To appear in the *Journal of Experimental & Theoretical Artificial Intelligence*
Vol. 00, No. 00, Month 20XX, 1–16

ORIGINAL ARTICLE

Normative and Descriptive Rationality: From Nature to Artifice and Back

T.R. Besold[a]* and S. L. Uckelman[b]

[a]*Department of Computer Science, City, University of London, UK*
[b]*Department of Philosophy, Durham University, UK*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Rationality plays a key role in both the study of human reasoning and Artificial Intelligence (AI). Certain notions of rationality have been adopted in AI as guides for the development of intelligent machines and these notions have been given a normative function. The notions of rationality in AI are often taken to be closely related to conceptions of rationality in human contexts. In this paper, we argue that the normative role of rationality differs in the human and artificial contexts. While rationality in human-focused fields of study is normative, prescribing how humans *ought to reason*, the normative conception in AI is built on a notion of human rationality which is descriptive, not normative, in the human context, as AI aims at building agents which *reason as humans do*. In order to make this point, we review prominent notions of rationality used in psychology, cognitive science, and (the history of) philosophy, as well as in AI, and discuss some factors that contributed to rationality being assigned the differing normative statuses in the differing fields of study. We argue that while 'rationality' is a normative notion in both AI and in human reasoning, the normativity of the AI conception of 'rationality' is grounded in a descriptive account of human rationality.

Keywords: Artificial Intelligence; Cognitive Systems; Philosophy; Human Reasoning; Rationality; Normativity

1. Introduction

A central component of any discussion of Artificial Intelligence (AI)[1] is what is meant by 'Intelligence'. Intelligence can be, and has been, cashed out in many ways, and unfortunately, there is currently no plausible, complete theory or implementable model of intelligence. Instead, there is controversy about different the so-called "theories of intelligence", each of which encompasses a different taxonomic structure of abilities (cf., e.g., (Sternberg, 2000)). In this article, we do not attempt to answer the question "what is intelligence?" Instead, we focus on one feature that is commonly included in conceptions

*Corresponding author. Email: tarek-r.besold@city.ac.uk

[1]Throughout this article, when referring to 'AI' as an object of study and development we refer to work towards systems performing at the upper end of Nilsson's continuum: "[A]rtificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. According to that definition, lots of things... are intelligent... At the other end [of the extended continuum along which entities with various degrees of intelligence are arrayed] are humans, who are able to reason, achieve goals, understand and generate language, perceive and respond to sensory inputs, prove mathematical theorems, play challenging games, synthesize and summarize information, create art and music, and even write histories" (Nilsson, 2009, p. xiii).

of intelligence, namely, rationality.[2]

Rationality itself incorporates both a theoretical or epistemic dimension, concerned with what is rational to believe and with rational degrees of belief, and practical manifestations, regarding what it is rational to do, or to intend or desire to do. Historically, the theoretical study of rationality, conducted within philosophy and different branches of economics (cf., the first chapters of (Mele & Rawling, 2004)), has been largely considered to be domain-independent. Recently, the situation has changed as rationality as a research topic has been introduced into several experimental and engineering sciences, including AI, taking a more empirical and domain-grounded approach to conceptualizing, studying, and implementing models of rationality. This integration brings with it a need to explain in what ways the concepts of 'rationality' used in these different fields differ. In particular, we must consider whether the concept is being used normatively or descriptively uniformly across the disciplines or not.

In this paper we argue that rationality in AI plays a primarily normative role and that we design intelligent systems, agents, robots, etc.,[3] in the way that we do because we want them to reason or act in a particular way. When we speak of them acting rationally, we are not simply describing the way that they act, but comparing the way that they act against some specific benchmark. This benchmark is closely linked to human- or human-style rationality. It is not merely that we wish to have agents who can act correctly, according to some abstract characterization of correct action, but we would like them to be able to achieve these actions via the use of reason, and not only that, but to reason in a human-like fashion.[4] This mans that underlying the normative standard in AI is a descriptive conception of human rationality.

We begin in §2 by looking at how rationality is conceived in AI as a type of utility maximization.[5] We argue that utility maximization is too limited to be an account of rationality, and that it is better called 'ethicality' (the reasons for this choice of term become clear in that section). We compare the AI conception of rationality with the more human-oriented conceptions used in psychology, cognitive science, and philosophy in §3. Once we have established these different conceptions, we argue in §4 that the roots of the normative account of rationality used in AI are grounded in a descriptive account of rationality in nature (that is, humans). We summarize in §5.

[2]As a result, we almost entirely omit discussion of a number of related and interesting questions, such as "What does it mean to attribute intelligence to an artificial entity?" "What differentiates intelligence from (mere) knowledge—can something know many things but still not be 'intelligent' in the right sense?" "Can intelligence be explained in terms of the possession of certain faculties, such as rationality, goal-directedness, utility maximization, creativity, etc.?"

[3]In the following we use any of these (and related) terms interchangeably, usually referring to the entire category of systems, construed very broadly as entities implementing a perception-action cycle. Exceptions will be clear from context.

[4]There clearly are important parts of AI research either operating completely independent from any results on or models of human intelligence or cognition processes, or aiming to implement forms of general reasoning precisely avoiding certain characteristics of human reasoning (sometimes even presented as 'flaws in reasoning'). Examples of the former type of AI work are current systems for statistics-based or neural machine translation (Koehn, 2009; Wu et al., 2016) or video summarization (Song, Vallmitjana, Stent, & Jaimes, 2015), and a perfect reasoner based on algorithmic probability like AIXI (Hutter, 2007) can be seen as an instance of the latter class of research. Still, conceptually complementing these lines of work there are several major efforts with human-like reasoning at their core, for instance, in modelling commonsense reasoning, which is necessarily human-centric (Davis & Marcus, 2015), or in cognitive architectures and cognitive systems engineering (Langley, 2006). In these contexts 'rationality' is often taken as proxy for the notoriously hard-to-grasp notion of 'intelligence'.

[5]Throughout the article, selected notions of rationality from different disciplines are discussed. While many of the most prominent paradigms from the relevant fields are mentioned, the resulting collection is by no means complete. Instead, emphasis has been put on covering especially popular and/or paradigmatic accounts of rationality which usually can be taken as examples for an entire class of conceptualizations.

2. Rationality in AI: Rationality as utility maximisation

We begin by first making precise how 'rationality' is commonly understood by AI practitioners. For many researchers in AI, the long-term scientific aspiration of their activities still is what can be called "the mechanization of 'human-level' intelligence" (Nilsson, 2006, p. 68), by (re-)creating intelligence within an artificial system by computational means. Consequently, the corresponding field of study occupies an intermediate status between an engineering endeavour and a scientific discipline (cf., e.g., the discussions in (Besold, 2013a; Cassimatis, 2012)). Given the lack of a precise and implementation-friendly definition of 'intelligence', AI researchers often use proxies such as rationality. For example, Russell (2016), in his quest to find a suitable formal definition of intelligence, or the best proxy thereof, identifies rationality as a promising candidate, and Doyle (1992) in his survey of work at the intersection between (the economic theory of) rationality and AI claims that a theory of rationality might at some point equal mathematical logic in its importance for mechanising reasoning (cf. (Besold, 2013b) for a more detailed discussion).

If rationality is a necessary component of intelligence, it follows that this process of mechanization will require the re-creation of rationality within an artificial system. In order both for this goal to be reached and to know whether it has been reached, we must have an understanding of 'rationality' that is apt for mechanization. Certain conceptions of rationality have played an important role at different turning points within the development of AI as a science, sometimes emerging from new paradigms within the field, sometimes directly contributing to the creation of new stances and perspectives. Two especially relevant positions for AI as a whole are epitomized in proposals by Clark (2003) and Russell (2016).

Clark (2003) defines rationality in a general way as 'reason-governed behaviour' and provides an overview of such behavior. He lists as cases in point both robotics (Can reasoned action be explained without making appeal to inner, form-based vehicles of meaning? Can there be something like representation-free rationality, i.e., can phenomena like 'deliberative reasoning' or 'abstract thought' be explained by a complex of reflex-like mechanisms alone?) and global reasoning (How can 'non-classical' forms of human reasoning, such as non-monotonic, abductive, or analogical reasoning, be accounted for in AI systems?), among others.

Russell (2016) evaluates four different approaches to rationality from the history of AI: (1) *perfect rationality* (the capacity to generate maximally successful behaviour given certain background information); (2) *calculative rationality* (the theoretical capacity to compute a perfectly rational decision given certain background information); (3) *meta-level rationality* (the capacity to select an optimal combination of computation sequence and action, with the action selection performed by the computation); and (4) *bounded optimality* (perfect rationality additionally accounting for limitations on computational resources) (Russell, 2016, p. 8). All four notions rely on the basic mechanism of maximizing the expected utility—as captured by a performance measure on sequences of environment states—of an agent's actions, making rational behavior a well-defined function of a task environment. This account of rationality is teleological, and requires that there be some end goal that can be successfully attained. Then, the extent to which an AI's action directed at a certain goal conforms with our desires, expectations, or requirements of their actions is the extent to which we say that the AI is acting rationally.

While many AI researchers working on topics relating to rational agency deem the identification of rationality in (possibly artificial) intelligent agents with utility maximization as a perfectly sufficient characterization of the concept of rationality in general

(cf., e.g., (Bostrom, 2014)), in this section we show that it is not reasonable to consider an agent acting in this fashion 'rational'; instead agents who maximize utility should be called 'ethical' rather than 'rational'. This will pave the way for us to look at broader conceptions of rationality, centered around human rationality, in §3.

The idea of utility maximization as a normative guide to behavior underlies utilitarian (and other broadly consequentialist) moral and ethical theories, in contrast with Kantian/deontological theories (which give priority to the following of certain types of rules) or virtue ethics (which emphasizes the role of virtues or moral character). Utility maximization by itself need not have any moral or ethical quality to it, but as soon as a normative dimension is introduced and we use it to judge whether a person has acted in a good or correct manner, it takes on this ethical quality. In a normative context, equating rational behavior with utility-maximizing behavior makes it impossible for one to act rationally but unethically, or ethically but irrationally. Thus, the distinction between rational and ethical behavior collapses. If we wish to maintain a distinction between agents (natural or artificial) who act rationally and ones who act ethically (and we do), we must find a way to reject the identification of rationality with utility maximisation.

One way to object to this line of reasoning is to argue that it assumes that ethical activity in both the artificial and natural domain is at root utilitarian, and that in order to call the behaviour of a utility-maximizing robot 'ethical' we would have to call the same utility-maximizing behavior in a person 'ethical'. Not everyone admits consequentialist theories of ethics as correct, advocating deontological or aretaic alternatives instead. This is a reasonable position to take, but it is not an objection to the equation of rational agents and ethical agents. For in such a case, not only would the utility-maximizing robot not count as an ethical agent, it is unclear whether it would count as a rational agent—to allow it to do so would be to allow that it can be possible to act unethically but rationally.[6] Thus, we might be happy to admit that rationality entails ethicality, but there is more to acting rationally (or being rational) than merely acting ethically (or being ethical), as we can see by transferring the notions from the artificial domain to the natural one.

If rationality is nothing more than utility maximization, in both artificial and natural domains, then rationality in people will also be nothing more than utility maximization. This is the approach to human rationality that is often taken by economists—in the so-called *homo economicus*—in which case rationality then becomes nothing more than the imposition of certain types of consistency constraints on preferences (cf. (Hammond, 1997)). But to call these consistency axioms the sum of human rationality is to redefine the term beyond recognition. It simply ignores the broad way in which we use the term 'rational' in connection with human behavior, both descriptively and normatively, and we would be well within our rights to say that if rationality is to be defined as utility maximization or adherence to certain preference-consistency axioms, then we are not interested in rationality but rather in whatever we wish to call that faculty or concept which we call rationality in ordinary contexts. In particular, we would like to have an account which retains connections with both reason and rationalizability, regardless of whether we end up calling this phenomenon 'rationality' or not. Let us be clear here: This is not a dispute about terminology, about what it is we should call 'rationality', but about the underlying concept. We are interested in the phenomenon out there in the world that this word can (or should) pick out, regardless of whether or not we end up calling it rationality or something else.

[6]Whether one would wish to allow this possibility is beyond the scope of our discussion; it is worthwhile pointing out that there are at least *prima facie* reasons for not wanting to allow it.

We therefore relabel agents who act merely in a means-ends fashion as 'ethical' ones, and call agents 'rational' who function under a broader concept of rationality which also encompasses further facets (Cowen, 2004). If we retained the concept 'rational' for the utility-maximizing agents, we would then still need to find another label for the broader notion; and it turns out that some people already have: In AI, a distinction is often made between 'rational' agents, which act in a utility-maximizing way, and 'intelligent' agents in general, with the former being taken as a sub-class of the latter. Following the hierarchy of agent classes introduced in (Russell & Norvig, 2003), agent architectures in AI can be ordered on a scale of increasing degree of (perceived) intelligence and overall capability level starting from reflex agents and ranging through purely goal-based agents to utility-based agents and learning agents at the upper end. When speaking of 'rational' agents, reference is usually made to utility-maximizing agents as the second highest class in the hierarchy. Utility-maximizing agents in general are required to come equipped with a utility function as an internal performance measure, which in the case of 'rational' agents additionally has to match the performance measures of the environment. In this sense, while a utility function might not be necessary for rationality in general, 'rational' agents in AI can be modelled (and are expected to act) as if they are governed by one—thus, for all relevant purposes, becoming equal to ethical agents.

By placing rational agents/utility-based agents on the second highest level in their hierarchy, Russell and Norvig (2003) make it clear that rationality is not sufficient for full intelligent agency. Additionally, subsuming rational agents under the broader class of utility-based agents also provides evidence that the corresponding description of the lower-level of agents as 'rational' is inappropriate unless rationality would have to be taken as a sub-notion of utility maximization. This ailment is cured by instead opting for an ethical description. Artificial rationality *is* generally goal- or action-oriented: If that goal is utility maximization, then we call the artificial agents merely 'ethical'. If, however, the goal-directed behavior is aimed at a goal other than simple utility maximization, we will call it rational.

3. Concepts of rationality outside of AI

We have shown that a minimal condition for a conception of 'rationality' in AI is that it allows us to distinguish agents who are rational from ones who are merely ethical. We now look at how rationality is conceptualised outside of AI, focusing on human-oriented conceptions found cognitive science, psychology, and philosophy. These conceptions include both normative and descriptive ones.

3.1. *Rationality in cognitive science and psychology*

In the natural or human context rationality is not generally explicated in an exclusively goal-oriented way, but also involves notions such as cognitive consistency or adherence to values. Since Simon's foundational work on rationality resulting in the notion of *bounded rationality* (Simon, 1959), and his contributions to the study of human problem-solving and high-level reasoning (Newell & Simon, 1972; Simon, 1990), many distinct frameworks for modelling rationality (and establishing a normative theory) have been proposed in psychology and cognitive science (cf. (Besold, 2013b) for more detail). Breaking these distinct approaches down to their underlying theoretical foundations, we can identify four main types of models, together with corresponding normative interpretations for what counts as rational: (1) logic-based models, (2) probability-based models, (3) game-

theoretic-based models, and (4) heuristic-based models. *Logic-based models* (cf., e.g., (Evans, 2002)) consider a belief as rational if there is a logically valid reasoning process to reach this belief relative to the corresponding available background knowledge (putting significant emphasis on logical consistency between beliefs and knowledge). *Probability-based models* (cf., e.g., (Griffiths, Kemp, & Tenenbaum, 2008)) accept a belief as rational if the expected value of the belief is maximized relative to given probability distributions of background beliefs (which can either reflect a notion of goal-oriented utility, or refer to the likelihood of a certain belief being accurate given the background beliefs). For *models based on game theory* (cf., e.g., (Osborne & Rubinstein, 1994)) a belief is rational if the expected payoff of maintaining the belief is maximized relative to other possible beliefs (which at first sight comes closest to an exclusively utility-maximizing approach, but as discussed in (Cowen, 2004) can encompass numerous additional facets). These models clearly involve normative conceptions of rationality. In *heuristic-based models* (cf., e.g., (Gigerenzer, Hertwig, & Pachur, 2011)) classical ideas of normativity within a model of rationality become obsolete and in many cases are not anymore part of the respective frameworks.[7,8]

Although each type of frameworks has gained merit in modelling certain aspects of human rationality and providing a foundation for experimental studies, the generality of each class has at the same time been challenged by psychological experiments and theoretical objections. On the one hand, studies by Wason et al. question the human ability of reasoning in accordance with the principles of classical logic (Wason, 1966; Wason & Shapiro, 1971). Byrne (1989)'s findings on human reasoning with conditionals also indicate severe deviations from this classical paradigm. Similarly, when considering probability-based models, Tversky and Kahneman (1983)'s Linda problem illustrates a striking violation of the rules of probability theory by human reasoners. On the other hand, game-based frameworks are questionable due to the lack of a (reasonably) unanimously accepted concept of optimality in game-theory. There are numerous proposed variants of the Nash equilibrium (cf., e.g., (Halpern, 2008)), and it is not clear which one—if any—should be taken as 'the most rational one' in a given situation. As a result, we are left with a tension between the norms advanced by the theoretical models and the descriptive data the psychological evidence provides.

Heuristic approaches to judgment and reasoning follow a different approach, and are often conceptualized as approximations of a rational ideal, rather than an instantiation of the ideal itself. In some scenarios, heuristics work well and yield surprising results, but when heuristics are applied in psychology and cognitive science they still often lack formal transparency and explanatory power. In this respect, this distinguishes the actual use of heuristics in psychology and cognitive science from the crucial role heuristics play in AI. In AI, heuristics are frequently derived in a top-down manner from theoretical

[7]The question for classical normativity as standard for 'good' or 'correct' reasoning (as still underlying Kahneman's and Tversky's heuristics-and-biases program) becomes mostly obsolete in the case of current heuristic frameworks, and is replaced by an effort to build models which describe and predict actual human judgments and decision-making behavior (cf., e.g., (Gigerenzer, 1996)), placing heuristic models in the family of descriptive models (cf., e.g., (Baron, 2012)). When talking about the normative approach in this context, the term usually describes efforts which try to understand under which conditions a given heuristic performs optimal.

[8]These four families of models for rationality (though partly with exception of the fourth type) can actually be found within existing AI systems and theories. Still, the underlying notions of rationality have stayed close to their fields of origin and corresponding deficits and shortcomings have commonly been brought along. In consequence, AI systems mostly fall short in tasks such as predicting or exhibiting behavior resembling human-like rationality or rational action—which would be a crucial pragmatic need in all domains concerned with close interaction (and possibly even cooperation) between a human user and an AI system. Example scenarios for these are numerous, ranging from rational agents communication for cooperative dialogues (Sadek, Bretier, & Panaget, 1997) to adaptive and cooperative wheelchairs (Galluppi, Urdiales, Sandoval, & Olivetti, 2009).

foundations or from practical knowledge about an application domain or task, resulting in an explicit rule or evaluation function with strong theoretical underpinnings. Such theoretical underpinnings allow for the analysis of limitations or required preconditions of the respective heuristic. On the contrary, heuristics in the context of human reasoning often have the character of empirically learned 'rules of thumb' that lack a strong theoretical basis. Moreover, the status of heuristics as commonly used mechanisms of cognition—both across subjects as well as across tasks and individual cases—has been experimentally challenged (cf., e.g., (Oppenheimer, 2003)) and theoretical problems relating to complexity and perceived intractability in human decision-making have been pointed out (cf., e.g., (van Rooij, Wright, & Wareham, 2010)). Additionally, from a more general methodological or philosophical point of view, a quite fundamental criticism can be stated: Due to the open nature of the collection of heuristics propagated in most current accounts (i.e., whenever a phenomenon cannot be covered or described by an existing heuristic, a new one specifically fit to the task is introduced), the possibility of falsification and refutation of modeling assumptions and theories is not guaranteed, and a (reasonable) completion of the model can neither be checked for, nor feasibly assumed at any point.

Some researchers in cognitive science and decision theory question the completeness and suitability of the classical approaches to rationality on an even more fundamental level (echoing many of the intuitions and insights already articulated by Simon in his theory of bounded rationality). Gilboa (2010)—cf. also (Gilboa & Schmeidler, 2001)—defines rationality to crucially depends on the subject executing the reasoning and (allegedly) rational behavior. In his view "a mode of behavior is rational for a given decision maker if, when confronted with the analysis of her behavior, the decision maker does not wish to change it" (Gilboa, 2010, p. 3). Rationality then becomes subject-centered, as what is considered rational varies with the reasoner in question. If the decision maker does not understand the analysis or the reasons for which her behavior is considered not rational, she cannot be judged as irrational on these grounds. Gilboa argues that where limited cognitive capacities prevent comprehension of the expected rational rules, but would always make the reasoner take the same decision again, rationality on the reasoner's own terms has to be conceded, notwithstanding an external observer's dissenting judgment. This of course does not exclude the possibility of a reasoner acting irrationally, or committing an "error in reasoning", as in many cases individuals will still fall short of deciding at the best of their respective capacities. The novelty compared to previous accounts is instead introduced by personalizing the notion of rationality to that degree that the rational norm has to be achievable for the reasoning subject (given available cognitive resources, supporting formal tools and corresponding regulated reasoning styles such as logical reasoning, etc.) if the subject is to be held accountable under the norms of rationality. Kokinov (2003) goes even further in proposing that the concept of rationality as a theory in its own right ought to be replaced by a multilevel theory based on mechanisms and processes involved in decision-making, due to its failures as a descriptive theory of human decision-making, and as a normative theory for good decision-making. Utility maximization and rationality would then be rendered as contingent emergent properties. On such a view, humans are not *homo economicus*, not fundamentally.

3.2. *Rationality in philosophy*

The view of (human) rationality that is found in philosophy is broader even still. It still includes the instrumental dimension—as Cowen notes, "Philosophers ... commonly be-

lieve that economic logic focuses on instrumental rationality, as exemplified by a Humean ends-means logic" (Cowen, 2004, p. 213), but more than goal-oriented or ends-means behavior is required for a philosophical account of rationality. Alongside instrumental rationality we may also identify procedural rationality ("do our mental processes for forming values make sense?") and expressive rationality ("do we have the right ends or values") (Cowen, 2004, p. 213). Human rationality must encompass more than a simple teleological account because humans, unlike machines, are able to act without any clear goals in mind. Even in systems built upon BDI (belief, desire, intention) architectures (Rao & Georgeff, 1991)and the like, which provide artificial agents with a model of how human impulse/motivation works, equipping the system with the ability to get from (models of) beliefs to (models of) intentions and thus to develop its own goals, the notion of 'goal' still remains much more explicit and less 'subconscious' than for humans. This only serves to underscore that there is more to human reasoning and rationality than simple attainment of goal. Further components that play a role in human rationality include curiosity, free association, creativity, inspiration, as well as more structured components such as reasoning, rule-following, analogy, etc.[9]

Stanovich distinguishes two ways in which the concept of rationality is used with regard to human cognition, a weak sense and a strong sense. The weak sense is descriptive in nature and derives "from a categorical notion of rationality tracing to Aristotle (humans as the only animals who base actions on reason)" (Stanovich, 2012, p. 344). This weak sense is contrasted with a strong sense, how the term is used in cognitive science, and this sense—which Stanovich never explicitly defines—is "a normative notion" (Stanovich, 2012, p. 344). The strong sense of rationality is subdivided into two types: epistemic rationality, which "concerns how well beliefs map onto the actual structure of the world" (Stanovich, 2012, p. 344), and instrumental rationality, which involves "behaving in the world so that you get exactly what you most want, given the resources (physical and mental) available to you" (Stanovich, 2012, p. 345). There is a clear parallel between this instrumental notion and the utility-maximizing account of rationality given in §2, and indeed Stanovich notes that when the concept of rationality is transferred from cognitive science to economics, the idea of goal-fulfillment is cashed out in terms of utility maximization (Stanovich, 2012, p. 345).

Stanovich's epistemic rationality, the idea that our beliefs can map onto the "actual structure of the world" in a better or worse fashion, involves not merely how individual beliefs are mapped onto the world, but how well the structure of beliefs is mapped onto the structure of the world. Thus, we must inquire into what sort of structure our beliefs have that is relevant to a corresponding structure in the world. The most natural answer is that it is the logical structure of beliefs—their inferential relations with each other—that is at stake. Understanding epistemic rationality this way, the logical structure of beliefs plays a central role, and the way in which epistemic rationality is demonstrated is via appropriate performance in reasoning tasks that reflect this structure. If utility-maximising is the benchmark for human (instrumental) rationality, then performance of certain reasoning tasks is the benchmark for human (epistemic) rationality.

This emphasis on reasoning tasks in connection with rationality has a long history

[9]Notions of creativity, curiosity, etc., raise interesting questions (that we cannot pursue here) connected to 'interestingness', a constraint that is often put on things like automated theorem provers. We don't merely want automated theorem provers to churn out new results, we want these results to both be ones we can understand *and be ones we find interesting* (with the caveat that 'interesting' is itself a relative term: A mathematical result can be interesting in one context but be boring or trivial in another). This, unlike questions of rationality which we have shown above to be normative in nature, is a descriptive requirement. We want theorem provers to give us results that we *are* in fact interested in, not ones that we *should* find interesting. This is a subject which deserves further exploration beyond the scope of this paper.

in philosophy, rooted in Aristotle. Recall that Stanovich connects the weak, descriptive sense of rationality with Aristotle, not his stronger, normative and epistemic sense of rationality. But this is problematic in two respects: First, though the idea of rationality being specifically or uniquely associated with humans is indeed usually attributed to Aristotle in truth it is difficult to find any such claim explicitly in his writings.[10] Second, this 'Aristotelian' conception of rationality is normative, not descriptive.

When discussing human virtue in the *Nichomachean Ethics* in terms of the soul, Aristotle points out "that one element in the soul is irrational and one has a rational principle" (Aristotle, 1941a, Bk. I, Ch. 13, p. 950). While the irrational part of the soul is "common to all species and not specifically human" (Aristotle, 1941a, Bk. I, Ch. 13, p. 951), Aristotle does not explicitly say here that the rational part of the soul is *not* common to all species and *is* specifically human. He is only marginally more explicit elsewhere. In the *Metaphysics*, he gives a number of ways in which men are distinguished from other animals, including living not only with the aid of appearances and memory but also "by art and judgment" (τέχνῃ καὶ λογισμοῖς) (Aristotle, 1979, Bk. A, ch. I, p. 12).

The use of rationality or reason as a 'specific difference' of humanity, that is, the very property by which the species 'human' is speciated, is found, instead, in Porphyry. Porphyry distinguishes three ways in which things can differ: (1) commonly, (2) properly, and (3) most properly (Porphyry, 2003, §3, p. 8). One object differs from another in the most proper way when the first is distinguished from the other by means of a specific difference, and the example that Porphyry gives is "as a man differs from a horse by a specific difference, that of 'rational' " (Porphyry, 2003, §3, p. 9). Such a difference is 'most proper' because, unlike common and merely proper differences, most proper differences make two things not only 'otherlike' but actually 'other'. As Porphyry puts it, "when the difference 'rational' approaches 'animal', it makes it other and makes a species of animal ... it is in virtue of those differences which make a thing other that divisions of genera into species are made" (Porphyry, 2003, §3, p. 9). For Porphyry, not only is rationality what distinguishes the species humanity from all the other items in the genus 'animal', it is *only* the species humanity that is characterised by this difference: Any animal which does not possess rationality is other than man.

But what is this 'rationality' that both Aristotle and Porphyry are discussing? It is the ability to exercise reason, or indeed simply 'to reason'. In the *Topics*, Aristotle says that "reasoning is an argument in which, certain things being laid down, something other than these necessarily comes about through them" (Aristotle, 1941b, Bk. I, p. 188), and such arguments are divided into four types: demonstrations, dialectical reasoning, contentious reasoning, and mis-reasonings arising from premises specific to an individual science (for example when a geometer reasons from incorrect diagrams or definitions). An argument is a demonstration when it begins from premises which are "true and primary, or are such that our knowledge of them has originally come through premisses which are true or primary" (Aristotle, 1941b, Bk. I, p. 188). Dialectical reasoning is when the starting premises are "opinions that are generally accepted" (Aristotle, 1941b, Bk. I, p. 188), while contentious arguments begin from premises which seem to be generally admitted but are not in fact. Thus, reasoning as Aristotle uses it in both the *Topics* and the *Prior Analytics* is defined in normative contexts, telling us how one *should* reason in a scientific context, how one *should* reason is a dialectical context, and how one *should* reason in a contentious or sophistical context. If one excludes contentious reasoning and 'mis-reasoning' from the scope of reasoning (and there is evidence that Aristotle does),

[10]Further, as an anonymous referee points out, it is not clear that rationality as Aristotle defines it is unique to humans.

then we are left with a definition of reasoning which is equivalent to *good* reasoning—a clearly normative conception of rationality, not a descriptive one as Stanovich assumes. That the Aristotelian conception of reasoning was understood to be normative can also be seen in the different way in which medieval philosophers characterized the sciences of logic and grammar; the former is normative while the latter is descriptive. For example, in this Aristotelian tradition, Anselm of Canterbury distinguishes the *usus commune* 'common usage' of words from their *usus proprie* 'proper usage', associating the former with the study of grammar and the latter with the study of logic (of Canterbury, 1938–1961; Uckelman, 2009). As a result, it is problematic for Stanovich to connect his weak notion of rationality with the Aristotelian one, for it is neither specific to humans nor descriptive in nature. Instead, the conception of rationality as a capacity to reason correctly encompasses both the instrumental and epistemic aspects, and provides a normative grounding.

4. The normative and descriptive roles of rationality

Having identified various ways in which 'rationality' has been explicated in both natural and artificial contexts, we are now in a position to discuss whether these notions are best used normatively or descriptively. We aim to show that concepts of rationality that are predominantly normative in the artificial context are descriptive in the human context.

There are a number of different factors that influence how we design rational machines, including flexibility, reliability, rectitude, and user-friendliness. One of the most basic is a desire for familiarity: We desire the familiar and fear the unknown. When faced with the unknown, we seek to make it familiar.[11] An agent who does not act in a 'rational' fashion (however 'rational' ends up being cashed out) but instead acts in a chaotic and unpredictable fashion scares us.

This familiarity is, however, desirable only up to a certain point, after which we enter the so-called 'Uncanny Valley' (Mori, 2012), the descent into eeriness that characterizes the responses people have to human-like robots, which "abruptly shift from empathy to revulsion as it [the robot] approached, but failed to attain, a lifelike appearance" (Mori, 2012, p. 98).[12] Mori introduced the concept of the uncanny valley with respect to people's responses to the appearances of robots (cf. also (Kätsyri et al., 2015, pp. 2–3)), and the concept has been extended to the feelings of eeriness that arise when there is a mismatch between the appearance of an AI (such as a virtual character) and how it behaves (Tinwell, Nabi, & Charlton, 2013). A visual component is not necessary: We can also consider how people respond to the actions of robots and other AIs, regardless of their appearance. On one end of the spectrum, if the actions of a system are too far removed from human activity, such that they cannot be predicted or understood, then that very unpredictability will induce fear. But on the other hand, if these actions approach but do not attain a human-like characterizability, then we may find ourselves yet again in the uncanny valley. We return to this complication below.

A further component in our desire to create artificial agents that are rational in a human-like way is that the only commonly acknowledged examples for rationality—usually taken to be indicated indirectly but reliably through an agent's corresponding 'intelligent' or 'rational behavior' (Newell, 1982; Russell, 2016)—can be found in hu-

[11] Such as, for instance, when we impute a concept of causality onto the world, as Nietzsche argues we do (Nietzsche, 1967, p. 551).

[12] The empirical evidence for the uncanny valley hypothesis remains inconsistent (Kätsyri, Förger, Mäkäräinen, & Takala, 2015, p. 1).

man reasoning. From this observation stems the widespread assumption, often taken for granted without further justification, that (re-)creating human-style reasoning within an artificial system will also result in a (re-)creation of human-level capacities and mental faculties (Langley, 2012). Indeed, this recreation of human-like reasoning has been part of the definition of AI since the beginning. In using re-interpretations of the 'imitation game' as basis for his foundational proposal of how to characterize machine intelligence, Turing (1950) placed human-likeness of the machine's manifested behaviour and (observable) reasoning processes and outcomes at the core of AI.[13] In equating (or rather replacing) the question "Can machines think?" (Turing, 1950, p. 433) by the well-known imitation game counterparts involving a human and an artificial player, Turing established indistinguishability of the computer's game playing from its human opponent's performance—of which rationality and corresponding reasoning behavior form an important element that can easily be tested given the language-based nature of the games. And indeed, for example Besold (2013c) explicitly proposes human-like rationality as one of the four constitutive "SubTuring challenges" in a deconstruction of the original Turing Test into more specific subtasks.

The desire for intelligent machines to behave like ourselves is rooted in more than just a desire to minimize fear. There are practical reasons to require human-like rationality in agents when it comes to human-robot or human-computer interaction. In order for us to successfully interact with machines, they must be sufficiently like ourselves to be comprehensible. The importance of human comprehensibility of AI systems has very recently been emphasized by Forbus (2016), who makes a case for AI as a research endeavor being equivalent to learning how to create smart *software social organisms* which exhibit increasing abilities to participate in human culture and daily life. The comprehensibility of the system's behavior and outputs is paramount in this context, since efficient communication is a necessary condition for enabling participation in human society. In more technical contexts, the importance of making a system's decision and reasoning procedures recognizable and comprehensible for humans has been emphasized. For instance in the study of classification models (Freitas, 2014; Letham, Rudin, McCormick, & Madigan, 2015) improved comprehensibility of the model is considered to have positive impact on the trust users have in the resulting predictions in domains such as medical decision-making. Tying these general observations back into a rationality context, Butterworth and Blandford (1999) propose that an approach to reasoning about interactive behavior can be based on the assumption that computer users are rational, and that the behavior of the interactive system as a whole results from the rational behavior of the users in combination with the programmed behavior of the system's parts. Of course, the reliability of such an approach depends on the quality of the rationality model, and the latter's suitability for such a use—where comprehensibility of the system's reasoning and the resulting behavior can safely be considered one of the crucial factors.

The foregoing discussion then gives us an insight as to where the normativity of rationality in the context of AI comes from. It is rooted in some concept of human rationality and the desire for machines to act in a way sufficiently similar to humans to enable productive interactions with them as well as to remove the fear of the unknown. We want AI agents to be not only able to make good arguments but also to be able to listen to them.[14] Now that we have a normative concept of rationality in humans, and an under-

[13]This holds true for both modifications of the original game setting proposed by Turing (1950), the Original Imitation Game Test and the Standard Turing Test (using the terminology introduced by Sterrett (2000)). On the relation of both re-interpretations and the respective role of human-like performance in this context, cf. Besold (2013c).

[14]Thanks to an anonymous referee for this lovely analogy.

standing of how the desire for normativity comes about in the context of AI, we are in a position to see how the human counterpart matches up to the normative concept of rationality in artificial agents. We now argue for a surprising conclusion, namely that the normative concept of artificial rationality is rooted not in the normative concept of human rationality, but in a descriptive one.

First, it is important to note that the Aristotelian-medieval tradition, wherein humans are *by definition* rational, still nevertheless leaves open the possibility for humans to act irrationally. For example, Paul of Venice (early 15th C) describes the view of one of his opponents as "irrational" without thereby implying that his opponent is inhuman[15], and half a century earlier Gregory of Rimini also describes a view of his opponent as "entirely irrational" (Maierù, 1981, p. 483). That humans can fail to act in an entirely rational way has long been admitted: That they *often* fail to act in an entirely rational way began to be increasingly demonstrated in the 1960s and on (cf., e.g., the familiar results of Wason's selection task mentioned in §3), with various results showing that how people *do* reason and how people *should* reason often fail to match up.[16] It was then, not in Aristotle or his commentators, that a shift towards a descriptive approach to rationality, the type of rationality which Stanovich describes as "weak", arose, but rather in more modern attempts to characterise how it is that humans *actually* reason.

What then, is the relationship between these different accounts of human rationality and artificial rationality? In particular, are both conceptions of rationality to be understood normatively? If not, then from what does the asymmetry between artificial rationality and natural rationality arise? Recall the notion of the 'uncanny valley' that we discussed above: Part of what underpins the normativity of rationality in artificial contexts is the desire that our artificial agents work in a way that is similar to us so as to escape falling into that valley. If, however, humans do not, generally, attain the normative standard of rationality that is proposed, then the concept of human rationality which is being used as the benchmark for the normative standard of artificial rationality *must* be a descriptive one: Because humans fall short of perfect rationality, a perfectly rational machine would almost immediately fall victim to the uncanny valley.[17] This holds for any normative concept of human rationality that humans in actuality fall short of. If we aim to stay out of the uncanny valley by ensuring that our AIs reason in a human-like fashion, then the AIs will also fall short of that normative standard. If they didn't, then they wouldn't be reasoning in a human-like way. In summary, in order to call a machine 'rational' (or, by extension, 'intelligent'), we want it to reason *as we do* rather than *as we ought*.[18] This difference, thus, is mostly pragmatically motivated by the differing needs

[15] "This reply seems irrational to me" (of Venice, 1978, pp. 142–143).

[16] Various explanations of this fact have been produced, such as that humans are 'bad' at reasoning logically because we haven't actually identified the right logical constraints or logical system, perhaps due to the fact that human interaction and reasoning is so contextualized that taking into account all the relevant contextual factors in a general way in order to capture 'good' and 'bad' human reasoning is extremely difficult (an assumption underlying, for instance, the work of Varga, Stenning, and Martignon (2015)), because there are evolutionary factors at play (cf. Nowak, Page, and Sigmund (2000)'s explanation of empirical evidence for how humans play the Ultimatum Game), or because the function of reasoning itself might have to be reconsidered and re-framed (cf., for example, Mercier and Sperber (2011)'s proposal for conceptualizing reasoning as argumentation-centric in nature).

[17] Accordingly, for instance the already mentioned SubTuring rationality task proposed by Besold (2013c) explicitly targets rational reasoning as observed in humans.

[18] In advocated this, note that we are happy to exclude the case of, say, a pocket calculator, which we do not care to call either rational or intelligent, and thus do not care how it arrives at its results, so long as it gives us the correct results. More problematically, perhaps, this excludes projects explicitly aiming at implementing 'rational agents' in the strong, dedicatedly mathematical sense—rather than general human-level intelligent systems as encompassed by the definition borrowed from (Nilsson, 2009) at the beginning of §1—like, for example, the already mentioned AIXI model (Hutter, 2007).

and purposes a concept of rationality has to serve in either context.

Recognizing this helps us account for a perhaps puzzling fact, namely that when we speak of 'Artificial Intelligence', we are appealing to some, perhaps ill-defined, notion of *human* intelligence. Setting aside the question of metrics (how to measure the intelligence of a non-human subject), we might think it strange that the normativity of rationality in the artificial context is essentially rooted in the fact that we defined artificial intelligence to be "intelligent *like us*". There is no in-principle reason why this should be so; in fact, there is antecedent reason to think that perhaps the aim of AI researchers is to develop a concept of intelligence that is *supra*-human, going beyond our current capacities (Yampolskiy & Fox, 2012).[19] However, the mismatch between the normative and descriptive conceptions of rationality in humans provides us with an explanation: The very fact which motivates the normativity of rationality in AI is the same fact which requires that we use the descriptive account of human rationality as the benchmark. This also provides an explanation for the set-up of the various types of Turing tests, which are designed to incorporate how people actually act, rather than how they should act. A perfectly rational machine would almost certainly fail the Turing test.

5. Conclusion

We end this paper without giving any clear-cut definition of rationality. If the reader takes away nothing else from the paper, we hope that they agree that giving such a definition is extremely difficult. Beyond that, because of the different roles that the notion plays in different disciplines, it is important that one does not carelessly transfer definitions and models from one discipline into another: What may be suitable in the context of speaking of human rationality may not be suitable for discussing rationality in AI, for even if the concepts used are the same, whether they are being applied normatively or descriptively may differ. Similarly, when one is building a new model or a new implementation, care should be taken that the correct conception is being used in the correct way.

An asymmetry between the respective notions of rationality in a human and an artificial context has become apparent. While disciplines studying rationality in humans commonly apply a normative concept based on one of several theoretical rationality frameworks, the normative standard for artificial rationality arguably is—and has to be—a descriptive concept of human rationality. As pointed out in §4, 'human-likeness' has been a desideratum for artificially intelligent agents since the early days of AI research. This also is the case for the agent's reasoning capabilities and reasoning-based behavior, and thus for the notion of rationality for AI.

Acknowledgements

The authors would like to thank the organisers and participants of Dagstuhl Seminar 15221 ("*Multi-disciplinary approaches to reasoning with imperfect information and knowledge*"), May 2015.

The first author carried out parts of this work at the KRDB Research Centre of the Free University of Bozen-Bolzano (Italy), and at the Digital Media Lab of the University of

[19] After all: When leaving the potential explanatory dimension of AI with respect to cognitive phenomena aside (which some researchers as, for instance, Miłkowski (2013) anyhow consider as at best limited), there seems to be no urgent need to construct more agents whose capacities are equal to ours. We can create those through natural means.

Bremen (Germany).

References

Aristotle. (1941a). Nichomachean ethics. In *The basic works of Aristotle* (pp. 935–1112). Random House. (Edited with an introduction by Richard McKeon)

Aristotle. (1941b). Topics. In *The basic works of Aristotle* (pp. 187–206). Random House. (Edited with an introduction by Richard McKeon)

Aristotle. (1979). *Aristotle's metaphysics: Translated with commentaries and glossary by Hippocrates G. Apostle*. Peripatetic Press.

Baron, J. (2012). The point of normative models in judgment and decision making. *Frontiers in Psychology*, *3*(577).

Besold, T. R. (2013a). Human-Level Artificial Intelligence Must Be a Science. In K.-U. Kühnberger, S. Rudolph, & P. Wang (Eds.), *Artificial general intelligence—6th international conference, AGI 2013, proceedings* (Vol. 7999, pp. 174–177). Springer.

Besold, T. R. (2013b). Rationality in/for/through AI. In J. Kelemen, J. Romportl, & E. Zackova (Eds.), *Beyond artificial intelligence* (Vol. 4, pp. 49–62). Springer.

Besold, T. R. (2013c). Turing Revisited: A Cognitively-Inspired Decomposition. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (p. 121-132). Springer.

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

Butterworth, R., & Blandford, A. (1999). The principle of rationality and models of highly interactive systems. In M. A. Sasse & C. Johnson (Eds.), *Human-computer interaction—INTERACT '99*. IOS Press.

Byrne, R. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83.

Cassimatis, N. I. (2012). Human-Level Artificial Intelligence Must Be an Extraordinary Science. *Advances in Cognitive Systems*, *1*, 37–45.

Clark, A. (2003). Artificial intelligence and the many faces of reason. In S. Stich & T. Warfield (Eds.), *Blackwell guide to philosophy of mind*. Blackwell.

Cowen, T. (2004). How do economists think about rationality? In M. Byron (Ed.), *Satisficing and maximizing: Moral theorists on practical reason* (pp. 213–236). Cambridge University Press.

Davis, E., & Marcus, G. (2015, August). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103.

Doyle, J. (1992, May). Rationality and its roles in reasoning. *Computational Intelligence*, *8*(2), 376–409.

Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*, 978–996.

Forbus, K. D. (2016). Software social organisms: Implications for measuring ai progress. *AI Magazine*, *37*(1), 85–91.

Freitas, A. A. (2014, March). Comprehensible classification models: A position paper. *SIGKDD Explorations Newsletter*, *15*(1), 1–10.

Galluppi, F., Urdiales, C., Sandoval, F., & Olivetti, M. (2009). A study on a shared control navigation system: Human/robot collaboration for assisting people in mobility. *Cognitive Processing*, *10*(2), 215–218.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundation of adaptive behavior*. Oxford University Press.

Gilboa, I. (2010). Questions in decision theory. *Annual Reviews in Economics*, *2*, 1–19.

Gilboa, I., & Schmeidler, D. (2001). *A Theory of Case-Based Decisions*. Cambridge University Press.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling*. Cambridge University Press.

Halpern, J. Y. (2008). Beyond Nash Equilibrium: Solution Concepts for the 21st Century. In *Proc. of the 27th Annual ACM Symposium on Principles of Distributed Computing.*

Hammond, P. J. (1997). Rationality in economics. *Rivista Internazionale di Scienze Sociali*, *CV*, 247–288.

Hutter, M. (2007). Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel & C. Pennachin (Eds.), *Artificial general intelligence* (pp. 227–290). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6, article 390*, 1–16.

Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

Kokinov, B. (2003). Analogy in decision-making, social interaction, and emergent rationality. *Behavioral and Brain Sciences*, *26*(2), 167–169.

Langley, P. (2006). Cognitive architectures and general intelligent systems. *AI Magazine*, *27*(2), 33.

Langley, P. (2012). Intelligent behavior in humans and machines. *Advances in Cognitive Systems*, *2*, 3–12.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015, 09). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371.

Maierù, A. (1981). Logica Aristotelica e teologia trinitaria Enrico Totting da Oyta. In A. Maierù & A. P. Bagliani (Eds.), *Studi sul XIV secolo in memoria di Anneliese Maier* (pp. 481–512). Edizioni di Storia e Letteratura.

Mele, A., & Rawling, P. (Eds.). (2004). *Oxford handbook of rationality.* Oxford University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.

Miłkowski, M. (2013). Limits of computational explanation of cognition. In C. V. Müller (Ed.), *Philosophy and theory of artificial intelligence* (pp. 69–84). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mori, M. (2012). The uncanny valley. *IEEE Robotics and Automation Magazine*, 98–100. (Karl F. MacDorman and Norri Kageki, trans.)

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104) (No. 9). Prentice-Hall Englewood Cliffs, NJ.

Nietzsche, F. (1967). *The will to power.* Knopf.

Nilsson, N. J. (2006). Human-Level Artificial Intelligence? Be Serious! *AI Magazine*, *26*(4), 68–75.

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence.* New York, NY, USA: Cambridge University Press.

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*, 1773–1775.

of Canterbury, A. (1938–1961). De grammatico. In F. S. Schmitt (Ed.), *S. Anselmi Cantuariensis archiepiscopi opera omnia* (Vol. 1). Seccovii: Abbatia.

of Venice, P. (1978). *Pauli Veneti: Logica magna, secunda pars: Tractatus de veritate et falsitate propositionis et tractatus de significato propositionis.* Oxford University Press. (Ed. by Francesco del Punta and trans. by Marilyn McCord Adams)

Oppenheimer, D. (2003). Not so fast! (And not so frugal!): Rethinking the recognition heuristic. *Cognition*, *90*(1).

Osborne, M., & Rubinstein, A. (1994). *A course in game theory.* MIT Press.

Porphyry. (2003). *Introduction.* Oxford University Press. (Trans. by Jonathan Barnes)

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 473–484). Morgan Kaufmann.

Russell, S. (2016). Rationality and intelligence: A brief update. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (Vol. 376). Springer.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach*. Prentice Hall.

Sadek, M. D., Bretier, P., & Panaget, F. (1997). ARTIMIS: Natural dialogue meets rational agency. In *Proceedings of IJCAI-97* (pp. 1030–1035). Morgan Kaufmann.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *American Economic Review*, *49*(3), 253–283.

Simon, H. A. (1990). *Reason in human affairs*. Stanford University Press.

Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5179–5187).

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 343–365). Oxford University Press.

Sternberg, R. J. (Ed.). (2000). *Handbook of intelligence*. Cambridge University Press.

Sterrett, S. G. (2000). Turing's two tests of intelligence. *Minds and Machines*, *10*(4).

Tinwell, A., Nabi, D. A., & Charlton, J. P. (2013). Perception of psychopathy and the uncanny valley in virtual characters. *Computers in Human Behavior*, *29*, 1617–1625.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, *90*(4), 293–315.

Uckelman, S. L. (2009). Anselm's logic of agency. *Logical Analysis and History of Philosophy*, *12*, 248–268.

van Rooij, I., Wright, C., & Wareham, H. T. (2010). Intractability and the use of heuristics in psychological explanations. *Synthese*.

Varga, A., Stenning, K., & Martignon, L. (2015). There is no one logic to model human reasoning: The case from interpretation. In U. Furbach & C. Schon (Eds.), *Proceedings of the workshop on bridging the gap between human and automated reasoning, 25th international conference on automated deduction (CADE-25)* (Vol. 1412, pp. 32–46). CEUR-WS.org.

Wason, P. C. (1966). Reasoning. In *New horizons in psychology* (Vol. 1, pp. 135–151). Penguin Books.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*(1), 63–71.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, *abs/1609.08144*. Retrieved from `http://arxiv.org/abs/1609.08144`

Yampolskiy, R. V., & Fox, J. (2012). Artificial general intelligence and the human mental model. In A. Eden, J. Soraker, J. H. Moor, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment*. Springer.