Artificial Intelligence as Expected Intelligence

Logic and the Philosophy of Science are disciplines that avoid stopping and wandering around in circles on the same topics. Their role in dedicating themselves to scientific and applicative subjects always makes them extraordinarily fruitful and productive, in step with the developments of scientific and technological knowledge. The writings contained in this volume are proof of this. The set of research selected for this publication, starting from the 2022 postgraduate conference of SILFS (the Italian Society for Logic and the Philosophy of Science), testifies to a lively and rich knowledge widespread among young scholars of this area who combine pure logic and philosophy to the analysis of cutting-edge knowledge produced by current scientific and technological disciplines.

The result is a kaleidoscope of research on updated logical and philosophical-scientific topics, which gives rise to promising developments on the part of young scholars who are contributing today to the logic and philosophy of science to create further perspectives in the future.

ISBN 978-1-84890-455-2
90000
9 781848 904552

6.140 x 9.210
234 mm x 156 mm

Current Topics in Logic and the Philosophy of Science
Editors: Francesco Bianchini, Vincenzo Fano, Pierluigi Graziani

0.383
9.73mm

Current Topics in Logic and the Philosophy of Science

Papers from SILFS 2022 postgraduate conference

Editors
Francesco Bianchini
Vincenzo Fano
Pierluigi Graziani

6.140 x 9.210
234 mm x 156 mm

Content Type: Black & White
Paper Type: White
Page Count: 180
ISBN: 978-1-84890-455-2
Trim Size: 6.14x9.21
File Type: InDesignCC
Request ID: CSS4346079

SILFS

Volume 4

Current Topics in Logic and the Philosophy of Science

Papers from SILFS 2022 postgraduate conference

Volume 1
New Essays in Logic and Philosophy of Science
Marcello D'Agostino, Guilio Giorello, Federico Laudisa, Telmo Pievani and Corrado Sinigaglia, eds.

Volume 2
Open Problems in Philosophy of Sciences
Pierluigi Graziani, Luca Guzzardi and Massimo Sangoi, eds.

Volume 3
New Directions in Logic and the Philosophy of Science
Laura Felline, Antonio Ledda, Francesco Paoli and Emanuele Rossanese, eds.

Volume 4
Current Topics in Logic and the Philosophy of Science. Papers from SILFS 2022 postgraduate conference
Francesco Bianchini, Vincenzo Fano and Pierluigi Graziani, eds.

SILFS Series Editor
Marcello D'Agostino marcello.dagostino@unimi.it

Current Topics in Logic and the Philosophy of Science

Papers from SILFS 2022 postgraduate conference

Edited by

Francesco Bianchini

Vincenzo Fano

Pierluigi Graziani

ISBN 978-1-84890-455-2

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr
Department of Computer Science

http://www.collegepublications.co.uk

Original cover design by Laraine Welch

Contents

ARTICLES

Artificial Intelligence as Expected Intelligence

Martina Bacaro, Francesco Bianchini
University of Bologna
martina.bacaro2@unibo.it, francesco.biachini@unibo.it

Abstract. The inquiry into the nature of intelligence within artificial intelligence (AI) has been a persistent pursuit since the inception of the discipline, notably evident in Turing's seminal works predating the formalization of AI itself. Turing sought to establish the viability of thinking machines, laying the groundwork for subsequent reflections on attributing intelligence to artificial artifacts, particularly in the form of software programs. This chapter deals with the issue of measuring intelligence in artificial artifacts, emphasizing the importance of considering the expected intelligence. From an interactive standpoint, the human expectation of intelligence in an artificial entity revolves around performances deemed cognitively suitable for meaningful interactions. This perspective advocates for the detection of varying degrees of intelligence that align with the requirements of reliable and cognitively relevant interactions, even at intermediate levels. The discussion delves into the inherent differences between the simulative nature of human intelligence in artificial artifacts and the distinct characteristics expected for genuine intelligence in interactive contexts, particularly within the realm of social robotics. Indeed, the analysis of specific aspects of intelligence in robotic artifacts may prove enlightening by showing how the measurement of intelligence is significantly contingent upon the expectations that the human participant retains during interaction with AI. The inherently collaborative and interactive nature of the relationship between humans and robots underscores particular dimensions within the analysis of intelligence that might remain latent in interactions with disembodied AI systems.

Keywords: Measure of Intelligence, Expected Intelligence, Social Robotics.

1 Introduction

The issue of intelligence in artificial intelligence (AI) has been investigated since the dawn of the discipline, indeed it can be said that it was a foundational question. Turing's papers on machine intelligence addressed the problem before AI was even born [1]. In fact, Turing pursued the goal of establishing the possibility of machines, i.e. suitably programmed digital computers, which were capable of thinking. However, this starting point immediately led the British thinker to question the ways in which intelligence can be attributed to an artificial artifact, which in his reflection takes the form of a software program. Over the decades, the question has changed many times, developing in different directions that have followed the many evolutions of AI approaches [2]. In more recent times, the theme has also been re-proposed in the form of a possible measurement of intelligence in AIs, giving rise to further questions concerning the possibility of identifying, even quantitatively or along a scale, the presence of intelligence in artificial systems.

In this chapter we will address the problem of measuring intelligence in AIs arguing that it concerns and should consider more the expected intelligence of artificial artifacts. This is because from an interactive point of view, what the human being expects from an artificial artifact defined as intelligent is a set of performances that can be recognized as cognitively suitable for the interaction itself, so as to be recognizable and reliable. In this sense, we can speak of the detection of a, more or less high, degree of intelligence, which satisfies the requirements of a reliable and cognitively relevant interaction, even at intermediate levels. The investigation of general or human-level AI is certainly an interesting research field. We believe that the quantity of expected, and therefore attributable, intelligence, which artificial artifacts defined as intelligent must show, without this being a deception, but at the same time independently of the simulative nature of the human intelligence of the artifact itself, is equally important. We will therefore try to argue how the two aspects are only partially overlapping and we will provide some examples taken from the world of robotics, which, due to its nature as a discipline aimed at building situated and interactive entities with human beings, is suitable enough for such a kind of investigation.

The structure of the chapter is as follows. We will first analyze some aspects of the notion of measuring intelligence in natural and artificial systems (§ 2). We will then discuss the relationship between actual and expected intelligence in AI, also providing a proposal for parameterizable characteristics to measure expected intelligence (§ 3). We will then review some aspects

of the development of AI to see its shortcomings and how this influences expectations (§ 4) and then move the discussion on robotics to interactive contexts (§ 5). Finally, after discussing some attempts to measure robotic intelligence (§ 6), we will draw some conclusions underlining possible future prospects for investigation (§ 7).

2 Measuring intelligence in natural and artificial systems

Unexpectedly, the issue of measuring intelligence has not been a dominant topic in AI from the start, as it has been the case with regard to the development of methodologies to measure human intelligence since the first steps in the early Twentieth century by psychologists such as Alfred Binet and Théodore Simon [3]. The issue of intelligence in AI was instead immediately addressed according to a very different scheme, which admitted only two possibilities: the presence or absence of intelligence [1]. The presence of intelligence, according to Turing, was what could be detected and attributed from the outside to a certain type of machines (suitably programmed discrete-state machines) in relation to their behavior. For this reason, the Turing Test in the following decades took the form of a yes/no test on the presence of intelligence in machines, i.e. programs which, in view of this possibility, fall within the vast and multifaceted field of AI (from software agents, to NLP applications, to problem solving algorithms, to robotics, and to neural networks, just to mention some of the most important AI areas). If this already marks the difference with human beings and the very goal of measuring their intelligence, the enterprise has proved to be even more difficult and somehow opaque as over the years there has been no effective definition of intelligence to be able to approach the intelligence of AI. The two problems seem to go hand in hand.

One of the aspects that has been underlined by those who have addressed this issue has been the difference between the types of intelligence involved: on the one hand, intelligence as a set of task-specific skills and, on the other, intelligence as a general ability of learning and open-endedly performing. This difference has been underlined for example by Chollet [4] in outlining his proposal for the measurement of AI intelligence. In fact, an AI focusing on a specific aspect and implementing a performance related to a particular skill seems more easily measurable in terms of intelligence, as it will eventually correspond to the degree of accuracy with which that performance is carried out. The measure of the success with which the system performs its

task becomes a good marker for measuring the intelligence of the system, provided that the system actually performs the task for which it is designed in the way in which this is presupposed by the goals of the system itself, or "if our measure of performance captures exactly what we expect of the system" ([4]: 9). It's not hard to think of many AI systems where this type of measurement can be accomplished. The problem is different when trying to measure an AI system's ability to generalize, both in terms of a system's ability to deal with situations it has not previously encountered (a narrow form of generalization) and in relation to the ability to deal with situations that not even the developer of the system, as well as the system itself, has previously encountered (a broader form of generalization).

This aspect has also been present in the AI field since its dawn, for example in the attempt to generalize the abilities of the Logic Theorist into a General Problem Solver equipped with problem solving techniques that could be adapted to different types of problems and therefore independently of a specific type of content (in the case of the Logic Theorist, issues of logical scope and content). This declared goal by the authors of the program [5] later proved to be much harder to achieve than initially believed due to various problems connected with the possibility of operating within not always totally defined domains (unlike totally defined ones, the so-called toy domains) as well as with the adequacy of the represented and available knowledge to the program, just to mention the most relevant obstacles. The search for the generality of AI, yet, was not only a problem, but also a driving pressure in the evolution of AI approaches, leading to numerous results also theoretically. Among them we can mention, as an almost direct consequence, the approach of cognitive architectures such as SOAR, ACT-R and others, which constituted one of the main attempts to build AI systems provided with a greater power of operational generalization [6].

Another important aspect from this point of view is that the generalization of AI systems, in many approaches and, in particular, in cognitive architecture approach, has linked the possibility of achieving this kind of generalization to the reconstruction of cognitive abilities in the broad sense, and therefore to a cognitive matrix. This could lead us to consider the problem of measuring intelligence in an AI system of cognitive inspiration more easily solvable, precisely because the same techniques that are used to measure human intelligence seem relatively equally easily applicable, at least in principle, to cognitive artificial systems, i.e. to cognitive AI. If on one hand this may be true, on the other it should not be forgotten that a large part of the AI systems that today show a performance considered as very intelligent are machine learning systems, which therefore include a specific reference to

learning, based on neural networks also of the deep kind. These systems, however, have given rise to a further explanatory problem. In many cases, they operate as a black box and it is not possible to recognize the reasons why they achieve very high results in terms of precision and accuracy, due exactly to the nature of the system itself, which is therefore not "explainable". This has led to the development of a continuously updated field of techniques to overcome this problem, that of Explainable AI or XAI [7]. These systems seem to inherently escape a measurement of their level of intelligence, yet, at least in a cognitive sense of the term. What can be measured in these AI systems, which can largely be traced back to decision-making performance in terms of proposed outputs, is the accuracy of the results compared to expectations. Thus, also in this case, what is measured is performance, without having any possibility of defining whether it is intelligent or not, i.e. the result of an activity that can be qualified as intelligent. In more abstract terms, it can be said that these algorithms are highly generalizable - a same technique that can be applied to many domains depending on the starting dataset - but this does make them opaque to an explanation and measurement of their intelligence.

An extensive analysis of the different attempts to measure the intelligence of AI systems shows that the ways of measuring (which usually boils down to a performance evaluation) of an AI system are many, but also very fragmented and specific for different systems, limiting its overall success. It seems difficult to obtain a reference standard for measuring intelligence that can be used for all AI systems. Furthermore, the distinction between a task-oriented assessment and a cognitive skills assessment, which is proposed by Hernádez-Orallo [8], seems to recall the aforementioned distinction between task-specific skills and intelligence as a general ability of learning and open-endedly performing. In this case, however, the distinction is motivated by a greater evaluation capacity of implemented methods and metrics. For example, in the case of assessing systems built to carry out a specific task, what is evaluated is performance rather than intelligence. Consequently, there is a risk that more than the AI system, those who designed and built/programmed it will be evaluated. This is certainly a useful evaluation, even if it is far from being a measurement of intelligence in AI. Hernádez-Orallo identifies three types of task-oriented evaluation: human discrimination, problem benchmarks, peer comparison. Without going too far into the details of each of these kinds of methods, we can draw attention to the fact that while the first is rather subjective, the others aim to be more objective by posing, in different ways, a parameter or several reference parameters, whether they are a standard or a result of a comparison (e.g. a match, a game, a compe-

tition).
An ability-oriented evaluation can lead to a better measurement of intelligence, even in progressive terms; for example, in relation to how much a system that is based on continuous updating due to machine learning techniques develops a progress in its intelligence over time. The reference to the measurement of abilities involves more general cognitive aspects such as reasoning skills, inductive learning abilities, motion abilities, etc. [8]. Even in this case, however, the problem oscillates between a greater anthropocentrism, such as for example with the use of psychometric tests, also those that are adapted to AI as in the case of Psychometric AI [9]; and, on the other hand, a stronger search for an objective and neutral quantification, with respect to the human capabilities, of the system's abilities, for example through the use of metrics derived from the algorithmic information theory, which is based, among other things, on the Kolmogorov notion of computational complexity [10]. The proposal of a universal psychometrics [11] seems to be a way to overcome the impasse due to the dichotomy between anthropocentric methods and the search for formal standards (encompassing general and neutral tests) for measuring performance in the case of general cognitive abilities. Nevertheless, even in this case it does not seem possible to completely escape a sort of anthropocentrism which, as regards general-purpose systems, moves towards the notion of cognition and the definition of cognitive ability. This becomes even more evident if we systematically consider the methods for measuring intelligence, combining those for natural intelligence and those for artificial intelligence [12], despite the many developments at least in terms of discussion on such a topic that have been accomplished in recent years [13]. Somehow, it seems impossible to measure intelligence without attributing such general ability, or the vehicles of that ability, to something external to the system (a performance in a specific task or a more general skill) and against an external standard to the system, but belonging to other intelligent systems (basically the human beings).

3 On actual and expected artificial intelligence

This view related to the evaluation and measurement of intelligence in AI systems is clearly problematic and it also seems hard to see how it is possible to go in a direction that constitutes an advance with respect to the impasse between specific skills evaluation and general ability evaluation. Such an impasse seems to give support to those who propose a deep transformation

of our expectations on AI [14]. According to this view, artificial intelligent systems do not have much intelligence, indeed they probably do not have any, at least in the human sense of the term. They are machines, of course, and they are useful machines, however not for their intelligence, but as tools that are capable of producing good or excellent results. In this way, AI becomes a set of tools acting in a surprising way to achieve the expected results (excellent execution of the task). As an ability to act, AI systems have their strength in being interactive, autonomous and in the capability to learn and self-learn.

According to this view, AI exists, but not as intelligence per se. It is a very deflationary, perhaps radical, framework of AI, but perhaps less implausible than the catastrophic standpoints connected to the most recent developments of AI. On the other hand, if it is reasonable to consider such a framework as radical, it is also true that *an intelligent result is not always an intelligent behavior, or the result of an intelligent behavior*. This discrepancy, which is at the basis of the dissociative vision between intelligence and action of AI systems, now seems to be no longer contestable. The most recent AI developments show that the engineering solutions to achieve the most accurate performances are also those that lead to the best systems to replace humans in carrying out tasks considered intelligent or requiring human intelligence, without the use of such intelligence. In this case too an impasse is created, which goes beyond the one generated by the impossibility of explaining how certain results are achieved by the system. It is the impasse one faces with respect to the concept of intelligence itself, when it is replaced by something that one fails to consider, in a fine-grained analysis or from a structural point of view, intelligent.

However, there is a perspective that can be considered complementary, which sees AI systems as artifacts or social machines endowed with the peculiar features of being autonomous, self-learning and capable of carrying out a teleological, goal-driven behavior in their social interaction [15]. The social aspect makes the goal-oriented nature of these machines relevant in general, since their purpose is not that of those who designed them, nor that of the users with whom they interact, but the one which arises from the interaction between system and user. If considered broadly, this interpretative framework embraces a multiplicity of AI systems, from software managing the performance of many online platforms, to search engines, to interactive humanoid and non-humanoid robotic systems, just to name a few. We argue that in all these cases, and therefore in most cases in which human users are involved with AI systems (almost everyone when AI is involved), *it is essential the attribution of intelligence the human user does, often in real-time,*

to obtain good results and an interactive behavior as optimal as possible. In other words, attributing intelligence on the part of human users seems an unavoidable requirement for the correct epistemic, applied and ethical functioning of these systems. From an epistemic standpoint, the attribution of intelligence can be seen as a cognitive necessity for users attempting to comprehend the system's responses and actions. It serves as a cognitive shorthand, allowing individuals to navigate the complexity of AI interactions by ascribing human-like understanding and intentionality to the machine. In this way, users can better anticipate and interpret the system's behavior, facilitating a smoother and more intuitive collaboration. On the other hand, applied functionalities of AI systems, especially in decision-making processes, benefit significantly from the user's attribution of intelligence. Furthermore from the ethical point of view, recognizing the user's role in attributing intelligence underscores the responsibility of designers and developers to create AI systems that align with human values and ethical standards. In other words, the attribution of intelligence serves as a crucial bridge between the designed purpose of the system and the dynamic, evolving nature of human interaction. Recognizing and attributing intelligence to AI systems fosters a symbiotic relationship where the machine's goal-oriented behavior aligns with user expectations and needs.
At this point one may wonder how it is possible to characterize this notion of intelligence attribution. Of course, there have been many well-known attempts, from Turing onwards, to define the attribution of intelligence to artificial systems, and, from Dennett [16] onwards, to define the attribution of intentionality. However, we would like to propose a pragmatic method for attributing intelligence, with the awareness that it also has complex theoretical and ethical implications. We therefore propose the following characteristics that an artificial intelligent system should/could have according to the user's perspective:

1. intelligence is attributed to the system before its use or interaction with it (on the basis of the preliminary knowledge that the user has of it);

2. intelligence is attributed to the system during its use or interaction with it;

3. the qualification of intelligent is attributed to the results of the system performance or of the interaction with it;

4. the repeated uses or repeated interactions with the system over time allow an attribution of intelligence subject to time variability.

If each of these characteristics is parameterized using a scale of values we can obtain an overall measurement of the users' expected intelligence related to the AI system they are using or with which they are interacting. This measurement can vary over time (especially in consideration of the parameter derived from item 4) and can also be refined with more detailed scales of values, thus producing a finer-grained differentiation, depending on the aims to be achieved. The overall goal remains to create a tool that is as flexible as possible for attributing intelligence to AI systems. The limit that can be recognized in this tool is that of being user-dependent. However, this is also one of its strengths, when such a tool is used to help draw up policies for using or engaging with AI systems, an increasingly pressing problem from a political, social and normative point of view. In addition, it has further strengths. For example, it is not a tool that focuses on the idea of purpose or goal-oriented behavior, however not excluding this characteristic from the evaluated AI system.
Furthermore, from the point of view of aggregated data (of individual user evaluations) it can allow an overall measurement of the intelligence expected by AI systems that allows an assessment not only in terms of policies, but also related to the general notion of intelligence, which could be reconstructed starting from the aggregated data of many users by making hypotheses on the various motivations that lead to certain attributions of intelligence. Our assumption is that expected intelligence (on the part of human being) is coupled with something in the system allowing intelligent behavior in general. This does not imply that such a thing is the same as in human beings, but at least is something that gives rise to intelligence, otherwise everything could potentially be described as, or ascribed with, intelligence without any particular reason. The reason (a mechanism, a technique, a dynamical interaction, a mathematical/statistical function or whatsoever) should be, to not lose the general notion of "artificial intelligence". In fact, this is not the case in the real world we live.
Finally, the user would in any case be safeguarded in his interaction with AI systems by a more adequate and at the same time transparent use of an artifact which calls into question the explicit awareness, on the part of the user, of the fact that he is interacting with AI systems equipped of different recognizable or explicit levels of intelligence.
A proposal of this kind is only preliminary and certainly needs more in-depth investigations. In principle, however, it can be a first step towards overcoming the impasse on the actual evaluation of the intelligence of an artificial system, which allows to consider intelligent artificial artifacts that are classifiable as such, without committing oneself to the metaphysical, episte-

mological or mechanistic reasons of such intelligence and, at the same time, without completely discarding the notion of intelligence in analyzing these systems. Furthermore, the more complex theoretical-epistemological investigations could in any case be carried out ex post, downstream of this type of measurement.
The analysis of specific aspects of intelligence in robotic artifacts may prove enlightening by showing how the measurement of intelligence is significantly contingent upon the expectations that the human participant retains during interaction with AI. Indeed, the inherently collaborative and interactive nature of the relationship between humans and robots underscores particular dimensions within the analysis of intelligence that might remain latent in interactions with disembodied AI systems. In the following sections, we will try to show that the issue of human expectations towards artificially intelligent systems explicitly emerged from contemplation on what first examples of AI lacked and how the demand for a collaborative, human-like form of intelligence gave rise, in addition to philosophical and technical inquiries, to novel modes of intelligence assessment and evaluation.

4 What is lacking in artificial intelligence and how it affects expectations

The initial years of AI were marked by a prevalent notion that human intelligence could, in principle, be comprehended and replicated in a machine [17]. This idea was first conceived during the notable gathering held at Dartmouth in 1956 [18]. At the time, intelligence was predominantly perceived as the ability to process data through computation, involving the deliberate application of suitable inferences for a given purpose [1][19]. Consequently, the field of AI has primarily focused on enhancing the logical and inferential capabilities of machines, with the belief that machines would achieve human-like results once their computational capacity reached a level of complexity equivalent to human intelligence.
Despite a clear agenda and initial successful applications, the field of AI encountered increasingly complex challenges, eventually entering a period of setbacks known as the AI Winter [20]. During the late 60s, several factors contributed to the negative impact on this research domain. Firstly, the recognition of the complexity of many problems posed difficulties as the frameworks proposed by Marvin Minsky oversimplified the issues, resulting in early systems excelling only in simplistic tasks. Real-world problems proved to be too intricate to solve, and scaling up the capabilities of AI

systems went beyond the realm of faster hardware and memory [21]. In addition, the "thinking humanly" approach that characterized early AI endeavors proved inadequate for solving the problems at hand. This approach, known as the symbolic approach [22], involved attempting to replicate human problem-solving methods without breaking down the main problem into possible solutions and formulating an algorithm. Moreover, in many cases, positive results have been prone to over-interpretation, leading the public to believe that AI systems possess much greater intellectual capabilities than they actually did. For instance, the renowned program Eliza, developed by Weizenbaum [23], was often perceived by participants as having achieved a genuine understanding of human problems, akin to that of a psychotherapist. However, the program lacked any true knowledge or comprehension of its interlocutors or their issues. The enthusiasm surrounding Eliza can be attributed to the significantly low expectations people had for AI systems, beyond purely logical performance, which allowed for quick excitement to arise.

During this period, the debate surrounding the reasons why intelligent machines failed to achieve the expected accuracy in tasks envisioned by computer scientists was primarily divided into two positions. On one hand, Minsky attributed the mistakes to a naive preconception held by computer scientists and AI practitioners regarding the nature of the mind and its various aspects. In a seminal paper, Minsky [24] addressed the prevalent questions of that time, which continue to be relevant today, regarding the capabilities (or limitations) of intelligent machines. Characteristics such as creativity, non-logical thought, and self-awareness were deemed by computer scientists, brain researchers, and cognitive scientists to be inherently human and beyond the realm of implementation in machines. This notion profoundly influenced the development of AI. The expectations for machine performance were confined to logical tasks and the successful execution of programmed instructions, devoid of any consideration for creative problem-solving or deviations from what the programmer had explicitly taught the machine. However, Minsky argued that “all those beliefs which set machine intelligence forever far beneath our own are only careless speculations, based on unsupported guesses on how human minds might work” ([24]: 15). He suggested that it was necessary to redefine our conception of intelligence, not only with regards to machines but also in relation to ourselves as humans, and to place greater trust in the power of our intuition, which had been instrumental in constructing the initial AI models.

On the opposing side of the debate, in stark contrast to Minsky’s stance, certain philosophers with alternative perspectives on mind and cognition ex-

pressed skepticism regarding the potential achievements of AI as a whole. This skepticism stemmed from ontological disparities between humans and machines and the fact that computer scientists embarked on their work without any knowledge of the philosophical attempts to address the same problems. Drawing on the phenomenology of Heidegger and Merleau-Ponty, as well as Gestalt psychology, Dreyfus argues that "what distinguishes persons from machines [...] is not a detached, universal, immaterial soul but an involved, self-moving, material body" ([25]: 149). The core argument posits that the problems confronted by computer scientists had already been contemplated by philosophers, and the arguments developed by the latter could be applied to the discourse on meaning and AI. The problem of meaning is a longstanding one in philosophy, and the manner in which computer scientists approached it traversed the history of philosophical thought on the subject, from Plato onwards. Dreyfus anticipated the failure of proposed plans for general problem solvers and automatic translation machines due to computer scientists' naive conception of mental functioning. As a result, he recommended that researchers familiarize themselves with modern philosophical approaches to human beings and intelligence if they aimed to replicate the characteristic aspects of human intelligence.

It is possible to argue that the positions put forth by Minsky and Dreyfus originated from the same recognition of the naivety of computer scientists, yet they were approached in divergent manners, resulting in distinct solutions and future agendas. In essence, Dreyfus believed that these limitations were practically insurmountable, starting from the phenomenological perspective on the body and consciousness. Conversely, Minsky advocated for a revaluation of the concept of intelligence and maintained that the desired outcome could be achieved with appropriate adjustments. However, Minsky's proposals proved to be more problematic than anticipated and sparked a new wave of contemplation on the feasibility of implementing mechanisms of signification and constructing machines that genuinely exhibit concern for their actions, as exemplified by the frame problem [26] and the symbol grounding problem [27].

Starting from these problems and recognizing the significant role of the body in the emergence of intelligent behavior, Brooks initiated a revolution in the field of AI and robotics, giving rise to a new paradigm known as "nouvelle AI" [28]. Indeed, in the 1980s, Brooks charted a course for redefining the approach to building artificial machines that could exhibit human-like intelligence and behavior through embodiment, which had profound implications for the subsequent development of ideas in cognitive sciences and philosophy of mind [29]. His proposal emphasized the necessity of constructing

complete agents that operate in dynamic environments using real sensors in order to truly test the concepts of intelligence [28]. This required moving beyond the symbolic artificial intelligence model that had characterized previous approaches. Consequently, a new framework emerged, in which intelligence was no longer viewed solely in terms of accomplishment of isolated tasks but primarily in terms of interaction with the environment. The central idea was that "the true details of interacting with the world are not the same as abstract thinking has led many workers in Artificial Intelligence to believe" ([30]: 1). Brooks' work significantly propelled the field of robotics, which had previously progressed at a slower pace compared to disembodied intelligent systems. This breakthrough yielded numerous advancements, leading to the development of robots capable of interacting with both the environment and humans in a social manner, such as Kismet [31]. The most notable transformation brought about by this new paradigm in AI was the abandonment of the "human thinking" approach in favor of a perspective in which intelligence arises from more fundamental processes involving the interaction between an organism's body and its environment.

This reconfiguration of the concept of intelligence represents a significant model shift: intelligence shifts from being perceived as demonstrable solely through the correct execution of tasks to a conception wherein the foundational aspect of humanoid intelligence lies in how an agent interacts with its world. Most importantly, the work on robotics emphasized that the pursuit of disembodied tasks and the singular focus on developing computational capabilities, detached from immersion in an environment, were insufficient for achieving AI on par with human intelligence. However, the shift in focus from task performance to interaction with the world has presented challenges in evaluation, particularly when human agents are involved. Indeed, in the context of interactive AI, especially embodied AI over the past three decades, evaluation metrics have become more nuanced and now consider elements that were previously deemed irrelevant. As highlighted by Minsky, while the expectations of AI performance within the specialized field of computer science influenced internal considerations, *the desire to deploy robots in real-world settings and have them interact with human users without prior knowledge of their inner workings has brought the issue of expectations to the forefront.* Regardless of the level of accuracy with which robots can solve various tasks, even surpassing human capabilities, the perceived level of intelligence in robots depends on multiple variables, often rooted in embodied communication and extending beyond the mere explicit accomplishment of the task at hand.

5 Expected intelligence for robots in interactive contexts

One of the main factors that has been demonstrated to influence human expectations in interactions with robots is the degree of human-likeness in the robot's appearance. Extensive research has shown that when robots exhibit a high degree of human-likeness, it elicits anthropomorphic tendencies in humans, leading them to engage with the robots in a manner that resembles social interactions with other humans [32][33]. However, this anthropomorphic tendency can be seen as a double-edged sword. While it promotes social engagement, it also creates a potential for misleading expectations regarding the cognitive capabilities of the robots.
In the field of Human-Robot Interaction (HRI), this challenge of understanding the capabilities of robotic agents in interactive contexts has been identified as the Perceptual Belief Problem (PBP). As highlighted by Thellman and Ziemke [34], the PBP appears to be unique to the field of HRI and does not arise in standard Human-Computer Interaction. Unlike computers, social robots have physical bodies that enable them to navigate and engage with a complex and unstructured physical world, and at the same time bodies make robots able to physically interact with humans. The PBP in HRI arises because humans, when interacting with social robots, need to assess the robots' abilities and limitations based on their perceptual and cognitive capabilities. Since social robots often exhibit human-like appearances and behaviors, humans may attribute human-like cognitive capacities to them, creating potential misconceptions about the robots' actual cognitive abilities. This anthropomorphic tendency can lead to the *expectation that the robot possesses comprehensive perceptual beliefs similar to those of humans.* However, fulfilling such expectations is challenging given the current technology and robot architectures. Although a humanoid robot may have human-like eyes, ears, arms, and legs, it does not necessarily imply that it can see, hear, grasp objects, shake hands, or walk side by side in the same way as a human would in interaction. Thus, when confronted with a humanoid robot, human agents may be disappointed to realize that the robot's capabilities are limited. Moreover, *these expectations are difficult to measure* because assumptions about the actual cognitive and behavioral capabilities of robots are not always transparent to individuals. For instance, even if people treat the robot as an intentional agent during interaction, when asked whether they believe the robot has a mind, they typically respond in the negative [35]. Consequently, new methodologies have been developed to evaluate the PBP without relying solely on verbal self-assessment [36]. Finding a way

to address the gap posed by the PBP is crucial to ensure a high degree of collaboration and *to prevent disappointment in humans.*
A high degree of human likeness can also have negative effects on interaction when combined with a low level of affinity toward the robot. This phenomenon is known as the Uncanny Valley Effect [37], wherein a negative feeling arises in human agents interacting with robots that closely resemble humans but lack the same level of human-like behavior. Various explanations have been proposed to account for this effect, but consensus has not yet been reached. One of the most widely accepted explanations [38] attributes the effect to a mismatch between human agents' expectations, influenced by the humanoid appearance, and the actual behavior exhibited by the robot in interaction. Recently, this explanation has also been examined through the lens of the theory of predictive coding [39][40], which posits that the fundamental functional structure of the brain, across all levels of organization, involves the comparison of observations with predictions and strives to operate in a way that minimizes any discrepancies between them.
Not only do humanlikeness and anthropomorphism play a crucial role in shaping expectations regarding robot intelligence in interaction, but the cultural background and representations of robots in cultural settings also significantly impact human expectations. As Kamide and Mori [41] emphasize, culture and philosophy have a profound influence in the context of Human-Robot Interaction (HRI). In their study, they compare the philosophical systems of the West (e.g., Europe and the Americas) to those of the East (e.g., Asia and the Middle East). The Western tradition seeks a systematic, consistent, and comprehensive understanding of the universe, while the Eastern tradition adopts a more holistic or circular view of the world. These differences in philosophical orientations can lead to varying attitudes towards robots, suggesting that cultural and philosophical leanings may foster greater readiness for acceptance among Eastern populations, as also suggested by MacDorman et al. [42] in the context of Japanese culture. Indeed, cultural narratives in which individuals engage from childhood significantly influence the acceptance of robots in society. In Japanese culture, for example, where robots have been portrayed as companions and allies in manga adventures since the 70s, the majority of people tend to exhibit a positive outlook towards the integration of robots engaging in daily activities in close proximity to the general public [43]. Conversely, individuals from Western cultures often harbor more negative attitudes regarding the inclusion of robotic companions for assistive and care tasks, influenced by apocalyptic narratives found in cinema and literature, such as the works of Asimov or the Terminator movies [44][45]. On the same topic, Horstmann

and Krämer [46] conducted a study revealing that individuals' exposure to media portrayals of social robots significantly influences their expectations of robots' abilities and capabilities, subsequently reinforcing and amplifying those expectations. Furthermore, the study found that individuals' knowledge of negatively depicted fictional social robots contributes to the development of negative expectations, perceiving robots as potential threats. Conversely, individuals who possess a greater understanding of the capacities and limitations of robot technology, based on non-fictional knowledge, exhibit reduced levels of anxiety towards robots. In a follow-up study [47], they examined how a negative violation of expectations caused by a social robot, coupled with the valence of the subsequent reward, could influence participants' desirability of interacting with a social robot and its impact on HRI. Interestingly, the study revealed that when the robot violated participants' expectations, they evaluated the robot's competence, sociability, and interaction skills more negatively [1].
Therefore, the framework for evaluating robotic intelligence needs to account for these expectations, both prior to actual interactions and during online interactions with robots. In the following paragraph, we will explore how incorporating these new insights can further enhance the understanding and assessment of expectations in HRI. Although no study, to our knowledge, specifically investigates the role of these expectations in influencing the evaluation of robot intelligence, certain tests have been developed to measure perceived intelligence in robot interactions, along with related metrics that facilitate the overall assessment of HRI quality.

6 Measuring robotic intelligence

In the interactive context, the measurement of a robot's intelligence is an assessment of how effectively a human and a robot collaborate [50]. As discussed in the previous paragraph, human expectations regarding intelligence are not only relevant in terms of preconceived beliefs and assumptions, but also during the actual interaction with the robot. Negative attitudes, in particular, can hinder the potential benefits of HRI, such as therapeutic practices.
To evaluate these assumptions and attitudes towards robots, Nomura et al.

[1]The influence of cultural background on expectations towards robots is an extensively explored aspect in the Human-Robot Interaction literature. Delving deeper into this topic is beyond the scope of the present chapter; see [48; 49] for a general overview.

[51] developed the Negative Attitude towards Robots Scale (NARS). This scale consists of a questionnaire that explores various aspects of human-robot interaction, including situations, social influence, and emotions. In their experimental study, participants engaged with Robomovie, a robot designed for human communication [52], and they were asked to interact verbally with the robot and to touch it. The findings revealed that negative attitudes towards robots can indeed impact the interaction. Some individuals displayed resistance to touching or speaking with the robot, which undermined and compromised the quality of interaction. Additionally, the study showed that these attitudes can vary depending on the participants' gender and culture. Furthermore, previous encounters with robots were found to influence the results of the NARS test. Participants with prior direct experience with robots were more inclined to engage in verbal interactions and touch the robot during the study. Subsequent studies examined the efficacy of the NARS scale in assessing changes in expectations during actual interactions [53] and explored cultural differences among participants [54].
Other assessment methods, often used in conjunction with the NARS test, have been developed to evaluate the overall quality of HRI after the initial validation stage. One widely employed approach is the Godspeed test [55], which aims to ensure comparability across different experimental settings and the replicability of results in various languages and contexts. This test encompasses several factors, including anthropomorphism, animacy, likeability, perceived safety, and perceived intelligence, all of which are assessed based on participants' experiences during the interaction. Of particular relevance to our study is the assessment of perceived intelligence. The researchers conceptualize perceived intelligence in robots as a combination of robotic competence and the potential elicitation of the perception of intelligence through "random behavior" during interaction. Participants were asked to rate their perceived intelligence on a scale ranging from 1 to 5 of different dichotomies, i.e., Incompetent/Competent, Ignorant/Knowledgeable, Irresponsible/Responsible, Unintelligent/Intelligent, and Foolish/Sensible. Although the test has undergone statistical validation, it is important to highlight two key considerations.
Firstly, attempting to capture perceived intelligence through strict dichotomies, as described above, may not provide a comprehensive assessment of quality. Regarding perceived intelligence, certain terms in the Godspeed test appear to be redundant (Incompetent, Ignorant, Unintelligent) and may not be easily discernible. Other terms (Irresponsible/Responsible, Foolish/Sensible) seem unrelated to the evaluation of perceived intelligence but are more relevant to ethical considerations. Dichotomies such as Intelligent/Unintel-

ligent and Ignorant/Knowledgeable align better with the general concept of perceived intelligence, as they reflect stereotypical conceptions of intelligence. This makes us think that this list can be restricted or enlarged at will, also because the authors did not give any explanation for having chosen these terms and no others. The second issue to emphasize is that the Godspeed test alone may not adequately capture the overall quality of HRI. As previously mentioned, the embodied factor plays a crucial role in robot interactions and is not considered in this test. Therefore, the integration of implicit measures of perceived intelligence, which assess the bodily attitudes people exhibit during robot interactions, could enhance the evaluation process. Such measures have been adopted in various experimental settings [56][57] and would contribute to a more comprehensive framework for assessing interactions while also enhancing internal test validity.
In relation to the impact of expectations on human-robot interaction, Rosén et al. [58] developed an evaluation framework aimed at providing a comprehensive understanding of the social robot expectation gap. This term describes a disparity in which expectations, whether excessively high or too low, can result in disconfirmation. The authors illustrate this phenomenon by considering two scenarios: firstly, when individuals interact with a social robot and hold the belief, influenced by depictions in science fiction movies, that the robot could feel and express pain when it falls, yet it does not, leading to a contradiction of expectations and the emergence of high expectations; secondly, when users do not assume the robot to be able to engage in verbal communication, but it initiates a conversation, leading to a contradiction of expectations and the development of low expectations. Consequently, for the authors it is possible to say that the quality of interaction can be perceived as high, irrespective of the robot's capabilities, as long as expectations are confirmed. To assess the implications of high and low expectations on social robots prior to, during, and after interaction, the authors proposed an evaluation framework encompassing affect, cognitive processing, and behavior and expectations. Moreover, they highlighted that expectations are dynamic and change over time - in line with our previous proposal - but they do not provide experimental evidence for the effectiveness of this evaluation model.
However, it is crucial to develop a means of evaluating robotic intelligence in order to facilitate suitable human-robot interactions and design effective robot behaviors. To gain a comprehensive understanding of the extent to which robots can exhibit intelligence, Winfield [59] proposed a framework in which robotic intelligence arises from the integration and interaction of four distinct types: morphological intelligence, swarm intelligence, individ-

ual intelligence, and social intelligence. Using a star diagram, the author compared different organisms based on their exhibited intelligence. For example, humans may demonstrate high levels of individual and social intelligence, surpassing other animal species, but exhibit a lower degree of swarm intelligence, in contrast to certain animal species like ants. This framework can also be applied to robots, allowing for an analysis of their diverse forms of intelligence. Winfield further suggested that one possible reason for human disappointment in terms of robotic intelligence is that "none of the intelligence graphs for the robots score on more than two axes, whereas all of the exemplar animals score on at least three" ([59]: 6). This highlights a distinction between the types of intelligence exhibited by living beings and those demonstrated by artificial artifacts.

Despite the fundamental difference that Whitfield's conclusions highlight between living beings and artificial artifacts, the disparity in terms of the intelligence exhibited by these entities appears to be once again focused on restricted aspects of intelligence. In the majority of cases, intelligence is conceptualized as a quality demonstrated through abstract reasoning for specific tasks and subsequently evaluated by comparing it to the best possible outcomes, which are ultimately determined by how humans accomplish such tasks.

These examples taken from HRI show how the problem of expected intelligence is widely present in interactive robotics, a field in which intelligence can be assessed on multiple dimensions including physical ones (performance in specific tasks, acting in the real world, verbal and non-verbal interactions with human users, etc.). Every test shows some of the characteristics that we have included in our general list, although not all at the same time. However, it should appear evident, by means of the HRI, as a general tool for assessing expected intelligence and for drawing up a shared and aware metric, is a theoretical and a practical problem that can no longer be postponed.

7 Conclusions

In conclusion, two final considerations can be made. The first pertains to the often inadequate consideration of expectations in most tests aimed at assessing or evaluating intelligence in artificial artifacts. These tests often fail to fully acknowledge the role of expectations held by both human users and scientists regarding their overall understanding of the interaction and the task at hand. While expectations may sometimes be examined from the perspectives of anthropomorphism or cultural background, the assess-

ments rarely, if ever, take into account the variety of expectations that users may have when engaging with artificial artifacts. Furthermore, in the case of robots, apart from the factors we have previously discussed as influencing expectations, issues concerning the attribution of intentionality [57] and perceived agency [60] are also considered to be influential factors that affect the actual interaction with robots and consequently impact the evaluation of their intelligence.
At the same time, these attempts for assessing and evaluating intelligence strive to break it down into subsections, thereby seeking to simplify the problem. Indeed, mainstream methods separate different abilities and evaluate them individually, subsequently aggregating the singular evaluations. Consequently, the overall evaluation of the artificial artifact's intelligence is understood as the sum of the separate evaluations of its performances, whether accomplished or not. As emphasized by Anzalone and colleagues [61], the challenge in evaluating AI lies not in reducing it to the efficacy of individual algorithms, but in offering a more comprehensive account that encompasses broader forms of intelligence, namely social and interactive intelligence [62]. Therefore, what we propose is to complement this purely methodological evaluation approach with aspects derived directly from interaction studies, in order to facilitate a more robust evaluation that takes into consideration various factors that we have observed to contribute to the assessment of a robot's intelligence to varying extents, such as expectations and the overall quality of interaction.
Finally, if the conception of intelligence that we wish to apply to artificial artifacts is derived from or must be comparable to human intelligence, it is imperative to carefully consider the precise basis on which we are aligning them. The response to the question "how can an artificial artifact be deemed intelligent?" will always be influenced by the particular concept of intelligence that is embraced.
It is not possible to say whether a final answer will ever be given on what intelligence is and on what intelligence is from the point of view of the artificial. However, we can consider it reasonably certain that the attribution of intelligence by the human user is an unavoidable characteristic, both from a cognitive and a practical point of view. This also has positive motivations and implications, as we have tried to argue. The proposal we have made is that taking into account the intelligence expected in artificial artifacts and attempting to measure it through metrics based on the various aspects of user interaction or use can lead to better results in understanding the reality of the artificial and, at the same time, to better designs of artificial artifacts. And perhaps to a more well-structured answer to the question of intelligence

in AI. The list of characteristics proposed to measure expected intelligence is in line, as we have tried to show, with the measurement attempts made in social robotics, a field that seems to exemplify more than others in AI the strong implications that the question on the intelligence brings with it, constituting an area of investigation whose results can also be extended to the interaction and use with other AI systems which are more distant from an embodied configuration.

References

[1] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59, 433–460, (reprinted in J. Copeland (ed.), *The essential Turing*, Oxford University Press, 2004, pp. 441–464).

[2] Moor, J. H. (ed.) (2003), *The Turing Test. The Elusive Standard of Artificial Intelligence*, Dordrecht, Springer. https://doi.org/10.1007/978-94-010-0105-2

[3] Esping, A., Plucker, J.A. (2015), Alfred Binet and the Children of Paris, in S. Goldstein, D. Princiotta, J. A. Naglieri (eds.), *Handbook of Intelligence, Evolutionary Theory, Historical Perspective and Current Concepts*, Springer, pp. 153-161.

[4] Chollet, F. (2019). On the Measure of Intelligence. *arXiv*:1911.01547v2, https://doi.org/10.48550/arXiv.1911.01547

[5] Newell, A., Shaw, J.C., Simon, H.A. (1959). Report on a general problem-solving program. https://exhibits.stanford.edu/feigenbaum/catalog/sy501xd1313

[6] Lieto, A. (2021), Cognitive Design for Artificial Minds. London, UK: Routledge, Taylor & Francis.

[7] Saeed, W., Omlin, C. (2023), Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, *Knowledge-Based Systems*, 263, 110273, https://doi.org/10.1016/j.knosys.2023.110273.

[8] Hernández-Orallo, J. (2017a), Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, *Artificial Intelligence Review*, 48, pp. 397–447 DOI 10.1007/s10462-016-9505-7

[9] Bringsjord, S, (2011), Psychometric artificial intelligence, *Journal of Experimental and Theoretical Artificial Intelligence*, 23, pp. 271–277.

[10] Li, M, Vitányi, P. (2008), *An introduction to Kolmogorov complexity and its applications*, Springer, New York.

[11] Hernández-Orallo, J., Dowe, D. L. (2010), Measuring universal intelligence: Towards an anytime intelligence test, *Artificial Intelligence*, 174, pp. 1508–1539.

[12] Hernández-Orallo, J, (2017b), *The Measure of all Minds. Evaluating Natural and Artificial Intelligence*, Cambridge University Press, New York.

[13] Hernández-Orallo, J., Loe, B.S., Cheke, L., Martínez-Plumed, F., Ó Éigeartaigh, S. (2021), General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific Report*, 11, 22822. https://doi.org/10.1038/s41598-021-01997-7

[14] Floridi, L. (2023),*The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities*, Oxford University Press, Oxford.

[15] Cristianini, N., Scantamburlo, T. and Ladyman, J. (2023). The social turn of artificial intelligence. *AI & Society*, 38, pp. 89–96. https://doi.org/10.1007/s00146-021-01289-8

[16] Dennett, D. C. (1987). *The intentional stance*. Cambridge (MA): The MIT Press.

[17] Cordeschi, R. (2002). *The discovery of the Artificial. Behavior, Mind and Machines before and beyond Cybernetic*s. Dordrecht/Boston/London: Kluwer Academic Publishers.

[18] Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*. Cambridge (MA): Oxford University Press.

[19] Russell, S.J, and Norvig, P.R. (1995) *Artificial Intelligence: A Modern Approach*. London: Pearson.

[20] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., and Rahmim, A. (2021). A Brief History of AI: How to Prevent Another Winter (A Critical Review). *PET clinics*, 16(4), 449–469. https://doi.org/10.1016/j.cpet.2021.07.001

[21] Dennett, D. (1984). Cognitive wheels: the frame problem of AI. *Minds, Machines and Evolution*, 129-151.

[22] Newell, A. (1980) Physical symbol systems. *Cognitive Science*, 4, 135-183.

[23] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.

[24] Minsky, M. L. (1982). Why People Think Computers Can't. *AI Magazine*, 3(4), 3. https://doi.org/10.1609/aimag.v3i4.376

[25] Dreyfus, H. L. (1972). *What computers can't do: The limits of artificial intelligence*. New York: MIT Press.

[26] Shananan, M. (2004) The frame problem. The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), URL <https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>.

[27] Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42: 335-346.

[28] Brooks, R. (1991) Intelligence without Reason. *Proceedings of the 12th international joint conference on Artificial intelligence* - Volume 1 (IJCAI'91). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 569–595.

[29] Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. Cambridge (MA): MIT Press.

[30] Brooks, R. (1989). The whole iguana. *Robotics science*, 432-456.

[31] Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge (MA): MIT Press.

[32] Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, 9, 2136.

[33] Damiano, L., and Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9, 468.

[34] Thellman, S. and Ziemke, T. (2021). The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics. *Trans. Hum.-Robot Interact.* 10, 3, Article 29 (July 2021), 15 pages. https://doi.org/10.1145/3461781

[35] Fussell, S., Kiesler, S., Setlock, L., and Yew, V. (2008). How people anthropomorphize robots. In Proceedings of the 3rd ACM/IEEE international conference on Human-robot interaction (HRI '08). Association for Computing Machinery, New York, NY, USA, 145–152. https://doi.org/10.1145/1349822.1349842

[36] Thellman, S. and Ziemke, T. (2020). Do You See what I See? Tracking the Perceptual Beliefs of Robots. iScience, Volume 23, Issue 10, 2020, 101625, https://doi.org/10.1016/j.isci.2020.101625

[37] Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, 7, 33-35.

[38] Kätsyri, J., Förger, K., Mäkäräinen, M., and 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, 390.

[39] Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, 7(4), 413-422.

[40] Urgen, B. A., Li, A. X., Berka, C., Kutas, M., Ishiguro, H., and Saygin, A. P. (2015). Predictive coding and the Uncanny Valley hypothesis: Evidence from electrical brain activity. *Cognition: a bridge between robotics and interaction*, 15-21.

[41] Kamide, H., and Mori, M. (2016) One being for two origins - a necessary awakening for the future of robotics. In 2016 *IEEE workshop on advanced robotics and its social impacts (ARSO)*. IEEE, Piscataway, NJ.

[42] MacDorman, K.F., Vasudevan, S.K., Ho, C-C. (2009). Does Japan really have robot mania? Comparing attitude by implicit and explicit measures. *AI & Society*, 23: 485-510.

[43] Han, J., Hyun, E., Kim, M., Cho, H., Kanda, T., and Nomura, T. (2009). The Cross-cultural Acceptance of Tutoring Robots with Augmented Reality Services. *J. Digit. Content Technol. its Appl.*, 3, 95-102.

[44] Haring, K.S., Mougenot, S., Ono, F., Watanabe, K. (2014). Cultural Differences in Perception and Attitude towards Robots. *International Journal of Affective Engineering*, 2014, Volume 13, Issue 3, Pages 149-157, https://doi.org/10.5057/ijae.13.149

[45] Dumouchel, P., and Damiano, L. (2017). *Living with robots.* Cambridge (MA): Harvard University Press.

[46] Horstmann, A. C., and Krämer, N. C. (2019). Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology*, 10, 939.

[47] Horstmann, A. C., and Krämer, N. C. (2020). Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations. *PloS one,* 15(8), e0238133.

[48] Papadopoulos, I., and Koulouglioti, C. (2018). The influence of culture on attitudes towards humanoid and animal-like robots: An Integrative Review. Journal of Nursing Scholarship, 50(6), 653-665.

[49] Lim, V., Rooksby, M., and Cross, E. S. (2021). Social robots on a global stage: establishing a role for culture during human–robot interaction. International Journal of Social Robotics, 13(6), 1307-1333.

[50] Crandall, J. W., and Goodrich, M. A. (2003). Measuring the intelligence of a robot and its interface. In *NIST's Performance Metrics for Intelligent Systems Workshop*, Arlington, VA, 2003.

[51] Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006a). Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 7(3), 437-454.

[52] Ishiguro, H., Ono, T., Imai, M., and Kanda, T. (2003). Development of an interactive humanoid robot "Robovie"—an interdisciplinary approach. In *Robotics Research: The Tenth International Symposium* (pp. 179-191). Springer Berlin Heidelberg.

[53] Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE transactions on robotics*, 24(2), 442-451.

[54] Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2005). Cultural Differences in Attitudes Towards Robots. Companions: Hard Problems and Open Challenges in Robot-Human Interaction, 1.

[55] Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1, 71-81.

[56] Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. (2013). Robots can be perceived as goal-oriented agents. *Interaction Studies*, 14(3), 329-350.

[57] Thellman, S., and Ziemke, T. (2019) The Intentional Stance Toward Robots: Conceptual and Methodological Considerations. *CogSci'19. Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 1097–1103.

[58] Rosén, J., Lindblom, J., and Billing, E. (2022). The Social Robot Expectation Gap Evaluation Framework. In *International Conference on Human-Computer Interaction* (pp. 590-610). Cham: Springer International Publishing.

[59] Winfield, A. F. (2017). How intelligent is your intelligent robot?. *arXiv preprint arXiv*:1712.08878.

[60] van der Woerdt, S., and Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, 54, 93-100.

[61] Anzalone, S. M., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7, 465-478.

[62] Barchard, K. A., Lapping-Carr, L., Westfall, R. S., Fink-Armold, A., Banisetty, S. B., and Feil-Seifer, D. (2020). Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(4), 1-29.