I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16000

Artificial Intelligence and the Economics of Decision-Making

Wim Naudé

MARCH 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16000

Artificial Intelligence and the Economics of Decision-Making

Wim Naudé
RWTH Aachen University and IZA

MARCH 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Artificial Intelligence and the Economics of Decision-Making

Artificial Intelligence (AI) scientists are challenged to create intelligent, autonomous agents that can make rational decisions. In this challenge, they confront two questions: what decision theory to follow and how to implement it in AI systems. This paper provides answers to these questions and makes three contributions. The first is to discuss how economic decision theory – Expected Utility Theory (EUT) – can help AI systems with utility functions to deal with the problem of instrumental goals, the possibility of utility function instability, and coordination challenges in multi-actor and human-agent collectives settings. The second contribution is to show that using EUT restricts AI systems to narrow applications, which are "small worlds" where concerns about AI alignment may lose urgency and be better labelled as safety issues. This papers third contribution points to several areas where economists may learn from AI scientists as they implement EUT. These include consideration of procedural rationality, overcoming computational difficulties, and understanding decision-making in disequilibrium situations.

JEL Classification: D01, C60, C45, O33

Keywords: economics, artificial intelligence, expected utility theory, decision-theory

Corresponding author:
Wim Naudé
RWTH Aachen University
Kackertstraße 7
52072 Aachen
Germany

E-mail: naude@time.rwth-aachen.de

1 Introduction

The challenge facing Artificial Intelligence (AI) scientists is to create intelligent, autonomous agents that can make rational decisions. This challenge has confronted them with two questions (Oesterheld, 2021, p.2): "What decision theory do we want an AI to follow? and how can we implement such a decision theory in an AI?"

This paper provides a critical overview of how the economic theory of decision-making has helped to answer these two questions, and how it can benefit from the practical solutions that AI scientists are working on. The main contribution is to identify how economists can contribute to solving the AI alignment problem, and provide a fresh perspective on the alignment problem. AI systems are said to be aligned when they do what they are supposed to do, and no harm. They are said to be value-aligned when they share human values. The alignment problem has so far largely attracted computer scientists, programmers and philosophers. Economists have so far contributed little (Gans, 2018).

The paper is structured as follows. In section 2 the first question above, what decision theory should AI follow? is answered. The foundation of both modern data-based AI and economics, Expected Utility Theory (EUT), is outlined, examples are given of its adoption in AI, including in sequential decision-making, and the challenges to EUT discussed. In section 3, the second question, how can we implement EUT in AI systems? is answered. Here, the alignment problem is stated, and three ways in which to approach AI alignment in the field of economics explained: instrumental goals, utility function instability and utility function coordination.

From the discussions in sections 2 and 3 it follows that the application of EUT to current, narrow, AI results in artificial smart agents in very simple situations, so that concerns about AI alignment may, from this perspective lose some urgency - the problems that AI do pose, are perhaps better labelled safety issues, rather than alignment issues.

Finally, this paper concludes in section 4 that it is not only AI that benefits from the economic theory of decision-making: economics may also benefit from how AI scientists implement EUT - AI may help economics to model human decision-making better.

2 What decision theory do we want an AI to follow?

> "*AI researchers aim to construct a synthetic homo economicus, the mythical perfectly rational agent of neoclassical economics*" (Parkes and Wellman, 2015, p. 267)

AI scientists want AI systems to make rational decisions. Thus, they have resorted to rational choice theory - on which economics is based. Rational choice theory is not a single unified theory but consists of variants. For a review of these, see Herfeld (2020). In economics, rational choice theory, and specifically Expected Utility Theory (EUT), is used to model human decision-making. The decision-makers who strictly follow EUT have been labelled *homo economicus*[1] to make clear that real human decision-making tend to depart from some of the assumptions of the theory. As will be discussed below, AI systems may more closely resemble homo economicus than homo sapiens.

2.1 Expected utility theory

EUT is attractive for AI scientists because "All tasks that require intelligence to be solved can naturally be formulated as a maximization of some expected utility in the framework of agents" (Hutter, 2007, p.33). In economics, intelligent human agents are modelled as goal-oriented, rational agents acting to maximize their subjective utility subject to resource

[1]For a discussion of the concept and origin of *homo economicus*, the economic human, see Persky (1995).

constraints. The field AI has adopted this approach, which is reflected in a standard definition of the field of AI as "the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions [...]" (Russell and Norvig, 2021, p.vii).

The functions that map percept sequences (perceptions) to actions should help agents to select actions to achieve their goals (Parkes and Wellman, 2015). In the case of genes, for instance, the goals are survival and gene transmission (Kamatani, 2021). The human phenotype, including its brain, is the expression or action of its genes, which aims at survival and transmission (reproduction) (Williams, 1966; Dawkins, 1976). In the terminology of AI, survival and reproduction are the supergoals of genes, and the human brain is a subgoal (or instrumental subgoal) (Yudkowsky, 2001). Below we will return to the topic of subgoals/instrumental goals.

Expected Utility Theory (EUT) and qualitative variants thereof (Dastani et al., 2005; Gonzales and Perny, 2020; Russell, 2019) is a formal model of rational decision-making set out by Von Neumann and Morgenstern (1944) (vNM) and generalized by Savage (1954).

The foundations of EUT go back however to Bernoulli (1738) and his solution to the St. Petersburg Paradox (List and Haigh, 2005). The St.Petersburg Paradox arises in gamble where a fair coin is tossed n-times, until it lands on heads, with the gambler then receiving a prize of $\$2^n$. The paradox is that even though the expected value of the gamble being $\sum_{n=1}^{\infty}(\frac{1}{2})^n \times 2^n = \infty$, no one would pay very much to take this gamble. Bernoulli (1738) solved this by showing that what is important is not to maximize expected *value*, but expected **utility**. Utility should also be maximized after ignoring outcomes with very small probabilities - otherwise this would lead to the problem of "Pascal's Mugger[2] (Monton, 2019).

In economics, and based on Von Neumann and Morgenstern (1944), EUT justifies the spec-

[2]See e.g. Yudkowsky (2007).

ification of an *Utility Function* which allows an agent to compare different outcomes from actions - the utility function reflects the agent's preferences.[3] Consequently, the agent will choose actions that maximises the (expected) value of the utility function. Note that in this approach, agents maximise *expected* utility because each possible future outcome is subject to probability - an outcome is a lottery. Actions and their outcomes can thus be compared to playing a lottery.

A lottery can be denoted by $L = [p_1(C_1), p_2(C_2), p_3(C_3), ..., p_k(C_k)]$ where the $C_k's$ are the outcomes and $\sum_{i=1}^{k} p_i = 1$ the probabilities of each outcome. The expected value (E) of this lottery is $E(L) = \sum_{i=1}^{k} p_i C_i$, the mathematical average. If the set of lotteries available to an agent i is Λ then the agent's utility function[4] U_i represents the preferences of the agent over various lotteries, with $U_i(L_1) \geq U_i(L_2) \iff L_1 \succsim L_2, \forall L_1, L_2 \in \Lambda$ (Maschler et al., 2013). Thus, if lottery L_1 is preferred to lottery L_2 the utility from L_1 will be greater or equal than the utility from L_2.

In following EUT to make decisions, an agent will do best to choose the lotteries L_i^* (or consumption bundles, as in household economics) to maximise expected utility, $\mathbb{E}[U(L_i)]$. This decision can be written as

$$L_i^* = \arg\max_{L} \mathbb{E}[U(L_i)] \tag{1}$$

This choice of L_i^* can be found by solving for $\dfrac{\partial \mathbb{E}[U(L_i)]}{\partial L_i} = 0$

In vNM, human agents will maximize expected utility, choosing the lotteries from Λ that will generate the most utility. This will, say in the example of a consumer aiming to maximize utility from buying various bundles of goods, lead them to goal-directed decisions and pursuit

[3] For a survey of how preferences are incorporated in the utility function of AI agents, see Pigozzi et al. (2016).

[4] A linear utility function is a von Neumann-Morgenstern utility function and implies a risk-neutrality. If u_i is concave then the agent is risk averse and if U_i convex, risk-seeking (Maschler et al., 2013).

of instrumental subgoals, such as acquiring money or income. Just like not all lotteries can be played, not all bundles of goods can be afforded. Von Neumann and Morgenstern (1944) proved that if individual agents' preferences[5] are characterized by completeness, transitivity, and continuity, then they will behave as if they were maximizing expected utility (Moscati, 2016).

In the decision calculus, so far, the outcomes of the decision on L_i accrues only to the agent making the choice: it is implicitly assumed that there are no externalities. In reality, however, and in the case of concerns about AI , the unintended consequences of agents' decisions need to be considered. This is a formidable problem. Consider for instance that, if we denote external costs/benefits of a decision or action on L_i by $C(L_i)$ then the decision in (1) can be re-written as

$$L_i^* = \arg\max_L \mathbb{E}[U(L_i) - C(L_i)] \tag{2}$$

Gauchon and Barigou (2021) discusses the general complexity of this problem, noting that finding the optimum requires various assumptions and somewhat tenuous interpretations of the terms in the first-order conditions.

2.2 Examples in AI

Equivalents to utility functions[6] used in AI systems include value functions, objective functions, loss functions, reward functions (especially in Reinforcement Learning), and preference orderings (Eckersley, 2019). The concepts of utility function and goal are often used interchangeably in the AI literature (Dennis, 2020). Where loss functions (gradients) are used,

[5]In economics one does not require direct knowledge of an agent's preferences - it can be inferred from their choices - their revealed preferences, a notion introduced by Ramsey (1931) and Samuelson (1947).

[6]For an extensive overview of the mathematics used in Machine Learning (ML) see Gallier and Quaintance (2022).

which is the case when some objective function is minimized (for e.g. minimizing the error of wrongly predicting what is in an image) the sign on the above utility function would be negative.

An example from ML is the ubiquitous use of artificial neural networks (ANN), such as the Multilayer Perceptron (MLP)[7] to perform classifications (say classifying images or text).

Formally, an ANN aims, given a (data) set of N samples $D = \{[x_1, x_2, ...x_n], [y_1, y_2, ...y_n]\}$ to find the best approximation for the function describing $f(x_i) = y_i$ which maps the inputs (x) to the outputs (y), where the outputs would be the classification (Lichtner-Bajjaoui, 2020). The objective function is to minimize the expected value of incorrect classifications, which is equivalent to maximizing the utility or goal of the ML (typically back-propagation[8]) algorithm (García-García et al., 2022). In the case of the MLP the probabilities attached to each element x to be classified belonging to a class k is a vector $P(y_k|x)$ that can be written as

$$P(y_k|x) = s'\Big\{\sum_{j=1}^{n_2} \omega_{jk} \times s\Big\{\sum_{i=1}^{n_1} \omega_{ij} \times x_i + b_{0j}\Big\} + b_{0k}\Big\} \tag{3}$$

Where s and s' are respectively known as the activation functions of the hidden and output layers of the neural network, the n the number of neurons in each of these layers, the ω's weights on the connections between the layers and neurons, and b_0's threshold values (activation functions). The back-propagation algorithm will adjust the weights and threshold values to minimize the loss function in classification (and maximize the probability that a classification is accurate). For a more detailed discussion and examples, see García-García et al. (2022).

[7] A MLP is a supervised learning algorithm consisting of various layers of perceptrons, where a perceptron is a program that tries to mimic a biological neuron to perform binary classifications. See Rosenblatt (1958) and Schmidhuber (2015).

[8] See Rumelhart et al. (1986) who introduced back-propagation as a supervised learning technique.

Deep Learning using ANNs to perform classifications has found wide use in Recommender Systems, such as is used to suggest what movies or songs subscribers on Netflix or Spotify may want to watch, or what consumers on Amazon or other platforms may want to buy (Ricci et al., 2022). Jenkins et al. (2021) have shown how these systems can be made more accurate by being explicitly based on micro-economics based utility functions, which they term *Neural Utility Functions*. They pointed out that Recommender Systems, which minimize an objective function of choice prediction, do not use quasi-concave utility functions, as economics typically do. Accordingly, they cannot evaluate trade-offs in decisions, such as taking into account that choices are affected by whether there are complements or substitutes. They also show that if they augment DL models with quasi-concave Neural Utility Functions,[9] that the choice-prediction loss is smaller - using utility functions with a foundation in economic theory improves AI (Jenkins et al., 2021).

A final example of utility functions in the field of AI is that deep convolutional neural networks (CNNs), a class of ANNs used for image classification, can be interpreted as utility maximizers subject to costly learning , i.e. they face informational costs (this is elaborated below under bounded rationality) (Pattanayak and Krishnamurthy, 2021). Applying this idea to deep CNN, Pattanayak and Krishnamurthy (2021) found that they could predict the image-classification performance of 200 deep CNNs with an accuracy $> 94\%$, removing the need to re-train the models.

2.3 Sequential Decision-Making

Many decisions replying on EUT are not one-off decisions but sequential. This is particularly, but not exclusively so in multi-agent settings - where both economics and AI science rely on Game Theory. Gonzales and Perny (2020) discusses the use of graphical models such as Decision Trees to analyse such sequential decision-making. In complex sequential

[9] They use CES and Cobb-Douglas utility function specifications.

decision-making under uncertainty, economists use stochastic dynamic programming (Bellman, 1957a,b) and its Markov Decision Process (MDP) model (Howard, 1960). A typical example is of inventory management (Ahiska et al., 2013).

Effective sequential decision-making by AI agents is vital in virtually all AI applications - from playing games like Chess and Go, to autonomous robots and vehicles, health planning and chatbots. In all of these, the sequence in which decisions are made are important for the overall maximization of the utility function. Reinforcement Learning (RL) is the branch of AI that focuses mainly on sequential decision-making. In most RL[10] an AI agent learns about the underlying MDP "through execution and simulation, continuously using feedback from the past decisions to learn the underlying model and reinforce good strategies" (Agrawal, 2019, p.2). To speed up learning, recourse may be taken, given the nature of the goal, to Supervised Learning or Imitation Learning[11] (Hutter, 2000; Ding et al., 2019). For detailed discussions of RL the reader is referred to Arulkumaran et al. (2017) and Sutton and Barto (1998). Charpentier et al. (2021) describes how RL is used in development of autonomous vehicles. Salimans et al. (2017) refers to the success of RL for developing AI systems that can excel in games such as Atari and Go.

2.4 Challenges to EUT

EUT is subject to at least two challenges. One is that experiments have found that humans may violate the EUT under certain conditions, thus apparently not acting rationally (List and Haigh, 2005); a second relates to evaluating possible future outcomes where there is no objective probability distribution (LeRoy and Singell, 1987) - also known in economics as Knightian uncertainty after Knight (1933).

[10]Exceptions are so-called black-box optimization or direct policy search, which includes a class of optimization algorithms known as Evolution Strategies (ES) (Salimans et al., 2017) and the AIXI model of universal AI of Hutter (2000, 2007) which dispenses with the Markov assumption that the the future only depends on the present.

[11]Particularly useful in robotics (Ding et al., 2019).

Regarding the first challenge, an example is the Ellsberg Paradox or Ellsberg's Urn - see Ellsberg (1961). According to Binmore (2017) a way around the Ellsberg Paradox, which reflects humans' ambiguity aversion, is to screen agents beforehand using another rationality criterion.[12] In the case of using EUT to model the decision-making of AI-agents, such screening is implicitly done - by the selection of AI agents who are not human, to begin with. Thus AI agents more fully inhabit the world of neoclassical economics meaning that economic theory can usefully be applied to AI (Caplin et al., 2022). Other approaches that have been tried in AI to avoid the Ellsberg Paradox is to model AI agents' behaviour indeed closely on that of humans, and to do so by relying on models from behavioural economics - see for instance Tamura (2009).

Regarding the second challenge, to deal with uncertainty, both economics (Harsanyi, 1978) and AI (Pearl, 1985) revert to Bayes' Theorem (Harré, 2021). Von Neumann-Morgenstern's EUT (Von Neumann and Morgenstern, 1944) is based on objective probabilities. (Savage, 1954) generalised this to subjective probabilities. Here,"Bayesian" agents form subjective probabilities based on their priors (beliefs). As new information comes to light, they update their subjective probabilities - and accordingly modify their actions (Savage, 1954; Harsanyi, 1978). *How* their priors are established is a question of some contention and highly relevant to the agenda of AI scientists (Binmore, 2008, 2017).

2.5 Bounded Rationality

Bayesian expected utility maximizers - subjective utility maximizers - are the Mythical Agents of both economics and the field of AI. The problem is that rational decision-makers often make poor - less than optimal- choices (Binmore, 2017). This is because, unlike in

[12]Several generalizations of EUT have been proposed to deal with this and other shortcomings of EUT to be a good descriptive model of human decision-making. A discussion of these falls outside the scope of this paper. The reader is referred to Gonzales and Perny (2020), who discusses amongst others Rank Dependent Utility (RDU) and decision models outside the probabilistic framework; and to Schoemaker (1982) who discusses nine variants of EUT - from expected monetary value to Prospect Theory.

the theoretical idealized world of economics, in reality agents - both human and AI - face informational and computation limits. As Simon (1955, p.114) put it, it may be more useful to replace the mythical agent of economics with an " of limited knowledge and ability" which does not have the "global" rationality of the mythical agent. In the case of humans, mistakes in decision-making are often "predictably irrational" (Ariely, 2009) - which has been ascribed to cognitive biases (Kahneman et al., 2021). Computational limitations and cognitive biases have been used by behavioral economists to argue that Human Sapiens differ from *Homo Economicus* (Thaler, 2000). Intelligence agents' rationality is thus *bounded*.

Bounded rationality is not only applicable to humans, but also to AI agents - even though they may have vastly better computational abilities (Dennis, 2020). As Wagner (2020, p.114) point out "whilst the new species of 'machina economicus' [...] behaves more economic than man, it too is faced with bounded rationality. Algorithms work with finite computational resources which in practice means that they cannot achieve Turing completeness and are limited to linear bounded automation." The reference here to *Turing Completeness* is to the theoretical possibility that AI can be globally rational as the mythical agent of neoclassical economics (Lee, 2019). A *Universal Turing Machine* (UTM) is a "computing machine", proposed by Turing (1936) that can "be used to compute any computable sequence" (Turing, 1936, p.241). It is Turing Complete. However, it is subject to the *Halting Problem*, which is how to determine if and when the UTM will find a solution (Lee, 2019). Turing (1936) proved that there is no general algorithm for solving this problem in all cases.

Finite computation resources, as described in the previous paragraph, implies that there are information costs involved in making a decision as described in equation (1). These information costs can be further specified to come from the updating of an agent's Bayesian priors. If, in the example of equation (1) in section 2.1 the agents' probability distribution over the choice of L_i is $p(L)$ then computational resources (costs) will be expended to change from a prior probabilistic strategy $p_0(L)$ to a posterior probabilistic strategy $p(L)$ in the

process of decision-making (Leibfried and Braun, 2016). This informational cost is known as the Kullback-Leibler divergence (D_{KL}) and can be specified as $D_{KL} = (p(L)||p_0(L)) \leq B$ where $B \geq 0$ is the upper bound of available computational resources (Leibfried and Braun, 2016). With these informational costs, the expected utility maximization problem in (1) can be modified to

$$L_i^* = (1 - \beta) \arg\max_L \mathbb{E}[U(L_i)] - \beta D_{KL}(p(L)||p_0(L)) \tag{4}$$

Where $\beta \in (0, 1)$ is the trade-off between expected utility and informational cost (Leibfried and Braun, 2016). It is also consistent with Sims (2003)'s "rationally inattentive utility maximization" where paying more attention to making a decision implies high attention costs. The point is that learning is costly, it bounds rationality, and needs to be taken into account in models of bounded rational decision-making (Lipnowski and Doron, 2022; Pattanayak and Krishnamurthy, 2021).

The difference though between AI bounded rationality and human bounded rationality is that while human agents are subject to both computational limits and cognitive biases, AI agents will face fewer computational limits and can unlearn cognitive biases. Learning may be less costly. AI agents can be programmed with error correction mechanisms and these will inevitably drive them to Homo Economicus, or more appropriately as Parkes and Wellman (2015) have suggested, *Machina Economicus*.

As Omohundro (2008b) explains, the nature of AI systems such as Deep Learning[13] and (Deep) Reinforcement Learning[14] as "learning" agents means that they are self-improving systems. They will thus learn where they have been making sub-optimal decisions or have

[13]The dominant approach to unpack functions that maps precepts from the environment into actions is Deep Learning (LeCun et al., 2015; Sarker, 2021).

[14]For more comprehensive explanations of Reinforcement Learning (RL) see Arulkumaran et al. (2017) and Sutton and Barto (1998). Charpentier et al. (2021) describes how RL is used in development of autonomous vehicles. Salimans et al. (2017) refers to the success of RL in games such as Atari and Go.

been deviating from their goals, and correct for it in a way "*discover and eliminate their own irrationalities in ways that humans cannot*" (Omohundro, 2008b, p.4).

Thus, the state of the art in the fields of economics and AI is the modelling of intelligent agents that rely on Bayesian probability theory to inform beliefs (priors), and utility theory to inform their preferences. With beliefs and wants determined, and with limits on their computational abilities and resources, by aiming to maximize expected utility, intelligent agents will act in a boundedly rational manner under uncertainty (Benya, 2012; Riedel, 2021). Intelligent agents with beliefs and preferences but who fail to attempt to maximize expected utility may be vulnerable and subject to exploitation (Shah, 2019a). Eventually evolutionary pressures will lead to the disappearance of such agents from a population.

This bounded mythical agent of rational decision theory has much to commend it, as its voluminous application in the economics literature, and its dominance in explaining decision-making, attest to (Binmore, 2008; Dixon, 2001; Moscati, 2016). It has even be found to be applicable to decision-making in other primates, not only humans (Pastor-Berniera et al., 2017). Indeed, it is due to its strengths that it has come to underpin the development of AI.

But, there are also kinks in the armour, which, in light of the continued advancement in AI, poses a number of challenges for AI scientists in the practical implementation thereof. These are discussed in the next section.

3 How can we best implement EUT in AI systems?

EUT is thus, as the previous section discussed, a very attractive theoretical decision-making theory for an AI to follow. How to best implement the EUT in AI systems faces an important challenge: the alignment problem.

3.1 The alignment problem

Although AI agents are also rationally bounded, they have fewer cognitive biases than humans, and can unlearn these. Learning and eliminating its biases imply that AI may become recursively self-improving. In such a situation, where AI learns, self-correct and recursively self-improve, one would need to avoid the eventually that an AI emerges and pursue goals (utility) that conflict with human interest (Bostrom, 2014), or even if these do not conflict with human interest, nevertheless can have unintended negative consequences. These may follow because its utility function may not capture all the considerations relevant in a situation - "humans care about many features of the environment that are difficult to capture in any simple utility function" (Taylor, 2016, p.125).

The challenge is, as formulated by Omohundro (2008b, p.36), that AI systems following EUT

> "will maintain utility functions which encode their preferences about the world. In the process of acting on those preferences, they will be subject to drives towards efficiency, self-preservation, acquisition, and creativity. Unbridled, these drives lead to both desirable and undesirable behaviors. By carefully choosing the utility functions of the first self-improving systems, we have the opportunity to guide the entire future development."

It is therefore essential, as Riedel (2021) and others have argued, to understand the utility functions (goals) of AI agents - not only for participating with and competing against AI agents, but eventually for constraining and aligning AI's goals (Bostrom, 2014). Constraining and aligning AI systems' goals or utility functions is a topic that has generated a large and growing literature under the headings of AI alignment and AI ethics (Hauer, 2022), which ultimately aims to ensure that AI "benefits humans" and humans do not lose control over AI (Kirchner et al., 2022, p.1). Note that this is a very human-centric agenda based on a

view of human exceptionalism (Murphy, 2022).

Discussing all the risks in AI goal design falls outside the scope of the present paper - a recent washing list of such AI goal design risks identified 26 different risks (Kokotajlo and Dai, 2019). Further reviews on aligning AI are contained in Christian (2020), Everitt et al. (2018), Hubinger (2020) and Kirchner et al. (2022). The AI alignment problem arises in the eventually that AI recursively self-improve - something which it cannot do at present.

Economists have so far contributed little to this topic. As Gans (2018) pointed out, "The underlying ideas behind the notion that we could lose control over an AI are profoundly economic. But, to date, economists have not paid much attention to them."

How could economists contribute? At least three possible ways to approach AI alignment in the field of economics stand out: instrumental goals, utility function instability and utility function coordination. Explaining how AI s scientists approach these will enrich the field of economics, and addressing the challenges that each pose for the operationalization of EUT in AI systems are fertile areas for further research.

3.2 Instrumental goals

The first is to tackle the problem of instrumental goals. Any smart AI system with a goal-oriented utility function will, given the "AI drives" listed in the quote above, develop instrumental (or sub) goals (Gabriel and Ghazavi, 2021; Omohundro, 2008a). How this creates a problem for value alignment is illustrated by the *Paperclip Maximiser* thought experiment.[15] It describes an ASI with a top goal[16] to manufacture paper clips:

[15]The same point is made by The King Midas problem which is cited by Russell (2019) who refers to the classical story of King Midas who, when he had the opportunity to be granted any wish, wished that anything he touches turn to gold. When he subsequently touched his daughter, she also turned to gold.

[16]In implementing utility functions in AI a distinction is made between top goals (or super goals) and sub-goals (Yudkowsky, 2001), for example, if the top goal of an AI system is to drive a vehicle from point to A to point B, a sub-goal may be to ensure that the vehicle is operational.

> "starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities. More subtly, it could result in a superintelligence realizing a state of affairs that we might now judge as desirable but which in fact turns out to be a false utopia, in which things essential to human flourishing have been irreversibly lost. We need to be careful about what we wish for from a superintelligence, because we might get it"- Bostrom (2003, p.17).

These concerns have on the one hand led to proposals to constrict the utility function that AI agents optimize - for instance to try to implement the 1956 suggestion of Herbert Simon not to try to build use utility maximizers, but utility satisficers (Simon, 1956) - which essentially engages only in some limited form of optimization (Armstrong et al., 2012). Shortcomings of these proposals are discussed by Taylor (2016).

On the other hand, concerns have led to the use of approaching the design of AI Systems by building in uncertainty about the utility function, and letting the AI system discover the utility function by learning from humans. This is where Reinforced Learning (RL)[17] with its reward function optimization and supplemented recursive reward modeling is an example[18] (Gabriel and Ghazavi, 2021). The aim is that this search will lead to alignment with human values.

In essence, RL has been argued to reflect the evolutionary process which has given rise to much of society's current laws and regulations: "I read the history of Western law and the simple rules that emerged from it as decentralized RL. Jurists and agents, through a combination of reasoning and experience, saw what worked and what did not. Those rules that led to Pareto improvements survived and thrived. Those that did not, dwindled" (Fernandez-Villaverde, 2020, p.15). According to Shah (2019b) the uncertainty of AI with

[17] Also including Cooperative Inverse Reinforcement Learning (CIRL) - see Hadfield-Menell et al. (2016)

[18] Apprenticeship Learning is another proposal, based on the idea that an AI system tries to imitate a human expert in performing a particular task where the utility function is uncertain (Abbeel and Ng, 2004). The shortcoming of this proposal is that AI systems may then never become smarter than humans at a task, leaving us bereft of their potential advantage (Taylor, 2016).

RL will lead to AI systems that are "deferential, that ask for clarifying information, and that try to learn human preferences." See also Hadfield-Menell et al. (2016) and Shah (2019b)).

A weakness of getting AI to learn about human preferences and its utility function, and a weakness in general of RL, is that learning itself is an endogenous and imperfect process (Kuriksha, 2021). Two typical problems in learning are cognitive limitations and that learning in one environment does not necessarily carry over to a different environment. Kuriksha (2021) applied an RL model to explore the economic implications of such imperfect learning for AI agents that have to make savings-consumption decisions. He found that agents agents who learned to optimize saving in an environment withlow levels of wealth would not save optimally if transferred to an environment with high levels of wealth, and vice versa.

Another weakness of getting AI to learn about human preferences and its utility function is as Turchin and Denkenberger (2020, p.159) point out, that "if AI extracted human values from the most popular TV series, it could be "Game of Thrones" [...] and then the'paradise' world it created for us would be utter hell." Consequently, it may be preferential to ensure that the AI systems that can learn about human preferences are Artificial Moral Agents (AMA) (Allen et al., 2005). AMAs are "artificial agents capable of making ethical and moral decisions" (Cervantes et al., 2020, p.503). However, the design of such AMA remains an elusive goal (Cervantes et al., 2020).

3.3 Utility Function Instability

Another angle from which economics can contribute to the AI alignment problem is to address the problem of utility function instability, in other words, the problem that an aligned utility function becomes mis-aligned. There are two major aspects that involve utility function instability.

One is the related problems of wireheading and self-delusion. Wireheading, or reward-hacking, refers to agents directly stimulating their reward centres, thus interfering in reward provision (Cohen et al., 2022). In the case of AI agents it is particularly a problem in Reinforcement Learning (RL) (Everitt and Hutter, 2016). Various methods are being tested and developed to avoid wireheading. These include imitation learning, rewarding agents that maximise their impact on the environment instead of the signals they receive, value reinforcement learning (VRL), inverse reinforcement learning (IRL), apprenticeship learning (AL), myopia and quantilization (Cohen et al., 2022; Everitt and Hutter, 2016). Related to wireheading is that AI-agents may self-delude. This would occur when AI agents deliberately change their own observations of the impacts of their actions so that it may seem that they are maximizing their utility functions, while in fact, they are not (Hibbard, 2012; Ring and Orseau, 2011).

The second cause of utility function instability is that an AI may change its own utility function autonomously (Dennis, 2020; Totschnig, 2020). It may do so to compensate for being boundedly rational. For instance the AI-agent may perform an action in pursuit of a goal and fail to achieve the goal due to differences in its (imperfect) model of the world and the actual world. It may therefore change the goal. The question is, what informs the direction of change? Here AI developers have been exploring the programming of values, based on the argument that if the values of AI are aligned, then it will not change its goals in a direction that will be potentially harmful to humans. There is, however, at present no clear understanding of how to program this into AI systems: "our current lack of understanding about how to adequately program behaviour that can flexibly adopt and drop goals is one of the key limitations to our ability to take artificial intelligence to the next level" (Dennis, 2020, p.2493).

3.4 Utility Function Coordination

The third angle from which economics can contribute to the AI alignment problem is to address the challenge for rational goal-directed decision-making posed by multi-actor and human-agent collectives (HAC) settings (Wagner, 2020).

So far the discussion has been about a single decision-maker agent. In an economy with many agents the challenge is how the individual AIs' decisions should be modelled? How do individual decisions add up to aggregate outcomes? And how can humans effectively and safely interact with AI agents?

Here, another basic methodological foundation that AI scientists and economists share, is Game Theory (Russell, 2019). In multi AI-agent environments, AIs rely on a game-theoretic view of the world, "where agents rationally respond to each others' behavior, presumed (recursively) to be rational as well. A consequence is that agents would expect their joint decisions to be in some form of equilibrium, as in standard economic thinking" (Parkes and Wellman, 2015, p.269). We know however from Game Theory - the Prisoners' Dilemma is an example - that decisions that are rational and optimal on the individual level do not necessarily aggregate to the best outcome for society. The Prisoner's Dilemma exist because of a lack of coordination.

In the field of AI, where, as was discussed in section 2.1, RL has been successful in sequential optimization as is evident in the success of AI agents using RL to play GO, Atari games or steer autonomous vehicles, most of these applications of RL involve multiple agents. It requires AI agents to interact with other intelligent agents, and more generally, the external environment. Therefore, multi-agent RL (MARL) has become popular. As Zhang et al. (2021) discuss, MARL's theoretical foundations are provided by Game Theory - specifically Markov games and extensive-form games. In these, because each agent may have a different utility function, and all agents are continuously adjusting their policies/ actions given feed-

back from the environment, the environment facing all agents is changing all the time. i.e. becomes non-stationary, violating the Markov assumption followed in single-agent RL. This non-stationarity property of MARL remains a challenge in the development of AI systems (Zhang et al., 2021).

In such situations, it is not only that AI developers need to get the utility functions of AI agents to be optimal or appropriate, but they must also specify the institutional features of the environment - the "rules of the game" and the "play of the game," to use the terminology from institutional and transaction cost economics (North, 1991). This, in effect, can facilitate utility function coordination amongst the various interacting agents. The field of mechanism design, which has been richly applied in economics, has been shown to be useful in this regard for AI systems, though not without challenges (Parkes and Wellman, 2015; Varian, 1995).

Although multi-agent environments are populated by AI agents with different utility functions poses challenges, this feature has been found to be nevertheless useful in ML, particularly in driving learning using models from evolutionary game theory. Thus the latter underpins the use of Generative Adversarial Networks (GANs), which exploit the fact that agents (typical neural network based) may have opposing goals and divergent utility functions, to train AI systems (Goodfellow et al., 2014; Guo et al., 2020).

Another problem that arises in AI multi-actor and HAC environments when there is no utility function coordination is a familiar one in economics: the Principal-Agent problem (Grossman and Hart, 1983). Drawing on multilateral Principal-Agent models (e.g. (Bernheim and Whinston, 1986)) it can be seen that with AI the simple bilateral Principal-Agent problem becomes one with three agents - the human user of AI is the principal, the AI system as the agent, and the provider of the AI as another agent (Wagner, 2020, p.118). Wagner (2020) explores the implications of the Principal-Agent problem in such a setting, pointing out that there are likely to be substantial and increasing information asymmetries between the

human principal and the AI agent - and AI provider - given the superior speed of information processing of AI, the continued background tracking of humans online, and the black-box nature of many current AI decisions. Without utility function coordination between these agents, it is possible that the interest of AI systems and those of their principal agents will increasingly divergence.

As far as the interaction between humans and AI agents in multi-agent situations is concerned, this has been gathering scrutiny under the rubric if human-agent collectives (HACs) (Wagner, 2020). These HACs have been described as starting to exhibit properties of a collective mind or even supermind, and has led to observations that the close integration of AI agents and human agents in a collective mind could improve the strength of institutions and weaken the relevance of methodological individualism, a central plank of neoclassical economics (Wagner, 2020; Arrow, 1994). Modelling institutions, including markets, without invoking the assumptions of methodological individualism, remain a challenge for economists. A recent review of the case for methodological individualism in the social sciences by Neck (2021) for instance, omits to consider the implications of the rise of HACs and the growing autonomous economy (Arthur, 2021).

4 Concluding Remarks

> '*Neoclassical theory involves very smart people in incredibly simple situations, while the real world entails very simply people in incredibly complex situations*" - Axel Leijonhufvud as quoted by Daneke (2020, p.28).

The challenge facing AI scientists is to create intelligent, autonomous agents that can make rational decisions. In this challenge, they confront two questions: what decision theory to follow? and how to implement such a decision theory in AI systems? This paper provided answers to these questions by providing a critical overview the appropriateness of Expected

Utility Theory (EUT), and to outline how economics can contribute to implement EUT in AI Systems, taking into account the alignment problem.

It was shown that modern data-based AI has economics' EUT at its very basis - the *Homo Economicus* - which is a boundedly-rational Bayesian expected utility maximizer. The challenges of endowing AI agents with utility functions (goals and sub-goals) were discussed, which include the problem of instrumental goals, the possibility of utility function instability, and coordination challenges in multi-actor and human-agent collectives settings. These challenges complicates the design of future AI systems whose values and actions need to be aligned with human interests. A first contribution of this paper was to outline how economic decision-making theory can address these.

A second contribution of this paper can now be formulated. This is that using EUT as a decision theory constrains AI to "small worlds," in the terminology of Savage (1954). Savage used two proverbs to explain the difference between small and large worlds, and argued that Bayesian rationality is applicable to the former, and less useful in the latter. One proverb is "Look before you leap" and another "Cross that bridge when you come to it." As Binmore (2007, p.25) explains, "You are in a small world if it is feasible always to look before you leap. You are in a large world if there are some bridges that you cannot cross before you come to them."

The "narrow" AI systems that currently (2023) exist, all inhabit small worlds. They have no choice. As the discussion referring to the Ellsberg Paradox pointed out, implicitly, all AI-systems are therefore "screened", meaning that we restrict the class of agents who apply EUT. We also restrict the class of agents by the data and algorithms we endow them with. And second, all current narrow AI systems are, being based on the Bayesian approach, of the "look before you leap" type. Unlike humans, AI still cannot cross that bridge when it comes to it. This establishes the constrained domain of what we call *narrow* AI systems, which "are extremely bounded in that they are highly specialized on specific tasks and thus

might not behave rationally beyond their dedicated domain" (Wagner, 2020, p.114).

In practice therefore, based on homo economicus, the domain of AI systems are restricted to narrow applications, e.g. chatbots or search engines, which are the small worlds in terms of the EUT approach. In the world of narrow AI therefore, concerns about AI alignment may loose urgency. The risks for the safe use of narrow AI are therefore perhaps better labelled safety issues, rather than alignment issues.

In conclusion, a third contribution of this paper was, not only to describe how AI as field benefits from the economic theory of decision-making, but to point to several areas where economists may from learn from AI scientists as they implement EUT. One such area is procedural rationality. Economic theory tends to ignore the reasoning process by which agents make rational decisions, i.e., how agents find the optimum of their expected utility functions (Dixon, 2001). Economics have preferred *substantive rationality* over *procedural rationality* (Simon, 1978). It amounts to an approach where "what decisions are made is more important than how they are made" (Harré, 2021, p.12).

As Dixon (2001) has suggested, there is a need in economics to consider the process of reasoning itself, because human decision-making is prone to mistakes - even beyond those due to being bounded by computational ability. He concludes[19] that "the role for artificial intelligence in economics would then seem primarily to be in situations where economic agents make mistakes, and possibly bad mistakes."

AI research, such as in the sub-field of RL, may also help economists overcome computational difficulties in understanding reasoning under bounded rationality (Charpentier et al., 2021) and help to model human behaviour in disequilibrium situations (Dixon, 2001). One may also interpret non-rational quirks in human judgment as the result of AI learning techniques (e.g. DL or RL) that are "inappropriately applied" (Camerer, 2019).

[19]Dixon (2001) that AI research, by highlighting the mechanisms of reasoning, may also throw light on strategic behaviour, where economic agents may face incentives to intentionally make mistakes.

In sum, economics helps AI to model rational decision-making by artificial intelligent agents and to constrain these decisions to small worlds; and AI may help economics to model human decision-making better.

Acknowledgement

This paper has been extracted and elaborated from my more comprehensive review of the economics of AI, published as an earlier IZA Discussion Paper. See Naudé, W. (2022). The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest. IZA DP No. 15713.

References

Abbeel, P. and Ng, A. (2004). Apprenticeship Learning via Inverse Reinforcement Learning. *Proceedings of the Twenty-First International Conference on Machine Learning, ICML 04. New York, NY, USA*.

Agrawal, S. (2019). Reinforcement Learning Lecture 1: Introduction. *University of Minnesota*.

Ahiska, S., Appaji, S., King, R., and Warsing Jr, D. (2013). A Markov Decision Process-Based Policy Characterization Approach for a Stochastic Inventory Control Problem with Unreliable Sourcing. *International Journal of Production Economics*, 144(2):485–495.

Allen, C., Smit, I., and Wallach, D. (2005). Artifcial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3):149–155.

Ariely, D. (2009). Predictably Irrational: The Hidden Forces That Shape Our Decisions. *London: HarperCollins*.

Armstrong, S., Sandberg, A., and Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4):299–324.

Arrow, K. (1994). Methodological Individualism and Social Knowledge. *American Economic Review, Papers and Proceedings*, 84(2):1–9.

Arthur, W. (2021). Foundations of Complexity Economics. *Nature Reviews Physics*, 3:136–145.

Arulkumaran, K., Deisenroth, M., Brundage, M., and Bharath, A. (2017). A Brief Survey of Deep Reinforcement Learning. *arXiv:1708.05866v2 [cs.LG]*.

Bellman, R. (1957a). Dynamic Programming. *Princeton: Princeton University Press*.

Bellman, R. (1957b). A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684.

Benya (2012). Why You Must Maximise Expected Utility. *AI Alignment Forum*, 13 Dec.

Bernheim, B. and Whinston, M. (1986). Common Agency. *Econometrica*, 54(4):923–942.

Bernoulli, D. (1738). Commentarii. *Acad. Scientiarum Imperialis Petropolitanae 5 175-192; English translation (1954) Econometrica*, 22:23–36.

Binmore, K. (2007). Rational Decisions in Large Worlds. *Annales d'Économie et de Statistique*, 86:25–41.

Binmore, K. (2008). Rational Decisions. *Princeton University Press: Princeton, NJ.*

Binmore, K. (2017). On the Foundations of Decision Theory. *Homo Oeconomicus*, 34:259–273.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. *In I. Smit et al. (eds.) Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. 2nd ed. International Institute of Advanced Studies in Systems Research and Cybernetics*, pages 12–17.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford: Oxford University Press.*

Camerer, C. (2019). Artificial Intelligence and Behavioral Economics. *In A. Agrawal and J.Gans and A. Goldfarb (eds.) The Economics of Artificial Intelligence: An Agenda. Chicago: University of Chicago Press*, pages 587–610.

Caplin, A., Martin, D., and Marx, P. (2022). Modeling Machine Learning. *NBER Working Paper No. 30600.*

Cervantes, J., Lopez, L., Rodriguez, L., Cervantes, S., Cervantes, F., and Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26:501–532.

Charpentier, A., Élie, E., and Remlinger, C. . (2021). Reinforcement Learning in Economics and Finance. *Computational Economics.*

Christian, B. (2020). The Alignment Problem: Machine Learning and Human Values. *New York: W.W. Norton.*

Cohen, M., Hutter, M., and Osborne, M. (2022). Advanced Artificial Agents Intervene in the Provision of Reward. *AI Magazine*, 43(3):282–293.

Daneke, G. A. (2020). Machina-Economicus or Homo-Complexicus: Artificial Intelligence and the Future of Economics? *Real-World Economics Review*, 93:18–39.

Dastani, M., Hulstijn, J., and van der Torre, L. (2005). How to Decide What to Do? *European Journal of Operational Research*, 160:762–784.

Dawkins, R. (1976). The Selfish Gene. *Oxford: Oxford University Press*.

Dennis, L. (2020). Computational Goals, Values and Decision-Making. *Science and Engineering Ethics*, 265:2487–2495.

Ding, Y., Florensa, C., Phielipp, M., and Abbeel, P. (2019). Goal-Conditioned Imitation Learning. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*.

Dixon, H. (2001). Surfing Economics. *Red Globe Press London*.

Eckersley, P. (2019). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *SafeAI 2019: Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75:643–699.

Everitt, T. and Hutter, M. (2016). Avoiding Wireheading with Value Reinforcement Learning. *arXiv:1605.03143v1 [cs.AI]*.

Everitt, T., Lea, G., and Hutter, M. (2018). AGI Safety Literature Review. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 5441–5449.

Fernandez-Villaverde, J. (2020). Simple Rules for a Complex World with Artificial Intelligence. *PIER Working Paper No. 20-010*.

Gabriel, I. and Ghazavi, V. (2021). The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. *arXiv:2101.06060 [cs.CY]*.

Gallier, J. and Quaintance, J. (2022). Algebra, Topology, Differential calculus, and Optimization Theory For Computer Science and Machine Learning. *Mimeo: University of Pennsylvania*.

Gans, J. (2018). AI and the Paperclip Problem. *VoxEU Column*, 10 June.

García-García, J., García-Ródenas, R., López-Gómez, J., and Martín-Baos, J. (2022). A Comparative Study of Machine Learning, Deep Neural Networks and Random Utility Maximization Models for Travel Mode Choice Modelling. *Transportation Research Procedia*, 62:374–382.

Gauchon, R. and Barigou, K. (2021). Expected Utility Maximization with Stochastically Ordered Returns. *Mimeo: University of Lyon*.

Gonzales, C. and Perny, P. (2020). Decision Under Uncertainty. *In P. Marquis and O. Papini and H. Prade (eds.). A Guided Tour of Artificial Intelligence Research, I, Springer*, pages 549–586.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, pages 2672–2680.

Grossman, S. and Hart, O. (1983). An Analysis of the Principal-Agent Problem. *Econometrica*, 51:7–45.

Guo, I., Langrené, N., Loeper, G., and Ning, W. (2020). Robust Utility Maximization Under Model Uncertainty via a Penalization Approach. *arXiv:1907.13345v5 [math.OC]*.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and S.Russell (2016). Cooperative Inverse Reinforcement Learning. *arXiv:1606.03137 [cs.AI]*.

Harré, M. (2021). Information Theory for Agents in Artificial Intelligence, Psychology, and Economics. *Entropy*, 23:310.

Harsanyi, J. (1978). Bayesian Decision Theory and Utilitarian Ethics. *American Economic Review*, 68(2):223–228.

Hauer, T. (2022). Importance and Limitations of AI Ethics in Contemporary Society. *Humanities and Social Sciences Communications*, 9(272):1–8.

Herfeld, C. (2020). The Diversity of Rational Choice Theory: A Review Note. *Topoi*, 39:329–347.

Hibbard, B. (2012). Model-Based Utility Functions. *Journal of Artificial General Intelligence*, 3(1):1–24.

Howard, R. (1960). Dynamic Programming and Markov Processes. *Cambridge MA: The MIT Press*.

Hubinger, E. (2020). An Overview of 11 Proposals for Building Safe Advanced AI. *arXiv:2012.07532 [cs.LG]*.

Hutter, M. (2000). A Theory of Universal Artificial Intelligence Based on Algorithmic Complexity. *arXiv:cs/0004001 [cs.AI]*.

Hutter, M. (2007). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. *In B. Goertzel and C. Pennachin (eds). Artificial General Intelligence. Cognitive Technologies. Springer, Berlin, Heidelberg*, pages 227–290.

Jenkins, P., Farag, A., Jenkins, J., Yao, H., Wang, S., and Li, Z. (2021). Neural Utility Functions. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 35(9):7917–7925.

Kahneman, D., Sibony, O., and Sunstein, C. (2021). Noise: A Flaw in Human Judgment. *London: HarperCollins*.

Kamatani, N. (2021). Genes, the Brain, and Artificial Intelligence in Evolution. *Journal of Human Genetics*, 66:103–109.

Kirchner, J., Smith, L., and Thibodeau, J. (2022). Understanding AI Alignment Research: A Systematic Analysis. *arXiv:2206.02841v1 [cs.CY]*.

Knight, F. (1933). Risk, Uncertainty and Profit. *Houghton Mifflin Co, Boston*.

Kokotajlo, D. and Dai, W. (2019). The Main Sources of AI Risk? *AI Alignment Forum*, 21 March.

Kuriksha, A. (2021). An Economy of Neural Networks: Learning from Heterogeneous Experiences. *PIER Working Paper 21-027, University of Pennsylvania*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature*, 521:436–444.

Lee, C. (2019). The Game of Go: Bounded Rationality and Artificial Intelligence. *Yusof Ishak Institute Working Paper no. 2019-04*.

Leibfried, F. and Braun, D. A. (2016). Bounded Rational Decision-Making in Feedforward Neural Networks. *arXiv:1602.08332v2 [cs.AI]*.

LeRoy, S. and Singell, L. (1987). Knight on Risk and Uncertainty. *Journal of Political Economy*, 95(2):394–406.

Lichtner-Bajjaoui, A. (2020). A Mathematical Introduction to Neural Networks. *Advanced Mathematics Master Thesis, Universitat de Barcelona*.

Lipnowski, E. and Doron, R. (2022). Predicting Choice from Information Costs. *arXiv:2205.10434*.

List, J. and Haigh, M. (2005). A Simple Test of Expected Utility Theory Using Professional Traders. *Proceedings of the National Academy of Science USA=*, 102(3):945–948.

Maschler, M., Solan, E., and S.Zamir (2013). Game (t)heory. *2nd edition. Cambridge: Cambridge University Press*.

Monton, B. (2019). How to Avoid Maximizing Expected Utility. *Philosophers Imprint*, 19(8):1–25.

Moscati, I. (2016). Retrospectives: How Economists Came to Accept Expected Utility Theory: The Case of Samuelson and Savage. *Journal of Economic Perspectives*, 30(2):219–236.

Murphy, T. (2022). Human Exceptionalism. *Do the Math Blog*, 16 February.

Neck, R. (2021). Methodological Individualism: Still a Useful Methodology for the Social Sciences? *Atlantic Economic Journal*, 49:349–361.

North, D. (1991). Institutions. *Journal of Economic Perspectives*, 5(1):97–112.

Oesterheld, C. (2021). Approval-Directed Agency and the Decision Theory of Newcomb-Like Problems. *Synthese*, 198(27):6491–6504.

Omohundro, S. (2008a). The Basic AI Drives. *Proceedings of the First AGI Conference*.

Omohundro, S. (2008b). The Nature of Self-Improving Artificial Intelligence. *Mimeo*.

Parkes, D. and Wellman, M. P. (2015). Economic Reasoning and Artificial Intelligence. *Science*, 6245:267–272.

Pastor-Berniera, A., Plott, C., and Schultz, W. (2017). Monkeys Choose as if Maximizing Utility Compatible with Basic Principles of Revealed Preference Theory. *PNAS*, 114(10):E1766–E1775.

Pattanayak, K. and Krishnamurthy, V. (2021). Rationally Inattentive Utility Maximization Explains Deep Image Classification. *arXiv:2102.04594 [cs.LG]*.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*.

Persky, J. (1995). Retrospectives: The Ethology of Homo Economicus. *Journal of Economic Perspectives*, 9(2):221–231.

Pigozzi, G., Tsoukias, A., and Viappiani, P. (2016). Preferences in Artificial Intelligence. *Annals of Mathemathical Artificial Intelligence*, 77:361–401.

Ramsey, F. (1931). Truth and Probability. *In F. Ramsey (ed.). Foundations of Mathematics and other Logical Essays. New York: Harcourt.*

Ricci, F., Rokach, L., and Shapira, B. (2022). Recommender Systems: Techniques, Applications, and Challenges. *In F. Ricci and L. Rokach and B. Shapira (eds.). Recommender Systems Handbook (3 ed.). New York: Springer*, pages 1–35.

Riedel, C. J. (2021). Value Lock-in Notes. *Mimeo*, 25 July.

Ring, M. and Orseau, L. (2011). Delusion, Survival, and Intelligent Agents. *In J. Schmidhuber and K.R. Thórisson and M. Looks (eds.) AGI 2011. LNCS (LNAI)*, 6830:11–20.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65:386–407.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning Representations by Back-Propagating Errors. *Nature*, 323:533–536.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. *Viking Books*.

Russell, S. and Norvig, P. (2021). Artificial Intelligence: A Modern Approach. *4th Edition. Pearson Education, Inc.*

Salimans, T., J., Chen, X., Sidor, S., and Sutskever, I. (2017). Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *arXiv:1703.03864v2 [stat.ML]*.

Samuelson, P. (1947). Foundations of Economic Analysis. *Cambridge: Harvard University Press*.

Sarker, I. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):420.

Savage, L. (1954). The Foundations of Statistics. *New York: Dover*.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 6:85–117.

Schoemaker, P. (1982). The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations. *Journal of Economic Literature*, 20(2):529–563.

Shah, R. (2019a). AI Safety without Goal Directed Behaviour. *Alignment Forum*, 7 January.

Shah, R. (2019b). Stuart Russell's New Book on Why We Need to Replace the Standard Model of AI. *Alignment Newsletter no. 19*.

Simon, H. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1):99–118.

Simon, H. (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, 63(2):129–138.

Simon, H. (1978). On How to Decide What to Do. *The Bell Journal of Economics*, 9(2):494–507.

Sims, C. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3):665–690.

Sutton, R. and Barto, A. (1998). Introduction to Reinforcement Learning. *1st ed. Cambridge, MA: MIT Press*.

Tamura, H. (2009). Modeling Ambiguity Averse Behavior of Individual Decision Making: Prospect Theory under Uncertainty. *In Torra, V., Narukawa, Y., Inuiguchi, M. (eds.). Modeling Decisions for Artificial Intelligence. MDAI 2009. Lecture Notes in Computer Science, vol 5861. Springer, Berlin, Heidelberg*.

Taylor, I. (2016). Dependency redux: Why Africa is not rising. *Review of African Political Economy*, (43):8–25.

Thaler, R. (2000). From Homo Economicus to Homo Sapiens. *Journal of Economic Perspectives*, 14(1):133–141.

Totschnig, W. (2020). Fully Autonomous AI. *Science and Engineering Ethics*, 26:2473–2485.

Turchin, A. and Denkenberger, D. (2020). Classification of Global Catastrophic Risks Connected with Artificial Intelligence. *AI & Society*, 35:147–163.

Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. 42:230–265.

Varian, H. (1995). Economic Mechanism Design for Computerized Agents. *Proceedings of the First USENIX Workshop on Electronic Commerce New York, New York*, July.

Von Neumann, J. and Morgenstern, O. (1944). Theory of Games and Economic Behavior. *Princeton NJ: 1st Ed. Princeton University Press.*

Wagner, D. (2020). Economic Patterns in a World with Artificial Intelligence. *Evolutionary and Instutional Economics Review*, 17:111–131.

Williams, G. (1966). Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. *Princeton, NJ: Princeton University Press.*

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *The Singularity Institute, San Francisco, CA*, June 15.

Yudkowsky, E. (2007). Pascal's Mugging: Tiny Probabilities of Vast Utilities. *Less Wrong Blog*, 19 October.

Zhang, K., Yang, Z., and Basar, T. (2021). Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv:1911.10635v2 [cs.LG].*