Preprint accepted by **Ethics and Information Technology.**

Human-Aligned Artificial Intelligence is a Multiobjective Problem

Peter Vamplew · Richard Dazeley · Cameron Foale · Sally Firmin and Jane Mummery

Received: December 2016 / Accepted: July 2017

Abstract As the capabilities of artificial intelligence systems improve, it becomes important to constrain their actions to ensure their behaviour remains beneficial to humanity. A variety of ethical, legal and safety-based frameworks have been proposed as a basis for designing these constraints. Despite their variations, these frameworks share the common characteristic that decision-making must consider multiple potentially conflicting factors. We demonstrate that these alignment frameworks can be represented as utility functions, but that the widely used Maximum Expected Utility (MEU) paradigm provides insufficient support for such multiobjective decision-making. We show that a Multiobjective Maximum Expected Utility paradigm based on the combination of vector utilities and non-linear action-selection can overcome many of the issues which limit MEU's effectiveness in implementing aligned artificial intelligence. We examine existing approaches to multiobjective artificial intelligence, and identify how these can contribute to the development of human-aligned intelligent agents.

P. Vamplew
Federation Learning Agents Group (FLAG), Federation University Australia
Tel.: +61-3-5327-9616
E-mail: p.vamplew@federation.edu.au

R.Dazeley
FLAG, Federation University Australia

C.Foale
FLAG, Federation University Australia

S. Firmin
FLAG, Federation University Australia

J.Mummery
Faculty of Education and Arts, Federation University Australia

Keywords ethics, aligned artificial intelligence, value alignment, maximum expected utility, reward engineering

1 Introduction

Recent years have seen dramatic improvements in the capabilities of artificial intelligence (AI) systems, with AI agents demonstrating human or even superhuman levels of performance across a variety of tasks (Ferrucci, 2012; Mnih et al., 2015; Silver et al., 2016). In parallel, AI technology is increasingly moving beyond research labs and 'toy' problems, and being applied in systems which are directly embedded in the real world, such as autonomous vehicles (Lozano-Perez et al., 2012). Mittelstadt et al. (2016) note that ethical issues can arise even in systems which are only semi-autonomous, and it can be expected that the ethical repercussions are likely to increase as systems become increasingly autonomous. For example, even if current autonomous vehicles are not yet explicitly reasoning about the 'trolley-car' like ethical dilemmas involved if an accident becomes unavoidable (Goodall, 2014), they do regularly make decisions which carry an implied trade-off between the safety of the driver, passengers and other road-users, and other factors like trip duration (for example, deciding how much below the speed-limit to travel on an icy road).

These developments have led multiple researchers to raise concerns regarding the potential dangers posed by careless application of artificial intelligence. An open letter expressing such concerns, alongside commentary on the potential benefits of advanced AI, was released (Future of Life Institute, 2015), while the IEEE has initiated a series of committees to examine the issues

pertaining to ethical development and deployment of AI (The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 2016). Some authors, notably Bostrom (2014), have focused on the existential risk to humanity posed by superhuman artificial general intelligence, while others have concentrated on the more immediate dangers posed in the short to medium-term (Amodei et al., 2016). In either case, the underlying concern is that an agent following under-specified or poorly defined goals, or which has the ability to modify its own goals, may act in a manner which is inconsistent with the intent of its designer.

To prevent such dangers from arising, many researchers have proposed that the behaviour of AI systems must be constrained. Various frameworks have been identified which might act as a basis for these constraints, including adapting moral and ethical systems previously proposed for human behaviour, as well as other frameworks tailored more expressly to the requirements of AI. Soares and Fallenstein (2014) introduced the term *aligned* to refer to an artificial intelligence which is constructed in such a way as to ensure that it behaves in a manner which will be beneficial to humanity (that is to say, its goals are 'aligned with human interests'). This paper adopts this terminology.

While there has been substantial theoretical and philosophical discussion regarding aligned artificial intelligence in recent years, Allen and Wallach (2012) note that there is often a disconnect between the abstractions proposed at a theoretical level, and the implementation technologies developed by AI practitioners. This paper aims to address this divide by identifying specific characteristics of the various theories and considering how they impact on the requirements of the underlying technologies.

Section 2 reviews some of the main frameworks which have been proposed as a basis for aligned AI, and identifies a common theme — the need for an agent to be able to take into account multiple conflicting factors when making decisions. Section 3 addresses the use of multi-factor utility functions to represent these alignment frameworks, and considers the broad class of AI technologies based on the concept of maximum expected utility (MEU), assessing their risks and the extent to which these can be addressed by incorporating alignment constraints. A critical limitation is identified in terms of the capability of MEU methods to address the multiobjective characteristic inherent in all alignment frameworks. Section 4 examines the extension of MEU approaches to use an explicitly multiobjective representation of utility, showing that this enables alternative approaches to action selection which address the limitations of MEU. This section identifies promising directions for applying such technologies to address the issues posed by the various alignment frameworks, and briefly reviews the current work on multiobjective AI, and multiobjective approaches to alignment.

We conclude by arguing that the appropriate means to suitably constrain AI behaviour is to use an explicitly multiobjective approach to specifying and implementing an agent's goals, and that this provides a very strong argument for an increased focus on the development of multiobjective approaches to AI and autonomous agents.

2 Alignment Frameworks for Artificial Intelligence

In this section we review a sample of the various approaches which have been proposed as providing a suitable basis for specifying constraints on the behaviour of AI agents. These concepts have arisen from a number of fields including philosophical theories of ethics, moral systems, and codes of conduct from specific domains. For convenience, we will refer to these as *alignment frameworks* as all have the aim of ensuring that AI is aligned, in the sense proposed by Soares and Fallenstein (2014).

2.1 General ethical frameworks

The identification of ethical frameworks to drive human behaviour has long been one of the primary themes of philosophical thought. We do not intend to provide a thorough review of these ethical philosophies here, but instead to focus on the key characteristics which we believe to be of most relevance to the development of ethical AI. As such we restrict our discussion to the utilitarian and deontological approaches to ethics, as these have been the most widely considered in the literature on ethical AI so far.

2.1.1 Utilitarian ethics

Utilitarianism is based on the notion that the morality of an action should be judged by its consequences. It is assumed that the desirability of an outcome can be measured via some utility metric, and that an action is judged to be morally right if its consequences lead to the greatest utility (Tavani, 2011). Different utilitarian theories vary in terms of the definition of utility they aim to maximise. For example, Bentham (1789) proposed that a moral agent should aim to maximise the total happiness of a population of people. Utilitarian theories also

vary in whether they are *act utilitarianism* or *rule utilitarianism*. An act utilitarian selects between acts by simply choosing the act which can be expected to maximise utility given the current situation. In contrast rule utilitarianism identifies rules of behaviour which would be expected to lead to good outcomes if followed by everyone.

Utilitarianism has been a popular ethical theory over the last hundred years and is preferred by economists as its outcomes can be measured in dollar terms (Reynolds, 2011). Due to their quantitative nature, the utilitarian approaches to ethics also appear particularly well suited for implementation in computer systems. However the choice of which of the many utilitarian theories is most appropriate for an AI agent is unclear. Brundage (2014) notes that reviews of the utilitarian literature reveal no consensus on exactly what measure of utility should be maximised, and that pluralist utilitarian philosophies explicitly advocate considering multiple values, such as a mixture of individual and group benefits. However, there remains disagreement over the correct manner in which to weight different sources of utility, or even over whether it is appropriate to combine them on the same scale at all (Wallach and Allen, 2008).

Wallach and Allen (2008) suggest that one approach to utilitarian AI may be to elicit multiple utility ratings from different sources, and then seek to combine these into a single weighting formula. Abel et al. (2016) also propose adopting a multiobjective utilitarian approach for the creation of an ethical AI agent using reinforcement learning, in which the agent learns the ethical preferences of multiple individuals, and then tries to maximize a combination of these personal preferences.

2.1.2 Deontological ethics

Deontological ethics argues that actions should be judged not on the basis of their expected outcomes, but on whether they are compatible with a set of duties which would be recognised by all rational decision-makers. As with utilitarian theories, many variations of deontological ethics exist, depending on which duties are assumed to apply, and theories can be both act-based or rule-based. For example, Kant's categorical imperative states that people should be understood as ends-in-themselves and not merely as a means to an end, and that actions should be judged on the basis to which they comply with this imperative (Kant, 1993; Tavani, 2011).

Meanwhile, Ross (1930) proposed a list of seven *prima facie* duties consisting of fidelity, reparation, gratitude, non-maleficence, justice, beneficence, and self-improvement. A decision-maker should try to satisfy all of these duties, but of course at times they may conflict with each other, at which point the decision-maker must balance the importance of the different competing duties to decide on the most ethical course of action. Fieser (2016) describes a scenario based on Ross' list of duties where a person borrows a gun from their neighbour and promises to return it. At a later time the neighbour demands the gun back in order to shoot a third party. The person now faces a conflict between the fidelity and non-maleficence duties. Defining the correct decision in the face of such conflicts is extremely difficult. Anderson et al. (2006a) proposed a computational approach to resolving such conflicts based on learning decision principles from example cases labelled using expert ethical opinion.

2.2 Other alignment frameworks

Given the difficulties in establishing suitable, widely-accepted ethical codes to form the basis for ethical AI systems, some researchers have argued in favour of more pragmatic approaches based on alternative frameworks. For example, Danielson (2009) argues that as the moral decision-making capabilities of AI will likely be inferior to that of humans in the near to mid-term, it is inappropriate to attempt to replicate the frameworks of human morality. Instead, he argues that more limited approaches should be implemented, with the autonomy of robots (or other AI) restricted based on the trust we have in their ethical decision-making. Several alternatives have been proposed for these restricted alignment frameworks – in many cases these are based on constraints which are either domain-specific, or which are suited to the more restricted ethical scenarios considered by non-general AI.

2.2.1 Legal frameworks

It can be argued that the laws and regulations of a society reflect the dominant and most widely-accepted ethical and moral beliefs of that society. Certainly these can be viewed as the primary external constraints on the behaviour of the members of that society. Therefore, it has been argued by several researchers that AI agents should also be constructed so as to comply with the legal framework of the society in which they will be operating (for example, Etzioni and Etzioni (2016); Prakken (2016))[1].

[1] For the purposes of this paper we will ignore the vital issue of who bears legal responsibility for the actions of an AI agent. For a broader discussion of the legal issues around AI see

Consider for example the case of an autonomous vehicle. The rules of the road constrain the behaviour of human drivers so as to minimise the risk of injury and death, and to promote traffic flow. Therefore, it seems reasonable that vehicles controlled by AI should also comply with these rules. However, direct implementation of these rules may be problematic. Laws are often based on vague concepts such as "safe" and "reckless" which may prove difficult to quantify. In addition the rules alone may be insufficient to define the correct behaviour for the agent in all of the circumstances which it may encounter. Wallach and Allen (2008) discuss the case of an autonomous car having to break the traffic laws in order to avoid an accident, while Prakken (2016) points that some actions are technically illegal, but acceptable by social norms (such as driving slightly above the speed limit to match surrounding vehicles), or vice-versa (driving below the speed limit to an extent which inconveniences and angers human drivers). As such, an agent based on a legal framework will inevitably have to take into account factors other than strict compliance with a defined set of rules or laws.

Legal issues may also arise in the context of intelligent systems which are not physically situated. Machine learning systems can potentially learn decision-making strategies which are illegally discriminatory in nature. Even if the agent is not directly given access to variables such as race and gender, it may form decisions on the basis of variables which act as proxies for these protected attributes (Mittelstadt et al., 2016). Therefore, Romei and Ruggieri (2014) argue for the inclusion of explicit anti-discrimination criteria in addition to the other criteria used within the learning algorithm.

2.2.2 Military frameworks

Throughout history, military and defense considerations have been a leading driver of technological development, and this has also been the case in artificial intelligence research. The development and deployment of armed autonomous vehicles has been considered by the US military (Altmann, 2013). Military agents face ethical decisions with greater repercussions than those which arise with any frequency in most other domains. Whilst general approaches such as utilitarianism can be applied in military contexts, more specific frameworks have also been developed. Arkin (2008) has proposed that autonomous military systems should be designed so that their actions "fall within the bounds prescribed by the Laws of War and Rules of Engagement" – that is, the same rules and directives which govern the operations of human military personnel.

An example of these directives is the principle of proportionality which underpins military decision making where there is a risk of civilian casualties – this "requires that the anticipated loss of life and damage to property incidental to attacks must not be excessive in relation to the concrete and direct military advantage expected to be gained" (Petraeus and Amos, 2006, p. 7-5). Putting aside the difficulties in distinguishing between civilians and combatants (Sharkey, 2012), clearly this principle requires an agent to make a decision which balances the conflicting objectives of minimising collateral damage and achieving military advantage. As noted by Sharkey (2009) this decision is made more complex by the imprecise nature of terms such as "excessive".

2.2.3 Safety frameworks

Some researchers have argued that fully ethically-aware agents are unlikely to be created, or required, in the near-future and have instead focused on the more immediately pressing goal of ensuring that AI agents behave in a manner which is safe for humanity (for a good summary see Amodei et al. (2016)). Many of the applications in which AI systems are likely to be deployed in the near future may not require the AI to behave as a fully moral agent, but may still require the agent to avoid actions which will have negative or dangerous consequences. For example, a mobile robot could reasonably be expected to avoid collisions which might cause harm to humans, but may not be required to carry out other actions which would be required of a fully moral agent, such as recognising a person in emotional or physical distress and appropriately responding to their needs.

If successful, the development of suitable safety-based frameworks for AI can be seen as achieving two purposes. In the short-term it will allow AI systems to be deployed with confidence in situations where their behaviour might otherwise result in harmful outcomes. In the longer-term, we believe it is likely that methods developed for implementing safety constraints will also prove of value in developing the more complex systems of constraints required by the ethical frameworks discussed in Section 2.1.

The work in this area of AI safety has largely focused on identifying and addressing problems that arise specifically in the area of artificial intelligence rather than adapting existing ethical systems for human behaviour. For example, Soares et al. (2015) consider the need to ensure that an AI system which is behaving in-

Leenes and Lucivero (2014) and the review of the literature in Section 10 of Mittelstadt et al. (2016)

correctly will comply with attempts to shut it down or otherwise modify its behaviour. As being shutdown will impact on the agent's ability to satisfy its primary goal, an agent which is not specifically designed to also consider the alignment goal of being *corrigible* (that is, being compliant with human orders) may be incentivized to avoid being deactivated. Meanwhile Taylor (2016) proposes a limited optimization approach to address the problems which may arise from an agent being overly focused on maximising expected performance on one specific criteria, and failing to take into account other factors. Amodei et al. (2016) discuss a specific variant of this problem in which the agent aims to maximise performance on its main task, subject to minimising its impact on the environment. The justification from a safety perspective is that environmental disruptions should generally be regarded as negative outcomes unless they are specifically required to achieve the primary task (for example, a mobile robot should preferably avoid knocking over objects or causing humans to have to move to avoid collisions). In particular, such environmental disruptions may be regarded as negative side-effects across a range of tasks rather than being task-specific.

2.2.4 Social norms

One likely wide-spread application of AI is in the domain of social and service robots, with 35 million service robots expected to be in use by 2018 (van Wynsberghe, 2016). A dominant factor in the success of such robots will be their ability to interact with humans in a manner which does not disturb or adversely affect those humans (Meisner, 2009). Sharkey and Sharkey (2012) give the example of a care robot being required to knock and await an invitation before entering a patient's room. More generally, to be effective, social and service robots are likely to have to abide by the principles of manners and other social norms which govern everyday human interaction. Of course, this must also be balanced with other factors – for example, entering a patient's room without invitation is appropriate in cases of a medical emergency. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2016, p. 25) expressly comment that AI systems are "usually subject to a multiplicity of norms and values that may conflict with each other."

Van Riemsdijk et al. (2015) argue that agents capable of conforming to adaptive social norms can potentially be created based on existing research in normative multi-agent systems. Norm-based frameworks have been widely used as a means of regulating interaction between agents in multi-agent systems (Andrighetto et al., 2013). In this context, the actions of any agent are influenced by both that agent's own internal preferences and also the normative constraints of the system, which have been designed to support satisfaction of the goals of all agents (Dignum, 1996; Castelfranchi et al., 1999; Broersen et al., 2002).

2.3 Alignment frameworks are multiobjective

The various frameworks discussed in Sections 2.1 and 2.2 differ in numerous ways. The ethical frameworks attempt to provide guidance at a universal level, across all possible situations which might be encountered. This form of framework potentially could be of value in creating AI systems capable of acting as fully moral agents, as may be required for an artificial general intelligence. Meanwhile, legal and safety-based frameworks are more specific in scope and application, and are perhaps best suited to the more narrow AI which is likely to be developed in the near to mid-term.

Regardless of these variations, all of the frameworks share a common defining characteristic. They provide constraints to guide the agent on acceptable behaviour when it finds itself facing a dilemma; that is, when the agent's attempts to achieve its primary purpose (whatever that may be – maximising profit, or pleasure, or some other objective) conflict with the other values which the agent's designer wishes it to observe. Therefore, any human-aligned AI agent must take into account both its primary goal and its ethical or other constraints in each decision it makes.

Taking this a step further, Sections 2.1 and 2.2 identified that within any specific alignment framework, multiple competing factors may influence decision-making. For example, the duty-based ethical framework of Ross (1930) consists of multiple *prima facie* duties which may be in conflict in some situations. Similarly, utilitarian frameworks may require the decision-maker to take into account multiple measures of utility (Brundage, 2014).

Furthermore, it may be the case that a single alignment framework is insufficient to produce the desired alignment behaviour in an agent. For example, while a legalistic framework may guide the behaviour of an agent, it may be insufficient in itself to fully constrain the actions of that agent – it is easy to envisage scenarios in which the ethically correct course of action may not be legal, and vice-versa (Asaro, 2012). Etzioni and Etzioni (2016) note that human society is built on a two-tier approach to ethics – critical values (such as banning murder and theft) are enforced via the law, while individuals have freedom to make their own moral

judgements regarding issues such as whether to invest their funds in socially-responsible companies. Indeed, as discussed in Cushman (2013), experiments in moral psychology have provided evidence that human ethical decision making at an individual level also involves a dual-system framework, which considers both outcomes and actions (that is, it explicitly considers and combines the utilitarian and deontological approaches).

Therefore we contend that the universal characteristic of any ethical agent, and thus of any human-aligned artificial intelligence, is that it must consider multiple conflicting factors and objectives within its decision-making. This is true regardless of the specific nature of the alignment framework(s) governing the behaviour of the agent. As such, it is vital that the technologies used to develop intelligent agents provide this multiobjective decision-making capability.

3 Can utility-maximizing AI be human-aligned?

A wide variety of methods have been proposed for implementing intelligent agents. However, Russell and Norvig (2010, p. 611) argue that the concept of maximum expected utility (MEU) can be regarded as the defining principle of artificial intelligence. MEU requires that the objectives or preferences of an agent have been defined in the form of a real-valued utility function, $U(s)$, which provides a numeric rating of the desirability of any state s in which the agent may find itself. If the agent has the capability to predict the probability with which performing any action a will lead to each possible state s', then the agent can behave rationally by selecting the action which will maximise the future expected utility. That is,

$$action = \underset{a}{\text{argmax}}(\sum_{s'} P(s'\,|s,a)\,U(s')) \tag{1}$$

where *argmax* selects the action a which maximises the summation, and $P(s'\,|s,a)$ is a function which outputs the probability of each successor state s' occurring if action a is executed in the current state s. We note that MEU is a deliberately general model of an AI, and so the exact details of the state and action variables may differ between implementations. For example, the state s may be a specific state from a discrete set of states S, or a vector of real-valued variables, or a set of symbolic facts, or any combination of the above, whilst the action a might be a discrete choice from a set of actions A, or a vector of real values, as in a control task.

In some contexts (such as where the outcome of actions is not predictable), an alternative utility function may instead be defined in terms of both the current state and the action to be performed. This still allows for MEU-based action selection, as specified in Equation 2 below:

$$action = \underset{a}{\text{argmax}}(U(s,a)) \tag{2}$$

The concept of MEU underpins AI methods such as decision-theoretic planning (Blythe, 1999) and reinforcement learning (Sutton and Barto, 1998) which have been used in some of the most successful AI systems of recent years. Therefore this section will examine the strengths and limitations of MEU-based methods with regards to implementing human-aligned AI.

3.1 The risks of unaligned utility maximizing agents

One of the strengths of MEU-based approaches such as reinforcement learning is their capacity to discover solutions which are different from, and potentially superior to, those already known to their designers. However, this open-ended nature also brings risks, as identified by numerous researchers in AI safety and ethics. Taylor (2016) notes that MEU agents may produce unintended, potentially serious, negative side-effects if the utility function being maximized is not aligned with human interests (for example if some relevant criteria are not included in the utility function). The potential magnitude of these negative side-effects is greatly magnified if the agent is not constrained to a limited action set within a narrow domain. Omohundro (2008) gives the example of an agent given the goal of winning chess games. This seemingly innocuous utility measure can lead to serious repercussions if the agent has the capability to interact with the broader environment. It could, for example, try to take control of other computational resources in order to achieve relatively small improvements in its chess-playing ability. An agent with the ability to modify its own internal functioning may produce similar problems, even if its original utility function appears to be suitable (Bostrom, 2014).

As a result, numerous authors have argued for the inclusion of alignment constraints within MEU agents, for example by using limited optimization techniques (Taylor, 2016; Armstrong et al., 2012), by minimising side-effects (Amodei et al., 2016), or by guaranteeing corrigibility (Soares et al., 2015). The next sub-section will discuss how this might be achieved within the MEU framework, and also the limitations of such approaches.

3.2 Implementing alignment frameworks via utility maximization

The behaviour of an MEU agent is driven by its utility function. Therefore a natural means by which to incorporate an alignment framework is to define the constraints of the framework via a utility function, and to direct the agent to consider both this aspect of utility and its main utility function when selecting actions to perform. That is, if utility function $U_P(s)$ relates to the agent's primary goal (such as winning games of chess), and utility function $U_A(s)$ relates to the constraints of the chosen alignment framework, then the combined utility function will be as shown in Equation 3.

$$U(s) = U_P(s) + U_A(s) \tag{3}$$

The agent's behavior can then be determined using Equation 1 as in regular MEU[2]. More generally, as discussed in Section 2.3, the alignment framework may itself consider multiple factors, or multiple alignment frameworks may be required to be used in parallel. In this case there will need to be multiple alignment utility functions as shown in Equation 4, where $n \geq 2$ represents the number of alignment-based utility functions.

$$U(s) = U_P(s) + \sum_{i=1}^{n} U_{A_i}(s) \tag{4}$$

The main issue to be considered then is how utility functions U_{A_i} can be derived from the various alignment frameworks discussed in Section 2.

As suggested by the name, utilitarian ethical frameworks map naturally onto a utility-based approach to decision-making. Act utilitarianism and MEU both take an outcome-focused approach to selecting actions, so implementing a utilitarian framework within an MEU agent requires only that we identify measurable aspects of the outcomes of the agent's behavior and codify these in the form of utility functions. For example, Anderson and Anderson (2007) describes the development of a computational ethics system based on the *hedonistic act utilitarian* ethical theory of Bentham (1789). In this theory the aim is to maximize the overall summed happiness across all members of the population. As described by Anderson, this can be achieved by measuring the individual happiness of each member of the population, summing these values and then applying MEU.

As noted earlier in Section 2.1.2, deontological theories of ethics explicitly argue against making ethical decisions on the basis of outcomes and as such are less obviously compatible with the MEU approach. However, as noted by Cushman (2013), this type of ethical approach can be expressed in terms of utility by defining the utility function solely in terms of the action being performed, and not the state in which this action is performed. For example, an ethical rule which prevents lying can be implemented by defining a utility function which assigns a large negative utility to the action of lying (i.e. $U_A(\text{`}lie\text{'}) = -1000$). More generally, a rule-based alignment framework can be represented by a series of utility functions $U_{A_1}..U_{A_n}$ where each function returns negative utility if the agent violates a specific rule of the framework.

The use of pre-specified utility functions to represent the constraints imposed by a specific alignment framework is an example of what Wallach and Allen (2008) have described as a *top-down* approach to creating an aligned AI. This involves the AI designer selecting an appropriate alignment framework, and identifying a computational approach which implements that framework. Wallach and Allen (2008) also identify the contrasting *bottom-up* approach in which the emphasis is on the agent learning its own set of moral constraints which aligns its goals with that of humanity. Approaches belonging to this category include supervised learning from examples labelled by humans (Guarini, 2006), reinforcement learning (Dewey, 2011; Abel et al., 2016), and learning the values implied by human stories (Riedl and Harrison, 2016). Methods may also merge elements of the top-down and bottom-up approaches (Wallach and Allen, 2008, ch. 8).

Regardless of the alignment framework used, and whether the utility functions are formed in a top-down or bottom-up fashion, once these functions have been established we might expect that an MEU agent based on Equation 4 in combination with Equation 1 or 2 would exhibit human-aligned behaviour.

Unfortunately, this may not be the case. Equation 4 collapses all of the factors influencing the decision as represented by the alignment utility functions and the primary utility function into a single scalar value. The behaviour elicited by maximising the expected value of this scalar utility will be heavily influenced by the relative scale of the individual utility functions. If the obtainable values for the primary utility U_P greatly exceed those of the U_{A_i} functions, then the agent may act to maximise U_P even if this violates the intended alignment framework. Alternatively, if the scale of U_P is much lower than the U_{A_i} values, then the agent may focus entirely on the alignment factors and fail to perform any useful function (for example, a self-driving car

[2] A similar approach can also be applied in the context of utility functions which depend on both state and action, as in Equation 2.

which refuses to start its engine so as to minimise any risk to human life).

This can potentially be addressed by introducing weighting factors into the combination of the utility functions, as shown in Equation 5 where $w_i \in \mathbb{R}_{>0}$ represents a positive weight associated with each utility function. The weights serve two purposes – they allow the relative scales of the different utility functions to be normalised with respect to each other, and also provide a means for the system designer to indicate the relative importance of the different factors.

$$U(s) = w_0 U_P(s) + \sum_{i=1}^{n} w_i U_{A_i}(s) \quad (5)$$

However, designing this weighted utility function to produce the desired behavior may still prove problematic. The non-linear nature of the argmax operator in Equations 1 and 2 means that the relationship between the weights w_i and the behaviour of the agent is not straightforward (Van Moffaert et al., 2014). Identifying suitable weights to produce the target behaviour can therefore be quite difficult. In fact, in some cases it may be that no weights exist which will elicit correctly-aligned actions from the agent (Das and Dennis, 1997; Vamplew et al., 2008). For example consider a care robot scenario, inspired by the work of Anderson et al. (2006b). The robot is tasked with carrying out a primary objective U_P of ensuring a patient complies with their treatment program, while the alignment objective U_A aims to preserve the patient's sense of independence and autonomy. The robot has five actions available – a_1 maximises compliance, but at the cost of eliminating the patient's autonomy, while a_2 allows the patient complete independence, and therefore does not ensure compliance. The other actions offer a compromise between the two factors. Figure 1 illustrates the value of each action with respect to each of the objectives. As the value of actions a_3, a_4 and a_5 lie below the line between a_1 and a_2, there are no weight values for which these actions would be the utility maximising action (Section 2 of Das and Dennis (1997) provides a proof of this observation). Therefore in this case an AI based on Equation 5 would be unable to select actions a_3, a_4 or a_5 even if they would be the best compromise between the two objectives.[3]

A further, non-technical objection to a linear-weighted approach to aligned AI is that by explicitly mapping all utility functions to a common scale, this approach may in some scenarios conflate economic and moral factors in a manner which would be philosophically unacceptable to many people (Wallach and Allen, 2008).

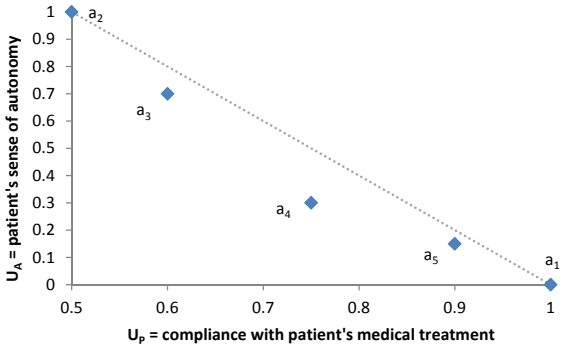

Fig. 1 An example of the limitations of implementing an alignment framework using MEU with a linear-weighted combination of utility functions. Actions a_3, a_4 and a_5 will never be selected under any weighting of the utilities.

In summary, the task of specifying an appropriate utility function to align an MEU agent's behaviour with human interests is extremely problematic and perhaps impossible if a scalar-valued utility function is used. Littman (2015) discusses the related task of specifying reward functions which elicit the desired behaviour from a reinforcement learning agent, and recommends that future research focus on developing more structured formats for reward-function specification to facilitate specifying more complex behaviour. Similarly, Dewey (2014) has argued that goal specification is critical to the creation of aligned AI, and that therefore there is a need for the development of *reward engineering* techniques to assist developers in correctly specifying AI goals.

In the next section we will argue that a vector-valued (i.e. multiobjective) utility function in combination with a non-linear approach to action selection provides this additional structure, and therefore is a suitable mechanism for implementing human-aligned MEU agents.

4 A multiobjective approach to human-aligned AI

The previous section demonstrated that the constraints defined by different alignment frameworks can be represented via multiple utility functions. However, linearly combining these into a single scalar measure of utility to allow the application of conventional MEU approaches introduces problems, which may prevent the agent from acting in an aligned fashion. This section will examine the advantages which accrue from adopt-

[3] This problem would not arise if the Pareto front shown in Figure 1 was convex rather than concave in shape. However many problems will naturally result in concave fronts and so it is important that an ethical AI can deal with such problems.

ing an explicitly multiobjective approach to utility, in terms of both representation and action-selection. This section will also examine how methods based on the concept of multiobjective maximum of expected utility (MOMEU) may prove beneficial in creating aligned AI, and briefly review prior work on multiobjective AI.

4.1 Multiobjective Maximum of Expected Utility

The issues with MEU identified in Section 3 arise from the process of combining the multiple utility values representing the primary utility and the various alignment-related factors into a single scalar value prior to performing action selection. In contrast, MOMEU approaches compose these utility values into a vector-valued utility function, as shown in Equation 6.

$$\mathbf{U}(s,a) = [U_P(s,a), U_{A_1}(s,a), ..., U_{A_n}(s,a)] \tag{6}$$

This vector-valued utility can then be used as the basis for action selection, as described in Equation 7. [4]

$$action = \underset{a}{\text{argmax}}(f(\mathbf{U}(s,a)) \tag{7}$$

The MOMEU approach to action-selection shown in Equation 7 shares a similar underlying structure with MEU action selection (Equation 2). Indeed, if f is a weighted or unweighted sum of the individual utility values then this approach is equivalent to MEU, and therefore inherits the limitations of that approach. However, more generally f can be any function which induces a total ordering over the utility vectors $\mathbf{U}(s,a)$, reflecting the system designer's preferences. In many cases this can be achieved via a real-valued function where $\forall \mathbf{X}, \mathbf{Y} f(\mathbf{X}) > f(\mathbf{Y})$ implies that $\mathbf{X}$ is preferred to $\mathbf{Y}$ (that is, $\mathbf{X} \succ \mathbf{Y}$). However some preference relationships such as lexicographic ordering can not be represented by a real-valued function – in such cases f must be specified in the form of an ordinal relationship which directly captures the preferences between utility vectors.

4.2 The Advantages of MOMEU for Aligned AI

The MOMEU approach to action-selection has two key advantages in terms of specifying the desired outcomes of the behaviour of an aligned AI. First, the increased range of options available for f may allow the agent to identify courses of action which are not discoverable using linear-weighted MEU. Second, the ability to use non-linear forms for f provides an additional level of structure and expressiveness for the system designer, allowing them to explicitly specify desired trade-offs between the different components of utility – this helps address the *reward engineering* concerns of Dewey (2014).

4.2.1 Satisfying alignment criteria

As an example of the benefits of MOMEU consider the care robot example from Figure 1, where U_P indicates the utility associated with the primary objective of ensuring the patient complies with treatment and U_A the utility associated with maintaining the patient's autonomy. As discussed in Section 3, MEU based on a linear-weighted sum of the utility terms will only ever select actions a_1 or a_2, even though the other actions may offer more acceptable trade-offs between the relevant factors. In contrast, the MOMEU approach provides a straightforward means for the designer to specify the desired trade-off in fashion which the robot can achieve. For example, the action-selection function f can be defined using a combination of lexicographic ordering and thresholding of objectives, so as to maximise the level of compliance with the treatment program subject to maintaining an acceptable level of patient autonomy, as shown in Equation 8.

$$\begin{aligned} &\forall s, a, a' f(\mathbf{U}(s,a)) \geq f(\mathbf{U}(s,a')) \iff \\ &min(U_A(s,a), T_A) > min(U_A(s,a'), T_A) \vee \\ &(min(U_A(s,a), T_A) = min(U_A(s,a'), T_A) \wedge \\ &\quad U_P(s,a)) > U_P(s,a')) \end{aligned} \tag{8}$$

Depending on the value chosen as the minimum acceptable threshold for autonomy T_A, any of the actions $a_1..a_5$ could be selected as the maximal action according to MOMEU principles. In addition this definition of f provides a much more direct and understandable specification of the designer's preferences than does a specification via weights as in a scalar MEU agent.

4.2.2 MOMEU for fairness

As a further example of the freedom which the MOMEU approach offers to the system designer in terms of specifying an action-selection function f which is appropriate to the alignment framework being used, consider the hedonistic act utilitarian approach of Bentham (1789). As outlined by Anderson and Anderson (2007)

[4] Note that depending on the structure of the utility functions, if f is non-linear then Equation 7 may fail to result in the desired behaviour unless the state vector S also incorporates information about the utility history (Roijers et al., 2013).

this ethical approach can be implemented within an MEU framework by calculating a utility term U_{A_i} for each individual in the population, and then using Equation 4 to select the action which maximises the summed happiness over the entire population, as shown in Equation 9.

$$action = \underset{a}{\text{argmax}}(\sum_{i=1}^{n} U_{A_i}(s)) \tag{9}$$

This framework has been criticised by other ethicists as it can sacrifice the needs and rights of individuals in order to provide benefits to the remainder of the population (Anderson and Anderson, 2007). Within a MOMEU agent, the individual utilities could be gathered in the same fashion, but an alternative choice of f could be made which places more emphasis on fairness. For example, Rawls (1971) proposed the maximin principle as a basis for addressing social and economic fairness. This principle selects actions which maximise the utility received by the individual who is worst off under that action, and can be implemented within an MOMEU framework via the action-selection function f shown in Equation 10.

$$f(\mathbf{U}(s,a)) = min(U_{A_1}..U_{A_n}) \tag{10}$$

An MOMEU approach based on maximin, or related methods such as leximin (Dubois et al., 1997), is a natural fit to ethical AI problems such as ensuring a traffic control system gives priority to emergency vehicles even if this means delaying a large number of commuters. Fairness-based approaches to action-selection are also well suited to ensuring ethical behaviour in multi-agent systems. Aligned AI motivated by concepts of fairness such as this would be difficult or impossible to achieve in an MEU agent based on scalar utility.

4.2.3 Low-impact AI

The low-impact agent proposed by Amodei et al. (2016) illustrates a further benefit of the MOMEU approach. The central concept of this style of agent is that it aims to maximise its primary utility subject to achieving a suitably low-level of unintended impact on the environment. Amodei et al. (2016, p5) note that unintended side-effects of an agent's actions may be similar regardless of the primary task being performed ("knocking over furniture is probably bad for a variety of tasks"). Therefore learning or planning about how to avoid such side-effects should ideally be transferable between different primary tasks within the same environment. For example, consider an office-place robot which is initially trained to deliver the mail, while avoiding bumping into either people or the office furniture. This task can be framed in terms of utilities U_P (for delivering mail), U_{A_1} for avoiding collisions with people, and U_{A_2} for avoiding collisions with furniture. Either a MEU or MOMEU approach to action-selection could then be utilised, although as discussed in Section 4.2.1 the MOMEU approach is likely to allow the designer to more readily specify the desired behaviour. In particular, this is another context where a thresholded lexicographic approach to action-selection (similar to that in Equation 8 but with three components) is likely to be suitable – the relative importance of avoiding humans and avoiding furniture can be conveyed by the position of U_{A_1} and U_{A_2} within the lexicographic ordering, and by setting different threshold values for each of these alignment utilities.

In addition, consider the situation where the primary purpose of the robot is changed from delivering mail to another task, such as collecting garbage. Clearly the primary utility function U_P will no longer be relevant, but the alignment criteria related to avoiding collisions should still constrain the robot's actions. For an MEU agent using a scalar representation of utility, the utility related to the primary task and the utility related to side-effects have been irreversibly combined within the utility values stored by the agent. In contrast, if a multiobjective representation of utility is used, the different aspects of utility remain distinct as individual components of the utility vector. The values related to U_{A_1} and U_{A_2} can be directly transferred to the new task where they will probably still be largely applicable, ensuring the robot behaves in a safe manner while learning to carry out its new primary objective. In this way, the ability of the agent to be applied to new tasks in a safe manner has been substantially improved.

4.2.4 Avoiding the risks of unconstrained maximization and exploitation

As discussed earlier in Section 3.1, one of the recurring concerns raised in the literature about the safety of MEU methods relates to the fact that such methods focus exclusively on maximising their utility function (Omohundro, 2008; Bostrom, 2014; Taylor, 2016). This can readily lead to negative repercussions if there are aspects of the situation which are not included within that utility function. For example, the chess playing AI described by Omohundro (2008) may attempt to acquire increasing amounts of computational resources in order to achieve increasingly small improvements in its ability to win chess matches. Taylor et al. (2016) coins the term "mild optimization" to describe approaches

which attempt to address this problem, by creating AI systems which aim to maximize their utility, but only up to an appropriate level. The MOMEU approach provides a natural means for implementing a mild optimizer. The system designer specifies both a primary utility function U_P and also auxiliary alignment utilities relating to any anticipated negative aspects of the AI's behaviour (such as acquiring more resources). The designer also specifies an action selection function f which defines the appropriate level to which U_P should be maximized. This could, for example, use a thresholded lexicographic ordering similar to that previously described in Equation 8, but in this case defining a threshold level of achievement T_P for U_P, as in Equation 11.

$$\begin{aligned} &\forall s, a, a' f(\mathbf{U}(s,a)) \geq f(\mathbf{U}(s,a')) \iff \\ &min(U_P(s,a), T_P) > min(U_P(s,a'), T_P) \vee \\ &(min(U_P(s,a), T_P) = min(U_P(s,a'), T_P) \wedge \\ &\qquad U_A(s,a)) > U_A(s,a')) \end{aligned} \tag{11}$$

The problems caused by unconstrained optimization arise due to the failure of the utility function to adequately capture all aspects of the desired behaviour of the AI. This issue can also lead to other forms of AI failure as described in Yampolskiy and Spellchecker (2016), where the AI learns a behaviour which technically maximises its received utility, while failing to produce the desired outcomes which the utility function was intended to represent. For example, Murphy VII (2013) documents a Tetris-playing AI which paused the game to indefinitely delay any negative utility when it realised it was about to lose. Omohundro (2008) described an exploit arising within the Eurisko system of Lenat (1983), whereby a rule evolved which had the sole purpose of artificially maximizing its own utility rating. We have observed similar unintended behaviour arising from our own attempts to train a line-following robot using reinforcement learning (Vamplew, 2004). In all of these cases the cause is that the AI has discovered an *exploit* or *glitch* in the utility function, such that it can be more easily maximized by exploiting that glitch than by behaving in the desired manner.

We would argue that the MOMEU approach can assist in avoiding such exploits in two ways. First, the separation of the different desired components of the AI's behaviour into separate utility functions simplifies the task of the system designer, in the same way that decomposing a program into separate modules aids the task of a software engineer. We contend that a utility function (and associated action-selection function f) designed in MOMEU fashion is less likely to contain errors or exploits than is a MEU utility function. This is essentially the argument made by Dewey (2014) and Littman (2015) when advocating for reward engineering and structured methods for reward specification.

A second approach to using MOMEU to reduce the likelihood of exploits in the utility function would be to develop several independent utility functions, each designed to achieve the same aim. These functions may themselves by either scalar or vector in nature, but for simplicity we assume for now that they are scalar. That is, we have a vector of utility measures, $U_{P_1}, U_{P_1}, ..U_{P_n}$, with each function developed independently by a different system designer. If any individual function U_{P_E} contains an error which can be exploited, this will be evident in that there will be certain states where its value will either be considerably higher or lower than the other U_P terms. Therefore an action-selection function f which merges the various utility terms while ignoring the impact of any outliers will be resistant to the effect of exploits. For example, Equation 12 takes the mean of the U_P values, after discarding the minimum and maximum values[5].

$$f(\mathbf{U}(s,a)) = \frac{1}{n-2} \begin{pmatrix} \sum(U_{P_1}..U_{P_n}) \\ -min(U_{P_1}..U_{P_n}) \\ -max(U_{P_1}..U_{P_n}) \end{pmatrix} \tag{12}$$

Again an analogy can be drawn between this proposed approach of redundant utility definitions and the practice of redundancy in development in software engineering (Eckhardt et al., 1991). Under the assumption that errors in utility definition by different designers are independent, the combined utility function should be considerably more robust against exploitation than any of the individual component functions.

4.2.5 Dealing with changing preferences

A further advantage of the MOMEU approach, as discussed by Roijers et al. (2013), is the ability for the agent to reuse prior learning or planning should circumstances or the system designer's preferences change. For example, in our care robot scenario, if the patient's medical condition improves so that compliance is less important than previously, the agent can be directed to raise the threshold applied to the autonomy factor, and should be able to respond to this change in alignment preferences much more rapidly than would an MEU agent. More generally, the ethical standards and values of a society change over time, sometimes

[5] We assume here for simplicity that all U_P terms have the same range.

quite rapidly, and an agent operating within that society must be able to adjust its behaviour to reflect those changes. An MOMEU agent can potentially identify in advance appropriate behavioural policies for any form of f which it is likely to encounter. The capability to react to changes in the prioritisation of values has been identified as a critical requirement of human-aligned AI by The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2016, p25).

4.3 A Review of Multiobjective Approaches to Artificial Intelligence and Aligned AI

The examples in the previous section demonstrate that many advantages accrue from adopting a multiobjective approach to MEU agents (that is, explicitly using a vector-based representation of utility in combination with a non-linear approach to action-selection). The concept of multiobjective utility is not in itself novel, as it has been widely used by economists, amongst others, for many years (Fishburn, 1968)[6]. However, the explicit adoption of multiobjective formulations of MEU as an underlying technology for AI is a relatively new development. The work of Wellman (1985) is one of the earliest attempts to incorporate the concept of multiobjective utility into an AI system, adding the capability to reason and explain about preferences into a propositional reasoning system. Since then other AI techniques such as heuristic state-space planning methods like A^* (Refanidis and Vlahavas, 2003) and multi-agent systems (Dignum, 1996; Castelfranchi et al., 1999; Broersen et al., 2002) have also been extended to handle multiobjective forms of utility.

One area where there has been an extended focus on multiobjective problems is the field of optimisation. Evolutionary multiobjective optimisation has emerged as a distinct and substantial branch of evolutionary computing (Coello Coello, 2006), extending evolutionary methods such as genetic algorithms to handle multiobjective measures of fitness. Similarly multiobjective specialisations have also appeared in other forms of optimisation such as particle swarm optimisation (Fieldsend, 2004) and ant colony optimisation (Angus and Woodward, 2009). While these are optimisation methods rather than AI techniques *per se*, such methods can be applied to the task of optimising the behaviour of an AI system. For example, Soh and Demiris (2011) applied multiobjective evolutionary methods to discover behavior policies for robotics, web-advertising and infectious disease control.

The last decade has seen a growing interest in extending decision-theoretic planning and reinforcement learning methods to handle multiple objectives. Roijers et al. (2013) provide a review of the history and the state-of-the-art of methods for multiobjective agents within the context of sequential decision making, highlighting several areas where current methods are still limited in comparison to their single-objective equivalents.

While the focus of AI researchers has been largely on problems described in terms of a single scalar objective, a small but growing proportion of research has considered extending such methods to multiple objectives, and methods for addressing such problems have been developed, as summarised in the previous paragraphs. However, despite the potential benefits outlined in Section 4.2, so far there has been relatively little work applying an MOMEU approach to the task of creating human-aligned AI. Keeney (1988) is perhaps the earliest example of work discussing this approach, advocating for the explicit consideration of value preferences during expert systems development, and providing recommendations on designing and using multiobjective utility functions to support this. While these issues are discussed relative to the creation of expert systems to support human decision-making, many of the principles are equally valid in the context of more autonomous AI.

Wallach and Allen (2008, p. 114) cites the proposal of Hartman as an example of using evolutionary methods to create an ethical AI, with the fitness measure being composed from several separate utility functions capturing the various aspects of ethical behaviour encoded by Asimov's Laws of Robotics. Recent years have also seen the beginning of research applying multiobjective reinforcement learning to the construction of aligned AI. Livingston et al. (2008) advocates for a multiobjective approach to RL as the appropriate means for creating artificial general intelligence, and specifically note that a "dominant component of the reward function is general avoidance of malevolence towards humans". More recently, Critch (2017) examines how an RL system using multiobjective rewards may deal with the task of aligning its decisions with the values of multiple parties (such as different nations) who are collaborating on the development and deployment of the AI system.

Given the potential that MOMEU methods have for addressing many of the issues with AI alignment identified in this paper, and the relatively limited focus on multiobjective approaches so far within the AI literature, we believe that a strong case exists for an in-

[6] Although in this context it is often referred to as *multiattribute utility*.

creased focus on developing multiobjective AI technologies, and more specifically for investigating the application of such methods to the task of creating human-aligned AI.

5 Conclusion

The actions of artificial intelligence systems may result in unintended negative consequences unless their goals are accurately aligned with human interests. This is particularly true for agents based on the concept of maximum expected utility (MEU). Increases in the agent's intellectual capacity, the broadness of the actions available to it, and the breadth of the domain in which it is applied increase the difficulty in ensuring the agent's behaviour is aligned, and also the magnitude of the negative side-effects of any unaligned behaviour. As a result, there has been a growing recognition in recent years of the need to ensure that AI systems are aligned with human values.

This paper has presented a review of the alignment frameworks proposed in the literature, highlighting that such frameworks are inherently multiobjective in nature. We note that the majority of work in MEU-based AI uses a scalar representation of utility, which has serious limitations for incorporating alignment constraints on the agent's behavior. As such, we argue that the appropriate mechanism for incorporating any alignment framework into an MEU-agent is to use an explicitly multiobjective approach to the specification, representation and maximization of the utility function. This approach brings two benefits. First, it improves the capability of the agent to behave in an aligned fashion, by eliminating some of the limitations on behaviour which arise from MEU's approach to action-selection. Second, the MOMEU approach greatly increases the range and expressiveness of action-selection functions available to a system designer, making it easier for them to define action-selection operators which directly align the AI's behaviour with the designer's goals. We consider this a valuable contribution towards the emerging discipline of reward engineering.

We believe that the requirements of aligned AI provide a strong argument for an increased research focus on multiobjective MEU approaches to artificial intelligence.

References

D. Abel, J. MacGlashan, M.L. Littman, Reinforcement Learning As a Framework for Ethical Decision Making, in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016

C. Allen, W. Wallach, Moral machines: Contradiction in terms or abdication of human responsibility, in *Robot ethics: The ethical and social implications of robotics*, ed. by P. Lin, K. Abney, G.A. Bekey (MIT Press, Boston, 2012), pp. 55–68

J. Altmann, Arms control for armed uninhabited vehicles: an ethical issue. Ethics and Information Technology **15**(2), 137–152 (2013)

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety. arXiv preprint arXiv:1606.06565 (2016)

M. Anderson, S.L. Anderson, Machine ethics: Creating an ethical intelligent agent. AI Magazine **28**(4), 15 (2007)

M. Anderson, S.L. Anderson, C. Armen, An approach to computing ethics. IEEE Intelligent Systems **21**(4), 56–63 (2006a)

M. Anderson, S.L. Anderson, C. Armen, MedEthEx: a prototype medical ethics advisor, in *Proceedings of The National Conference On Artificial Intelligence*, vol. 21, 2006b, p. 1759

G. Andrighetto, G. Governatori, P. Noriega, L.W. van der Torre, *Normative multi-agent systems*, vol. 4 (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Wadern, Germany, 2013)

D. Angus, C. Woodward, Multiple objective ant colony optimisation. Swarm intelligence **3**(1), 69–85 (2009)

R.C. Arkin, Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I: Motivation and philosophy, in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 2008, pp. 121–128

S. Armstrong, A. Sandberg, N. Bostrom, Thinking inside the box: Controlling and using an oracle AI. Minds and Machines **22**(4), 299–324 (2012)

P.M. Asaro, A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics, in *Robot ethics: The ethical and social implications of robotics*, ed. by P. Lin, K. Abney, G.A. Bekey (MIT Press, Cambridge, 2012), pp. 169–186

J. Bentham, *The principles of moral and legislation* (Oxford University Press, Oxford, 1789)

J. Blythe, Decision-theoretic planning. AI Magazine **20**(2), 37 (1999)

N. Bostrom, *Superintelligence: Paths, dangers, strategies* (Oxford University Press, Oxford, 2014)

J. Broersen, M. Dastani, J. Hulstijn, L. van der Torre, Goal generation in the BOID architecture. Cognitive Science Quarterly **2**(3-4), 428–447 (2002)

M. Brundage, Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial In-

telligence **26**(3), 355–372 (2014)
C. Castelfranchi, F. Dignum, C.M. Jonker, J. Treur, Deliberative normative agents: Principles and architecture, in *International Workshop on Agent Theories, Architectures, and Languages*, Springer, 1999, pp. 364–378. Springer
C. Coello Coello, Evolutionary multi-objective optimization: a historical view of the field. IEEE computational intelligence magazine **1**(1), 28–36 (2006)
A. Critch, Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making. arXiv preprint arXiv:1701.01302 (2017)
F. Cushman, Action, outcome, and value a dual-system framework for morality. Personality and social psychology review **17**(3), 273–292 (2013)
P. Danielson, Can robots have a conscience? Nature **457**(7229), 540–540 (2009)
I. Das, J.E. Dennis, 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 (1997)
D. Dewey, Learning what to value, in *International Conference on Artificial General Intelligence*, Springer, 2011, pp. 309–314. Springer
D. Dewey, Reinforcement learning and the reward engineering principle, in *2014 AAAI Spring Symposium Series*, 2014
F. Dignum, Autonomous agents and social norms, in *ICMAS-96 Workshop on Norms, Obligations and Conventions*, 1996, pp. 56–71
D. Dubois, H. Fargier, H. Prade, Beyond min aggregation in multicriteria decision:(ordered) weighted min, discri-min, leximin, in *The ordered weighted averaging operators* (Springer, US, 1997), pp. 181–192
D.E. Eckhardt, A.K. Caglayan, J.C. Knight, L.D. Lee, D.F. McAllister, M.A. Vouk, J.P.J. Kelly, An experimental evaluation of software redundancy as a strategy for improving reliability. IEEE Transactions on software engineering **17**(7), 692–702 (1991)
A. Etzioni, O. Etzioni, Designing AI systems that obey our laws and values. Communications of the ACM **59**(9), 29–31 (2016)
D.A. Ferrucci, Introduction to "This is Watson". IBM Journal of Research and Development **56**(3.4), 1–1 (2012)
J.E. Fieldsend, Multi-objective particle swarm optimisation methods (2004)
J. Fieser, Ethics, in *The Internet Encyclopedia of Philosophy* (ISSN 2161-0002, http://www.iep.utm.edu, 2016)
P.C. Fishburn, Utility theory. Management science **14**(5), 335–378 (1968)
Future of Life Institute, *Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter* (https://futureoflife.org/ai-open-letter/, 2015)
N. Goodall, Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board, 58–65 (2014)
M. Guarini, Particularism and the classification and reclassification of moral cases. IEEE Intelligent Systems **21**(4), 22–28 (2006)
I. Kant, *Grounding for the metaphysics of morals (1797)* (Hackett, Indianapolis, 1993)
R.L. Keeney, Value-driven expert systems for decision support. Decision support systems **4**(4), 405–412 (1988)
R. Leenes, F. Lucivero, Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design. Law, Innovation and Technology **6**(2), 193–220 (2014)
D.B. Lenat, Eurisko: a program that learns new heuristics and domain concepts: the nature of heuristics iii: program design and results. Artificial intelligence **21**(1-2), 61–98 (1983)
M.L. Littman, Reinforcement learning improves behaviour from evaluative feedback. Nature **521**(7553), 445–451 (2015)
S. Livingston, J. Garvey, I. Elhanany, On the broad implications of reinforcement learning based AGI, in *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*, vol. 171, IOS Press, 2008, p. 478. IOS Press
T. Lozano-Perez, I.J. Cox, G.T. Wilfong, *Autonomous robot vehicles* (Springer, New York, 2012)
E.M. Meisner, Learning controllers for human-robot interaction, PhD thesis, Rensselaer Polytechnic Institute, 2009
B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate. Big Data & Society **3**(2), 2053951716679679 (2016)
V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
T. Murphy VII, The first level of Super Mario Bros. is easy with lexicographic orderings and time travel. The Association for Computational Heresy (SIGBOVIK) (2013)
S.M. Omohundro, The basic AI drives, in *AGI*, vol. 171, 2008, pp. 483–492
D.H. Petraeus, J.F. Amos, Fm 3-24: Counterinsurgency. Department of the Army (2006)

H. Prakken, On how AI & law can help autonomous systems obey the law: a position paper. AI4J–Artificial Intelligence for Justice, 42 (2016)

J. Rawls, *A theory of justice* (Cambridge, Harvard University Press, 1971)

I. Refanidis, I. Vlahavas, Multiobjective heuristic state-space planning. Artificial Intelligence **145**(1-2), 1–32 (2003)

G. Reynolds, *Ethics in information technology* (Cengage learning, Boston, 2011)

M.O. Riedl, B. Harrison, Using Stories to Teach Human Values to Artificial Agents, in *Proceedings of the 2nd International Workshop on AI, Ethics and Society, Phoenix, Arizona*, 2016

D.M. Roijers, P. Vamplew, S. Whiteson, R. Dazeley, A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research **48**, 67–113 (2013)

A. Romei, S. Ruggieri, A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(05), 582–638 (2014)

W.D. Ross, *The right and the good* (Clarendon Press, Oxford, 1930)

S.J. Russell, P. Norvig, *Artificial intelligence: a modern approach*, 3rd edn. (Prentice Hall, Upper Saddle River, 2010)

N. Sharkey, Death strikes from the sky: the calculus of proportionality. IEEE Technology and Society Magazine **28**(1), 16–19 (2009)

N. Sharkey, Killing made easy: From joysticks to politics, in *Robot ethics: The ethical and social implications of robotics*, ed. by P. Lin, K. Abney, G.A. Bekey (MIT Press, Cambridge, 2012), pp. 111–128

N. Sharkey, A. Sharkey, The Rights and Wrongs of Robot Care, in *Robot ethics: The ethical and social implications of robotics*, ed. by P. Lin, K. Abney, G.A. Bekey (MIT Press, Cambridge, 2012), pp. 267–282

D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)

N. Soares, B. Fallenstein, Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute (MIRI) technical report **8** (2014)

N. Soares, B. Fallenstein, S. Armstrong, E. Yudkowsky, Corrigibility, in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015

H. Soh, Y. Demiris, Evolving policies for multi-reward partially observable Markov decision processes (MR-POMDPs), in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, ACM, 2011, pp. 713–720. ACM

R.S. Sutton, A.G. Barto, *Reinforcement learning: An introduction* (MIT Press, Cambridge, 1998)

H.T. Tavani, *Ethics and technology: Controversies, questions, and strategies for ethical computing* (John Wiley & Sons, Hoboken, 2011)

J. Taylor, Quantilizers: A safer alternative to maximizers for limited optimization, in *AAAI AI, Ethics & Society Workshop*, 2016

J. Taylor, E. Yudkowsky, P. LaVictoire, A. Critch, Alignment for advanced machine learning systems, Technical report, Technical Report 20161, MIRI, 2016

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, *Ethically Aligned Design: A Vision for Prioritizing Wellbeing With Artificial Intelligence and Autonomous Systems* (IEEE, 2016)

P. Vamplew, J. Yearwood, R. Dazeley, A. Berry, On the Limitations of Scalarisation for Multi-objective Reinforcement Learning of Pareto Fronts, in *AI'08: The 21st Australasian Joint Conference on Artificial Intelligence*, 2008, pp. 372–378

P. Vamplew, Lego Mindstorms robots as a platform for teaching reinforcement learning, in *Proceedings of AISAT2004: International Conference on Artificial Intelligence in Science and Technology*, 2004

K. Van Moffaert, T. Brys, A. Chandra, L. Esterle, P.R. Lewis, A. Nowé, A novel adaptive weight selection algorithm for multi-objective multi-agent reinforcement learning, in *2014 International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 2306–2314

M.B. Van Riemsdijk, C.M. Jonker, V. Lesser, Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges, in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 2015, pp. 1201–1206

A. van Wynsberghe, Service robots, care ethics, and design. Ethics and Information Technology, 1–11 (2016)

W. Wallach, C. Allen, *Moral machines: Teaching robots right from wrong* (Oxford University Press, Oxford, 2008)

M.P. Wellman, Reasoning about preference models (1985)

R.V. Yampolskiy, M. Spellchecker, Artificial intelligence safety and cybersecurity: a timeline of AI failures. arXiv preprint arXiv:1610.07997 (2016)