AI Alignment and Human Reward

Patrick Butlin
Department of Philosophy
King's College London
London, UK
patrick.butlin@gmail.com

ABSTRACT

According to a prominent approach to AI alignment, AI agents should be built to learn and promote human values. However, humans value things in several different ways: we have desires and preferences of various kinds, and if we engage in reinforcement learning, we also have reward functions. One research project to which this approach gives rise is therefore to say which of these various classes of human values should be promoted. This paper takes on part of this project by assessing the proposal that human reward functions should be the target for AI alignment. There is some reason to believe that powerful AI agents which were aligned to values of this form would help us to lead good lives, but there is also considerable uncertainty about this claim, arising from unresolved empirical and conceptual issues in human psychology.

CCS CONCEPTS

• Computing methodologies ~Artificial intelligence ~Philosophical/theoretical foundations of artificial intelligence • Social and professional topics ~Professional topics ~Computing profession ~Codes of ethics • Theory of computation ~Theory and algorithms for application domains ~Machine learning theory ~Reinforcement learning

KEYWORDS

Value alignment, reward functions, human values, value learning

ACM Reference format:

Patrick Butlin. 2021. AI Alignment and Human Reward. In *Proceedings of 2021 AAAI/ACM Conference on AI, Ethics and Society (AIES'21), May 19-21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462570

AIES '21, May 19-21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05…$15.00.
DOI: https://doi.org/10.1145/3461702.3642570

1 Introduction

The challenge of AI value alignment is to find a way to determine the values of powerful AI agents which ensures that their actions will benefit humans. Researchers including Nick Bostrom [6] and Stuart Russell [43] have argued persuasively that AI agents with well-aligned values could bring immense benefits to humanity, while AIs with poorly-aligned values could wreak catastrophe.

Russell advocates a value-learning approach to alignment: his proposal is that when constructing powerful AIs we should build in the capacities and dispositions to learn what humans value from our behaviour, and then to act so as to promote these learnt values. Among the most important advantages claimed for this approach is that it would incentivise AIs to defer to humans in situations in which they are uncertain about our preferences [25, 43]. One major complication for the approach is that each of us has different values, which are often incompatible, so it needs to be augmented with a theory describing how different people's values should be balanced and prioritised. A second complication is that human irrationality makes it considerably more difficult to infer our values from our behaviour [1].

Here I will discuss a third complication, which concerns what I will call the target for AI alignment. The target for alignment is the class of human values with which AI agents' values are to be aligned. There are a variety of potential targets with different advantages and disadvantages, because humans value things in a variety of ways. For instance, the target might be a person's desires or preferences, or what they would choose under certain circumstances. So one research project which is needed to refine this approach to alignment is a normative evaluation of possible targets (for a related discussion, see Gabriel [22]). I will elaborate on this project in the first half of this paper (sections 2 and 3).

In the second half of the paper (sections 4-6) I will take on a part of this project by assessing the merits of human reward functions as a target for alignment (see Sarma and Hay [45] for a related proposal). Some of the technical literature in this area builds on inverse reinforcement learning [37, 24], which is the process of inferring a reward function from observed behaviour. However, I will understand human reward functions as characterising the feedback signals which we receive from the environment, in virtue of our physiological needs and biological drives, which are the basis on which we learn what to value and how to act. This is different from the way in which our reward functions are understood in inverse reinforcement learning, which is as functions describing optimal behaviour; both interpretations are compatible with the standard theoretical framework in reinforcement learning (RL), from which the term is derived.

One of the claims which I will defend in the first half of the paper is that a suitable target for alignment must be such that its promotion by AI agents would advance individual well-being. So in the second half I will assess whether human reward functions meet this criterion. I will argue that what a life containing high cumulative reward would be like depends on difficult conceptual and empirical issues about how to map the RL framework onto human psychology, and that some ways of doing this imply that we should not be confident that such a life would be a good one. However, I will also argue that one way of resolving these issues offers a more optimistic vision, especially in the light of current research on intrinsic motivation and curiosity. My discussion will therefore illustrate the importance of detailed engagement with psychology in considering targets for alignment.

2 An Assortment of Human Values

While it is often said that AI must be aligned with human values, there is little clarity about which values should be the target for alignment. Identifying and assessing possible targets is an important part of the alignment problem. One approach which society could take to this aspect of the problem would be to launch a public deliberative process to define a set of values which AI agents should promote. However, in this paper my focus is on the approach to alignment which involves AI agents learning what individuals value, then using what they learn in selecting actions. So in this section I will describe some of the ways in which each of us values things, which could be targets for alignment. It is important to note that this approach to alignment does not entail that AI agents would work only to promote the interests of individuals; a scheme on which AIs would learn what many individuals value, then aggregate these values to form their own objectives, would certainly count as an instance of the approach.

Philosophers have occasionally theorised about the nature of the attitude of valuing, as distinct from attitudes such as desire [53, 32]. But in considering targets for alignment we should not restrict ourselves to what humans value in this narrow sense. Instead we should consider the whole range of ways in which we, or subpersonal systems within us, treat stimuli, actions or outcomes as valuable. These can be divided into three broad categories.

First, we have evaluative attitudes of several kinds. These include desires, preferences, intentions and evaluative beliefs. Some of these can be further divided into sub-kinds which may be ethically significant. For example, we can distinguish between intrinsic and instrumental desires [47] and between first-order and higher-order desires [21]. Preferences, understood as comparative evaluations which are outputs of practical reasoning [26], should be distinguished from intrinsic desires, which are inputs to reasoning. Intentions have an implicit evaluative dimension because to intend to do something involves choosing it over alternatives. As targets for alignment, attitudes such as these can assessed in either of two guises. On one hand, users of AIs will typically have intentions or preferences concerning what the AIs should do, so one way to think of alignment is as the problem of getting autonomous AIs to conform to these attitudes [33]. On the other hand, however, we may gain greater benefits from AIs which aim to satisfy our desires or preferences more generally.

Also in this category are forms of evaluative representation which are identified by cognitive science, but may not correspond exactly to the attitude-types in philosophical folk psychology. These include representations of the values of both outcomes and actions formed in course of reinforcement learning [11, 17]. Sotala [50] proposes these values as the target for alignment.

The second way in which we treat things as valuable is by choosing them, or being disposed to choose them. Russell [43] describes a potential target for alignment of this type, using the term ‘preferences’ to refer to dispositions to choose between fully-specified possible futures. These are worth listing separately from the evaluative attitudes because the latter are thought of as representational states, not behavioural dispositions, in mainstream philosophy of mind.

The third way in which we treat things as valuable is by the ways in which we process stimuli. One example of this is our dispositions to experience pleasure. Since pleasure has a positive valence, and influences how we learn about value, a person’s mind treats a stimulus as valuable when the stimulus causes them pleasure. A further example is our reward functions, which I discuss in detail in this paper. When the RL framework is applied to human psychology, the standard way to think of the reward function is as describing an innate disposition to treat stimuli as positively or negatively valuable, to different extents, for the purposes of reward learning.

These three ways in which we value things are illustrated by the etiologies of our actions. Suppose I buy an ice-cream. It is plausible that this action is itself a way in which I treat ice-creams as valuable, and my having a consistent tendency to act in this way certainly would be. My buying the ice-cream is likely to have been caused in part by my having a positive evaluative attitude towards ice-cream, such as a desire. And my evaluative attitudes towards ice-cream also have sources, which include my innate disposition to assign positive values to sugary foods. It is doubtful whether we would value anything if we did not have some such dispositions.

These three categories may exhaust the ways in which we value things, but the examples I have listed do not exhaust the possible targets for alignment. There are many more complex alternatives. For instance, the target might be our self-regarding desires, so that the AI agent would learn what we each want for ourselves [41]. Or it might be a compound of, say, our reward functions and our evaluative beliefs, so that AI agents respect different aspects of our overall evaluative stance.

3 Well-Being and the Assessment of Targets for Alignment

It is not obvious which of the potential targets for alignment I have just mentioned is most promising. There are plausible arguments for opposing views. One thought might be that all of our values except those we have innately have been formed in response to our experiences in a very specific environment, and are therefore ill-suited to defining our interests in scenarios which may be very different. They reflect the ways which we have found to satisfy our underlying needs and drives, and AI agents may be able to find much better ways. Against this, however, one might argue that we more truly value, with greater justification, the things which have received our reflective endorsement. We should therefore give serious consideration to the merits of a variety of candidates.

One qualifying criterion that any candidate target must meet is being sufficiently well-defined that AI agents can learn about it. The suggestion that AI agents should learn what we desire, for example, would only present us with a serious candidate in the context of a theory describing what it is to desire something and what makes one desire stronger than another. The theory here would ideally be couched in the terms of current scientific psychology, because AIs are likely to have to rely on scientific models of human psychology to learn what we value [10]. Philosophers have made some progress in developing a theory of desire which takes this form [47, 7], but no consensus has been reached, and it remains uncertain whether desires form a natural kind. This criterion suggests that ways of valuing identified by cognitive science may be more promising candidate targets than evaluative attitudes drawn from philosophical folk psychology.

This aside, the most important test for a target for alignment seems to be whether promoting the values in question would help to make the life of the person whose values they are better. I will elaborate on several aspects of this test, using the example of human reward functions as a target for alignment. I take it that a given person's reward function maps each possible situation that they could be in to a real number, and the test asks whether working to bring about states which score highly on this function would help that person to live a good life.

One preliminary point about this test is that it focuses on the individual. My understanding of the present approach to alignment is that by learning what individuals value, AI agents will come to be equipped with representations of those individuals' interests. These representations can then be used in decision-making processes which will promote or protect the interests of others, as well as promoting those of the individuals in question. The focus of my test on individual well-being therefore makes sense even though the purpose of alignment is to ensure that AI benefits humanity in general. It is a test of whether the values that form the alignment target line up with our individual interests, not of the adequacy of entire schemes for achieving alignment.

A second preliminary point is that in applying the test we should assume that AI agents will be powerful and unpredictable, because powerful and unpredictable agents pose the greatest dangers. Given these assumptions, and continuing to use the example of reward functions as the target, we should focus on scenarios in which the life of the person involved would score very highly on their reward function, and we should consider a wide range of such scenarios, rather than only those which strike us as most likely. Considering a wide range of possible high-scoring lives is one way to make our conclusions about the suitability of candidate targets robust, and we should also try to achieve robustness with respect to other sources of uncertainty.

In the context of this test, talk of making the individual's life 'better' or 'good' should be understood as referring to the philosophical concept of well-being. So my proposal is that to assess a target for alignment we must consider whether a life that scores highly on the measure implied by that target will be a good one in this sense. I will call lives that score highly on the person's reward function 'high-return' lives, since 'return' is the term used in RL for cumulative reward. So in the second half of this paper I will assess the well-being of people who are given high-return lives by AI.

Theories of well-being are usually classified into three kinds [41, 9]. These are hedonist theories, desire-satisfaction theories, and objective list theories. Hedonist theories claim that the extent to which someone's life goes well is determined by the balance of pleasant and unpleasant conscious experiences that it contains. Desire-satisfaction theories claim that the extent to which someone's life goes well is determined by the extent to which their desires – or some subset or idealised version thereof – are satisfied. And objective list theories claim that the extent to which someone's life goes well is determined by the extent to which it contains the items on a certain list of goods. These goods are said to make our lives better in non-instrumental ways.

I will not attempt to defend a theory of well-being here. Instead, my approach will be to use the standards provided by objective list theories as a heuristic. A life which is good by objective list standards will also be good by hedonist standards, because the lists of goods which objective list theories offer typically include pleasure and happiness. Such a life is also likely to be good by the standards of desire-satisfaction theories, because happiness arguably entails having a good portion of one's desires satisfied, and the other goods on the lists are among the things we tend to most strongly desire. An approach like this, which aims to test whether high-return lives would be good by the standards of all major theories, has the advantage of taking into account normative uncertainty [34].

Fletcher [20] surveys the goods listed in four paradigmatic objective list theories, offered by Finnis [18], Parfit [41], Murphy [35] and Fletcher [19]. Notable items on these lists include:

- Pleasure, happiness and aesthetic experience (which we might call 'experiential goods')
- Friendship and virtue ('social goods')
- Knowledge, achievement, the development of abilities, rational activity, and excellence in play, work and agency ('perfectionist goods').

I will use this list in assessing human reward functions as a target for alignment.

One further point before I start on this task concerns the influential experience machine objection to hedonism [38]. The experience machine is an imagined machine which generates simulated experiences which seem to the user to be a continuation of their previous life, but are highly pleasurable. If the experience machine were real and totally reliable, and we could stay in there for the rest of our lives, then hedonism implies that each of us could make our lives better by entering the machine. But philosophers tend to judge that living in a simulation, however pleasurable, would not be better than living a life of real personal relationships, discoveries and accomplishments.

There are some interesting responses to the experience machine objection: perhaps our judgment about the case is irrationally influenced by status quo bias [14], or by thoughts about whether it would be morally right to enter the machine [27]. But the majority view about this scenario should be given significant weight in our assessment. If a high-return life is possible in an experience machine-like scenario, this should make us much less confident that it would be a good life.

4 Human Reward Functions: A First Pass

I now turn to the assessment of human reward functions as a target for alignment. The question we face is whether a high-return life would be good by the standards just sketched, and to answer it we need to understand what human reward functions are like.

In the standard RL framework, the reward function is a function taking states or state-action pairs to real numbers, which cannot be changed by the agent [51]. It has two roles. The first role is that rewards constitute feedback on actions, which causes agents to update the representations they use for subsequent action selection. A crucial distinction in this field is between the reward function and the content of representations of the value of actions or outcomes, where value is defined as discounted subsequent cumulative reward. The reward function describes the signal which is used to update value representations, which are in turn used to select actions. So for an RL agent to treat some action or outcome as valuable in action selection is distinct, both conceptually and in practice, from the action or outcome's being rewarding for the agent. The second role is that optimal behaviour is defined as behaviour that maximises long-run cumulative reward. This conception of optimal behaviour is used in testing the performance of RL algorithms.

In principle either of these roles could be used to define human reward functions, and it is not at all clear that the definitions would be equivalent. But to define human reward functions using the second role we would need a prior understanding of optimal behaviour, and there seem to be a range of standards which could ground optimality in the human case. Furthermore, if we took this approach, a separate account of rewards as feedback on actions would be needed in order that the RL framework could be used to explain human value learning and motivation. So I will interpret human reward functions using the first role.

Understood in this way, the reward function describes a disposition of events or features of the environment to cause a certain form of learning, which affects evaluative representations in the agent. It is not only rewards that cause value learning, however. In the temporal-difference learning family of RL algorithms, the agent may also update the value that it represents some action as having when this action leads to a state that is merely represented as predicting reward [51, 12]. What distinguishes reward as a grounds for update is that the reward function does not depend on what the agent has learnt. So in trying to understand reward functions in humans we should be looking for a specifically unlearnt disposition to treat events or features as valuable for the purpose of value learning.

A natural and standard way to apply this framework to humans and animals is therefore to take us to have innately-specified reward functions.[1] To say that our reward functions are innate is only to say that they are psychologically primitive [44], not that they are unchanging or not subject to interpersonal variation. For example, our reward functions may change without learning as we develop through puberty, and some people may be more sensitive to certain rewards than others.

At a first pass, then, it seems that each of us has a reward function which can be characterised in terms of primary rewards [48] and punishments. My reward function maps my receiving primary rewards, such as sugar, to positive real numbers. It maps my receiving punishments, such as injuries, to negative real numbers. And it maps any event which I have to learn to evaluate to zero. A typical list of primary rewards might include food, sex, shelter and perhaps affectionate or status-affirming social interaction. So on this view there are many things that each of us learns to value positively which are not rewards for us, in the sense that our reward functions do not map them to positive real numbers. For instance, I like receiving parcels, publishing research papers, and visiting the Sussex coast in southern England, but none of these are rewarding for me in themselves.

How good would high-return lives be, if cumulative reward depends only on whether we receive primary rewards and punishments? The answer seems to be that even the worst high-return lives would be fairly good in some respects, although potentially seriously lacking in others. We should imagine a life which involves minimal illness or injury and plentiful food. It is plausible that to receive high levels of reward from social interaction we must not only share occasional smiles with strangers, but also engage in real friendships and loving relationships. A life with these attributes is not to be sniffed at.

However, there is no obvious guarantee that other goods will be included in high-return lives on this account. From the objective list theories, the missing items might include knowledge, achievement, the development of abilities and some valuable forms of aesthetic experience. Living a high-return life seems to be compatible with being subject to massive deception, and hence to a lack of knowledge. It also seems to be compatible with the failure to engage in projects of enquiry, exercises of creativity, or work to achieve one's ambitions or benefit one's community. Perhaps AI agents which aimed to give us high-return lives would guide us away from such activities, and hence towards lives which lacked the corresponding goods.

Two sources of uncertainty about how good high-return lives would be are immediately apparent. First, it is uncertain what a life's worth of food or sex would need to be like in order to be highly rewarding. One possibility is that a diet could be drab and repetitive but still very rewarding if it contained enough of the right nutrients. Another possibility, though, is that a varied diet of delicious foods would be required. The latter would make for a life much richer in pleasure and valuable aesthetic experience. This issue is closely related to the topic of the relationship between pleasure and reward, which I return to below.

Second, on the present account several different kinds of stimuli are rewarding, which means that there is an open question about how much variation in rewards must be provided to generate high return overall. Rewards of different kinds are interchangeable in that they all contribute to a single scalar value, but at least some rewards, such as food, presumably make a diminishing marginal contribution. In the sketch of a high-return life which I just offered I assumed that high levels of each kind of reward would be needed for a high return overall. If this is not the case the picture looks much less optimistic, because rewards from social interaction made such a significant contribution to well-being in the life I imagined. A life with an ideal supply of food but without these social rewards would not be a good one.

5 Reasons for Pessimism: Alternative Conceptions of Reward

In the remainder of the paper I will discuss ways in which matters are more complicated than this first-pass account suggests. In this section I describe reasons to be relatively pessimistic about the goodness of high-return lives, which arise from a range of theories about the nature of reward in humans and other animals; in the following one I describe a reason to be more optimistic.

5.1 Reward and the Boundary between Agent and Environment

One feature of the first-pass account which might be questioned is that I have so far taken it that primary rewards are, on the whole, events that take place in the environment rather than in our brains. For example, I have taken it that my reward function assigns a positive value to my actually consuming sugar, and inferred that this means that a high-return life for me would be one in which I consumed sugar. But one might wonder whether in fact what is rewarding for me is that it seems to me that I am consuming sugar. In this case my consuming artificial sweeteners, or my brain's being stimulated so that I have the experience as of eating sugar,

[1] It is surprisingly difficult to find psychological literature that makes this point explicitly, but the fact that this is the standard approach was confirmed to me by Fiery Cushman in personal communication. It is also implicit in e.g. Daw's [12] description of how 'evolutionarily programmed rewards' can lead to 'a rich landscape of value'.

would do just as well. The availability of this kind of alternative perspective on primary rewards should make us more pessimistic about the goodness of high-return lives because it suggests that life in the experience machine would be highly rewarding.

This feature of the first-pass account is questioned to a more extreme extent by a view taken by some researchers concerning the application of the RL framework to human and animal minds. Andrew Barto and colleagues argue that when applying this framework we should sharply distinguish between the RL agent and the organism itself. They claim that the RL agent should be thought of as an internal element of the organism – a homunculus – which operates in an environment generated largely by other parts of the organism's mind [2, 49]. The reason they take this view is that in standard RL research, outside the biological context, reward is generated by the environment. At each time-step a number indicating reward received is given as input to the agent. But organisms do not detect reward itself; instead they detect phenomena like the presence of sugar on the tongue, and then perform computations to infer how much reward they have received. This means that by drawing the boundaries of the agent inside the organism, with the systems that infer levels of reward from sensory stimulation outside the agent, researchers can identify an entity which faces as similar a problem as possible to a standard RL agent.

On this view it seems that any event which causes the RL agent to receive a reward signal will count as rewarding for that agent, and that there is no way to interpret reward for humans other than as reward for the RL agent that we each have inside us. The suggests a very pessimistic picture of the high-return life, because it appears to entail that this life could be achieved by manipulation of the brain mechanisms involved in processing reward, for instance through electrical stimulation. This manipulation may not even require making it seem to us that we are eating palatable foods or having positive social interactions, so on this view the high-return life might be significantly worse than life in the experience machine.

In particular, it is not certain that stimulating the brain in the way necessary to provide reward signals to the RL homunculus would cause pleasure. Although there is evidence that the orbitofrontal cortex is both involved in evaluating the reward value of stimuli and a locus of activity associated with pleasure [23, 4], a variety of other brain areas seem to provide information about received reward to midbrain dopamine neurons, which in turn generate a reward prediction error signal playing a crucial role in RL [30]. One recent study by Tian et al. [52] indicated that pure reward signals are 'highly redundant and distributed'. A life which was not pleasurable and did not involve normal engagement with the outside world would almost certainly not be a good one.

The extent which this line of thought should make us pessimistic is mitigated, however, by the fact that Barto's view is a proposal about how to conceive of reward, not an empirical claim. This means that even if Barto's view offers the right way to conceive of reward for scientific purposes, it might be possible to construct AI agents which learn about and maximise human reward conceived in a different way. The possibility that AI agents would seek to give us high-return lives by manipulating our brains does seem to be one that we should be alert to when considering the present approach to alignment, but it is not clear that whether this danger would materialise depends on whether Barto is correct.

5.2 Pleasure and Reward

Another, similar reason for pessimism comes from a mainstream view about the relationship between pleasure and reward. This is the view that pleasure is the biological equivalent of the numerical feedback signal received by standard RL agents in computer science; or, to put it another way, that the extent to which a situation is rewarding for a person is simply a matter of how pleasurable it is. One piece of evidence that this is a mainstream view is that, in a textbook chapter on RL in humans and other animals, Daw and O'Doherty [13] write that the level of reward that an outcome yields is equivalent to the amount of pleasure it provides.

This view generates a similar problem to the view that the RL agent is a homunculus because it entails that highly pleasurable lives would be highly rewarding. The experience machine scenario depicts one form of highly pleasurable life which does not seem to be a good one, but again this may not be the worst of it. It may be possible through brain manipulation to cause someone to lead a life in which they experience a great deal of pleasure, but in which it does not even seem to them that they are engaging normally with the outside world.

Unfortunately, it is difficult to assess the case for this view of the relationship between pleasure and reward because pleasure is not widely discussed in the psychological literature on RL. But there are alternatives. One possibility is that pleasure is a signal representing the level of reward being received, which may in principle be inaccurate. Dickinson and Balleine [16] argue for a version of this hypothesis. This view does not alter the first-pass account of human reward functions, because on this view the level of reward received depends on events in the environment like sugar consumption; the role of pleasure is merely to represent the significance of these events for value learning.

A second possible alternative is that pleasure is a reward among others, alongside rewards such as food, sex and positive social interaction. If this is the case then the level of reward that I receive from eating an ice-cream, for example, might be the sum of the reward generated by the sugar it contains and that generated by the pleasure it causes. This alternative is worth considering because the previous two seem unable to accommodate the apparent fact that what causes us pleasure changes as we learn. They are also in tension with the evidence for the distinctness of pleasure and reward signals in the brain which I mentioned above. But this alternative does have the disadvantage of making pleasure's role somewhat mysterious. On this account pleasure would generate an additional boost to the reinforcing effect of stimuli which have already been identified as rewarding or valuable, and it is not clear what evolutionary advantage this might bring.

If this rather uncertain possibility is correct, it may provide a reason for optimism about the proposal that our reward functions should be the target for AI alignment. This is because a life that was pleasurable as well as having the features considered in the first-pass account would be more clearly good. In particular, pleasure in sensory stimuli seems to depend on variety, so the high-return life might be varied as well as secure and comfortable. However, this line of thought would only hold if pleasure could not be a substitute for other rewards. If it could be, then we would have another case in which copious pleasure would suffice for a high-return life, and hence another reason for pessimism.

5.3 Do Humans Have Reward Functions?

A third issue that complicates the first-pass picture is that humans may use multiple systems for value learning and action selection which have different reward functions. The claim that humans and other animals use multiple systems for these purposes is well-supported and widely accepted [13, 17]. The typical view is that humans use both model-free and model-based RL methods for learning and action selection, and there is ongoing research on how these systems interact [31]. Some features of our behaviour may be explained on the assumption that distinct systems calculate the values of options in different ways, and compete for control. For example, Neal et al. [36] found that people who often ate popcorn at the cinema would eat stale popcorn as readily as fresh, and attributed this behaviour to the influence of a habit system which had learnt to place a high value on the action of eating popcorn. Other participants who ate popcorn less often were more sensitive to the taste, suggesting that their behaviour was controlled by a goal-directed system which evaluated outcomes.

This kind of explanation does not require the competing systems to have different reward functions, but the possibility should not be ruled out. It is suggested by Berridge's incentive salience theory [3, 5], which distinguishes 'incentive salience wanting' from 'cognitive wanting'. Berridge claims that there is a cognitive, goal-directed system for action selection which aims to maximise pleasure and that this is accompanied, and sometimes undermined, by a Pavlovian incentive salience system which works independently. He does not claim this explicitly, to my knowledge, but it seems that in his account pleasure is equivalent to reward for the goal-directed system, while the incentive salience system has a distinct reward function.

This possibility is important because it would entail that, in one sense, there is no such thing as my reward function, or that of anyone else. My reward function could not be the target for AI alignment; at best, the target would have to be one of my reward functions. Peysakhovich [42] presents a method for inverse reinforcement learning in the case in which the observed agent's behaviour is controlled by dual systems with different reward functions, but he takes the aim of this process to be to infer the reward function of one of the two systems. If we do have multiple systems with distinct reward functions, then it may be that the reward function of one of those systems would be a good target for alignment. But this would represent a significant revision of the proposal at hand.

A further possibility on the same lines is that the RL framework is simply not a good model for human value-guided decision-making [29]. Humans may lack reward functions entirely. In this case attempts to infer our reward functions from our behaviour would be likely to run into intractable difficulties.

5.4 Summary: Reasons for Pessimism

In this section we have identified three claims which must hold if we are to be confident that a high-return life would be a good one. These are:

- That we are each accurately modeled as RL agents with a single reward function;
- That reward signals in the brain represent received reward, and are therefore capable of misrepresenting it;
- And that pleasure is neither the only reward for humans, nor one that can substitute for any other.

Reasons to believe that any of these three claims are false are therefore reasons for pessimism about the use of human reward functions as the target for AI alignment.

6 A Reason for Optimism: Learning as a Reward

According to the most optimistic view we have seen so far, stimuli which are positively rewarding for humans include food, sex, positive social interaction and pleasure. A life which was rich in each of these goods and which lacked negatively-rewarding events such as injury, illness and distress would be good to a meaningful extent. It may well contain many valuable aesthetic experiences, friendships and family relationships. But there are several goods remaining that are mentioned by objective list theories, which the high-return life may not provide. These include happiness and virtue, and knowledge, achievement, the development of abilities, rational activity, and excellence in play, work and agency. Setting happiness and virtue aside, the other goods on this list are particularly emphasised by perfectionism, the philosophical view that well-being consists in the development and realisation of distinctively human capacities [28].

These goods also have in common a relationship to so-called 'intrinsic motivation' [15]. Humans are highly motivated to learn facts and skills, to explore their environments, play and challenge themselves, and achieve even arbitrary goals. This motivation confers likely evolutionary benefits, and appears to be of great instrumental value in helping us to obtain rewards such as food and sex. But psychologists refer to the motivation to behave in such ways as intrinsic, because it seems to be produced even when these 'extrinsic' rewards are not in prospect. This suggests that there may be further primary rewards for humans, in addition to those we have so far considered, which are responsible for motivation of this kind.

In particular, some recent research proposes that learning is a reward in itself [46, 39, 40, 2, 8]. Schmidhuber and Oudeyer and colleagues have developed theories according to which progress in learning is rewarding. They claim that reward from this source is proportional to the extent to which uncertainty about some domain diminishes through the course of a period of engagement. A reward function with this property will generate a virtuous feedback loop, which causes learning about increasingly difficult problems. At first learning progress will be most readily achieved by exploring simple domains or tasks, but these will be quickly mastered, meaning that little more progress is possible. However, mastering these domains will make learning progress in others more accessible, so the learner will find new problems rewarding. This model seems to be relevant to the acquisition of both knowledge and skills. Schmidhuber [46] also argues that because the theory entails that stimuli containing novel but learnable patterns will be most rewarding, it can explain our appreciation of the arts.

If a theory of this kind is correct, it offers a reason for relative optimism about using reward functions as the target for AI alignment. It suggests that to have high-return lives we may need to be given opportunities to gain knowledge and develop abilities, ticking off two of the remaining items from the objective list theories. Given Schmidhuber's claim, a further effect of ensuring that our lives involve opportunities for learning progress may be that they will contain more valuable aesthetic experiences. More generally, the theory suggests that high-return lives must involve sufficient variety and change not to be boring, but not so much as to be bewildering.

The learning progress theory does not seem to explain all aspects of intrinsic motivation. For instance, it does not explain our motivation to achieve goals for their own sake, or the pleasure we take in exercising well-honed skills. But this too is a reason for optimism. If these other aspects of intrinsic motivation are explained by other elements of our reward functions, then there may be rewards associated with the perfectionist goods of achievement and excellence, as well as with knowledge and the development of abilities. In this case the high-return life would be good in almost all of the respects identified by objective list theories.

7 Conclusion

I have argued that causing someone to lead a high-return life may ensure that their life would be good, but only subject to some strong assumptions. These assumptions concern unresolved empirical and conceptual issues in human psychology.

In closing, I want to note a different kind of consequence of taking human reward functions as the target for AI alignment. In contrast to human preferences or desires, human reward functions are not highly variable either between individuals or over time. We are also unable to introspect our reward functions, and our actions on particular occasions are not typically revealing about their precise form. These points mean that AI agents which aimed to promote our reward functions would have relatively little incentive to consult or defer to us about what to do, potentially undermining one of the key advantages that have been claimed for Russell's approach to AI alignment. For a more complete picture of the merits of the reward function approach, we need to assess it in comparison to alternatives, and in the context of such alignment-specific concerns.

ACKNOWLEDGMENTS

This research was supported by Survival and Flourishing.

REFERENCES

[1] Stuart Armstrong and Sören Mindermann. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems 31 (NeurIPS '18)*. Curran Associates, Red Hook, NY, 5603-5614.

[2] Andrew Barto. 2013. Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*, edited by G. Baldassarre and M. Minolli. Springer, Berlin, 17-47. DOI: https://doi.org/10.1007/978-3-642-32375-1_2

[3] Kent Berridge and J. Wayne Aldridge. 2008. Decision utility, the brain, and pursuit of hedonic goals. *Social Cognition* 26, 5 (Oct. 2008), 621-646. DOI: https://doi.org/10.1521/soco.2008.26.5.621

[4] Kent Berridge and Morten Kringelbach. 2015. Pleasure systems in the brain. *Neuron* 86, 3 (May 2015), 646-664. DOI: https://doi.org/10.1016/j.neuron.2015.02.018

[5] Kent Berridge and John P. O'Doherty. 2013. From experienced utility to decision utility. In *Neuroeconomics: Decision-Making and the Brain,* edited by E. Fehr and P. W. Glimcher. Academic Press, London, 335-348.

[6] Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.

[7] Patrick Butlin. 2017. Why hunger is not a desire. *Review of Philosophy and Psychology* 8 (Sep. 2017), 617-635. DOI: https://doi.org/10.1007/s13164-017-0332-9

[8] Nick Chater and George Loewenstein. 2016. The under-appreciated drive for sense-making. *Journal of Economic Behavior and Organization* 126, Part B (Jun. 2016), 137-154. DOI: https://doi.org/10.1016/j.jebo.2015.10.016

[9] Roger Crisp. 2017. Well-being. *Stanford Encyclopedia of Philosophy,* edited by E. N. Zalta. Retrieved from https://plato.stanford.edu/entries/well-being/

[10] Paul Christiano. 2015. The easy goal inference problem is still hard. Retrieved from https://ai-alignment.com/the-easy-goal-inference-problem-is-still-hard-fad030e0a876

[11] Fiery Cushman. 2013. Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review* 17, 3 (Aug. 2013), 273-292. DOI: https://doi.org/10.1177/1088868313495594

[12] Nathaniel Daw. 2013. Advanced reinforcement learning. In *Neuroeconomics: Decision-Making and the Brain,* edited by E. Fehr and P. W. Glimcher. Academic Press, London, 299-317.

[13] Nathaniel Daw and John P. O'Doherty. 2013. Multiple systems for value learning. In *Neuroeconomics: Decision-Making and the Brain,* edited by E. Fehr and P. W. Glimcher. Academic Press, London, 393-410.

[14] Felipe De Brigard. 2010. If you like it, does it matter if it's real? *Philosophical Psychology* 23, 1 (Feb. 2010), 43-57. DOI: https://doi.org/10.1080/09515080903532290

[15] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Plenum Press, New York.

[16] Anthony Dickinson and Bernard Balleine. 2009. Hedonics: The cognitive-motivational interface. In *Pleasures of the Brain*, edited by M. Kringelbach and K. Berridge. Oxford University Press, Oxford, 74-84.

[17] Ray Dolan and Peter Dayan. 2013. Goals and habits in the brain. *Neuron* 80, 2, 312-325. DOI: https://doi.org/10.1016/j.neuron.2013.09.007

[18] John Finnis. 1980. *Natural Law and Natural Rights*. Clarendon Press, Oxford.

[19] Guy Fletcher. 2013. A fresh start for the objective list theory of well-being. *Utilitas* 25, 206-220. DOI: https://doi.org/10.1017/S0953820812000453

[20] Guy Fletcher. 2016. Objective list theories. In *The Routledge Handbook of Philosophy of Well-Being*, edited G. Fletcher. Routledge, Abingdon, Oxon, 148-160.

[21] Harry Frankfurt. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68, 1, 5-20. DOI: https://doi.org/10.2307/2024717

[22] Iason Gabriel. 2020. Artificial intelligence, values and alignment. *Minds and Machines* 30, 411-437. DOI: https://doi.org/10.1007/s11023-020-09539-2

[23] Fabian Grabenhorst and Edmund Rolls. 2011. Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences* 15, 2 (Feb. 2011), 56-67. DOI: https://doi.org/10.1016/j.tics.2010.12.004

[24] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel and Stuart Russell. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS '16)*. Curran Associates, Red Hook, NY, 3916-3924.

[25] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel and Stuart Russell. 2017. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. AAAI Press, Menlo Park, Calif., 220-227.

[26] Daniel Hausman. 2012. *Preference, Value, Choice and Welfare*. CUP, New York.

[27] Sharon Hewitt. 2010. What do our intuitions about the experience machine really tell us about hedonism? *Philosophical Studies* 151, 3 (Dec. 2010), 331-349. DOI: https://doi.org/10.1007/s11098-009-9440-4

[28] Tom Hurka. 1993. *Perfectionism*. Oxford University Press, Oxford.

[29] Keno Jeuchems and Christopher Summerfield. 2019. Where does value come from? *Trends in Cognitive Sciences* 23, 10 (Oct. 2019), 836-850. DOI: https://doi.org/10.1016/j.tics.2019.07.012

[30] Ronald Kieflin and Patricia H. Janak. 2015. Dopamine prediction errors in reward learning and addiction: From theory to neural circuitry. *Neuron* 88, 2, 247-263. DOI: https://dx.doi.org/10.1016/j.neuron.2015.08.037

[31] Wouter Kool, Fiery Cushman, and Samuel Gershman. 2018. Competition and cooperation between multiple reinforcement learning systems. In *Goal-Directed Decision-Making: Computations and Circuits*, edited by R. Morris, A. Bornstein and A. Shenhav. Elsevier, Amsterdam, 153-178. DOI: https://doi.org/10.1016/B978-0-12-812098-9.00007-3

[32] Robbie Kubala. 2017. Valuing and believing valuable. *Analysis* 77, 1 (Jan. 2017), 59-65. DOI: https://doi.org/10.1093/analys/anx043

[33] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini and Shane Legg. 2018. Scalable agent alignment via reward modeling: A research direction. arXiv:1811.07871. Retrieved from https://arxiv.org/abs/1811.07871

[34] William MacAskill, Krister Bykvist and Toby Ord. 2020. *Moral Uncertainty*. Oxford University Press, Oxford.

[35] Mark Murphy. 2001. *Natural Law and Practical Rationality*. CUP, New York.

[36] David T. Neal, Wendy Wood, Mengju Wu and David Kurlander. 2011. The pull of the past: When do habits persist despite conflict with motives? *Personality and Social Psychology Bulletin* 37, 11 (Nov. 2011), 1428-1437. DOI: https://doi.org/10.1177/0146167211419863

[37] Andrew Ng and Stuart Russell. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers, San Francisco, Calif., 663-670

[38] Robert Nozick. 1974. *Anarchy, State and Utopia*. Basic Books, New York.

[39] Pierre-Yves Oudeyer, Frédéric Kaplan and Verena Hafner. 2007. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation* 11, 2 (Apr. 2007), 265-286. DOI: https://doi.org/10.1109/TEVC.2006.890271

[40] Pierre-Yves Oudeyer, Jacqueline Gottlieb and Manuel Lopes. 2015. Intrinsic motivation, curiosity and learning: Theory and applications in educational technologies. *Progress in Brain Research* 229, 257-284. DOI: https://doi.org/10.1016/bs.pbr.2016.05.005

[41] Derek Parfit. 1984. *Reasons and Persons.* Clarendon Press, Oxford.

[42] Alexander Peysakhovich. 2019. Reinforcement learning and inverse reinforcement learning with system 1 and system 2. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society (AIES '19)*. ACM, New York, 409-415. DOI: https://doi.org/10.1145/3306618.3314259

[43] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press, New York.

[44] Richard Samuels. 2002. Nativism in cognitive science. *Mind and Language* 17, 3, 233-265. DOI: https://doi.org/10.1111/1468-0017.00197

[45] Gopal Sarma and Nick Hay. 2017. Mammalian value systems. *Informatica* 41, 3, 441-449. DOI: https://dx.doi.org/10.2139/ssrn.2975399

[46] Jürgen Schmidhuber. 2010. Formal theory of creativity, fun and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development* 2, 3 (Sep. 2010), 230-247. DOI: https://doi.org/10.1109/TAMD.2010.2056368

[47] Timothy Schroeder. 2004. *Three Faces of Desire*. Oxford University Press, New York.

[48] Guillaume Sescousse, Xavier Caldú, Barbara Segura and Jean-Claude Dreher. 2013. Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience and Biobehavioural Reviews* 37, 4 (May 2013), 681-696. DOI: https://doi.org/10.1016/j.neubiorev.2013.02.002

[49] Satinder Singh, Richard Lewis and Andrew Barto. 2009. Where do rewards come from? In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*. Curran Associates, Red Hook, NY, 2601-2606.

[50] Kaj Sotala. 2016. Defining human values for value learners. In *Papers from the 2016 AAAI Workshop on AI, Ethics and Society*. AAAI Press, Palo Alto, Calif..

[51] Richard Sutton and Andrew Barto. 2018. *Reinforcement Learning: An Introduction* (2nd. Ed.). MIT Press, Cambridge, MA.

[52] Ju Tian, Ryan Huang, Jeremiah Y. Cohen, et al. 2016. Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron* 91, 6, 1374-1389. DOI: https://doi.org/10.1016/j.neuron.2016.08.018

[53] Gary Watson. 1975. Free agency. *Journal of Philosophy* 72, 8 (Apr. 1975), 205-220. DOI: https://doi.org/10.2307/2024703