. For example, *a scalar reward signal can represent weighted combinations of objectives*, different trade-offs over time, and risk-seeking or risk-averse utilities. [70, p.4, emphasis added]

In addition, the authors later acknowledge the existence of other forms of reinforcement learning such as multi-objective RL or risk-sensitive methods, but dismiss these as being solutions to specialised cases, and contend that more general solutions (i.e. those based on maximising a cumulative scalar reward) are to be preferred. We will present an argument contesting this view in Sect. 3.

> 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. […] 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. [70, p.11]

## 3 The limitations of scalar rewards

### 3.1 Theoretical limitations of scalar rewards

The limitations of scalar rewards and the advantages of vector-based multi-objective rewards for computational agents have been extensively established in prior work [41, 60, 62]. In the interests of space and brevity, we focus here on the aspects that are of most relevance to the reward-is-enough hypothesis.

Clearly, many of the tasks faced by an intelligent decision-maker require trade-offs to be made between multiple conflicting objectives. For example a biological agent must aim to satisfy multiple drives such as reproduction, hunger, thirst, avoidance of pain, following social norms, and so on. A computational agent does not have the same physical or emotional motivations and so, when applied in the context of a highly-constrained task such as playing a board game like Go, there may be a single, clearly defined objective. However, it is likely that the agent will need to account for multiple factors in its decision making, when applied in less restricted environments. The ubiquity of multiple objectives is evident even in the cases presented by Silver et al. For example, in Sect. 3 they suggest that a squirrel's reward may be to maximise survival time, or reproductive success, or to minimise pain, while a kitchen robot may maximise healthy eating by its user, or their positive feedback, or some measure of the user's endorphin levels. An agent based on scalar rewards must either be maximising only one of these competing objectives, or some scalarised combination of them.

The prevalence of genuinely multi-objective tasks in the real-world is reflected in the thriving nature of research fields such as multi-criteria decision-making [18, 75, 80] and multi-objective optimisation [20, 27]. Furthermore, this is reflected in the broad range of application areas that involve multi-objective aspects, as identified in [19, 41], which span almost all aspects of human society. In particular, any decision which affects multiple stake-holders will require consideration of multiple objectives [21].

Silver et al. acknowledge that multiple objectives can exist, but state that "a scalar reward signal can represent weighted combinations of objectives". In the context of multiple objectives, the agent is concerned with maximising some measure of *utility* which captures the desired trade-offs between those objectives. While it is certainly true that rewards representing multiple objectives can be combined into a scalar value via a linear weighting, it is well known that this places limitations on the solutions which can be found [24, 79].[1] Furthermore, in many real world problems different objectives operate at significantly different time scales, making it practically impossible to find a meaningful trade-off among them in a single time step of learning (examples from biological intelligence are given in Sect. 4.2). This means that a scalar representation of reward may not be adequate to enable an agent to maximise its true utility [62]. In particular, some forms of utility such as lexicographic ordering cannot be represented as a scalar value [28]. In contrast, intelligence based on vector rewards and approaches that are explicitly multi-objective can directly optimise any desired measure of utility [41].

A second advantage of multi-objective reward representations is that they allow for a greater degree of flexibility in adapting to changes in goals or utility. A scalar reward representing a weighted combination of objectives directly encodes a single, fixed weighting of those objectives, and this restricts the agent to learning only about that weighting. In contrast, an agent using a multi-objective representation can follow behaviour which is optimal with respect to its current goal, while simultaneously performing learning with regard to other possible future goals or utility preferences. This allows for rapid or even immediate adaptation should the agent's goals or utility change, which we would argue is likely

[1] Specifically, there may be solutions which lie in concave regions of the Pareto front representing optimal trade-offs between objectives, and no linear weighting exists which will favour these particular solutions. For a practical example of the implications of this, see [14].

to arise in dynamic environments, particularly in the context of life-long learning. This approach, which is known in multi-objective reinforcement learning research as *multi-policy learning* [62], cannot readily be achieved by an agent observing only a scalar reward. Silver et al. themselves state that “Intelligence may be understood as a flexible ability to achieve goals", but scalar rewards are not sufficient to provide the degree of flexibility supported by multi-policy multi-objective methods.

Finally we wish to address Silver et al.’s comment that “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". We disagree with the implied assumption that maximisation of a cumulative scalar reward is the general case. Scalar rewards (where the number of rewards $n = 1$) are a subset of vector rewards (where the number of rewards $n \geq 1$). Therefore, intelligence developed to operate in the context of multiple rewards is also applicable to situations with a single scalar reward, as it can simply treat the scalar reward as a one-dimensional vector. The inverse is not true – while a vector reward can be mapped onto a scalar reward, this inevitably involves a loss of information, which will limit some capabilities of the intelligence, as discussed in the previous two paragraphs. Therefore it is clear that problems with scalar rewards are in fact the special case. A similar argument can be made regarding the generality of risk-aware decision-making compared to single-objective decision-making.

## 3.2 The multi-objective nature of intelligent abilities

Section 3 of Silver et al. identifies the following set of intelligent abilities which they assert could arise by maximising a scalar reward: Knowledge and learning, Perception, Social intelligence, Language, Generalisation, Imitation, and General intelligence.

For some abilities such as knowledge and learning, and imitation, we concur with the reward-is-enough hypothesis. While multi-objective rewards may provide benefits relating to these areas such as improving efficiency, they are not strictly required in order for an intelligent agent to develop these abilities. However, we contend that other abilities are clearly multi-objective in nature. In the following subsections we will address the benefits which multi-objective rewards may provide over scalar rewards for each of these aspects of intelligence. We believe that the issue of general intelligence merits a deeper discussion, particularly in regard to the creation of artificial general intelligence, and so we will defer discussion of this facet of intelligence until Sect. 6.

### 3.2.1 Perception

When discussing perception, Silver et al. note that there may be costs associated with carrying out actions to acquire information such as the energy and computational overheads involved in turning the head to look for a potential predator. As such, there is an implicit trade-off between the costs associated with misperception and the costs incurred in information gathering. Furthermore, there is no reason to assume that the relationship between these costs must be linear or that the relationship will remain fixed over time. For example, if the acuity of the eyesight of an organism diminishes with age, then the relationship between the time taken to obtain information visually and the accuracy of that information will change over the organism’s lifetime. This may in turn alter the optimal behaviour for that agent – whereas in its youth it may have been optimal to gather accurate information before deciding whether to fight or flee, as it ages the optimal decision may favour fleeing as the time and risk associated with more accurately assessing the potential threat increase due to its poorer visual capability. As discussed earlier, an intelligence based on vector rewards which decompose the factors such as time and accuracy of perception can more readily adapt its behaviour to changes in these factors compared to an intelligence based on a scalar reward which encodes a hard-wired trade-off between these factors.

### 3.2.2 Social intelligence

Silver et al. identify that social intelligence may arise from reward maximisation in an environment populated by multiple agents, by simply letting an agent treat other agents as part of the environment. However, learning in multi-agent systems, and the emergence of social intelligence, is more naturally expressed as a multi-objective decision-making problem. In a competitive setting, the agent should consider its main goals as well as the long-term impact an action has on the future choices of its opponents, which may require a trade-off between the two categories of objectives. In a cooperative setting, agents may have different capabilities, needs, and rewards. Therefore, a solution must be found that represents a trade-off among objectives acceptable to all agents, such that it allows for the cooperation to be established and maintained over time. Having a clear idea of the utility will help deciding whether to maintain or break partnerships when utility changes.

In multi-agent settings, the difference between vector-valued rewards – in multi-agent settings often called payoffs – and scalar-valued rewards is especially clear. For single-objective problems solutions can often be guaranteed, while in the multi-objective problem setting such guarantees cannot be obtained. This is even the case when the utility functions of the agents participating in the system are known upfront and are common knowledge. For example, while for single-criterion coalition formation games individually stable and even core stable partitions are guaranteed to exist, Igarashi et al. [44] show that even for *multi-criteria coalition formation games (MC2FGs)* with known *linear* utility functions this is not necessarily the case. Furthermore, while for single objective normal form games Nash equilibria are guaranteed to exist, this is not necessarily the case for *multi-objective normal form games (MONFGs)* [61]. These considerations demonstrate that the multi-objective multi-agent problem setting is fundamentally different from its single-objective counterpart.

### 3.2.3 Language

By supporting communication of information and coordination of actions, language is clearly a beneficial attribute for an intelligent agent operating within a multi-agent environment. Prior work has demonstrated that agents that are able to communicate can achieve mutually beneficial cooperative behaviour, which may not be possible without communication [23]. It has also been shown that reward-maximising agents can develop their own linguistic structures that exhibit advanced features such as compositionality and variability [40]. In this regard the arguments of Silver et al. are therefore correct; maximisation of a scalar reward is sufficient to give rise to the development and use of language.

However we contend that scalar rewards do not suffice to account for the full complexity of language displayed by humans. The use of language is intertwined with the role of social intelligence, and so the arguments from the previous section also apply to the development of linguistic capabilities. Harari [39] describes how the development of a social language was a principle driver of the cognitive revolution separating modern humans from earlier humans and animals. While many animals can communicate factual information, threat warnings, and even lie, there is no evidence of communicating for the purpose of higher levels of intentionality associated with human social and cultural behaviours [17, 30, 37].

Language plays multiple roles in human interactions; in addition to communication of factual information, it can also serve to strengthen relationships, display emotions or personality, persuade or mislead others, etc. As such the use of language by an intelligent agent may be driven by a variety of different conflicting factors – for example, an agent may wish to persuade another agent to carry out a particular action, while still preserving the long-term trust relationship with them. We argue that such sophisticated use of language can only emerge where there is a desire to go beyond simply achieving a particular reward and it is, therefore, far more likely to derive from a multi-objective approach to decision-making.

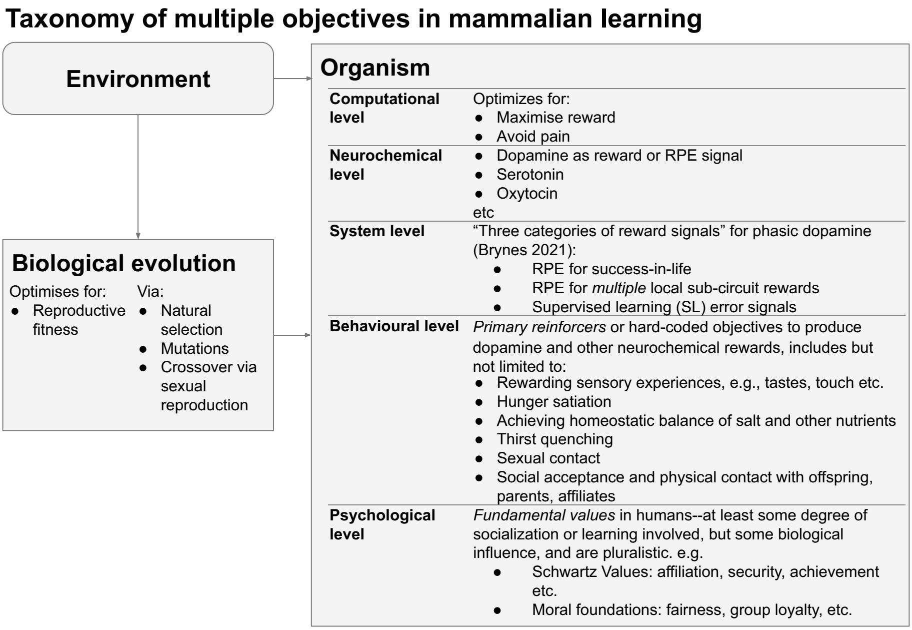
### 3.2.4 Generalisation

Silver et al. define generalisation as the ability required of an agent that is maximising its cumulative reward within ongoing interaction with a single complex environment. The agent will inevitably encounter situations which deviate from its past experience, and so to behave effectively in those novel situations it must be able to appropriately generalise from prior experience.

An agent maximising a scalar reward may exhibit some aspects of generalisation, such as generalising across variations in state. However other aspects of generalisation, such as generalising to new tasks or to changes in user preferences, will be problematic. As the agent has only observed its current reward signal, it has no basis for adapting its behaviour should the reward signal undergo a significant change. In contrast, as discussed earlier, a multi-policy multi-objective agent can learn with regard to all components of the vector reward they are receiving, regardless of the extent to which each component contributes to their current utility (it could even learn about factors which do not contribute at all to its current utility, but which may be beneficial to know about in the future [46]). Should the task or the user's preferences change, the agent can almost immediately adapt its behaviour [3, 78].[2]

For example, consider an autonomous vehicle which has learned a policy based on a scalar reward carefully handcrafted to address various factors such as travel time, passenger comfort, safety and fuel consumption. Now imagine that the cost of fuel rises considerably so that the current policy is too expensive. The reward signal will need to be redefined to place more emphasis on conserving fuel, and the agent controlling the vehicle will need to relearn an optimal policy based on this new reward. In contrast a multi-objective agent treats these factors as separate elements of a vector reward, and can use multi-policy learning to identify in advance various policies which are optimal for different preferences over those factors. When fuel costs change a multi-objective agent can simply update its utility function to place a greater weight on fuel consumption, and can immediately follow a

[2] We note similarities in this aspect of generalisation across tasks between multi-objective reinforcement learning (MORL) and other approaches such as multi-task reinforcement learning [66] or successor features [8]. However, as discussed in Sect. 6.3 of [41], there are also differences between these approaches, with MORL representing a more general class of methods. The relationship between MORL and successor features has recently been explored further in [4].



Taxonomy of multiple objectives in mammalian learning

Environment

Biological evolution

Optimises for:
- Reproductive fitness

Via:
- Natural selection
- Mutations
- Crossover via sexual reproduction

Organism

| | |
|---|---|
| **Computational level** | Optimizes for:<br>● Maximise reward<br>● Avoid pain |
| **Neurochemical level** | ● Dopamine as reward or RPE signal<br>● Serotonin<br>● Oxytocin<br>etc |
| **System level** | “Three categories of reward signals” for phasic dopamine (Brynes 2021):<br>● RPE for success-in-life<br>● RPE for *multiple* local sub-circuit rewards<br>● Supervised learning (SL) error signals |
| **Behavioural level** | *Primary reinforcers* or hard-coded objectives to produce dopamine and other neurochemical rewards, includes but not limited to:<br>● Rewarding sensory experiences, e.g., tastes, touch etc.<br>● Hunger satiation<br>● Achieving homeostatic balance of salt and other nutrients<br>● Thirst quenching<br>● Sexual contact<br>● Social acceptance and physical contact with offspring, parents, affiliates |
| **Psychological level** | *Fundamental values* in humans--at least some degree of socialization or learning involved, but some biological influence, and are pluralistic. e.g.<br>● Schwartz Values: affiliation, security, achievement etc.<br>● Moral foundations: fairness, group loyalty, etc. |

**Fig. 1** Biological evolution generates organism genotypes, optimising for the fit of an organism’s phenotype to its environment. Genotypes are the genetic instruction set that are decoded into into an organism that exhibits a set of phenotypical expressions. These include all of the learning and reward systems that make up an organism’s biological intelligence. Further details about each level of objectives in the organism are described in the text

policy which is optimal with respect to the changed conditions without the need for any further learning.

## 4 Multi-objective reinforcement learning in natural intelligences

If our arguments in favour of multi-objective representations of reward are correct, then it would be expected that naturally evolved intelligences such as those in humans and animals would exhibit evidence of vector-valued rewards. In fact, evolution has developed organisms that delegate learning not just into multiple objectives but even into multiple learning systems within an organism. There are multiple objectives at a basic biological regulatory level, and these are matched with multiple objectives at every level of analysis of the organism. In this section we show that these cannot be reduced to any single objective.

### 4.1 Distinguishing biological evolution and biological intelligence

When considering natural intelligences it is important to draw a clear distinction between the evolutionary process which has produced these intelligences, and the functioning of the intelligences themselves. Biological evolution optimises for a single objective (i.e. reproductive fitness), to an environment that varies over time and space (though its operationalisation depends greatly on the environment [42, 70]). In contrast, at at least five distinct levels of analysis for human organisms, an organism's computational processes include multiple objectives that must be balanced during action selection (see Fig. 1).

### 4.2 Biological intelligence has no single primary objective

In optimising for reproductive fitness, evolution generates an organism's genotype (Fig. 1), which in turn creates what we could call an *intellectual phenotype*, a set of innate intellectual capacities that an organism can use to learn about and interact with its environment. In mammals, as in most other organisms, the bio-computational processes that constitute that intellectual phenotype have no single objective; rather, they include multiple objectives including hunger satiation [29], thirst quenching, social bonding [59], and sexual contact [34], as described in Fig. 1.

Even if reproduction is regarded as a 'single objective' of the evolutionary process, broadly construed, at the organism level it is not a primary reinforcer at all. Rather, the organism's phenotype includes a set of features, including intelligent systems, which have been tuned by the evolutionary process because they tend to lead to environmental fitness and ultimately genetic reproduction. This includes sexual orgasm [34], social contact with conspecifics like offspring [59], and so on. These do not necessarily internally store any kind of explicit representation of reproduction as a goal. Rather reproduction is an emergent result of the organism fulfilling other multiple primary objectives. Understanding the multi-objective nature of motivation in biological systems is critical because different systems are more dominant at different times, depending on context. An example in this regard concerns the remediation of pain and hunger. On the one hand, pain warrants an immediate response, while on the other hand, for hunger this response can be delayed for hours up to days. These examples demonstrate that such objectives could not be aggregated in a single reward signal, due to the difference in time scale in which these rewards are relevant, and we note that distinct biological subsystems are in place to support these differences. Simply understanding the computational processes of the organism is insufficient to predict behaviour without also including an account of the state of the organism and its environment, and the relevant objectives that arise as a function of that state.

The clearest example of biological primary objectives might be what are called 'primary reinforcers' in behavioural psychology. These constitute behavioural goals that are innate [29], such as hunger satiation, thirst quenching, and sexual satisfaction (Fig. 1, 'behavioural level'). These are themselves entire families of objectives, because an organism needs a wide array of nutrients to survive and can pursue very specific objectives to ensure that each nutrient is obtained. Even for specific primary reinforcers, there may be multiple biological signals acting as proxy objectives to ensure those objectives are obtained. For instance, salt consumption alone activates taste receptors, interoceptive signals related to ingestion, and blood osmolality detectors designed to maintain homeostatic balance [72]. These receptors drive the value assigned to consuming salt during decision-making at any particular moment. Thus, even for just a single critical nutrient, there are multiple biological objectives [55, 82] on multiple level tuned to ensure an appropriate level of consumption.

At a psychological level (Fig. 1), humans appear to hold multiple irreducible and irreconcilable moral [38] and life [68] objectives. Moral objectives include preferences for equality, for harm avoidance, and upholding authority hierarchies and group loyalties [38] and may have a basis in distinct biological tendencies. Schwartz et al. [68] identified ten distinct human values, including benevolence, security, hedonism, and autonomy that appear to be distinct life objectives for people across cultures to varying degrees.

### 4.3 There's no plausible neurochemical or neuroanatomical single objective function

Proponents of a single-objective account of RL in biological organisms might look for a single neurotransmitter release that could conceivably underlie all the objectives outlined above – perhaps a global reward signal or brain region that combines all the objectives outlined in the previous section into a single process. However, even if such a mechanism was identified as, for instance, release of dopamine to indicate reward prediction error, the reward-is-enough hypothesis would not follow: that single signal is properly thought of as *hard-coded* to achieve multiple objectives, as outlined in Fig. 1.

Dopamine is often associated with reward, but it is more appropriate to characterize it as a 'reward prediction error' (RPE) signal than as a simple signal of positive reinforcement [54] (Fig. 1, 'Neurochemical level'). The reward system releases dopamine as a response to a signal indicating a reward is coming [67], but on release of a reward, dopamine is only delivered if the reward was unexpected. It might be said only a single neurotransmitter performs the RPE function in mammalian reward learning, but this is not the same as saying mammalian RL is single-objective, because the RPE signal is delivered to drive learning across a wide variety of domains. For instance, not only is dopamine responsible for modulating goal-directed behavior, but it also seems to play a role in learning within sensory systems including visual, auditory, olfactory, and taste cortices [53]. Additionally, other neurochemicals like oxytocin [52] or serotonin [81] are important in the experience of pleasure, attachment, and motivation, and although these may ultimately depend on the dopaminergic system for motivational power, their presence demonstrates the neurochemical hardcoding of a variety of objectives in animal behavior [52, 67, 81].

A number of neuroanatomical brain regions are important for the production and assessment of value calculation and reward (Fig. 1, 'Computational level'). A full survey of biological decision-making is outside of the scope of this article. But in brief, subcortical regions like the nucleus accumbens are thought to be involved in reward processing [45], while the ventromedial prefrontal cortex (vmPFC) appears to formulate value signals associated with potential rewards. However, these do not appear to function as parts of single-objective reinforcement learning systems because the values represented are context dependent [65]. When an organism is hungry, the value of food, as recorded in the vmPFC, is higher [74]; when an organism is thirsty, the value of drinking is higher. Something like a 'common currency' might indeed be represented in the vmPFC [51] and related regions, but to extend the metaphor, the 'exchange' rate between that currency and various physiological and psychological goals and drives changes currently based on context, limiting the applicability of any single-objective account.

### 4.4 The brain has multiple objective functions at a systemic level

Byrnes [15] attempts a description of the brain as an integrated reward learning system. The model described draws on Mollick et al.'s [54] description of the brain's phasic dopamine system as well as classic work describing the brain as an array of parallel loops [5]. Here, three separate categories of phasic dopamine signals are described: reward prediction error for basic universal goals, reward prediction error for motor action execution, and supervised learning error signals. These are distinct processes required for human intelligence, all with their own objectives, but they are all required for a human brain (or a mammalian brain more generally) to function correctly.

## 5 Internally-derived rewards

One could argue that an agent concerned with maximising a scalar reward may still develop the capabilities required to carry out multi-objective decision-making. Natural intelligence provides an example of this. As we discussed in Sect. 4, the evolutionary objective of reproduction has led to the development of organisms with specialised sensors, internally derived reward signals, and learning systems associated with those signals. Another example of this is the perception of fairness and inequality, which has been identified as a process embedded in the human brain [16]. Conceivably this could also arise in the context of computational intelligence, where agents based on evolutionary algorithms or reinforcement learning might construct their own internal reward signals to guide their learning and decision-making [32, 71, 76].

One benefit of internalised rewards is that they provide a less sparse reward signal. If the agent learns to identify events which are correlated with future occurrences of its external reward, then creating secondary reward signals for those events will provide more immediate feedback, thereby speeding up learning and adaptation. For example, developing taste sensors which respond to particular nutrients in food will provide immediate rewards to an animal which eats that food, correlating to the more delayed benefits which may accrue from the intake of those nutrients. Similarly, a team of autonomous rovers aiming to discover signs of life on another planet will find this task exceedingly difficult to learn if they are only provided with a single reward at the end of the task, but can learn far more effectively if provided with short-term rewards for activities which are correlated with the long-term goal [83]. It has been shown that multi-objectivisation or reward decomposition (where the primary scalar reward is decomposed into several distinct rewards) can be beneficial for a computational RL agent, particularly where the primary reward is sparse [13, 49, 83]. Similarly, agents may use internally derived rewards to drive aspects of the learning process itself such as exploration [9, 57].

Regardless of whether vector rewards are derived externally or internally, the agent still needs to make decisions based on those vector values. Silver et al. argue that an agent maximising a scalar reward could theoretically develop multi-objective capabilities. However this would require the agent to modify its own internal structure. Therefore, we believe that it is more practical to construct multi-objective agents through the explicit design of multi-objective algorithms. Similarly, we argue that where we can design multi-objective reward structures for computational agents, it makes sense to do so rather than relying on them to identify such structures themselves. In fact, we contend that it typically will be easier to specify multi-objective rewards directly than to design a scalar reward which captures all of the various factors of interest. We note that the reward-is-enough hypothesis is theoretically focused, and so these practical considerations do not constitute an argument against the hypothesis. Nevertheless we feel it is important to highlight these issues as they clearly impact on the pathway to the development of more powerful computational intelligence, as will be discussed further in the following section.

# 6 Reward maximisation and general intelligence

## 6.1 The risks of single-objective general intelligence

One of the main arguments presented by Silver et al. is that the maximisation of even a simple reward in the context of a suitably complex environment (such as those which occur in the natural world) may suffice for the emergence of general intelligence. They illustrate this via the following scenario:

> 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. [70, p.10]

While the development of open-ended, far-reaching intelligence from such a simple reward is presented positively by Silver et al., this scenario is strikingly similar to the infamous *paperclip maximiser* thought experiment from the AI safety literature [10]. In this example, a superintelligent agent charged with maximising the production of paperclips enslaves humanity and consumes all available resources in order to achieve this aim. While unrestricted maximisation of a single reward may indeed result in the development of complex, intelligent behaviour, it is also inherently dangerous [56]. For this reason, AI safety researchers have argued in favour of approaches based on satisficing rather than unbounded maximisation [73], or on multi-objective measures of utility which account for factors such as safety or ethics [77].

Even when safety is not at risk, it is vital to carefully consider which types of behaviour can arise when deploying learning agents in society. Depending on the application, elements such as bluffing or information hiding can potentially be harmful and undesirable. For example, consider a smart grid setting, in which autonomous agents decide on behalf of the homeowners how to best store or trade locally generated green energy. This is, in general, a competitive scenario. While the main goal of each agent is to ensure the comfort and reduce the costs of the household, it is not acceptable for agents to manipulate the market or submit bluff-bids, even if these behaviours would be optimal with respect to their reward signal. Since autonomous agents need to operate in the context of our society, we should give careful thought to and investigate what type of emergent behaviour is desirable. This will be important to avoid the development of selfish and harmful mechanisms, under the pretext of being optimal with respect to a single numerical feedback signal.

Therefore we argue that even if scalar rewards are enough for the development of general intelligence, they are not sufficient for the far more important task of creating human-aligned AGI. While safety and ethics are not the focus of Silver et al.'s paper, it is concerning that these issues are not acknowledged in a paper which is actively calling for the development of AGI. More broadly we note that Silver et al. do not provide guidance as to the actual objective or reward signal which may be suitable for the creation of human-aligned AGI.

Even if we were to accept the hypothesis that an intelligent agent's behaviour can arise from maximising a single scalar reward within the environment, that does not necessarily imply that the design and construction of such a scalar reward is feasible. In this sense, the pebble-collecting scenario is an outlier as the agent has access to a straightforward, dense source of reward ("+1 reward... each time a round-shaped pebble is collected"). In practice the sort of problems for which general intelligence is required will rarely lead to such simple and immediate feedback, and the design of a suitable reward will be a major challenge. In addition analysing the behaviour of a general intelligence in terms of a single reward is unlikely to provide sufficient insight into the drivers of that behaviour. For example, interpreting the behaviour of a squirrel as maximising a single scalar reward representing "survival" does not help to understand the mechanisms of that reward, nor does it assist in the construction of an equivalent reward signal that induces squirrel-like behaviour in an arbitrary intelligent agent.

As described by Amodei et al. [7], reward specification is difficult even in trivial systems, and reward misspecification and reward hacking often lead to surprising, unintended, and undesirable behaviour. In more complex systems with more general agents, the potential for reward misspecification is significantly increased [31]. We argue, then, that the application of a single scalar reward signal leads to significant risk of unpredictable and undesirable behaviour. Given the limitations of their human designers, scalar rewards will most likely not be enough for the development of AGI with guaranteed behavioural properties, and predictable reward design is better achieved using multi-objective methods.

### 6.2 A multi-objective approach to human-aligned general intelligence

In order for general artificial intelligence to be beneficial to humanity, it will need to be accountable and adaptable to human ethics, as well as human needs and aims. To be a part of society, an agent needs to adapt, and be accountable to others in this society, and we argue that agents that optimise for a single objective are severely handicapped in doing so. In this section, we will first explain the need for multiple objectives from an ethical and human-alignment point of view, and second explain our vision for future AI systems that form an integral part of society; agents that can be reviewed with respect to, and adapted to, a changing society.

Ultimately, all ethical capabilities that AGI can attain will have to come from humans, as well as all other goals and aims. There simply is no other available source for ethics and general goals than humans. However we do not agree amongst ourselves what ethically optimal behaviour is. Philosophers – arguably the closest there is to generalist experts in ethics – disagree on a wide range of ethical questions including meta-ethics, moral motivation, and normative ethics [12], and their positions on these disagreements correlate systematically with personal identities including race and gender. Psychologically, moral intuitions seem to arise from a pluralistic set of incommensurable and innate moral values that differ systematically across different political parties [36]. Fundamental human psychological values differ individually and across cultures [68]. Furthermore, even if it is not a question of ethics, the question of what to care about is not trivial either [35]. Therefore, we cannot expect people who need to specify the rewards for an AI system to get it right, especially not in one go. Moreover, priorities are likely to change over time. Any AI system – general or not – that is deployed for a long time, and that is possibly propagated to new generations of the systems, must be able to deal with this.

Humans who will have to specify what the agent must care about will likely identify multiple goals that are important. For example, a self-driving car will need to minimise travel time, minimise energy consumption, as well as minimise the expected harm resulting from a trip to both the occupants of the car, other humans in the environment, animals, and property. Furthermore, a reasonable trade-off between these objectives (which may well be a non-linear one, as we explained in Sect. 3.1) must be identified during training. However, the engineers of such a system are most likely not ethicists nor legal experts. Furthermore, they do not own the utility of the system's deployment, and may not ultimately bear the responsibility for the system being deployed in practice. Therefore, the human-alignment process will necessarily be in the hands of others, who need a clear explanation from the AI system about the available trade-offs between objectives in different situations. We therefore believe that an explicitly multi-objective approach is necessary for responsible agent design. For further arguments why human-aligned AI is inherently multi-objective, please refer to [77].

Once the design, training, and testing is completed, and the AI system is deployed, it will be equipped with a set of objectives, and a mechanism (e.g., a utility function [63]) for making online decisions about trade-offs between these objectives. However, it may well encounter situations that were not foreseen in training, and the trade-offs between objectives in these situations become unlike hitherto encountered trade-offs. Here we encounter another key benefit of an explicitly multi-objective approach – and in our opinion an ethically necessary capability, of any agent. In combination with probabilistic methods, the agent can identify when it becomes too uncertain about which trade-offs will be desired and therefore, if possible, defer the decision back to responsible humans, or shut down safely. When this is not possible, these situations can be identified and reported back. For example, if a parrot suddenly gets loose in an automated factory, and the AI system correctly identifies it as an animal, but has never had to make choices between animal safety, human safety, product damage, and production before, the system may try to opt for safe system shutdown if a responsible human cannot be reached. However, if the parrot flies straight into the production line, more drastic immediate action might be required, and there is no time to shut down safely. These immediate actions will necessarily be taken on the fly, and will have to be reported and later reviewed to see whether the taken actions were indeed what the responsible humans would have wanted the system to do.

Finally, when undesirable things have occurred, an agent needs to be able to explain the decision made. Single objective systems are only able to provide simple details such as what was the perceived state and that its chosen action maximised its reward [22]. Such explanations provide little understanding to the users. However, an explicitly multi-objective approach confers significant further benefits. Namely, it can help diagnose exactly what went wrong, such as: what was the trade-off between objectives; what alternative outcomes would have occurred with alternative trade-offs; or, was the selected policy providing an undesired trade-off between objectives [26]. Hence, a multi-objective approach allows explicitly attributed details, as well as contrastive and counterfactual explanations of the reasoning process behind behaviour rather than just the outcome [25]. Being able to explicitly identify the trade-off underlying an agent's basis for reasoning has long been argued as a key component of transparency in AGI [43].

One possible implementation of these concepts would be a review-and-adjust cycle [41]. Figure 2 demonstrates how a review-and-adjust cycle could be applied when developing an agent for a multi-objective Markov decision process (MOMDP). During the planning or learning phase, an AGI would utilise a multi-objective algorithm to compute a set of optimal policies for all possible utility functions. In the selection phase, a policy is selected to be executed, possibly with direct or indirect user feedback. The selected policy is then executed, during the execution phase. The outcome from the policy execution can then be reviewed by the overseer (either a human, the AGI itself, or another AGI), along with the

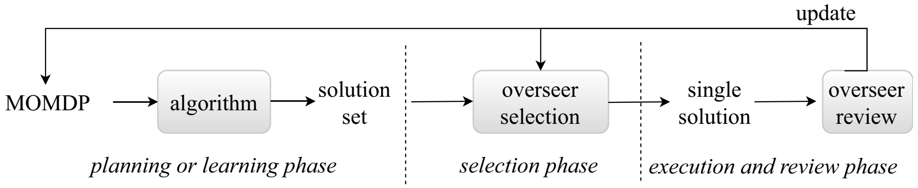


**Fig. 2** Our proposal for a responsible review-and-adjust scheme for future AI. During the learning/planning phase the agent identifies multiple policies which would be optimal under different utility functions. One of these policies is then selected and executed, and a subsequent review of the outcomes may lead to an adjustment in overseer selection (without a need to remodel or retrain), or other changes such as the introduction of new objectives

AGI's explanation of its policy selection. The MOMDP, utility function or set of solutions can then be updated based on this review. We note that such reviews can not only be triggered by incidents, but also by regular inspection.

We see such a review-and-adjust cycle as an essential feature of future AI. We, as AI researchers have to enable responsible deployment, and well-informed review of the systems we create is a key feature of this. It is our opinion that the above-mentioned benefits are not merely desirable, but that it is a moral imperative for AI designers to obtain them, in order to create AI systems that more likely benefit society. In addition to the mathematical, technical, and biological arguments for why scalar reward is not enough, we thus also point out that there are ethical and societal reasons why scalar rewards are not enough.

We acknowledge the difficulties which may arise in implementing human oversight of AGI if the latter achieves superhuman levels of intelligence [6, 11, 33]. Superintelligent AGI may be motivated to, and highly capable of, deceiving human overseers, or its behaviours and reasoning may simply be too complex for human understanding. Nevertheless we would argue that it is certainly preferable to attempt such oversight than not, and that a multi-objective AGI will provide greater transparency than a single-objective AGI. Nevertheless, we acknowledge that such an approach will not necessarily guarantee a manageable superhuman AGI, and therefore a careful ethical consideration of such research efforts is warranted.

## 7 Conclusion

Silver et al. argue that maximisation of a scalar reward signal is a sufficient basis to explain all observed properties of natural intelligence, and to support the construction of artificial general intelligence. However, this approach requires representing all of the different objectives of an intelligence as a single scalar value. As outlined in Sect. 3, this places restrictions on the behaviour which can emerge from maximisation of this reward. Therefore, we contend that the *reward-is-enough* hypothesis does not provide a sufficient basis for understanding all aspects of naturally occurring intelligence, nor for the creation of computational agents with broad capabilities.

In the context of the creation of AGI, a focus on maximising scalar rewards creates an unacceptable exposure to risks of unsafe or unethical behaviour by the AGI agents. This is particularly concerning given that Silver et al. are highly influential researchers and employed at DeepMind, one of the organisations best equipped to expand the frontiers of

AGI. While Silver et al. “hope that other researchers will join us on our quest", we instead hope that the creation of AGI based on reward maximisation is tempered by other researchers with an understanding of the issues of AI safety [48, 50] and an appreciation of the benefits of multi-objective agents [1, 2].

**Acknowledgements** This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program, and by the National Cancer Institute of the U.S. National Institutes of Health under Award Number 1R01CA240452-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or of other funders. Pieter J.K. Libin acknowledges support from the Research Foundation Flanders (FWO, fwo.be) (postdoctoral fellowship 1242021N). Johan Källström and Fredrik Heintz were partially supported by the Swedish Governmental Agency for Innovation Systems (Grant NFFP7/2017-04885), and the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Conor F. Hayes is funded by the National University of Ireland Galway Hardiman Scholarship. Gabriel Ramos was partially supported by FAPERGS (Grant 19/2551-0001277-2) and FAPESP (Grant 2020/05165-1).

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

## References

1. Abdolmaleki, A., Huang, S., Hasenclever, L., Neunert, M., Song, F., Zambelli, M., Martins, M., Heess, N., Hadsell, R., & Riedmiller, M. (2020). A distributional view on multi-objective policy optimization. In *International Conference on Machine Learning* (pp. 11–22). PMLR.
2. Abdolmaleki, A., Huang, S. H., Vezzani, G., Shahriari, B., Springenberg, J. T., Mishra, S., TB, D., Byravan, A., Bousmalis, K., Gyorgy, A., et al. (2021). On multi-objective policy optimization as a tool for reinforcement learning. arXiv preprint arXiv:2106.08199.
3. Abels, A., Roijers, D., Lenaerts, T., Nowé, A., & Steckelmacher, D. (2019). Dynamic weights in multi-objective deep reinforcement learning. In *International Conference on Machine Learning* (pp. 11–20). PMLR.
4. Alegre, L. N., Bazzan, A. L., & da Silva, B. C. (2022). Optimistic linear support and successor features as a basis for optimal policy transfer. arXiv preprint arXiv:2206.11326.
5. Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience, 9*(1), 357–381.
6. Alfonseca, M., Cebrian, M., Anta, A. F., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research, 70,* 65–76.
7. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. arXiv preprint arXiv:1606.06565. https://arxiv.org/pdf/1606.06565.pdf.
8. Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., & Silver, D. (2017). Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems* (pp. 4055–4065).
9. Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems* (pp. 17–47). Springer.
10. Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, pp. 12–17.
11. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies.

12. Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies, 170*(3), 465–500.
13. Brys, T., Harutyunyan, A., Vrancx, P., Nowé, A., & Taylor, M. E. (2017). Multi-objectivization and ensembles of shapings in reinforcement learning. *Neurocomputing, 263,* 48–59.
14. Brys, T., Van Moffaert, K., Van Vaerenbergh, K., & Nowé, A. (2013). On the behaviour of scalarization methods for the engagement of a wet clutch. In *2013 12th International Conference on Machine Learning and Applications* (Vol. 1, pp. 258–263). IEEE.
15. Byrnes, S. (2021). Big picture of phasic dopamine. Alignment Forum. https://www.alignmentforum.org/posts/jrewt3rLFiKWrKuyZ/big-picture-of-phasic-dopamine.
16. Cappelen, A. W., Eichele, T., Hugdahl, K., Specht, K., Sørensen, E. Ø., & Tungodden, B. (2014). Equity theory and fair inequality: A neuroeconomic study. *Proceedings of the National Academy of Sciences, 111*(43), 15368–15372. https://doi.org/10.1073/pnas.1414602111.
17. Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See The World: Inside the mind of another species*. University of Chicago Press.
18. Clemen, R. T. (1996). *Making hard decisions: an introduction to decision analysis*. Brooks/Cole Publishing Company.
19. Coello, C. A. C., & Lamont, G. B. (2004). *Applications of multi-objective evolutionary algorithms* (Vol. 1). World Scientific.
20. Coello, C. A. C., Lamont, G. B., Van Veldhuizen, D. A., et al. (2007). *Evolutionary algorithms for solving multi-objective problems* (Vol. 5). Springer.
21. Coyle, D., & Weller, A. (2020). "Explaining'' machine learning reveals policy challenges. *Science, 368*(6498), 1433–1434.
22. Cruz, F., Dazeley, R., & Vamplew, P. (2019). Memory-based explainable reinforcement learning. In *Australasian joint conference on artificial intelligence* (pp. 66–77). Springer.
23. Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). Tarmac: Targeted multi-agent communication. In: International Conference on machine learning (pp. 1538–1546). PMLR.
24. Das, I., & Dennis, J. E. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural optimization, 14*(1), 63–69.
25. Dazeley, R., Vamplew, P., & Cruz, F. (2021). Explainable reinforcement learning for broad-xai: a conceptual framework and survey. arXiv preprint arXiv:2108.09003.
26. Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299,* 103525.
27. Deb, K. (2014). Multi-objective optimization. In *Search methodologies* (pp. 403–449). Springer.
28. Debreu, G. (1997) On the preferences characterization of additively separable utility. In *Constructing Scalar-Valued Objective Functions* (pp. 25–38). Springer.
29. Delgado, M., & Rigney, A. (2009). Reward systems: Human. *Encyclopedia of Neuroscience, 8,* 345–352.
30. Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "panglossian paradigm'' defended. *Behavioral and Brain Sciences, 6*(3), 343–355.
31. Dewey, D. (2014). Reinforcement learning and the reward engineering principle. In *2014 AAAI Spring Symposium Series*.
32. Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2008). Co-evolution of shaping rewards and meta-parameters in reinforcement learning. *Adaptive Behavior, 16*(6), 400–412.
33. Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. arXiv preprint arXiv:1805.01109.
34. Fleischman, D. S. (2016). An evolutionary behaviorist perspective on orgasm. *Socioaffective neuroscience and psychology, 6*(1), 32130.
35. Frankfurt, H. (1982). The importance of what we care about. *Synthese*, pp. 257–272.
36. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Elsevier.
37. Griffin, D. R. (1976). *The Question Of Animal Awareness: Evolutionary Continuity Of Mental Experience*. Rockefeller University Press.
38. Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.
39. Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. Random House.
40. Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *31st Conference on Neural Information Processing Systems*.

41. Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A.A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., Roijers, D.M.: A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems **36** (2022)
42. Henrich, J. (2015). *The secret of our success*. Princeton University Press.
43. Hibbard, B. (2008). Open source AI. *Frontiers in Artificial Intelligence and Applications, 171,* 473.
44. Igarashi, A., & Roijers, D. M. (2017). Multi-criteria coalition formation games. In *International Conference on Algorithmic DecisionTheory* (pp. 197–213). Springer.
45. Ikemoto, S., & Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Research Reviews, 31*(1), 6–41.
46. Karimpanal, T. G., & Wilhelm, E. (2017). Identification and off-policy learning of multiple objectives using adaptive clustering. *Neurocomputing, 263,* 39–47.
47. Kilcher, Y. (2021). Reward is enough (machine learning research paper explained). https://www.youtube.com/watch?v=dmH1ZpcROMk &t=24s.
48. Krakovna, V., Orseau, L., Ngo, R., Martic, M., & Legg, S. (2020). Avoiding side effects by considering future tasks. arXiv preprint arXiv:2010.07877.
49. Kurniawan, B. (2021). Single- and multiobjective reinforcement learning in dynamic adversarial games. Ph.D. thesis, Federation University Australia.
50. Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). AI safety gridworlds. arXiv preprint arXiv:1711.09883.
51. Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027–1038.
52. Love, T. M. (2014). Oxytocin, motivation and the role of dopamine. *Pharmacology, Biochemistry and Behavior, 119,* 49–60.
53. Macedo-Lima, M., & Remage-Healey, L. (2021). Dopamine modulation of motor and sensory cortical plasticity among vertebrates. *Integrative and Comparative Biology, 61*(1), 316–336.
54. Mollick, J. A., Hazy, T. E., Krueger, K. A., Nair, A., Mackie, P., Herd, S. A., & O'Reilly, R. C. (2020). A systems-neuroscience model of phasic dopamine. *Psychological Review, 127*(6), 972.
55. Oka, Y., Butnaru, M., von Buchholtz, L., Ryba, N. J., & Zuker, C. S. (2013). High salt recruits aversive taste pathways. *Nature, 494*(7438), 472–475.
56. Omohundro, S. M. (2008). The basic AI drives. In *AGI* (Vol. 171, pp. 483–492).
57. Oudeyer, P. Y., & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics, 1,* 6.
58. Ouellette, S. (2021). Reward is enough – but not efficient. https://www.linkedin.com/pulse/reward-enough-efficient-simon-ouellette/.
59. Perret, A., Henry, L., Coulon, M., Caudal, J. P., Richard, J. P., Cousillas, H., et al. (2015). Social visual contact, a primary "drive'' for social animals? *Animal Cognition, 18*(3), 657–666.
60. Rădulescu, R., Mannion, P., Roijers, D. M., & Nowé, A. (2020). Multi-objective multi-agent decision making: A utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems, 34*(1), 1–52.
61. Rădulescu, R., Mannion, P., Zhang, Y., Roijers, D. M., & Nowé, A. (2020). A utility-based analysis of equilibria in multi-objective normal-form games. *The Knowledge Engineering Review,35*.
62. Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research, 48,* 67–113.
63. Roijers, D. M., & Whiteson, S. (2017). Multi-objective decision making. *Synthesis lectures on artificial intelligence and machine learning, 11*(1), 1–129.
64. Roitblat, H. (2021). Building artificial intelligence: Reward is not enough. https://bdtechtalks.com/2021/07/07/ai-reward-is-not-enough-herbert-roitblat/.
65. Rudorf, S., & Hare, T. A. (2014). Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. *Journal of Neuroscience, 34*(48), 15988–15996.
66. Schaul, T., Horgan, D., Gregor, K., & Silver, D. (2015). Universal value function approximators. In *International conference on machine learning* (pp. 1312–1320).
67. Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.
68. Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38*(3), 230–255. https://doi.org/10.1016/S0092-6566(03)00069-2.

69. Shead, S. (2021). Computer scientists are questioning whether Alphabet's DeepMind will ever make A.I. more human-like. https://www.cnbc.com/2021/06/18/computer-scientists-ask-if-deepmind-can-ever-make-ai-human-like.html.
70. Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, pp. 103535.
71. Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development, 2*(2), 70–82.
72. Smith, B. J., & Read, S. J. (forthcoming). Modeling incentive salience in pavlovian learning more parsimoniously using a multiple attribute model. Cognitive Affective Behavioral Neuroscience.
73. Taylor, J. (2016). Quantilizers: A safer alternative to maximizers for limited optimization. In: AAAI Workshop: AI, Ethics, and Society.
74. Thomas, J. M., Higgs, S., Dourish, C. T., Hansen, P. C., Harmer, C. J., & McCabe, C. (2015). Satiation attenuates bold activity in brain regions involved in reward and increases activity in dorsolateral prefrontal cortex: an fmri study in healthy volunteers. *The American Journal of Clinical Nutrition, 101*(4), 697–704.
75. Triantaphyllou, E. (2000). Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study* (pp. 5–21). Springer.
76. Uchibe, E., & Doya, K. (2008). Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. *Neural Networks, 21*(10), 1447–1455.
77. Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology, 20*(1), 27–40.
78. Vamplew, P., Issabekov, R., Dazeley, R., Foale, C., Berry, A., Moore, T., & Creighton, D. (2017). Steering approaches to pareto-optimal multiobjective reinforcement learning. *Neurocomputing, 263,* 26–38.
79. Vamplew, P., Yearwood, J., Dazeley, R., & Berry, A. (2008). On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Australasian joint conference on artificial intelligence* (pp. 372–378). Springer.
80. Velasquez, M., & Hester, P. T. (2013). An analysis of multi-criteria decision making methods. *International Journal of Operations Research, 10*(2), 56–66.
81. Weng, J., Paslaski, S., Daly, J., VanDam, C., & Brown, J. (2013). Modulation for emergent networks: Serotonin and dopamine. *Neural Networks, 41,* 225–239.
82. Wolf, G., Schulkin, J., & Simson, P. E. (1984). Multiple factors in the satiation of salt appetite. *Behavioral Neuroscience, 98*(4), 661.
83. Yates, C., Christopher, R., & Tumer, K. (2020). Multi-fitness learning for behavior-driven cooperation. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference* (pp. 453–461).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Peter Vamplew[1] · Benjamin J. Smith[2] · Johan Källström[3] · Gabriel Ramos[4] · Roxana Rădulescu[5] · Diederik M. Roijers[6,7] · Conor F. Hayes[8] · Fredrik Heintz[3] · Patrick Mannion[8] · Pieter J. K. Libin[6,9,10] · Richard Dazeley[11] · Cameron Foale[1]**

Benjamin J. Smith
benjsmith@gmail.com

Johan Källström
johan.kallstrom@liu.se

Gabriel Ramos
gdoramos@unisinos.br

Roxana Rădulescu
roxana.radulescu@vub.be

Diederik M. Roijers
diederik.roijers@vub.be

Conor F. Hayes
c.hayes13@nuigalway.ie

Fredrik Heintz
fredrik.heintz@liu.se

Patrick Mannion
patrickmannion@nuigalway.ie

Pieter J. K. Libin
pieter.libin@vub.be

Richard Dazeley
richard.dazeley@deakin.edu.au

Cameron Foale
c.foale@federation.edu.au

[1] Federation University Australia, Ballarat, Australia

[2] Center for Translational Neuroscience, University of Oregon, Eugene, OR, USA

[3] Linköping University, Linköping, Sweden

[4] Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, Brazil

[5] AI Lab, Vrije Universiteit Brussel, Brussel, Belgium

[6] Vrije Universiteit Brussel, Brussel, Belgium

[7] HU University of Applied Sciences Utrecht, Utrecht, The Netherlands

[8] National University of Ireland Galway, Galway, Ireland

[9] Universiteit Hasselt, Hasselt, Belgium

[10] Katholieke Universiteit Leuven, Leuven, Belgium

[11] Deakin University, Geelong, Australia