Artificial Intelligence Review (2025) 58:352
https://doi.org/10.1007/s10462-025-11341-4

(Ir)rationality in AI: state of the art, research challenges and open questions

Olivia Macmillan-Scott[1] · **Mirco Musolesi**[1,2]

Accepted: 29 July 2025

Abstract
The concept of rationality is central to the field of artificial intelligence (AI). Whether we are seeking to simulate human reasoning, or trying to achieve bounded optimality, our goal is generally to make artificial agents as rational as possible. Despite the centrality of the concept within AI, there is no unified definition of what constitutes a rational agent. This article provides a survey of rationality and irrationality in AI, and sets out the open questions in this area. We consider how the understanding of rationality in other fields has influenced its conception within AI, in particular work in economics, philosophy and psychology. Focusing on the behaviour of artificial agents, we examine irrational behaviours that can prove to be optimal in certain scenarios. Some methods have been developed to deal with irrational agents, both in terms of identification and interaction, however work in this area remains limited. Methods that have up to now been developed for other purposes, namely adversarial scenarios, may be adapted to suit interactions with artificial agents. We further discuss the interplay between human and artificial agents, and the role that rationality plays within this interaction; many questions remain in this area, relating to potentially irrational behaviour of both humans and artificial agents.

Keywords Artificial intelligence · Rationality · Irrationality · Decision-making

1 Introduction

The way machines ‘think’ and how close this can come to human reasoning has long been a topic of debate within artificial intelligence (AI) research. From Turing’s proposal of a simple game to identify a machine that could think (Turing 1950), to the 1956 Dartmouth Summer Research Project Proposal that sought to demonstrate that any aspect of learning or intelligence can be simulated by machines (McCarthy et al. 2006), the problem of under-

✉ Olivia Macmillan-Scott
olivia.macmillan-scott.16@ucl.ac.uk

[1] AI Centre, Department of Computer Science, University College London, London, UK

[2] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Published online: 22 August 2025

standing how far AI can push the boundaries of intelligence is inevitably intertwined with how machines reason. As such, the concept of rationality is central to the field of AI (Besold 2013), whether we are referring to the rationality of artificial agents or of humans. Crucially, the goal is seldom to make an agent as rational as possible—some degree of irrationality is often desired, both in interactions amongst machines but also with humans. A more human-like artificial agent may at times elicit more trust (Waytz et al. 2014), whereas in other scenarios simulating the fallibility of human reasoning can result in an artificial agent that appears untrustworthy. Despite the centrality of rationality to the field, there are still many open questions to be addressed.

Given the advent of large language models (LLMs) (Brown et al. 2020), as well as their increasing pervasiveness in our daily lives and their potential for societal impact (Eloundou et al. 2023; Kasneci et al. 2023; Li et al. 2023), this problem has become even more fundamental. Generative AI (GenAI) systems raise new questions when it comes to rationality, particularly because the content they generate is of a form that humans ascribe meaning to. Models like LLMs have been shown to appear irrational (Binz and Schulz 2023; Macmillan-Scott and Musolesi 2024; Hagendorff et al. 2023), and this is not because these models are incorrectly achieving their set goals, but because we are evaluating them by metrics that differ from their training goals. While LLMs are not the only type of relevant artificial agent, they have highlighted how quickly AI can become integrated in our decision-making. If artificial agents are to be used in a decision-making capacity, whether this be medicine, diplomacy, or in our day-to-day, we must be sure that their output satisfies some clearly defined criteria and that we align the objective functions with these criteria. When it comes to human–AI interaction, if we do attempt to design more 'human-like' machines, can we accept that they may at times mimic our own irrationality? These questions are crucial if we are to mitigate *accidents*: unintended but potentially harmful consequences of AI, a concern that is key within AI safety (Amodei et al. 2016; Hendrycks et al. 2022).

Conversely, are there ways to leverage *irrational* behaviour (both of humans and machines) in order to build more rational artificial agents? What are the best methods to identify and interact with irrational agents? As we develop increasingly complex and capable artificial agents, and as these become more integrated in everyday activities, an important question arises as to how the concept of rationality fits into their design. This may refer to assessing the rationality of the agent itself, but also to approaching interactions with other agents that may act in seemingly irrational ways, whether this is another artificial agent or a human.

A number of differing definitions of rationality have been proposed within the field of AI, meaning that it is often unclear what is meant by a *rational agent* (van der Hoek and Wooldridge 2003). Often, definitions have been adapted from our understanding of rationality in humans that has been developed in disciplines like economics, philosophy and psychology, rather than establishing a distinct definition of machine rationality. The notion of rationality plays a different role in different fields, and it is important to understand what conception is most suitable in each context (Besold and Uckelman 2018). Knauff and Spohn (2021a) provide a comprehensive overview of the history and current understanding of rationality in philosophy, cognitive psychology and other disciplines, highlighting the varying notions that persist and advocating for more integrated interdisciplinary approaches. We cannot determine what it means to act or reason irrationally if we do not have a clear definition of rationality. However, these concepts are two sides of the same coin: we cannot

define one without the other, and we will see that in certain scenarios, acting 'irrationally' can in fact be the most rational approach.

Certain types of artificial agents already employ 'irrational' behaviours in order to maximise reward, such as random actions taken as part of exploration in reinforcement learning (Ladosz et al. 2022), or profit non-maximising strategies that attempt to reduce potential loss (Ganzfried 2023). As we will see, there are cases where irrational behaviour can lead to optimal outcomes. Crucial in determining when each type of behaviour is most suitable is understanding the environment that an agent is operating in, as well as potential agents one may interact with.

However, it is often not an easy task to identify what type of agents one is interacting with, or to establish their level of rationality. A number of opponent modelling methods have emerged that attempt to make predictions about an opponent, including their goals and actions (Albrecht and Stone 2018). For existing techniques, including policy reconstruction (Ganzfried and Sandholm 2011; Chakraborty and Stone 2014; Silver et al. 2016; Mealing and Shapiro 2017) or classification (Abdul-Rahman and Hailes 2000; Schadd et al. 2007; Iglesias et al. 2010), restrictive assumptions often need to be made that require some knowledge about the agent that is being modelled. Further developments in the identification of irrational agents will require domain-specific research, where employing combinations of existing methods presents a promising avenue for study.

Similarly, once we have established that we are interacting with an irrational agent, what is the best strategy to follow? Current research has centred on interactions in adversarial domains (Bowling and Veloso 2002; Foerster et al. 2018; Letcher et al. 2019; Kim et al. 2021; Lu et al. 2022), or with human opponents (Uprety and Song 2018; McElfresh et al. 2021; Chan et al. 2021; Azaria 2022). It remains unclear how methods for adversarial environments can be adapted to employ in scenarios which may involve irrational agents. In particular, these methods may not be ideal when we want to achieve a cooperative outcome or are looking to obtain coordination between agents—one promising avenue of research that could be adapted to ensure cooperation are opponent shaping methods (Foerster et al. 2018; Letcher et al. 2019; Kim et al. 2021; Lu et al. 2022). Strategies must be adapted to the type of opponent, whether this be a human or artificial agent, therefore combining these methods with identification and classification techniques will likely be most effective.

While machines may be formulated to act as rationally as possible, humans have been shown to act and reason in irrational ways (Kahneman and Tversky 1982). We therefore need to design machines that are able to handle such scenarios and that do not assume perfect rationality of other agents. This is particularly important because we will see that perfect rationality is seldom the aim; instead, alternatives such as *bounded rationality* or *bounded optimality* are more realistic and often more desirable (Russell 1997), so artificial agents are more likely to encounter boundedly rational opponents.

In designing agents that achieve the optimal outcome, it may appear counter-intuitive to consider incorporating human cognitive biases. Nevertheless, we will explore how heuristics may be incorporated into decision-making in artificial agents to make them more efficient and improve their performance (Gulati et al. 2022). As well as achieving greater capabilities, the use of cognitive biases inspired by human reasoning can increase the explainability of AI processes (Newell et al. 1958; Simsek 2020). However, building 'irrational' machines raises questions as to how human–AI interactions may be impacted. While the incorporation of human physical attributes has been shown to have a positive impact on

human perception of artificial agents (Pizzi et al. 2020), questions remain as to whether the same effect is observed when concerning cognitive attributes, or even whether this is desirable in certain scenarios.

The aim of this article is to provide a comprehensive survey of rationality and irrationality within AI, as well as to set out the open questions in this area. Section 2 introduces the varying definitions of rationality in different disciplines: the section begins by setting out how it is understood in AI and how this may be assessed, then surveys the definition of rationality in three disciplines that have most heavily influenced the conception within AI: economics, philosophy, and psychology. As we will see, the interdisciplinary nature of AI means that developments in computer science have been heavily influenced by notions of rationality in other fields. A comparison of all four disciplines is presented in Table 1. Section 3 details *rationally irrational* behaviours, defined as behaviours that violate typical assumptions of rationality, but that can be shown to be rational under certain conditions (see Table 2). This section also includes a discussion of perceived irrationality in GenAI. Section 4 then surveys existing methods for identifying and interacting with irrational agents; literature included in this section is summarised in Tables 3 and 4. Section 5 focuses on the interplay between humans and AI: both how human irrationality can be incorporated into AI design in a beneficial way, as well as how the rationality of machines can have an impact on human–AI interactions. Section 6 explores the new challenges brought about by GenAI, including tensions between what we train models to maximise for and how we evaluate their output. Finally, Sect. 7 presents a number of open questions that remain to be addressed in this area—these constitute both fundamental theoretical questions, as well as potential avenues for research from a methodological standpoint.

Table 1 Rationality and irrationality in different disciplines

	AI	Economics	Philosophy	Psychology
How is rationality determined?	Optimal outcome	Utility maximisation	Logical reasoning	Contested (Rationality wars)
How does rationality relate to preferences?	Rationality requires consistent preferences	Rationality requires consistent preferences	Rational preferences are arrived at through justified reasoning	Rationality requires consistent preferences
What is meant by bounded rationality?	Limited computation/time	Finite recursion	Human computational limitations, limits on knowledge and attention	Human computational limitations, limits on knowledge and attention
What is considered irrational?	Taking an action that does not maximise expected utility	Taking an action that does not maximise expected utility, inconsistent preferences	Unjustified, inconsistent beliefs or actions	Human cognitive biases and heuristics (although contested)
Rational Reasoning	Correct inference	Consistent preferences	Justified beliefs (epistemic rationality)	Adherence to rules of probability and logic
Rational Behaviour	Rational agent—acts so as to achieve best expected outcome	Taking action that maximises expected utility	Justified actions / desires (practical rationality)	Acting according to rational reasoning

Table 2 Types of irrationality that may result in optimal outcomes

Type of irrationality	When is it optimal?	Literature
Bounded rationality	Limited resources	Simon (1957, 1982)
		Russell (1997)
		Lee (2021)
		Simsek (2020)
		Wen et al. (2020)
		Gigerenzer (2001)
		Hüllermeier et al. (2021)
Random behaviour	RL exploration Mixed Nash Equilibria	Shirado and Christakis (2017)
		Icard (2021)
		Ladosz et al. (2022)
		Still and Precup (2012)
		Fernàndez and Veloso (2006)
		Mnih et al. (2015)
		Polvara et al. (2018)
		Yu et al. (2020)
		Kobayashi and Hashimoto (2007)
Profit non-maximising	Irrational opponents Adversarial scenarios	Thomson (1979)
		Li and Faisal (2023)
		Ren et al. (2020)
		Ganzfried (2023)
		Goto et al. (2012)
Human irrationality	Human–AI interaction	Armstrong and Mindermann (2018)
		Chan et al. (2021)
		Chen et al. (2020, 2021)
		Ghosal et al. (2023)
		McElfresh et al. (2021)
		Skalse and Abate (2023)
		Stella and Bauso (2023)
		Uprety and Song (2018)
		Evans et al. (2015b, 2015a)
		Caliskan et al. (2017)

2 Defining & assessing rationality

2.1 Rationality in AI

2.1.1 Defining rationality

The concept of a *rational agent* has become an integral part of the AI discourse, although its precise meaning varies significantly between uses (van der Hoek and Wooldridge 2003). Whether considering an application in computer vision, LLMs, or reinforcement learning, the goal is invariably to arrive at a model that produces *rational* decisions (Wooldridge 2000). However, what is *rational* can be interpreted a number of ways (Wheeler 2020): for instance, it could mean making a decision closest to what a human would make, or achieving the optimal outcome, or even achieving a good outcome in the quickest time possible. Gershman et al. (2015, p. 273) define computational rationality as "identifying decisions

Table 3 Identifying irrational artificial agents summary table

Topic	Literature
Opponent modelling surveys	Fürnkranz (2001)
	Van Den Herik et al. (2005)
	Olorunleke and McCalla (2005)
	Albrecht and Stone (2018)
	Nashed and Zilberstein (2022)
Fictitious play	Brown (1951)
Inferring reward function considering irrationality	Armstrong and Mindermann (2018)
	Chan et al. (2021)
Rational verification	Abate et al. (2021)
	Hammond et al. (2021)
	Gutierrez et al. (2021)
Goal/plan recognition	Ramírez and Geffner (2011)
	Tian et al. (2016)
	Vered and Kaminka (2017)
	Masters and Sardina (2021)
Policy reconstruction	Ganzfried and Sandholm (2011)
	Chakraborty and Stone (2014)
	Silver et al. (2016)
	Mealing and Shapiro (2017)
Recursive reasoning	van der Hoek and Wooldridge (2002)
	Sonu and Doshi (2015)
	de Weerd et al. (2017)
	Wen et al. (2020)
Type-based reasoning	He et al. (2016)
	Albrecht and Stone (2017)
Multi-agent reinforcement learning	Zhang et al. (2021)
Assessing rationality in LLMs	Binz and Schulz (2023)
	Hagendorff et al. (2023)
	Macmillan-Scott and Musolesi (2024)
Safe exploration survey	Avrahami-Zilberbrand and Kaminka (2014)
Safe exploration	Masters and Sardina (2021)
	Tambe and Rosenbloom (1995)
	Mao and Gratch (2004)
	Sukthankar and Sycara (2005)
	Gan et al. (2019)
	Pawlick et al. (2019)

with highest expected utility, while taking into consideration the costs of computation." Within the field of AI, several differing approaches exist to the understanding of a rational agent. As we will see, developments within computer science have integrated conceptions of a rational agent from various fields (Zilberstein 2011).

Whereas AI was first intended to replicate human intelligence (Brooks 1991), it can be argued that pursuing this goal will not produce models that act in the most rational way (Wooldridge 2000), therefore we first need to establish what the goal of AI is. A useful categorisation distinguishes the goal of artificial agents across two dimensions: reasoning vs. behaviour, and human vs. rational (Russell and Norvig 2022). The latter distinction exem-

Table 4 Interacting with irrational artificial agents summary table

Problem/question	Algorithm/method	Literature
Behaviour of intelligent machines and their interaction with humans	–	Rahwan et al. (2019)
Minimisation of loss	Minimax	Osborne (2004)
Minimax extension for robust multi-agent RL	M3DDPG	Li et al. (2019)
Consider return distribution rather than expected return	Distributional RL	Bellemare et al. (2023)
Interaction with potentially irrational opponent	Safe equilibrium	Ganzfried (2023)
Accounting for errors in rational decision-making	Trembling hand perfect equilibrium	Selten (1975)
Optimisation of minimax	Alpha-beta pruning search	Knuth and Moore (1975)
Variable learning rate, ensuring convergence	WoLF	Bowling and Veloso (2002)
Opponent shaping	LOLA	Foerster et al. (2018)
	SOS	Letcher et al. (2019)
	Meta-MAPG	Kim et al. (2021)
	M-FOS	Lu et al. (2022)
Modelling (irrational) human behaviour	–	Uprety and Song (2018)
		McElfresh et al. (2021)
		Kwon et al. (2020)
		Chan et al. (2021)
		Azaria (2022)

plifies that often the 'ideal' reasoning or behaviour is not the one that most resembles human processes. In this definition, *rational thinking* is defined in line with the *laws of thought* approach, based in philosophy (Boole 1854; Leech 2015). This approach is grounded within the discipline of logic, however problems with defining a rational reasoning process often arise when it comes to the application, whether it be with expressing a problem formally or exhausting computational resources. On the other hand, an agent that exhibits *rational behaviour* is one that acts so as to achieve the best outcome (or expected outcome); this result may be obtained by, but is not restricted to, correct inference (in line with rational reasoning). The distinction between rational reasoning and behaviour in AI is closely linked to the categorisation of theoretical (or epistemic) and practical rationality that is often used to study human rationality (Wedgwood 2021). As we will see below, the study of epistemic rationality is more grounded within philosophy, whereas it is psychology that is most concerned with rational actions (Knauff and Spohn 2021b).

It is often easier to implement and assess the rational behaviour aspect, as opposed to rational reasoning, as we can more easily observe behaviour. Knauff and Spohn (2021b) frame this distinction using Marr's (1982) levels of analysis, where his computational and algorithmic levels are referred to as output-oriented and process-oriented respectively. We are also often more concerned with output-oriented, computational mechanisms being rational rather than processes; that is to say, the focus is generally placed on evaluating the results of reasoning rather than the reasoning itself. Depending on the model architecture, pitfalls in behaviour may appear in different ways—an important one being the introduction of bias. The distinction between statistical learning and symbolic AI allows us to separate models that incorporate human biases through the data input, as opposed to those where

an expression of 'irrationality' will come from the model architecture itself (Maruyama 2020)—both are subject to the incorporation of human bias, the variation arises in how the bias may be introduced. The former, statistical learning, can be defined as following a bottom-up approach. In this case, models generally learn from human data and inputs, and therefore the biases that appear are those that were introduced by the human input (Shin and Shin 2023). Numerous examples exist of cases where the introduction of such biases has led to problematic results, most notably in computer vision (Buolamwini and Gebru 2018; Cook et al. 2019; Wilson et al. 2019; Noiret et al. 2021; Papakyriakopoulos and Mboya 2023) and natural language processing (NLP) (Bolukbasi et al. 2016; Rozado 2020; Dawkins 2021; Hovy and Prabhumoye 2021; Stanczak and Augenstein 2021). On the other hand, biases in symbolic AI can be said to come from a more top-down approach. Given that these types of models or agents are generally designed to follow rules of logic and formal rationality, they do not generally exhibit the same biases as statistical learning (Maruyama 2020)—however, 'irrational' behaviour will present itself in different ways. In this case, it is the rules that underlie the model that are subject to human bias and may lead to irrational reasoning. There are increasingly examples of hybrid approaches that leverage a combination of the two, particularly advances in neuro-symbolic AI (Besold et al. 2022).

Some of the discussion within statistical learning centres around removing the presence of human biases, or mitigating their presence within models and results. However, some have argued that there exist cases where the use of heuristics similar to those used in human reasoning may make decision-making processes easier to explain (Newell et al. 1958; Simsek 2020). For instance, the use of tallying has been shown to be effective in many contexts, and to generalise to unseen problems (Dawes and Corrigan 1974; Einhorn and Hogarth 1975; Dawes 1979). Similarly, the use of heuristics in Tetris can make AI more explainable (Simsek et al. 2016). Simsek (2020) argues that in linear environments, there are three conditions under which heuristics can yield accurate decisions: simple dominance (Hogarth and Karelaia 2006), cumulative dominance (Kirkwood and Sarin 1985; Baucells et al. 2008), and noncompensatoriness (Martignon and Hoffrage 2002; Katsikopoulos and Martignon 2006; Simsek 2013). Developments in AI may not only render computational processes more explainable, but they may also reproduce processes in the human brain in such a way as to develop our understanding of how such processes occur. Until now, this research has predominantly focused on the use of deep neural networks (Cichy et al. 2016; Kubilius et al. 2016; Jozwik et al. 2017; O'Connell and Chun 2018; Bertolero and Bassett 2020), which have become in some cases a sort of 'model organism' (Scholte 2018; Firestone 2020). Through the development of the General Problem Solver (GPS), Newell and Simon (1961) presented a similar idea of using this program to aid in the construction of theories of human thinking.

The debate surrounding how we define rationality within AI also lends itself to the question of what we are trying to achieve. Russell (1997, 2016) identifies four types of goals: perfect rationality, calculative rationality, bounded rationality, and bounded optimality. Russell and Norvig (2022) define them as follows: *perfect rationality* occurs when an agent maximises expected utility with every action it takes—in most environments, this is not a realistic goal; *calculative rationality* describes the case when an agent would eventually return what would have been the utility-maximising choice, but it may return it too late; *bounded rationality*, as proposed by Simon (1957, 1982), follows the goal of *satisficing* that will be discussed below; and *bounded optimality* is more concerned with the decision-

making process rather than the output–it refers to a case where agents take decisions as rationally as possible given their available computational resources, achieving an expected utility at least as high as any other agent with access to the same resources. While perfect rationality appears the ultimate goal and is theoretically possible, such as in the case of the Universal Turing Machine (UTM), the Halting Problem means we cannot know whether, when or how UTM will stop computing (Lee 2021).

There exist additional intrinsic limitations in accordance to Gödel's Incompleteness Theorems that relate to the Halting Problem; authors like Lucas (1961) and Penrose (1989) have highlighted the consequences of the Gödel's theorem as a limiting factor for machine rationality. Given that as a consequence of Gödel's theorem there must always be a set of propositions that can neither be proven nor disproven, or a fact about the world that cannot be demonstrated as true, there will always exist problems that cannot be solved by machines. Therefore, it can be argued that striving for perfect rationality is a futile goal. In fact, the unsolvability of the Halting Problem has previously been used to prove Gödel's Incompleteness Theorem (Calude 2021; Tourlakis 2022).

Nevertheless, much of the debate in the field of AI centres around achieving perfect rationality. However, as we have seen there are constraints that, in many environments, make this an unrealistic goal. Instead of perfect rationality, Simon (1957, 1982) introduced the aforementioned notion of *bounded rationality* to explain how humans make decisions under uncertainty; he argued that individuals *satisfice* rather than maximise: due to limits of cognitive ability or computation, time constraints, or incomplete knowledge, humans choose the first result that satisfies a chosen 'aspiration level' rather than obtaining the optimal solution (Gigerenzer 2020; Schwarz et al. 2022). Further models of bounded rationality have emerged in the literature since (Rubinstein 1998; Gigerenzer and Selten 2002). Surprisingly, the concept of bounded rationality has largely remained separate from the field of AI (Simsek 2020; Lee 2021).

Arguably then, bounded optimality appears to offer the best framework: as opposed to calculative rationality, we can be sure of attaining a solution at the desired time, and we guarantee that this solution will have at least as high a utility as that achieved by a boundedly rational agent (Horvitz 1988; Russell 1997). Out of the four types of goals identified by Russell (1997, 2016), this comes closest to the above definition of computational rationality given by Gershman et al. (2015). There may even be cases where bounded optimality cannot be achieved, therefore Russell et al. (1993) define a weaker property, asymptotic bounded optimality, as a more robust and tractable alternative. Zilberstein (2011), who instead classifies bounded optimality as one amongst other approaches to bounded rationality, concurs that it is difficult to achieve in practice. Instead, he proposes *metareasoning* as a more promising approach—more specifically, optimal metareasoning. Inspired by a component of human decision-making, metareasoning can be defined as a way of allocating resources: it is a "mechanism to make certain runtime decisions by reasoning about the problem solving—or object-level—reasoning process" (Zilberstein 2011, p. 29).

It is clear that there is no comprehensive definition of rationality that serves every purpose. Although perfect rationality appears to be a sensible aim, we have seen there are many limitations to achieving this in practice (namely the Halting Problem). A similar issue arises with bounded optimality, which is often difficult to attain. Bounded rationality is a more realistic goal, although in implementing this we must accept that we will not always achieve the optimal outcome.

2.1.2 Assessing rationality

When assessing whether an agent has attained the level of rationality we are seeking, a crucial question lies in how we measure its performance. Taking inspiration from developmental and comparative psychology, the performance vs. competence debate can be applied to artificial agents (Firestone 2020; Lampinen 2023). The debate highlights the distinction between *competence*, defined as a system's internal knowledge and capabilities, and *performance*, which relates to demonstrations of this knowledge. A *species-fair* comparison between humans and machines requires ensuring that tests account for both human and machine constraints, and implement task alignment. If tasks are not defined with this notion in mind, poor performance may be taken as a sign of low competence, when in fact it arises from tests that are not well suited to the constraints of either the human or machine being evaluated.

The type of test we require to assess an agent's rationality also depends on what aspect we are interested in. Returning to the distinction between reasoning and behaviour, our method of assessment will vary for each of these. Often, the characteristic of interest is rational behaviour, as we are interested in the result. To assess whether behaviour is rational, we can apply the rational agent approach: *an agent that acts so as to achieve the best outcome, or best expected outcome under uncertainty, would be deemed a rational agent*. This is in contrast to a test for rational thinking, which would ascertain whether the agent reasons in accordance with the *laws of thought*. Recent work has also questioned classical definitions of rationality when it comes to conditional logic (Eichhorn et al. 2018), defining a set of non-monotonic inference patterns and assessing them based on the plausibility semantics of Ordinal Conditional Functions (OCF) (Spohn 1988). In so doing, the authors define a new way of evaluating what may be deemed rational according to plausible reasoning standards, with a requirement of consistency, applicable to both human reasoning and artificial agents. While here we are concerned with rational reasoning and behaviour, other types of assessment would be necessary to determine whether an agent is behaving or reasoning in a human way; the Turing test (Turing 1950) is seen as an example of this type of assessment, even though it has been criticised for its inherent limitations (Hernandez-Orallo 2000).

Assessing the rationality of GenAI systems poses a new set of challenges. Even more clarity about what and how we are evaluating is needed, as these systems encompass often conflicting types of rationality. Taking LLMs as an example, these models are trained with the goal of predicting the most likely next token. A model may exhibit high performance in this sense, but produce an output that appears irrational to a human interpreter. Because the output of LLMs are in a form that humans ascribe meaning to (language), we evaluate their responses as we would human speech. Therefore, there is a tension between the goal of the model (next-token prediction) and the way we evaluate its output under a human lens. We refer to this problem as the *GenAI Evaluation Paradox*. There is therefore a need to align the goals considered in training and evaluation methodologies. When evaluating LLMs, we often look at the correctness of the responses or the coherence of reasoning, but this is not what LLMs have been trained to maximise. As we will see below, alignment and safety fine-tuning is generally at odds with the model's attempt to minimise next-token prediction loss. This argument also holds for other types of outputs produced by GenAI systems, like images or audio. The question of GenAI and its implications of rationality will be further explored in Sects. 3.5 and 6.

2.2 Rationality in other disciplines

2.2.1 Economics

The concept of rationality is a central assumption in certain areas of economics. In its definition, particularly within normative decision theory, rationality has become analogous to consistency (Hammond 1997; Schilirò 2012)—as such, close parallels can be drawn to the definition of rationality in philosophy, as will be seen below. A rational action or decision is therefore one that maximises utility, or expected utility where there is incomplete information. While in other disciplines the focus generally lies within rational behaviour or reasoning, within economics, it is extended to questions of preferences, beliefs, expectations and knowledge (Hammond 1997). For instance, the consistency of ordinal preferences is often emphasised, where what matters is the order of these preferences (see Hicks and Allen (1934)).

Economic theory establishes a quantifiable way of assessing rationality based on measures of utility. However, real values of utility can seldom be measured in an objective way. As a response to this problem, Samuelson (1938) developed the theory of revealed preference, focusing on observed demand behaviour to infer underlying preferences. Ultimately, looking at demand behaviour to develop an understanding of the underlying preferences can be assimilated to the reasoning vs. behaviour distinction that was discussed above: where we cannot assess the rationality of reasoning, as we do not observe this process, we can instead evaluate the rationality of observed actions. The theory of revealed preference was later developed, building on Samuelson's work, to further establish how to infer preferences from demand functions (Houthakker 1950; Arrow 1959; Uzawa 1960; Afriat 1967).

Although it can be argued that economic theory ultimately seeks to study and predict human behaviour (Sugden 1991), its concept of rationality is, unlike human behaviour, perfect, logical and deductive (Arthur 1994). Gintis (2000) highlights this notion, arguing that following the traditional economics definition of rationality renders humans 'hopelessly irrational.' The disconnect between the theoretical and empirical evidence of behaviour (Tsetsos et al. 2016), grounded in the fact that humans cannot reason beyond given levels of complexity, is addressed by the aforementioned concept of bounded rationality. Thaler, who collaborated with and built on the work of Kahneman and Tversky [see for example Thaler (1980), Tversky and Thaler (1990), Kahneman et al. (1991)], incorporated findings from psychology into our understanding of economics. The field of behavioural economics was pioneered by Thaler, and there is now extensive research within economics surrounding cognitive biases and heuristics (Slovic et al. 2002; Grandori 2010; Zindel et al. 2014; Enke et al. 2023).

So far we have considered expected utility theory as a descriptive or predictive theory that sets out to either explain how humans make decisions or predict the choices they will make. However, another perspective offers a normative view of expected utility theory (Cave 2005), seeing it instead as a way of studying how humans *should* rather than *do* make decisions (Briggs 2023). From this perspective, the irrationality of observed behaviour in our decision-making does not contradict the theory; rather, expected utility theory may be more in line with the aforementioned perfect rationality approach. If we consider applications of expected utility theory to machine learning, in particular reinforcement learning, the interpretation as a normative theory becomes apparent (Parkes and Wellman 2015):

expected utility is used to decide what action an agent should take next, and in this way find an optimal policy to follow [see for example Charpentier et al. (2020)]. Learning algorithms have further incorporated data from human psychological experiments to model choice behaviour and human decision-making (Brian Arthur 1993).

2.2.2 Philosophy

When defining rationality within artificial agents, Aristotle's 'laws of thought' approach has been employed as a way of defining and assessing rational thinking in machines (Russell and Norvig 2022). For the ancient Greek philosophers, however, rationality was a uniquely human trait that sets us apart from other animals (Nozick 1993). The conception of rationality within philosophy can similarly be attributed to Aristotle, whose classification of the virtues led to the distinction between *epistemic* (or theoretical) rationality and *practical* rationality (Rysiew 2008; Okasha 2016; Wedgwood 2021). This categorisation of the two types of rationality, relating to beliefs and decisions respectively, is closely linked to the separation between reasoning and behaviour discussed above. As understood in the epistemic sense, rationality can often be equated to beliefs that are justified, and is closely linked to truth (Rysiew 2008); crucially, rational belief must be supported by reasons and be generated through a reliable process (Nozick 1993). Conversely, practical rationality concerns actions and desires (Battaly and Slote 2015), and is generally more the concern of psychological study. Foley (1987) points to issues with this distinction, and argues that inconsistent beliefs may still be rational from an epistemic standpoint.

The current debate within epistemology is centred between internalist and externalist perspectives. The former's origins are sometimes traced back to Hume (Meeker 2001), although some have argued that this attributed view points to an inconsistency in his moral theory (Brown 1988; Coleman 1992). In the internalist view, reason lies at the core of rational belief, and is as such internal to the person. A belief may be false, however if there is justification for this belief, it is not necessarily irrational to maintain said belief (Audi 2002). In contrast, externalists ascertain that rational thinking and the justification of a belief is not only a result of internal processes, but is also affected by external factors (Farkas 2003).

Other definitions of rationality within philosophy centre on reasoning with a logical grounding. Stein (1996) developed a *Standard Picture* of rationality, where the emphasis is on reasoning according to principles based on logic, probability theory and so forth. In trying to model how the mind works, Sloman (1993) used computers as a way to represent our own rationality, equating the mind to a *control system*. A similar framework to the *Standard Picture* is the 'classical' conception discussed by Chase et al. (1998), who argue for the ecological rationality of heuristics used by humans and organisms more generally. Evnine (2001) sets out a theory of the *Universality of Logic*, which contends that rational creatures must possess certain logical abilities. A more Kantian perspective has been suggested, defending the *logic-oriented conception of human rationality*, claiming that rational animals are intrinsically logical animals (Hanna 2006).

However, similar to economics, the concept of rationality within philosophy has also been impacted by empirical studies in psychology. The debate around cognitive biases and limits in human reasoning have given rise to questions surrounding how rationality is defined (Rysiew 2008): if humans do not follow the formal norms of rationality, does this render them irrational, or do we need to redefine our conception of rationality? Some see

rationality as an evolutionary adaptation, and in this conception, humans cannot be deemed irrational (Dennett 1981; O'Brien 1993). In this line of argument, an explanation can be found for any 'irrationalities' displayed by humans. Others argue that our definition of rationality should be based on human intuition (Cohen 1981), and that any mistakes are due to *performance* rather than *competence* errors. If we are to define rationality in accordance to the way humans reason, how would this then impact the conception of machine rationality? It may be that we need a definition of rationality exclusive to humans, and another for machines.

Others emphasise the distinction between *normative* and *descriptive* rationality, mirroring the discussion within economics. In this view, philosophy would be more concerned with normative rationality, whereas psychology centres on descriptive rationality (Knauff and Spohn 2021b). The two are intrinsically linked, as, for instance, rational action will typically depend on coherent beliefs. For this reason, Knauff and Spohn (2021b) argue for a more integrated study of normative and descriptive approaches, advocating for an interdisciplinary research agenda.

2.2.3 Psychology

With regards to defining rationality in psychology, much of the focus in the literature has been on understanding how humans make decisions, placing the emphasis on practical and descriptive rationality. Significant study in psychology has been on human cognitive biases and limitations in human reasoning, and whether this deems humans irrational, or whether we need to redefine what rationality means in relation to human behaviour. This debate has formed two major camps, denoted by Sturm (2012) as the 'heuristics and biases' approach versus the 'bounded rationality' approach, and has been referred to as the 'rationality wars' in the psychology of human reasoning (Samuels et al. 2002). The 'heuristics and biases' literature, where the most defining work was carried out by Tversky and Kahneman, contends that humans often do not reason according to rules of logic and probability [see for example Wason (1968), Tversky and Kahneman (1971, 1986), Kahneman and Tversky (1982), Kahneman et al. (1991); for collections of such works, see Nisbett and Ross (1980), Gilovich et al. (2002)]. While research in this area often emphasised the irrationality in human reasoning, others have proposed a novel way to evaluate rational reasoning in conditional logic, arguing that even though some inference patterns displayed by human reasoners are not valid according to classical logic, they are nevertheless plausible and should thus be considered rational (Eichhorn et al. 2018). The authors emphasise consistency in reasoning as necessary for rationality, rather than classical logic.

In contrast to the 'heuristics and biases' literature, the 'bounded rationality' scholarship argues that human beings are not irrational in the way they reason and make decisions—the way humans reason is rational given computational limitations, as well as limits in knowledge and attention (Gigerenzer 1993; Gigerenzer and Goldstein 1996). Lewis et al. (2014) develop a framework of *computational rationality* that looks at bounds on decision-making mechanisms within the brain and attempts to unify explanations of behaviour and mechanisms within psychology, tying in with the utility maximisation approach from economics. Whereas some have argued that the two perspectives can be reconciled, as it is a question of how much and when we follow certain norms (Samuels et al. 2002), others do maintain that the arguments are substantively different, although not wholly incompatible (Sturm 2012).

In their work on heuristics, Kahneman and Tversky (1982) identified a number of cognitive biases present in human reasoning that contradict the laws of thought and logic, such as the Linda Problem (conjunction fallacy) (Tversky and Kahneman 1983) or the Gambler's Fallacy (Tversky and Kahneman 1971). The results have been replicated empirically numerous times, showing under what conditions humans appear to violate these norms, as well as the conditions that reduce the presence of these biases (Nisbett and Borgida 1975; Clotfelter and Cook 1993; Donovan and Epstein 1997; Tentori et al. 2004; Barron and Leider 2010) [in recent years, similar experiments have been done with LLMs instead of humans (Macmillan-Scott and Musolesi 2024; Binz and Schulz 2023; Hagendorff et al. 2023)]. Although these biases seem to illustrate irrationality within human reasoning, some argue that these heuristics and biases are often useful when applied to the right domains—generally domains similar to ones in which they emerged—and can be remarkably effective (Kahneman and Tversky 1982; Gigerenzer and Goldstein 1996).

Given the existence of these heuristics that appear to show that human reasoning often contradicts formal norms of rationality, recent studies in psychology and neuroscience have argued that there exist two types of reasoning processes within human cognition. System-1 is associated with fast, automatic, and largely unconscious processes, and is thought to be a product of evolution; in contrast, System-2 is slower, rule-based, and subject to deliberate conscious control, it is more plastic and is shaped both by culture and education (Evans and Over 1996; Sloman 1996; Stanovich 1999; Rysiew 2008). The identification of these two processes has led to a discussion around how we define rationality within human reasoning, and the potential of splitting our understanding of rationality in such a way that it mirrors the two cognitive systems (Evans and Over 1996; Sloman 1996; Stanovich 1999; Saunders and Over 2009). The debate had expanded into the AI literature; Bengio (2019) assimilates current deep learning capabilities to System-1, and discusses how recent developments are paving the way for deep learning architectures with capabilities that include System-2 tasks, which require conscious reasoning.

Considering the differing conceptions of what rationality means within psychology, it is unsurprising that there is no unified test for rational thinking. One important distinction that is often made is that between rationality and intelligence (Sutherland 1992). On one hand, within the field of AI, Russell (1997) sets out potential definitions of intelligence as the types of rationality or optimality covered above, in so doing establishing machine intelligence as rationality. Similarly, Gershman et al. (2015) discuss a computational rationality framework as a way to study intelligence (both in brains and machines), again equating the two notions. On the other hand, within psychology there is a clear distinction. Stanovich et al. (2011) see rationality as a much broader notion than intelligence that encompasses more cognitive skills, particularly if we define intelligence by what is measured in commonly used intelligence tests. They see rationality as a superordinate concept to intelligence given that it encompasses "both the reflective mind and the algorithmic mind" (Stanovich et al. 2011, p. 814). In the same way that we have developed a way to measure intelligence through the use of the IQ test, Stanovich (2016) proposes an assessment of rational thinking called the Comprehensive Assessment of Rational Thinking (CART). Other types of assessment are used empirically for particular aspects of rationality, such as the use of dominance tests to study choice behaviour (Tervonen et al. 2018). Dominance tests are often used in discrete choice experiments (DCEs) (Ryan and Bate 2001; McIntosh and Ryan 2002), and have therefore been used predominantly within economics—as such, the definition of ratio-

nality in this type of assessment must be in line with axioms of rational choice, including completeness, transitivity, and monotonicity (Mas-Colell et al. 1995).

A similar debate around how to define rationality exists in neuroscience, where more emphasis has been placed on pathology and how this relates to rational reasoning. A crucial question concerns whether pathology indicates an impairment on rationality, or whether mental illness does not necessarily signify irrationality. While some psychopathology patients may exhibit irrational behaviour, many hold the view that irrationality is not sufficient or necessary for mental illness (Bortolotti 2013; Cardella 2020): insanity and irrationality are not intrinsically linked. If we take the goal of the brain as minimising surprise or uncertainty (Friston 2010), both healthy and diseased brains can be considered rational in so long as they pursue this goal (Fiorillo 2017). Conversely, research has shown that lesions to the orbitofrontal cortex may hinder the anticipation of negative emotional consequences (Kirman et al. 2010). Accounting for these consequences can lead to better informed decision-making and choice behaviour, therefore some pathologies may indeed affect rational reasoning. Perhaps counterintuitively, emotions influence and can be considered essential to rational thinking (Martins 2010). Therefore, it is neither pathology nor emotions that cause the biases in judgement that are observed in our decision-making.

Table 1 presents a comparative analysis of rationality and irrationality across the domains of AI, economics, philosophy, and psychology.

3 Rational irrationality

Having established that theories and definitions of rationality vary significantly across and within disciplines, it is a challenge to then establish what constitutes an irrational agent. Agents may deviate from rational reasoning or behaviour in different ways, and may be considered rational or not depending on the definition we are working off. As we have seen, it may be that we need to develop differing definitions for rationality in humans and machines, or that we need to reevaluate our understanding of rationality altogether. Eichhorn et al. (2018) demonstrate that some human inferences labelled as irrational can in fact reflect rational strategies when judged within a conditional plausibility framework rather than by means of classical logic. In this section, we will be adhering to the aforementioned notion of perfect rationality (Russell 1997); that is to say, we define a rational agent as one that invariably takes actions that maximise its expected utility, given the available information.

In this section, we will discuss types of irrationality—these are deviations from perfect rationality in different ways. However, as we will see below, the term ‘irrational’ must be taken with caution. Each type of irrationality that we discuss below can in fact constitute the optimal behaviour or reasoning in the right scenario. For example, we discuss bounded rationality in this section: bounded rationality is generally considered as rational behaviour, however we include it here as it deviates from perfect rationality. In fact, Gigerenzer (2001) distinguished between nonrational theories of decision-making and theories of irrational decision-making; this distinction emphasises that non-rational theories may not necessarily lead to suboptimal outcomes in the real world. The categorisation included here is not an exhaustive list, but presents the cases most studied in the literature and are all behaviours that commonly occur in practice. The cases presented here illustrate the notion that irratio-

nal behaviour can sometimes emerge as preferable or even optimal, and are summarised in Table 2.

The cases where irrational behaviour can, in some scenarios, be optimal are distinguished from perceived irrationality. Perceived irrationality is particularly relevant when it comes to GenAI, and as such is briefly discussed below. We return to a fuller consideration of GenAI in Sect. 6. Whereas as an agent may be perceived to be irrational when for instance their goals are unknown, irrational behaviour is denoted as that where no matter the goal, the behaviour is suboptimal (Masters and Sardina 2021). This may arise particularly in situations involving limited information—agents may not be aware of other agents' utility functions, and may have different reward structures in such a way that behaviour appears to be irrational but could in fact be optimal for another agent. In game theoretical frameworks, an agent may appear irrational when their model of the game differs from our own (Ganzfried 2023). Behaviour being perceived as irrational has been studied in the domain of human–computer interaction, where perceived irrationality may arise from the discrepancy between an individual's expectations and their perception of a smart system's actions (Abadie et al. 2019). As we will see in more detail in Sect. 5.2, the question of how rational a system should be when interacting with humans remains an open question.

3.1 Bounded rationality

Bounded rationality is one of the most interesting characterisations of possible types of rationality. Proposed by Herbert Simon, bounded rationality seeks to explain how humans make decisions under uncertainty, and puts forward the idea of *satisficing* rather than maximising (Simon 1957, 1982). As mentioned above, the principle behind *satisficing* is that due to given constraints, humans choose the first result that satisfies a chosen 'aspiration level' rather than obtaining the optimal solution (Gigerenzer 2020; Schwarz et al. 2022). Although it was developed to better understand human decision-making, bounded rationality can also be applied to artificial agents and their behaviour under uncertainty or computational constraints (Russell 1997).

However, until now, research on artificial agents has not greatly overlapped with the study of bounded rationality (Simsek 2020; Lee 2021), therefore open questions remain on the behaviour and interaction of agents with limited computation. The constraints that exist in human reasoning differ from those of artificial agents, as does the threshold for what can be deemed as a *satisficing* solution. With regards to artificial agents, Wen et al. (2020) study this problem in the context of recursive reasoning in multi-agent interactions. They introduce the Generalized Recursive Reasoning (GR2) framework to model the dynamics in interactions of agents with differing levels of rationality—in this case, the levels of rationality are interpreted as levels of recursive reasoning. The motivation for this research is that, in interactions with irrational (non-optimal) agents, the effectiveness of existing MARL models significantly decreases (Shoham et al. 2003). In their model, an agent with a given level of recursive reasoning takes the best response to all possible lower-level agents. In this sense, the boundedly rational agent is taking the best response to other agents whose rationality is bounded to a higher degree, meaning that each agent is determining the best response when interacting with agents that have a lower level of recursive reasoning and therefore whose rationality is more bounded. While this generates valuable results, questions remain surrounding the behaviour of agents when interacting with opponents whose

rationality is less bounded—in this example, looking at environments with agents operating at higher levels of recursion.

3.2 Random behaviour

Random behaviour is behaviour that appears to have no rationale or clear motivation, and as such is interpreted as irrational due to the lack of an identifiable goal. It can be distinguished from other types of irrationality that, although suboptimal, still adhere to a recognisable strategy. However, random actions are often crucial to the behaviour of artificial agents, particularly in situations of bounded rationality or limited resources. When interacting with human participants, artificial agents that incorporate some degree of randomness have also been shown to improve collective performance (Shirado and Christakis 2017). Icard (2021) highlights two compelling rationales for randomisation: costly computation, where establishing the best course of action requires resources, and finite memory. He concludes that while there may be Bayesian justifications against randomising behaviour, there will arguably always be cases where a *satisficing* solution, following Simon's definition, requires some randomising.

Within the field of reinforcement learning, random exploration has emerged as the most common technique for exploration (Ladosz et al. 2022)—in particular, ε-greedy exploration, which maintains a balance between *exploration* and *exploitation* (e.g., Kaelbling et al. 1996; Bloembergen et al. 2015). Whereas exploration employs random behaviour to look for potentially higher rewards, exploitation sticks to known high rewards. Therefore, random behaviour does not necessarily equate to irrational behaviour—where there is a longer-term motivation, taking random actions may in fact allow an agent to arrive at the optimal strategy (Still and Precup 2012). The existence of this long-term motivation can constitute the determinant of whether this behaviour can be attributed rationality: random behaviour can in fact lead to optimal decision-making in the long run.

Within research on reinforcement learning, and particularly deep reinforcement learning, there is evidence of the advantages of random exploration (Fernàndez and Veloso 2006; Mnih et al. 2015; Polvara et al. 2018; Yu et al. 2020). However, there are also downsides to this type of behaviour, namely its inefficiency as it often revisits the same states. Ladosz et al. (2022) consider three approaches to address this inefficiency: reduced states/actions for exploration methods, exploration parameters methods, and network parameter noise methods. Each of these presents its own improvements and disadvantages; for instance, exploration parameter methods are particularly useful in achieving a good balance between exploration and exploitation, but does not fully solve the question of inefficient exploration of previously seen states. Nevertheless, these methods based on random exploration exemplify that such a behaviour is not inherently irrational.

Similarly, randomisation holds an important role within stochastic games. In game theory, the Nash Equilibrium is often used to signify the optimal strategy given available information. While these solutions are sometimes constituted by pure strategy Nash Equilibria, often there also exists a mixed solution (Daskalakis et al. 2009; Bloembergen et al. 2015). Mixed solutions in game theory involve randomisation over a set of strategies with a given probability distribution (Leyton-Brown and Shoham 2008). Therefore, in such cases, randomness exists as part of the 'ideal' rational behaviour (Icard 2021). As with reinforcement learning, it is then the use of randomness with a broader motivation that renders this type of

behaviour ultimately rational in these scenarios. The motivation behind random actions may not always be known to other agents in the interaction, and as such behaviour that appears random may only be perceived irrationality or may be deceptive.

3.3 Profit non-maximising

Revisiting the definition of a perfectly rational agent as one that takes actions to maximise utility at every instance, an agent that, given the same information, does not take these actions seems to provide a clear indication of an irrational agent. There are situations where agents may follow a strategy in which the action with highest known utility is not chosen. One such case is the maximin strategy (Thomson 1979): taking the most unfavourable action that could be taken by the opponent, the maximin strategy determines how to maximise one's own utility in response to this action. Essentially, instead of maximising utility, this strategy seeks to minimise loss. Increasing attention has also been placed on the distributional reinforcement learning framework (Bellemare et al. 2023), which takes into account the full distribution of the reward or return and attempts to minimise loss in this way. Some approaches have also combined the minimax algorithm and distributional learning in order to improve performance (Ren et al. 2020; Li and Faisal 2023). While this behaviour appears irrational, cases have been studied where playing 'rationally' (according to Nash Equilibria) can be shown to result in suboptimal outcomes. Notably, when playing against irrational agents, or agents that are following strategies from a different Nash Equilibrium, employing the 'optimal' strategy may result in lower payoff (Ganzfried 2023). Employing a loss-minimising strategy may also be advantageous in adversarial settings; however, purely conflicting-interest scenarios are rare in the real world, instead we are generally operating in mixed-motive or common-interest environments (Dafoe et al. 2020).

Departing from the maximin strategy, Ganzfried (2023) proposes a *safe equilibrium* that balances between the maximin and Nash Equilibrium strategies by calculating the probability that the opponent is acting rationally. It is interesting to note that Ganzfried does not study this type of behaviour as a type of irrational behaviour, but as a strategy to use in interactions with potentially irrational agents (this will be discussed in more depth in the following sections). Other types of profit non-maximising behaviour can in a similar way be viewed as a strategy to minimise loss when faced with a potentially irrational opponent rather than as an irrational behaviour in itself.

In contrast, Goto et al. (2012) present a similar method, however they do denote this behaviour as *partially irrational*. In their approach, artificial learning agents take decisions with low utility with a specified probability—their rationale for these 'irrational' actions is to give the agent a possibility to escape from suboptimal Nash Equilibria. The approach they present can be closely assimilated to exploration within reinforcement learning, and similarly is shown to sometimes arrive at a more optimal solution than a strategy following perfect rationality would. Their model also exemplifies how game theoretic concepts can be integrated into learning algorithms to produce agents that learn to make decisions in multi-agent settings.

3.4 Human irrationality in artificial agents

Among the types of irrationality we consider, artificial agents that contain or model human irrationality have received most attention (e.g., Armstrong and Mindermann 2018; Uprety and Song 2018; Chen et al. 2020, 2021; Chan et al. 2021; McElfresh et al. 2021; Stella and Bauso 2023; Skalse and Abate 2023; Ghosal et al. 2023). Two categories can be identified among agents of this class: the first are agents where human irrationality or cognitive biases have been intentionally incorporated, and the second is ones where human biases arise unintentionally, generally due to the data inputs—this distinction will be discussed in more depth in Sect. 5. The former has been shown to be a useful way to understand human mechanisms and learn how to improve interactions between humans and artificial agents (Uprety and Song 2018).

Whereas reinforcement learning was inspired by the way humans and other animals learn and make decisions under uncertainty (Littman 2015), inverse reinforcement learning (IRL) methods have been used to try and infer a reward function from observed behaviour. Skalse and Abate (2023) look at the most common IRL methods, namely optimality (Ng and Russell 2000), Boltzmann rationality (Ramachandran and Amir 2007), and causal entropy maximisation (Ziebart 2010), and test how robust these models are to misspecification. They find the most robust to be the Boltzmann-rational model, an important finding given that IRL models of human behaviour will always be misspecified to some degree. Boltzmann rationality (Luce 1959; Ziebart et al. 2010) is defined as that which "predicts that a human will act out a trajectory with probability proportional to the exponentiated return they receive for the trajectory" (Laidlaw and Dragan 2022, p. 1). However, others have argued that IRL models that assume noisy rationality, including Boltzmann rationality, are not suitable as the way humans deviate from rational behaviour is not merely noisy, but is instead systematic (Evans et al. 2015a, b). Armstrong and Mindermann (2018) find noisy rationality to be too strong of an assumption as it fails to account for bias, whereas a weaker simplicity assumption is also insufficient, and suggest more work is needed between the two extremes. Chan et al. (2021) similarly find that models of human behaviour as noisy-rational rather than systematically irrational perform significantly worse than those incorporating cognitive biases; in this case, Boltzmann rationality is used to model a noisy-rational agent, which is compared to systematic deviations from the Bellman equation to model irrationality. They further show that correctly modelling an agent's irrationality can lead to higher performance than inference from rational ones.

Others have shown that when artificial agents adopt cognitive biases observed in humans, there are cases where the performance of the agents can be improved, particularly when interacting with humans. Chen et al. (2020) demonstrate that introducing option comparison when making decisions under uncertainty can outperform calculation-based methods in highly noisy environments. While option comparison is often regarded as an irrational behaviour, they show that this heuristic is attained by a learning agent that is maximising cumulative reward. In the realm of language learning, models that acquire human biases can also be useful in better understanding existing prejudices and stereotypes (Caliskan et al. 2017).

3.5 Perceived irrationality in GenAI

GenAI systems present a slightly different case compared to the types of irrationality outlined above. Instead of having a case where irrational behaviour can in fact be optimal, we have multiple ways of assessing the rationality of a system. There is a tension between the *generative rationality* and rationality that we perceive or ascribe to the system. Taking the example of LLMs, generative rationality is achieved by minimising loss in next-token prediction, whereas we interpret the output using a more human lens. As we have seen, because the output of GenAI systems such as LLMs is in a form that humans ascribe meaning to, we therefore evaluate the output not in terms of how well it achieved the goal set in the model's architecture; instead, we often evaluate these models using tasks designed to evaluate human reasoning (Binz and Schulz 2023; Macmillan-Scott and Musolesi 2024; Hagendorff et al. 2023). We may look for characteristics like coherence or accuracy in LLM outputs. If, instead of language, the model simply produced the embeddings as output, we would likely evaluate them as (boundedly) rational. Therefore, these models often appear irrational not because they are incorrectly minimising loss, but because there is a *rationality entanglement*: overlapping types of rationality that are sometimes at odds. In fact, these hallucinations, contradictions, or other observed behaviours may be rational under the model's objective function.

The issue of rationality entanglement arises due to our way of interpreting the model's output. We could refer to this as *interpreted* or *perceived* irrationality rather than actual irrationality, as we are in a sense evaluating these models incorrectly. The issue of interpreted irrationality can also be seen in other multi-agent settings that do not involve GenAI, where behaviour is merely interpreted as irrational, even though an agent is correctly maximising expected utility. This interpretation may arise for a variety of reasons, such as an incorrect assumption about the opponent's reward function, or noisy information about an opponent's actions.

The behaviour that we interpret as irrational in GenAI systems therefore arises because we are not evaluating how well the model achieves the goal it was designed to attain. There is a misalignment in the objectives rather than a failure in reasoning. Using a traditional definition, an LLM would display rational reasoning by predicting the most likely next token, but instead we want it to produce true and consistent texts. As we will further explore below, additional layers like alignment and safety fine-tuning often in fact reduce the rationality of a model in the sense of loss minimisation. These additional training layers generally attempt to ensure that these models produce outputs that conform more closely to human expectations of rationality, even if this is at the expense of the rationality of the model in a more traditional sense. Are we then training GenAI models to behave as rational agents, or merely to behave in ways that appear normatively rational to humans?

4 Dealing with irrational artificial agents

4.1 Identifying irrational artificial agents

We have seen that different types of seemingly irrational agents may in fact be pursuing the optimal policy when operating in the right environment. However, this creates complica-

tions for agents modelling other agents, as the typical assumption of other agents' rationality may lead to a suboptimal outcome. Albrecht and Stone (2018) note that accurate modelling of other agents is particularly important in the absence of coordination and communication protocols. The first step is therefore to identify whether an agent is rational or not, and if faced with an an irrational agent, how to identify the type of irrational behaviour.

Methods for identification of agent rationality fall under the category of opponent modelling [for surveys, see Fürnkranz (2001), Van Den Herik et al. (2005), Olorunleke and McCalla (2005), Albrecht and Stone (2018), Nashed and Zilberstein (2022)], an early example of which is fictitious play (Brown 1951). These techniques often need to make assumptions not only about the modelled agent itself, but also about the environment—in particular relating to the observability of other agents' action (Albrecht and Stone 2018). Building an accurate model of observability of the environment is especially relevant when interacting with potentially irrational agents, as we have seen that modelling irrational behaviour as simply noisy can lead to much worse performance (Armstrong and Mindermann 2018; Chan et al. 2021). The work on *rational verification* is also relevant here, particularly in its application to multi-agent systems (Abate et al. 2021; Hammond et al. 2021). Rational verification checks whether given temporal logic properties will hold in a system, assuming that all agents within this system will act rationally and so choose strategies that form a game-theoretic equilibrium (Gutierrez et al. 2021). In building these methods, a distinction must also be made between irrationality and unreliability—whereas in the former, observed behaviours cannot be ignored, the latter relates to noisy observations that did not in fact happen and as such should not be used in modelling an agent's intent or reward function (Masters and Sardina 2021).

Several opponent modelling techniques exist in the literature, including goal/plan recognition (Ramírez and Geffner 2011; Tian et al. 2016; Vered and Kaminka 2017; Masters and Sardina 2021), policy reconstruction (Ganzfried and Sandholm 2011; Chakraborty and Stone 2014; Silver et al. 2016; Mealing and Shapiro 2017), recursive reasoning (van der Hoek and Wooldridge 2002; Sonu and Doshi 2015; de Weerd et al. 2017; Wen et al. 2020) and type-based reasoning (He et al. 2016; Albrecht and Stone 2017)—for a comprehensive survey of these methodologies, see (Albrecht and Stone 2018). Some of these methods directly assume rationality of a certain type: recursive reasoning techniques, such as Wen et al.'s (2020) discussed above, assume the agent being modelled is boundedly rational, and reasons with a given level of recursion. Classification instead attempts to correctly predict a label or category for the opponent (Albrecht and Stone 2018)—although in this case the way the modelled agent deviates from perfect rationality is not assumed from the outset, the observed behaviour is classified within existing models. In contrast, some goal recognition techniques have been proposed that calculate a rationality measure based on the agent's action history, and use this to inform the formula used to model the agent (Masters and Sardina 2021).

Work on agent modelling techniques has seen many recent developments (Zhang et al. 2021). However, numerous open problems and questions remain, in particular regarding the rationality of agents in an interaction. Knowing in what scenarios we expect to encounter different types of irrational agents would allow for more efficient and targeted techniques. A reinforcement learning agent is likely to exhibit random behaviour as part of exploration, and it would be a mistake to interpret this as irrational; in contrast, when interacting with a human agent, observed random behaviour may be a more accurate signal of irratio-

nality. Similarly, the type of agent that is interpreting the behaviour is important: LLMs have been shown to display behaviour that appears irrational to humans (Macmillan-Scott and Musolesi 2024), but its output is correct according to the specifications of the model, and therefore is rational from a technical perspective. Therefore, developments in this area require domain-specific research. Other open questions pertain to the possible combinations of opponent modelling techniques, as well as safe probing and exploration techniques (Albrecht and Stone 2018).

Safe exploration is of particular importance as agents that appear irrational may in fact be rational but deceptive agents. It may be very difficult to distinguish between the two, and particularly costly if the opponent turns out to be deceptive. Existing methods for dealing with deceptive agents remain largely domain-specific (Masters and Sardina 2021); some have looked at ways to reveal an agent's true intent (Tambe and Rosenbloom 1995), others have focused on disambiguation of goals (Mao and Gratch 2004; Sukthankar and Sycara 2005), or have focused on Stackelberg security games where there is a potential for the attacker to be untruthful (Gan et al. 2019)—for a survey, see Avrahami-Zilberbrand and Kaminka (2014). A clear application of safe exploration is to problems in cyber-security, where deception plays a central role (Pawlick et al. 2019), as in this area a priority may be mitigating the potential harmful effects of a deceptive agent.

4.2 Interacting with irrational artificial agents

Having established that an interaction is with an irrational agent, the question then arises of determining the best strategy to maximise payoff in such a scenario. Research in this area has centred on two aspects: interacting with irrational agents in adversarial scenarios, and interactions with irrational human agents. How the presence of an irrational agent may hinder cooperation/coordination and the best strategies to achieve cooperation with such an agent has received less attention, and is an important avenue of research. Applications pertain not only to scenarios that involve artificial agents, but also within human–machine interactions. As we have seen, humans often reason and act in ways that deviate from perfect rationality, therefore machines must be able to account for this. Not only should machines be able to account for irrational human decision-making; we also need to gain a better understanding of how interactions with humans may alter the behaviour of machines (Rahwan et al. 2019). It may be the case that techniques used in adversarial scenarios can be adapted to non-adversarial interactions, such that we interpret an irrational agent as a type of opponent.

In adversarial interactions, techniques have emerged to minimise an agent's own loss, as well as techniques that attempt to shape the behaviour of the opponent(s). A notable example of the former approach is the minimax algorithm (discussed above), which establishes the best strategy to minimise loss (Osborne 2004). Others have built on this idea, and developed further algorithms to adapt the minimax theorem to situations involving learning agents (Li et al. 2019). Distributional reinforcement learning has also emerged as a technique that considers not only the optimal action, but instead takes into account the full distribution of the potential return (Bellemare et al. 2023). Whereas these approaches focus on minimising loss, Ganzfried's (2023) approach discussed above balances between the maximin algorithm and Nash Equilibrium strategies, resulting in a 'safe equilibrium.' The trembling hand perfect equilibrium (Selten 1975) similarly attempts to account for some

degree of uncertainty, although for very small values of ε, where ε is the probability of an agent 'making a mistake' and not following the optimal strategy. A more efficient version of the minimax approach is given by the alpha-beta pruning search algorithm, which eliminates provably irrelevant subtrees to arrive at the same optimal move as minimax (Knuth and Moore 1975). Bowling and Veloso (2002) propose an algorithm that builds on their Win or Learn Fast (WoLF) principle: whereas the safe equilibrium approach adapts to a measure of the opponent's rationality (Ganzfried 2023), the WoLF principle reacts to their perceived position in relation to other agents by increasing learning rate when losing, and erring on the side of caution when winning. They note that previous multi-agent learning algorithms offered either convergence or rationality, whereas the WoLF principle achieves both properties. These varying approaches propose ways to adapt to the opponent's behaviour in such a way as to maximise reward in highly uncertain scenarios.

Another recent line of research is that of opponent shaping techniques. These techniques attempt to leverage the learning of opponents to one's advantage. Examples of these methods include the Learning with Opponent-Learning Awareness (LOLA) algorithm (Foerster et al. 2018), the Stable Opponent Shaping (SOS) algorithm (Letcher et al. 2019), or the Meta Multi-Agent Policy Gradient (Meta-MAPG) theorem (Kim et al. 2021). Lu et al. (2022) argue that these previous methods are myopic, asymmetric and require the use of higher-order derivatives; they instead propose Model-Free Opponent Shaping (M-FOS), which formulates the problem as a meta-game and does not require a model of the opponent's underlying learning algorithm. However, the application of these models remains relatively limited, focusing on social dilemma settings like the Iterated Prisoner's Dilemma. Further research is needed into the wider application of these methods, and the potential to combine them with loss-minimising techniques.

Aside from approaches intended for adversarial scenarios, methods have emerged that address the interaction of artificial agents with irrational humans. As mentioned in the previous section, some of these methods first attempt to model the biases and irrational behaviour exhibited by human agents (Uprety and Song 2018; McElfresh et al. 2021), in particular as attempting to model human behaviour as noisy-rational has been shown to lead to lower performance (Kwon et al. 2020; Chan et al. 2021). Azaria (2022) highlights that existing models generally assume that interactions between artificial and human agents can be modelled as purely zero-sum or fully cooperative, as such ignoring the human's goals and potential for adaptation to the behaviour of the artificial agent. Future work can build on techniques covered in the previous section for identifying and modelling systematic irrational behaviour in human agents, and develop methods that optimise the interaction between artificial and human agents in scenarios that are closer to the real world.

5 Human and AI irrationality

5.1 Incorporating human irrationality into AI

Although we have seen that humans often do not act fully rationally, and human reasoning exhibits many cognitive biases, there are cases where we may want to incorporate aspects of human reasoning into artificial agents. Initially it may appear counterintuitive—if we have established that these biases often violate the laws of logic or probability, why would

we want machines to do the same? In answering this question, it is useful to reflect on why these biases appear in human reasoning. As we have seen, psychologists argue that in certain domains, heuristics and biases found in human reasoning can be remarkably effective (Kahneman and Tversky 1982; Gigerenzer and Goldstein 1996; Gigerenzer and Brighton 2009). For instance, sometimes we fail to consider all available information, opting instead for quicker decision-making—in many situations, this may be a desirable strategy for humans to adopt (Sutherland 1992). Similarly, we may sometimes want an artificial agent to prioritise a quicker decision as opposed to a more 'rational' one, in a sense following Simon's (1957) idea of *satisficing*, and Russell's (1997, 2016) argument that in many cases bounded rationality is a more desirable goal than perfect rationality. It is also important to note that this is nothing new, humans have long inspired the development of AI, with notable examples including neural networks (Gurney 2018) and reinforcement learning (Littman 2015).

A distinction must be made between artificial agents that inadvertently incorporate human biases (primarily due to the data they have been trained on), and agents that intentionally incorporate some of these biases and heuristics. There are many examples of the former in statistical learning, such as in computer vision (O'Neil 2016; Buolamwini and Gebru 2018; Cook et al. 2019; Wilson et al. 2019; Noiret et al. 2021; Papakyriakopoulos and Mboya 2023) and NLP (Bolukbasi et al. 2016; Rozado 2020; Dawkins 2021; Hovy and Prabhumoye 2021; Stanczak and Augenstein 2021), where the data used to train models contained biases and has resulted in negative or harmful consequences. Risks arise when machines that learn from human data miscategorise human irrationalities as human values and optimise for these irrationalities (Gorman and Armstrong 2022). With the advent of LLMs, we are now seeing agents that not only inadvertently incorporate human biases, but also mimic limitations in human reasoning and exhibit the same cognitive biases (Bubeck et al. 2023; Macmillan-Scott and Musolesi 2024; Binz and Schulz 2023; Hagendorff et al. 2023).

However, in this section we focus on the second case—on those instances in which it may be beneficial to purposefully integrate heuristics into the way artificial agents make decisions. Often, the benefits of incorporating heuristics arise in cases where there are time constraints, limited computational resources, or partial information (Simsek 2020). Gigerenzer (2001) presents an illustrative example of a robot attempting to catch a ball: following a rational approach, the robot would need information on all the possible trajectories the ball might follow, as well as ways to measure it's velocity and angle, among other factors. However, professional athletes use the simple heuristic of keeping their eyes on the ball as they run towards it. Using this heuristic, a robot would only need to know the angle of gaze to achieve a high performance—a much simpler, 'less rational' method that achieves the same result.

Gulati et al. (2022) propose a taxonomy of cognitive biases present in human decision-making that may be valuable in the development of AI systems. While they highlight that this is particularly important for the future of human–machine collaboration, we argue that the benefits of research in this area will also be present in interactions between artificial agents, as well as in single agent settings. One example of this are the symmetry and mutual exclusion biases, which have been incorporated into a Naïve Bayes classifier (Taniguchi et al. 2017) and neural networks (Taniguchi et al. 2019; Manome et al. 2021). The authors show that these methods can be used to produce models that learn from small and biased datasets, and that their models outperform more traditional machine learning methods. Oth-

ers have also demonstrated how decision heuristics can outperform more complex statistical methods when there is limited data and training examples (Brighton 2006; Katsikopoulos 2011; Simsek and Buckmann 2015; Buckmann and Şimsek 2017). Similarly, within natural language, this type of approach can remove the need for a human trainer and lead to agents that learn from their own mistakes and correct this in planning sequences (Shinn et al. 2023). The effective use of human cognitive biases is also evident under uncertainty: Chen et al. (2020) show that an agent that integrates calculative and comparative decision-making outperforms an agent that merely calculates the optimal choice.

The benefits of incorporating cognitive biases into artificial agents do not only relate to improving the capabilities of boundedly rational agents. The use of heuristics can also be used to enhance the explainability of AI processes (Newell et al. 1958; Simsek 2020), as well as to design agents that are more robust to interactions with non-rational or boundedly rational agents (Wen et al. 2020). Examples of the latter can be found in the study of population dynamics (Yang et al. 2018) and video game design (Hunicke 2005; Peng et al. 2017). Much of the research in this area is recent, so questions remain as to the potential benefits that can arise from incorporating heuristics derived from human decision-making into artificial agents; the application to boundedly rational agents is clearer, but it has been shown that the impact of such heuristics can extend beyond bounded rationality. Notably, research has shown that behaviours that are often thought to be irrational in humans can emerge in artificial agents from a learning process that seeks to maximise cumulative reward (Howes et al. 2016; Chen et al. 2020).

5.2 Human–AI interaction with irrational machines

Artificial agents are becoming more and more integrated into our lives, and the frequency of human–AI interactions is quickly rising (Amershi et al. 2019). Therefore, it is important to understand how the rationality of these artificial agents impact the interaction. At the same time, it is worth noting that this relationship is bidirectional: the behaviour of machines and humans will each impact the other, and it is this co-behaviour that required a deep understanding (Rahwan et al. 2019). Small errors in the algorithm or data may have large unpredictable effects on society, and could even have the potential to alter the social fabric (Rahwan et al. 2019). As underlined by the field of AI safety, flaws in the design of AI systems may result in emergent behaviours that are potentially harmful and may pose a risk to society (Amodei et al. 2016; Hendrycks et al. 2022).

Although the goal is often to create agents that are as close to perfectly rational as possible, humans may feel more comfortable in interactions with agents that are slightly irrational or unpredictable. This is particularly relevant with the widespread use of LLMs like ChatGPT. On the other hand, artificial agents that are less rational may be perceived as untrustworthy or incompetent. The discrepancy between an individual's expectations and their perception of a smart system's actions can lead to the human perceiving the system as irrational (Abadie et al. 2019). The threshold for losing trust in algorithms appears to be much lower than for humans—a phenomenon referred to as *algorithmic aversion* (Dietvorst et al. 2015). Studies have looked at humans' perception of machines as an aid to decision-making (Binns et al. 2018), and the use of machines has been shown in cases to amplify human biases (Glickman and Sharot 2022).

Some have argued that human–machine collaboration may require AI systems that replicate human cognitive biases (Gulati et al. 2022). However, it has also been shown that we do not necessarily perceive and evaluate machines in the same way as humans and other components of our environment (Hidalgo 2021). Therefore, some of the biases that have been studied in human interaction may not be present in human–AI interaction, such as the social desirability bias (Gulati et al. 2022). Studies have looked at how the incorporation of human physical attributes affect the dynamics of the interaction: Pizzi et al. (2020) show how anthropomorphic agents can have a positive impact on the human's perception of the artificial agent, increasing satisfaction and acceptance; a similar effect is found for increasing trust in agents with more anthropomorphic features (Waytz et al. 2014). Whether this is also the case with cognitive attributes is currently not well understood (Shanahan 2024). Another important aspect is human's perception of machine morality; Benn and Grastien (2021) propose an approach of conveying ethical understanding in human–robot interactions, highlighting that what matters is not only the morality of an agent, but also the appearance of ethical understanding indicated by its behaviour.

6 The GenAI evaluation paradox

GenAI is unique in that it generates content in a format that humans ascribe meaning to. This characteristic of generative models leads to the *GenAI Evaluation Paradox*. There is a tension between what models are being trained to maximise and what they are being evaluated on Goldstein (2024). This issue can be viewed as a different kind of performance vs. competence distinction (Firestone 2020). As LLMs are the most commonly used types of GenAI systems, we can take these as an example, although the same argument applies to other GenAI systems that produce different types of output like visual or audio output (or combinations of modalities). At their core, LLMs aim to achieve predictive or generative rationality: this can be measured by looking at how good the models are at predicting the most likely next token. Another way to assess LLMs is in accordance to their epistemic rationality: the definitions of rational reasoning in philosophy and psychology discussed above that is used to evaluate human reasoning. For instance, Jiang et al. (2025) determine four necessary axioms for a rational agent: information grounding, logical consistency, invariance from irrelevant information, and orderability of preference. They clearly state that these axioms have been derived from work in cognitive psychology, so are similar to principles we would use to evaluate rational reasoning in humans.

The evaluation of LLM outputs through epistemic rationality leads to anthropomorphisation of LLMs and often results in researchers concluding that these models reason in irrational ways, such as by displaying human biases or outputting hallucinations. There is a misalignment between what we are training the models to maximise: generative rationality, and the behaviour we in actuality want these models to exhibit: epistemic rationality. There may then be a need to reevaluate how we are setting the training goals of GenAI systems to ensure that these are more in line with what we want to assess them on, whether this be accuracy, morality, or another metric.

The behaviour exhibited by LLMs that is interpreted as irrational sometimes mirrors biases in human reasoning, whereas other times it emerges in other ways. Numerous studies have applied cognitive psychology tools to evaluate whether LLMs replicate human cog-

nitive biases and heuristics (Binz and Schulz 2023; Macmillan-Scott and Musolesi 2024; Hagendorff et al. 2023), treating these models like participants in a psychology experiment and evaluating their outputs as you would for humans. The presence of this mimetic irrationality in LLM reasoning arises from the data they are trained on, which contains our own biases. However, sometimes LLMs display emergent irrational reasoning that is not present in the training data. For instance, they may hallucinate facts, contradict themselves, or give incorrect explanations. The presence of such emergent behaviour can sometimes be a product of additional training mechanisms such as Reinforcement Learning from Human Feedback (RLHF). RLHF has been shown to induce sycophantic behaviour in LLMs (Sharma et al. 2024), even to the point of displaying manipulative and deceptive tendencies in order to maximise feedback from users (Williams et al. 2025). Again, this is an example where the model is performing rationally according to its goals, but not according to what we consider rational or desirable. Aside from RLHF, other types of alignment and safety fine-tuning can similarly be at odds with the model's initial objective function. These methods often alter the model's weights, and some even modify the objective function itself (e.g., RLHF or contrastive learning). These efforts generally aim to constrain the output of GenAI systems within ethical and social norms. There is also a tension between the different types of safety fine-tuning; for instance, some types of fine-tuning may attempt to improve truthfulness, whereas RLHF can counteract this by reinforcing more sycophantic tendencies, which leads to LLMs lying in order to obtain positive feedback (Williams et al. 2025).

It is not enough to merely reevaluate how we define the model's objectives or evaluation metrics. Depending on the application and scenario that a GenAI system is being used in, we will want to maximise or curtail certain behaviours. Therefore, working to develop more specialised systems that are better suited to particular applications may mean we are better able to define what we are trying to achieve. We have already discussed that there is no unified conception of rationality within computer science, and that different types of rationality are better suited to certain applications. Rapid advancements in GenAI compel us to answer these questions with more urgency. In critical decision-making scenarios we will want to maximise properties like truthfulness and accuracy, whereas in more creative applications we may want to maximise originality. It is worth noting that virtually none of the applications that LLMs are currently applied to, if any, seek to maximise solely the performance of next-token prediction. Nonetheless, all of its current applications require the next-token prediction to work at least at a satisfactory level, leading us back to Simon's idea of *satisficing* and bounded rationality (Simon 1957, 1982).

GenAI presents a difficult balancing act between overlapping and sometimes conflicting objectives that lead to a rationality entanglement: we need to find the right equilibrium between generative performance of next-token prediction, alignment to human preferences, epistemic coherence, and task specialisation. Disentangling the tension between these aims is crucial to inform model architectures and evaluations.

7 Open questions

Below we define a set of open questions in a number of areas of rationality in AI. The list is not exhaustive but sets out a series of promising research directions or areas that have much to be explored.

7.1 Defining rationality in AI

1. *Is a general, unified definition of rationality in AI achievable?*
2. *How should rationality be defined for GenAI systems?*
3. *Should we distinguish between definitions and assessments of rationality for symbolic AI and statistical learning?*

As we have seen, different disciplines define rationality in very different ways. The understanding of rationality within AI has been influenced by some of these, particularly by work within economics, philosophy, and psychology. As artificial agents become more complex and capable, in particular considering the emerging capabilities of foundation models, it is important to determine whether we can or should establish a unifying definition of what constitutes a rational agent. A general framework could provide common ground for evaluating reasoning and decision-making across systems, but may be too general to prove useful. It will likely not suffice to adapt a definition from human rationality; we need to develop a definition of machine rationality distinct to that of humans. Even a distinct definition for AI agents that is set apart from human rationality will likely need to be further categorised. For example, GenAI systems may need to be evaluated in light of the GenAI Evaluation Paradox. Additionally, distinctions between symbolic AI and statistical learning approaches may warrant separate classification frameworks.

7.2 Training and evaluation

4. (a) *What novel training or evaluation methodologies can be developed that account for the GenAI Evaluation Paradox?*
   (b) *Should we develop benchmarks or testing suites that integrate multi-dimensional rationality, particularly for GenAI?*
5. *How can we realign the training objectives of GenAI systems with human-evaluated outputs?*
6. *Should we evaluate rational processes as well as rational output?*

In order to avoid the apparent failures in *competence* being judged because of poor *performance* (Firestone 2020), we need to ensure that we are training AI agents with objectives that are aligned to what they will be evaluated on. This issue is particularly pertinent for GenAI, where there is often a misalignment between training and evaluation. We are generally interested in output characteristics like accuracy or lack of harmful content, whereas, at their core, GenAI systems are not trained to maximise for these objectives. The rationality entanglement present in GenAI systems raises the question of whether we can account for rationality in multiple dimensions. Similarly, we generally focus on measuring only outputs, whereas looking at the rationality of reasoning and decision-making processes could provide novel insights.

7.3 Conflicts and trade-offs

7. *What are the computational and ethical trade-offs between pursuing perfect vs. bounded rationality in AI?*
8. *What are the safety implications of building systems that deviate from perfect rationality?*
9. *Can a single system maximise multiple types of rationality at once, or will there inevitably be a trade-off?*

When setting our objectives in the design of AI agents, there are a number of conflicts and trade-offs that must be considered. Perfect rationality has been shown to be an unrealistic goal (Russell 1997, 2016; Lee 2021); in choosing to aim instead for goals like bounded rationality, we need to understand what the implications of this choice are. Systems that deviate from perfect rationality may introduce unpredictability. Also, as we saw with GenAI, depending on how we evaluate a system, there may be more than one type of rationality that can be contained and assessed within one system. In such a case, we will likely need to prioritise which type of rationality we are most interested in.

7.4 Understanding irrationality in AI

10. *Can we develop a generalisable framework to categorise types of irrationality in artificial agents?*
11. *In what scenarios can the incorporation of irrational behaviour improve capabilities relating to issues like trust, performance or cooperation?*
12. *Can the modelling of irrationality in AI provide insights into human reasoning and decision-making?*
13. *How can we distinguish between actual irrationality and deceptive behaviour in artificial agents, especially in adversarial settings?*

We need to deepen our understanding not only of rationality, but also of the various forms that irrationality can take. We have seen that irrational behaviour does not necessarily lead to sub-optimal outcomes. A more general and comprehensive categorisation of different types of irrationality would be beneficial in clearly establishing when these types of behaviours can be useful and improve capabilities of AI systems or help us to investigate characteristics of human behaviour. A distinction must also be made between irrational behaviour and behaviour that is perceived to be irrational, but may in fact be deceptive. For instance, LLMs outputting inaccurate information may be a symptom of manipulative or deceptive behaviour (Williams et al. 2025).

7.5 Human–AI interaction

14. *Can we integrate human heuristics into AI reasoning to improve interpretability?*
15. *Should AI aim to correct human irrationality or adapt to it in collaborative contexts?*

16. *How should AI systems balance exploration and exploitation in environments where human preferences are uncertain or evolving?*

Heuristics and biases in human reasoning may at times lead us to incorrect conclusion, but they are useful in making decisions under uncertainty. Similar mechanisms can potentially be integrated into the decision-making of AI systems, and these may even improve interpretability (Newell et al. 1958; Simsek 2020; Simsek et al. 2016).

Increasingly, AI agents will need to interact with humans, and there are many open questions around how to approach this. In some cases, like safety-critical scenarios, it may be beneficial to try to correct human irrationality, whereas in others the AI system may need to adapt to this behaviour. When artificial agents encounter human irrationality, we need to decide what the interaction should look like. In particular, learning from human behaviour may be more complex than learning from artificial agents due to our changing preferences and reward functions (Carroll et al. 2024).

7.6 Multi-agent and collective rationality

17. *What combinations of opponent modelling techniques yield the most robust results in environments with irrational agents?*
18. *Can we study collective rationality in AI systems in the same way that we do individual rationality?*

Multi-agent settings with potentially irrational agents require robust opponent modelling techniques. It remains to be seen what methods can be adapted from adversarial scenarios, and what additional methods can be developed for these setting.

In this paper, although we have discussed multi-agent scenarios, the focus has been on individual rationality. Much of the literature, not just in computer science, has focused on individual rationality (Knauff and Spohn 2021b). We will likely need to take a different approach to better understand collective rationality, distinguishing between types of agents, as well as between scenarios involving only artificial agents and those including humans. Collective or social rationality adds additional dimensions, like group dynamics, communication and shared goals.

8 Conclusion

Rationality and irrationality play an important role within AI. Designing the ‘rational agent’ has been the objective and the focus of research and practice in the field for many years. However, achieving this requires a common understanding of what it means to be rational. As we have seen, definitions of rationality within economics, philosophy, and psychology have informed our understanding of the concept within AI. We may aim for different types of rationality, as certain types will be more suitable in some contexts. Crucially, a distinction must be made between how we define rationality in humans and machines. This article provides a survey of the existing literature and methods in this area. Having established that particular irrational behaviour may in fact be optimal under certain conditions, it is

clear that the issue is complex. When it comes to GenAI, additional questions and tensions arise regarding how we train and evaluate these models. From a methodological perspective, techniques have emerged for the identification and interaction with irrational agents. However, these remain scarce—further research is needed in this area, both looking at how existing opponent modelling or shaping methods may be adapted to account for irrationality, as well as how they may be combined to better suit the problem at hand. The question of interacting with irrational agents is crucial not only among machines, but also because humans often act in irrational ways. Human–AI interaction is a key aspect of today's AI systems, namely with the case of systems based on LLMs and their widespread use. Cognitive biases may in some instances be leveraged to improve the performance of artificial agents, whereas in human–AI interaction the design of machines will need to account for the irrational behaviour of humans. At the same time, further understanding is needed around how the rationality of an artificial agent impacts the dynamics of human–AI interaction—there may be cases where it is beneficial for the artificial agent to appear not fully rational, whereas in other cases this can deem the agent untrustworthy. This article lays out open questions relating to rationality within AI. As we strive to create more capable agents that are increasingly integrated in everyday lives, these questions will need to be addressed.

Acknowledgements The authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions, which have helped to improve the quality and scope of this manuscript.

Author contributions M.M. and O.M.-S. conceptualised the paper. O.M.-S. wrote the manuscript text and M.M. supervised the work. M.M. and O.M.-S. reviewed the manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abadie A, Carillo K, Fosso Wamba S, Badot O (2019) Is Waze joking? Perceived irrationality dynamics in user-robot interactions. In: Proceedings of HICSS-19

Abate A, Gutierrez J, Hammond L, Harrenstein P, Kwiatkowska M, Najib M et al (2021) Rational verification: game-theoretic verification of multi-agent systems. Appl Intell 51(9):6569–6584

Abdul-Rahman A, Hailes S (2000) Supporting trust in virtual communities. In: Proceedings of HICSS-00. pp 6007–6016

Afriat SN (1967) The construction of utility functions from expenditure data. Int Econ Rev 8(1):67–77

Albrecht SV, Stone P (2017) Reasoning about hypothetical agent behaviours and their parameters. In: Proceedings of AAMAS-17. pp 547–555
Albrecht SV, Stone P (2018) Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif Intell 258:66–95
Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P et al (2019) Guidelines for human-AI interaction. In: Proceedings of CHI-19, New York, USA. pp 1–13
Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv: 1606.06565
Armstrong S, Mindermann S (2018) Occam's razor is insufficient to infer the preferences of irrational agents. In: Proceedings of NeurIPS-18
Arrow KJ (1959) Rational choice functions and orderings. Economica 26(102):121–127
Arthur WB (1994) Inductive reasoning and bounded rationality. Am Econ Rev 84(2):406–411
Audi R (2002) The architecture of reason: the structure and substance of rationality. Oxford University Press, New York
Avrahami-Zilberbrand D, Kaminka GA (2014) Keyhole adversarial plan recognition for recognition of suspicious and anomalous behavior. In: Sukthankar G, Geib C, Bui HH, Pynadath DV, Goldman RP (eds) Plan, activity, and intent recognition: theory and practice. Morgan Kaufmann, Boston, pp 87–119
Azaria A (2022) Irrational, but adaptive and goal oriented: humans interacting with autonomous agents. In: Raedt LD (ed) Proceedings of IJCAI-22. pp 5798–5802
Barron G, Leider S (2010) The role of experience in the gambler's fallacy. J Behav Decis Mak 23:117–129
Battaly H, Slote M (2015) Virtue epistemology and virtue ethics. In: Besser-Jones L, Slote M (eds) The Routledge companion to virtue ethics (chapter 19). Routledge
Baucells M, Carrasco JA, Hogarth RM (2008) Cumulative dominance and heuristic performance in binary multiattribute choice. Oper Res 56(5):1289–1304
Bellemare MG, Dabney W, Rowland M (2023) Distributional reinforcement learning. The MIT Press, Cambridge
Bengio Y (2019) The consciousness prior. arXiv: 1709.08568
Benn C, Grastien A (2021) Reducing moral ambiguity in partially observed human-robot interactions. Adv Robot 35(9):537–552
Bertolero MA, Bassett DS (2020) Deep neural networks carve the brain at its joints. arXiv: 2002.08891
Besold TR (2013) Rationality in|for|through AI. In: Kelemen J, Romportl J, Zackova E (eds) Beyond artificial intelligence: contemplations, expectations, applications. Springer, Berlin/Heidelberg, pp 49–62
Besold TR, Uckelman SL (2018) Normative and descriptive rationality: from nature to artifice and back. J Exp Theor Artif Intell 30(2):331–344
Besold TR, d'Avila Garcez A, Bader S, Bowman H, Domingos P, Hitzler P et al (2022) Neural-symbolic learning and reasoning: a survey and interpretation. In: Hitzler P, Sarker MK (eds) Neuro-symbolic artificial intelligence: the state of the art (chapter 1). IOS Press
Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'It's Reducing a Human Being to a Percentage': perceptions of justice in algorithmic decisions. In: Proceedings of CHI-18. pp 1–14
Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci 120(6):e2218523120
Bloembergen D, Tuyls K, Hennes D, Kaisers M (2015) Evolutionary dynamics of multi-agent learning: a survey. J Artif Intell Res 53:659–697
Bolukbasi T, Chang K-W, Zou J, Saligrama V, Kalai A (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of NeurIPS-16. pp 4356–4364
Boole G (1854) An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities. Walton and Maberly, London
Bortolotti L (2013) Rationality and sanity: the role of rationality judgments in understanding psychiatric disorders. In: Fulford KWM, Davies M, Gipps R, Graham G, Sadler JZ, Stanghellini G, Thornton T (eds) The Oxford handbook of philosophy and psychiatry. Oxford University Press
Bowling M, Veloso M (2002) Multiagent learning using a variable learning rate. Artif Intell 136(2):215–250
Brian Arthur W (1993) On designing economic agents that behave like human agents. J Evol Econ 3:1–22
Briggs RA (2023) Normative theories of rational choice: expected utility. In: Zalta EN, Nodelman U (eds) The Stanford encyclopedia of philosophy, Fall 2023 edn. Metaphysics Research Lab, Stanford University
Brighton H (2006) Robust inference with simple cognitive models. In: Proceedings of the 2006 AAAI Spring symposium. pp 17–22
Brooks RA (1991) Intelligence without representation. Artif Intell 47(1):139–159
Brown GW (1951) Iterative solution of games by fictitious play. In: Koopmans TC (ed) Activity analysis of production and allocation. Wiley, pp 374–376
Brown C (1988) Is hume an internalist? J Hist Philos 26(1):69–87

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P et al (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Proceedings of NeurIPS-20. Curran Associates, Inc, pp 1877–1901
Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E et al (2023) Sparks of artificial general intelligence: early experiments with GPT-4. arXiv: 2303.12712
Buckmann M, Şimsek Ö (2017) Decision heuristics for comparison: how good are they? In: Guy TV, Kárný M, Rios-Insua D, Wolpert DH (eds) Proceedings of the NeurIPS-16 workshop on imperfect decision makers. pp 1–11
Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of FAccT-18. pp 77–91
Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186
Calude CS (2021) Incompleteness and the halting problem. Stud Log 109(5):1159–1169
Cardella V (2020) Rationality in mental disorders: too little or too much? Eur J Anal Philos 16(2):13–36
Carroll M, Foote D, Siththaranjan A, Russell S, Dragan A (2024) AI alignment with changing and influenceable reward functions. In: Proceedings of icml'24. JMLR.org
Cave EM (2005) A normative interpretation of expected utility theory. J Value Inquiry 39:431
Chakraborty D, Stone P (2014) Multiagent learning in the presence of memory-bounded agents. Auton Agent Multi-Agent Syst 28(2):182–213
Chan L, Critch A, Dragan A (2021) Human irrationality: both bad and good for reward inference. arXiv: 2111.06956
Charpentier A, Elie R, Remlinger C (2020) Reinforcement learning in economics and finance. arXiv: 2003.10014
Chase VM, Hertwig R, Gigerenzer G (1998) Visions of rationality. Trends Cogn Sci 2(6):206–214
Chen H, Chang HJ, Howes A (2020) Implications of human irrationality for reinforcement learning. arXiv: 2006.04072
Chen H, Chang HJ, Howes A (2021) Apparently irrational choice as optimal sequential decision making. In: Proceedings of AAAI-21. pp 792–800
Cichy R, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6:27755
Clotfelter CT, Cook PJ (1993) Notes: the "Gambler's Fallacy" in lottery play. Manag Sci 39(12):1521–1525
Cohen LJ (1981) Can human irrationality be experimentally demonstrated? Behav Brain Sci 4(3):317–331
Coleman D (1992) Hume's internalism. Hume Stud 18(2):331–347
Cook CM, Howard JJ, Sirotin YB, Tipton JL, Vemury AR (2019) Demographic effects in facial recognition and their dependence on image acquisition: an evaluation of eleven commercial systems. IEEE Trans Biom Behav Identity Sci 1(1):32–41
Dafoe A, Hughes E, Bachrach Y, Collins T, McKee KR, Leibo JZ et al (2020) Open problems in cooperative AI. arXiv: 2012.08630
Daskalakis C, Goldberg PW, Papadimitriou CH (2009) The complexity of computing a Nash equilibrium. Commun ACM 52(2):89–97
Dawes RM (1979) The robust beauty of improper linear models in decision making. Am Psychol 34(7):571–582
Dawes RM, Corrigan B (1974) Linear models in decision making. Psychol Bull 81(2):95–106
Dawkins H (2021) Marked attribute bias in natural language inference. In: Proceedings of ACL-IJCNLP-21
de Weerd H, Verbrugge R, Verheij B (2017) Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. Auton Agent Multi-Agent Syst 31:250–287
Dennett DC (1981) True believers: the intentional strategy and why it works. In: Haugeland J (ed) Mind design II: philosophy, psychology, and artificial intelligence. The MIT Press
Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol 144(1):114–126
Donovan S, Epstein S (1997) The difficulty of the Linda conjunction problem can be attributed to its simultaneous concrete and unnatural representation, and not to conversational implicature. J Exp Soc Psychol 33(1):1–20
Eichhorn C, Kern-Isberner G, Ragni M (2018) Rational inference patterns based on conditional logic. In: Proceedings of the AAAI conference on artificial intelligence, vol 32, no 1
Einhorn HJ, Hogarth RM (1975) Unit weighting schemes for decision making. Organ Behav Hum Perform 13(2):171–192
Eloundou T, Manning S, Mishkin P, Rock D (2023) GPTs are GPTs: an early look at the labor market impact potential of large language models. arXiv: 2303.10130

Enke B, Gneezy U, Hall B, Martin D, Nelidov V, Offerman T, van de Ven J (2023) Cognitive biases: mistakes or missing stakes? Rev Econ Stat 105(4):1–15
Evans JSBT, Over DE (1996) Rationality and reasoning. Taylor & Francis, Oxford
Evans O, Stuhlmueller A, Goodman N (2015a) Learning the preferences of bounded agents. In: Proceedings of the NeurIPS-15 workshop on bounded optimality
Evans O, Stuhlmueller A, Goodman N (2015b) Learning the preferences of ignorant, inconsistent agents. In: Proceedings of AAAI-15
Evnine SJ (2001) The universality of logic: on the connection between rationality and logical ability. Mind 110(438):335–367
Farkas K (2003) What is externalism? Philos Stud 112:187–208
Fernàndez F, Veloso M (2006) Probabilistic policy reuse in a reinforcement learning agent. In: Proceedings of AAMAS-06, New York, NY, USA. pp 720–727
Fiorillo CD (2017) Neuroscience: rationality, uncertainty, dopamine. Nat Hum Behav 1:0158
Firestone C (2020) Performance vs. competence in human-machine comparisons. Proc Natl Acad Sci 117(43):26562–26571
Foerster JN, Chen RY, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I (2018) Learning with opponent-learning awareness. In: Proceedings of AAMAS-18
Foley R (1987) The theory of epistemic rationality. Harvard University Press, Cambridge
Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138
Fürnkranz J (2001) Machine learning in games: a survey. In: Fürnkranz J, Kubat M (eds) Machines that learn to play games. Nova Science Publishers, pp 11–59
Gan J, Guo Q, Tran-Thanh L, An B, Wooldridge M (2019) Manipulating a learning defender and ways to counteract. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Proceedings of NeurIPS-19. Curran Associates, Inc
Ganzfried S (2023) Safe equilibrium. In: Proceedings of the 62nd IEEE conference on decision and control (CDC), pp 5230–5236
Ganzfried S, Sandholm T (2011) Game theory-based opponent modeling in large imperfect-information games. In: Proceedings of AAMAS-11
Gershman SJ, Horvitz EJ, Tenenbaum JB (2015) Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. Science 349(6245):273–278
Ghosal GR, Zurek M, Brown DS, Dragan AD (2023) The effect of modeling human rationality level on learning rewards from multiple feedback types. In: Proceedings of AAAI-23
Gigerenzer G (1993) The bounded rationality of probabilistic mental models. In: Manktelow KI, Over DE (eds) Rationality: psychological and philosophical perspectives. Taylor & Frances/Routledge, pp 284–313
Gigerenzer G (2020) What is bounded rationality? In: Viale R (ed) Routledge handbook of bounded rationality (chapter 2). Routledge
Gigerenzer G (2001) Decision making: nonrational theories. Int Encycl Soc Behav Sci 5:3304–3309
Gigerenzer G, Brighton H (2009) Homo heuristicus: why biased minds make better inferences. Top Cogn Sci 1(1):107–143
Gigerenzer G, Goldstein D (1996) Reasoning the fast and frugal way: models of bounded rationality. Psychol Rev 103(4):650–669
Gigerenzer G, Selten R (2002) Bounded rationality: the adaptive toolbox. The MIT Press, Cambridge
Gilovich T, Griffin D, Kahneman D (2002) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, Cambridge
Gintis H (2000) Beyond homo economicus: evidence from experimental economics. Ecol Econ 35(3):311–322
Glickman M, Sharot T (2022) Biased AI systems produce biased humans. OSF Preprints
Goldstein S (2024) LLMs can never be ideally rational. PhilPapers
Gorman R, Armstrong S (2022) The dangers in algorithms learning humans' values and irrationalities. arXiv: 2202.13985
Goto T, Hatanaka T, Fujita M (2012) Payoff-based inhomogeneous partially irrational play for potential game theoretic cooperative control: convergence analysis. In: Proceedings of ACC-12, pp 2380–2387
Grandori A (2010) A rational heuristic model of economic decision making. Ration Soc 22(4):477–504
Gulati A, Lozano MA, Lepri B, Oliver N (2022) BIASeD: bringing irrationality into automated system design. arXiv: 2210.01122
Gurney K (2018) An introduction to neural networks. CRC Press, London
Gutierrez J, Hammond L, Lin AW, Najib M, Wooldridge M (2021) Rational verification for probabilistic systems. arXiv: 2107.09119
Hagendorff T, Fabi S, Kosinski M (2023) Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nat Comput Sci 3(10):833–838
Hammond PJ (1997) Rationality in economics. Rivista Internazionale di Scienze Sociali 105(3):247–288

Hammond L, Abate A, Gutierrez J, Wooldridge M (2021) Multi-agent reinforcement learning with temporal logic specifications. In: Proceedings of AAMAS-21. pp 583–592
Hanna R (2006) Rationality and logic. The MIT Press, Cambridge
He H, Boyd-Graber J, Kwok K, Daumé H III (2016) Opponent modeling in deep reinforcement learning. In: Proceedings of ICML-16, New York, USA. pp 1804–1813
Hendrycks D, Carlini N, Schulman J, Steinhardt J (2022) Unsolved problems in ML safety. arXiv: 2109.13916
Hernandez-Orallo J (2000) Beyond the turing test. J Logic Lang Inform 9(4):447–466
Hicks JR, Allen RGD (1934) A reconsideration of the theory of value. Part I. Economica 1(1):52–76
Hidalgo CA (2021) How humans judge machines. MIT Press, Cambridge
Hogarth R, Karelaia N (2006) "Take-the-Best" and other simple strategies: why and when they work "Well" with binary cues. Theor Decis 61:205–249
Horvitz EJ (1988) Reasoning about beliefs and actions under computational resource constraints. In: Proceedings of UAI-87, Arlington, Virginia, USA. pp 429–447
Houthakker HS (1950) Revealed preference and the utility function. Economica 17(66):159–174
Hovy D, Prabhumoye S (2021) Five sources of bias in natural language processing. Lang Linguist Compass 15(8):e12432
Howes A, Warren PA, Farmer GD, El-Deredy W, Lewis RL (2016) Why contextual preference reversals maximize expected value. Psychol Rev 123(4):368–391
Hüllermeier E, Mohr F, Tornede A, Wever M (2021) Automated machine learning, bounded rationality, and rational metareasoning. arXiv:2109.04744
Hunicke R (2005) The case for dynamic difficulty adjustment in games. In: Proceedings of ACE-05. pp 429–433
Icard T (2021) Why be random? Mind 130(517):111–139
Iglesias J, Angelov P, Ledezma Espino A, Sanchis de Miguel A (2010) Evolving classification of agents' behaviors: a general approach. Evol Syst 1:161–171
Jiang B, Xie Y, Wang X, Yuan Y, Hao Z, Bai X et al (2025) Towards rationality in language and multimodal agents: a survey. arXiv: 2406.00252
Jozwik KM, Kriegeskorte N, Storrs KR, Mur M (2017) Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. Front Psychol 8:1726
Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4(1):237–285
Kahneman D, Tversky A (1982) The psychology of preferences. Sci Am 246(1):160–173
Kahneman D, Knetsch JL, Thaler RH (1991) Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 5(1):193–206
Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F et al (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 103:102274
Katsikopoulos KV (2011) Psychological heuristics for making inferences: definition, performance, and the emerging theory and practice. Decis Anal 8(1):10–29
Katsikopoulos KV, Martignon L (2006) Naïve heuristics for paired comparisons: some results on their relative accuracy. J Math Psychol 50(5):488–494
Kim DK, Liu M, Riemer MD, Sun C, Abdulhai M, Habibi G et al (2021) A policy gradient algorithm for learning to learn in multiagent reinforcement learning. In: Proceedings of ICML-21. pp 5541–5550
Kirkwood CW, Sarin RK (1985) Ranking with partial information: a method and an application. Oper Res 33(1):38–48
Kirman A, Livet P, Teschl M (2010) Rationality and emotions. Philos Trans R Soc 365:215–219
Knauff M, Spohn W (2021a) The handbook of rationality. The MIT Press, Cambridge
Knauff M, Spohn W (2021b) Psychological and philosophical frameworks of rationality—a systematic introduction. The handbook of rationality. The MIT Press, Cambridge
Knuth DE, Moore RW (1975) An analysis of alpha-beta pruning. Artif Intell 6(4):293–326
Kobayashi S, Hashimoto T (2007) Analysis of random agents for improving market liquidity using artificial stock market. In: Proceedings of ESSA-07
Kubilius J, Bracci S, Op de Beeck HP (2016) Deep neural networks as a computational model for human shape sensitivity. PLoS Comput Biol 12(4):1–26
Kwon M, Biyik E, Talati A, Bhasin K, Losey DP, Sadigh D (2020) When humans aren't optimal: robots that collaborate with risk-aware humans. In: Proceedings of HRI-20
Ladosz P, Weng L, Kim M, Oh H (2022) Exploration in deep reinforcement learning: a survey. Inf Fusion 85:1–22
Laidlaw C, Dragan A (2022) The Boltzmann policy distribution: accounting for systematic suboptimality in human models. In: Proceedings of ICLR-22
Lampinen AK (2023) Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. arXiv: 2210.15303

Lee C (2021) The game of go: bounded rationality and artificial intelligence. In: Kiel L, Elliott E (eds) Complex systems in the social and behavioral sciences: theory, method and application. University of Michigan Press, pp 157–180
Leech J (2015) Logic and the laws of thought. Philosophers' Imprint, 15
Letcher A, Foerster J, Balduzzi D, Rocktäschel T, Whiteson S (2019) Stable opponent shaping in differentiable games. In: Proceedings of ICLR-19
Lewis RL, Howes A, Singh S (2014) Computational rationality: linking mechanism and behavior through bounded utility maximization. Top Cogn Sci 6(2):279–311
Leyton-Brown K, Shoham Y (2008) Essentials of game theory: a concise multidisciplinary introduction, vol 2. Morgan & Claypool Publishers, Williston
Li L, Faisal AA (2023) Adversarial distributional reinforcement learning against extrapolated generalization. In: Proceedings of EWRL-23
Li S, Wu Y, Cui X, Dong H, Fang F, Russell S (2019) Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In: Proceedings of AAAI-19. pp 4213–4220
Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW (2023) Ethics of large language models in medicine and medical research. Lancet Digit Health 5(6):e333–e335
Littman M (2015) Reinforcement learning improves behaviour from evaluative feedback. Nature 521:445–51
Lu C, Willi T, de Witt CS, Foerster J (2022) Model-free opponent shaping. In: Proceedings of ICML-22
Lucas JR (1961) Minds, machines and gödel. Philosophy 36(137):112–127
Luce RD (1959) Individual choice behavior: a theoretical analysis. Wiley, New Jersey
Macmillan-Scott O, Musolesi M (2024) (Ir)rationality and cognitive biases in large language models. R Soc Open Sci 11:240255
Manome N, Shinohara S, Takahashi T, Chen Y, Chung U-I (2021) Self-incremental learning vector quantization with human cognitive biases. Sci Rep 11:3910
Mao W, Gratch J (2004) A utility-based approach to intention recognition. In: Proceedings of AAMAS-04
Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. W.H Freeman, San Francisco
Martignon L, Hoffrage U (2002) Fast, frugal, and fit: simple heuristics for paired comparison. Theor Decis 52(1):29–71
Martins N (2010) Can neuroscience inform economics? Rationality, emotions and preference formation. Camb J Econ 35(2):251–267
Maruyama Y (2020) Rationality, cognitive bias, and artificial intelligence: a structural perspective on quantum cognitive science. In: Proceedings of EPCE-20. pp 172–188
Mas-Colell A, Whinston M, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
Masters P, Sardina S (2021) Expecting the unexpected: goal recognition for rational and irrational agents. Artif Intell 297:103490
McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag 27(4):12
McElfresh DC, Chan L, Doyle K, Sinnott-Armstrong W, Conitzer V, Borg JS, Dickerson JP (2021) Indecision modeling. In: Proceedings of AAAI-21. pp 5975–5983
McIntosh E, Ryan M (2002) Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: implications of discontinuous preferences. J Econ Psychol 23(3):367–382
Mealing R, Shapiro JL (2017) Opponent modelling by expectation-maximisation and sequence prediction in simplified poker. IEEE Trans Comput Intell AI Games 9(1):11–24
Meeker K (2001) Is Hume's epistemology internalist or externalist? Dialogue Can Philos Rev 40(1):125–146
Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
Nashed SB, Zilberstein S (2022) A survey of opponent modeling in adversarial domains. J Artif Intell Res 73:277–327
Newell A, Simon HA (1961) GPS, a program that simulates human thought. In: Billings H (ed) Lernende automaten. pp 109–124
Newell A, Shaw JC, Simon HA (1958) Elements of a theory of human problem solving. Psychol Rev 65(3):151–166
Ng AY, Russell S (2000) Algorithms for inverse reinforcement learning. In: Proceedings of ICML-00, San Francisco, CA, USA. pp 663–670
Nisbett RE, Borgida E (1975) Attribution and the psychology of prediction. J Pers Soc Psychol 32:932–943
Nisbett RE, Ross L (1980) Human inference: strategies and shortcomings of social judgment. Englewood Cliffs
Noiret S, Lumetzberger J, Kampel M (2021) Bias and fairness in computer vision applications of the criminal justice system. In: Proceedings of SSCI-21. pp 1–8
Nozick R (1993) The nature of rationality. Princeton University Press, Princeton

O'Brien DP (1993) Mental logic and irrationality: we can put a man on the moon, so why can't we solve those logical reasoning problems. In: Manktelow KI, Over DE (eds) Rationality: psychological and philosophical perspectives. Routledge, London, pp 110–135
O'Connell T, Chun M (2018) Predicting eye movement patterns from fMRI responses to natural scenes. Nat Commun 9:5159
Okasha S (2016) Biology and the theory of rationality. In: Smith DL (ed) How biology shapes philosophy: new foundations for naturalism. Cambridge University Press, Cambridge, pp 161–183
Olorunleke O, McCalla G (2005) A condensed roadmap of agents-modelling-agents research. In: Proceedings of the IJCAI-05 workshop on modeling other agents from observation
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Penguin Books Limited, New York
Osborne MJ (2004) An introduction to game theory. Oxford University Press, Oxford
Papakyriakopoulos O, Mboya AM (2023) Beyond algorithmic bias: a socio-computational interrogation of the google search by image algorithm. Soc Sci Comput Rev 41(4):1100–1125
Parkes DC, Wellman MP (2015) Economic reasoning and artificial intelligence. Science 349(6245):267–272
Pawlick J, Colbert E, Zhu Q (2019) A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. ACM Comput Surv 52(4):1–28
Peng P, Wen Y, Yang Y, Yuan Q, Tang Z, Long H, Wang J (2017) Multiagent bidirectionally-coordinated nets: emergence of human-level coordination in learning to play StarCraft combat games. arXiv: 1703.10069
Penrose R (1989) Truth, proof, and insight. The emperor's new mind: concerning computers, minds, and the laws of physics. Oxford University Press, Oxford
Pizzi G, Scarpi D, Pantano E (2020) Artificial intelligence and the new forms of interaction: who has the control when interacting with a chatbot? J Bus Res 129:878–890
Polvara R, Patacchiola M, Sharma S, Wan J, Manning A, Sutton R, Cangelosi A (2018) Autonomous quadrotor landing using deep reinforcement learning. arXiv: 1709.03339
Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C et al (2019) Machine behaviour. Nature 568:477–486
Ramachandran D, Amir E (2007) Bayesian inverse reinforcement learning. In: Proceedings of IJCAI-07, San Francisco, CA, USA, pp 2586–2591
Ramírez M, Geffner H (2011) Goal recognition over POMDPs: inferring the intention of a POMDP agent. In: Proceedings of IJCAI-11. pp 2009–2014
Ren Y, Duan J, Li SE, Guan Y, Sun Q (2020) Improving generalization of reinforcement learning with minimax distributional soft actor-critic. arXiv: 2002.05502
Rozado D (2020) Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. PLoS ONE 15(4):1–26
Rubinstein A (1998) Modeling bounded rationality. The MIT Press, Cambridge
Russell S (1997) Rationality and intelligence. Artif Intell 94(1):57–77
Russell S (2016) Rationality and intelligence: a brief update. In: Müller V (ed) Fundamental issues of artificial intelligence. Springer, Cham
Russell S, Norvig P (2022) Artificial intelligence: a modern approach. Pearson Education, London
Russell S, Subramanian D, Parr R (1993) Provably bounded optimal agents. In: Proceedings of IJCAI-93
Ryan M, Bate A (2001) Testing the assumptions of rationality, continuity and symmetry when applying discrete choice experiments in health care. Appl Econ Lett 8(1):59–63
Rysiew P (2008) Rationality disputes—psychology and epistemology. Philos Compass 3(6):1153–1176
Samuels R, Stich S, Bishop M (2002) Ending the rationality wars: how to make disputes about human rationality disappear. In: Elio R (ed) Common sense, reasoning and rationality. Oxford University Press, Oxford, pp 236–268
Samuelson PA (1938) A note on the pure theory of consumer's behaviour. Economica 5(17):61–71
Saunders C, Over D (2009) In two minds about rationality? In: Evans JSBT, Frankish K (eds) In two minds: dual processes and beyond. Oxford University Press, Oxford, pp 317–334
Schadd FC, Bakkes SCJ, Spronck P (2007) Opponent modeling in real-time strategy games. In: Proceedings of GAME-ON-07. pp 61–70
Schilirò D (2012) Bounded rationality and perfect rationality: psychology into economics. Theor Prac Res Econ Fields 3:101–111
Scholte H (2018) Fantastic animals and where to find them. Neuroimage 180:112–113
Schwarz G, Christensen T, Zhu X (2022) Bounded rationality, satisficing, artificial intelligence, and decision-making in public organizations: the contributions of Herbert Simon. Public Adm Rev 82(5):902–904
Selten R (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. Int J Game Theory 4:25–55
Shanahan M (2024) Talking about large language models. Commun ACM 67(2):68–79

Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR et al (2024) Towards understanding sycophancy in language models. In: Proceedings of ICLR-24
Shin D, Shin EY (2023) Data's impact on algorithmic bias. Computer 56(6):90–94
Shinn N, Cassano F, Labash B, Gopinath A, Narasimhan K, Yao S (2023) Reflexion: language agents with verbal reinforcement learning. arXiv: 2303.11366
Shirado H, Christakis N (2017) Locally noisy autonomous agents improve global human coordination in network experiments. Nature 545:370–374
Shoham Y, Powers R, Grenager T (2003) Multi-agent reinforcement learning: a critical survey. Unpublished survey
Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
Simon HA (1957) Models of man: social and rational. Wiley, Hoboken
Simon HA (1982) Models of bounded rationality. Volume 1: economic analysis and public policy. The MIT Press, Cambridge
Simsek O (2013) Linear decision rule as aspiration for simple decision heuristics. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Proceedings of NeurIPS-13
Simsek O (2020) Bounded rationality for artificial intelligence. In: Viale R (ed) Routledge handbook of bounded rationality (chapter 15). Routledge, London
Simsek O, Buckmann M (2015) Learning from small samples: an analysis of simple decision heuristics. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M,Garnett R (eds) Proceedings of NeurIPS-15. pp 3159–3167
Simsek O, Algorta S, Kothiyal A (2016) Why most decisions are easy in tetris–and perhaps in other sequential decision problems, as well. In: Proceedings of ICML-16. pp 1757–1765
Skalse J, Abate A (2023) Misspecification in inverse reinforcement learning. In: Proceedings of AAAI-23. pp 15136–15143
Sloman A (1993) The mind as a control system. In: Hookway C, Peterson DM (eds) Royal institute of philosophy supplement. Cambridge University Press, Cambridge, pp 69–110
Sloman SA (1996) The empirical case for two systems of reasoning. Psychol Bull 119(1):3–22
Slovic P, Finucane M, Peters E, MacGregor DG (2002) Rational actors or rational fools: implications of the affect heuristic for behavioral economics. J Socio-Econ 31(4):329–342
Sonu E, Doshi P (2015) Scalable solutions of interactive POMDPs using generalized and bounded policy iteration. Auton Agent Multi-Agent Syst 29(3):455–494
Spohn W (1988) Ordinal conditional functions: a dynamic theory of epistemic states. In: Harper WL, Skyrms B (eds) Causation in decision, belief change, and statistics: proceedings of the Irvine conference on probability and causation. Springer Netherlands, Dordrecht, pp 105–134
Stanczak K, Augenstein I (2021) A survey on gender bias in natural language processing. arXiv:2112.14168
Stanovich KE (1999) Who is rational?: Studies of individual differences in reasoning. Psychology Press, Hillsdale
Stanovich KE (2016) The comprehensive assessment of rational thinking. Educ Psychol 51:1–12
Stanovich KE, West RF, Toplak ME (2011) Intelligence and rationality. In: Sternberg RJ, Kaufman SB (eds) The Cambridge handbook of intelligence. Cambridge University Press, Cambridge, pp 784–826
Stein E (1996) Without good reason: the rationality debate in philosophy and cognitive science. Clarendon Press, Oxford
Stella L, Bauso D (2023) The impact of irrational behaviors in the optional prisoner's dilemma with game-environment feedback. Int J Robust Nonlinear Control 33(9):5145–5158
Still S, Precup D (2012) An information-theoretic approach to curiosity-driven reinforcement learning. Theory Biosci 131(3):139–148
Sturm T (2012) The "Rationality Wars" in psychology: where they are and where they could go. Inquiry 55(1):66–81
Sugden R (1991) Rational choice: a survey of contributions from economics and philosophy. Econ J 101(407):751–785
Sukthankar G, Sycara K (2005) A cost minimization approach to human behavior recognition. In: Proceedings of AAMAS-05. pp 1067–1074
Sutherland S (1992) Irrationality: the enemy within. Constable and Company, London
Tambe M, Rosenbloom PS (1995) Event tracking in a dynamic multi-agent environment. Comput Intell 12(3):499–522
Taniguchi H, Sato H, Shirakawa T (2017) Application of human cognitive mechanisms to Naïve Bayes text classifier. In: Proceedings of AIP-17. p 360016
Taniguchi H, Sato H, Shirakawa T (2019) Implementation of human cognitive bias on neural network and its application to breast cancer diagnosis. SICE J Control Meas Syst Integr 12(2):56–64

Tentori K, Bonini N, Osherson D (2004) The conjunction fallacy: a misunderstanding about conjunction? Cogn Sci 28:467–477
Tervonen T, Schmidt-Ott T, Marsh K, Bridges JF, Quaife M, Janssen E (2018) Assessing rationality in discrete choice experiments in health: an investigation into the use of dominance tests. Value Health 21(10):1192–1197
Thaler R (1980) Toward a positive theory of consumer choice. J Econ Behav Organ 1(1):39–60
Thomson W (1979) Maximin strategies and elicitation of preferences. In: Laffon J-J (ed) Aggregation and revelation of preferences. North-Holland, Amsterdam, pp 245–268
Tian X, Zhuo HH, Kambhampati S (2016) Discovering underlying plans based on distributed representations of actions. In: Proceedings of AAMAS-16. pp 1135–1143
Tourlakis G (2022) Gödel's first incompleteness theorem via the halting problem. In: Computability. Springer
Tsetsos K, Moran R, Moreland J, Chater N, Usher M, Summerfield C (2016) Economic irrationality is optimal during noisy decision making. Proc Natl Acad Sci 113(11):3102–3107
Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460
Tversky A, Kahneman D (1971) Belief in the law of small numbers. Psychol Bull 76(2):105–110
Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev 90(4):293–315
Tversky A, Kahneman D (1986) Rational choice and the framing of decisions. J Bus 59(4):S251–S278
Tversky A, Thaler RH (1990) Anomalies: preference reversals. J Econ Perspect 4(2):201–211
Uprety S, Song D (2018) Reconciling irrational human behavior with AI based decision making: a quantum probabilistic approach. arXiv:1808.04600
Uzawa H (1960) Preference and rational choice in the theory of consumption. In: Arrow K, Karlin S, Suppes P (eds) Mathematical methods in the social sciences (chapter 9). Stanford University Press, Stanford
Van Den Herik H, Donkers H, Spronck P (2005) Opponent modelling and commercial games. In: Proceedings of CIG-05. pp 15–25
van der Hoek W, Wooldridge M (2002) Tractable multiagent planning for epistemic goals. In: Proceedings of AAMAS-02, New York, USA. pp 1167–1174
van der Hoek W, Wooldridge M (2003) Towards a logic of rational agency. Log J IGPL 11(2):135–159
Vered M, Kaminka GA (2017) Heuristic online goal recognition in continuous domains. In: Proceedings of IJCAI-17. pp 4447–4454
Wason PC (1968) Reasoning about a rule. Q J Exp Psychol 20(3):273–281
Waytz A, Heafner J, Epley N (2014) The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol 52:113–117
Wedgwood R (2021) Practical and theoretical rationality. In: Knauff M, Spohn W (eds) The handbook of rationality. The MIT Press, Cambridge
Wen Y, Yang Y, Wang J (2020) Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. In: Proceedings of IJCAI-20
Wheeler G (2020) Bounded rationality. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, Fall 2020 edn. Metaphysics Research Lab, Stanford University
Williams M, Carroll M, Narang A, Weisser C, Murphy B, Dragan A (2025) On targeted manipulation and deception when optimizing LLMs for user feedback. arXiv: 2411.02306
Wilson B, Hoffman J, Morgenstern J (2019) Predictive inequity in object detection. arXiv: 1902.11097
Wooldridge M (2000) Reasoning about rational agents. The MIT Press, Cambridge/London
Yang Y, Wen Y, Yu L, Zhang W, Bai Y, Wang J (2018) A study of AI population dynamics with million-agent reinforcement learning. In: Proceedings of AAMAS-18. pp 2133–2135
Yu C, Liu J, Nemati S (2020) Reinforcement learning in healthcare: a survey. arXiv: 1908.08796
Zhang K, Yang Z, Başar T (2021) Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Vamvoudakis KG, Wan Y, Lewis FL, Cansever D (eds) Handbook of reinforcement learning and control. Springer, Cham, pp 321–384
Ziebart BD (2010) Modeling purposeful adaptive behavior with the principle of maximum causal entropy (Unpublished Doctoral Dissertation). Carnegie Mellon University
Ziebart BD, Bagnell JA, Dey AK (2010) modeling interaction via the principle of maximum causal entropy. In: Proceedings of ICML-10. pp 1255–1262
Zilberstein S (2011) Metareasoning and bounded rationality. Metareasoning: thinking about thinking. The MIT Press, Cambridge
Zindel ML, Zindel T, Quirino MG (2014) Cognitive bias and their implications on the financial market. Int J Mech Mechatron Eng