AI

MDPI

Article

Shifting Perspectives on AI Evaluation: The Increasing Role of Ethics in Cooperation

Enrico Barbierato * and Maria Enrica Zamponi

Department of Mathematics and Physics, Catholic University of Sacred Heart, 25133 Brescia, Italy; mariaenrica.zamponi01@icatt.it

* Correspondence: enrico.barbierato@unicatt.it

Abstract: Evaluating AI is a challenging task, as it requires an operative definition of intelligence and the metrics to quantify it, including amongst other factors economic drivers, depending on specific domains. From the viewpoint of AI basic research, the ability to play a game against a human has historically been adopted as a criterion of evaluation, as competition can be characterized by an algorithmic approach. Starting from the end of the 1990s, the deployment of sophisticated hardware identified a significant improvement in the ability of a machine to play and win popular games. In spite of the spectacular victory of IBM's Deep Blue over Garry Kasparov, many objections still remain. This is due to the fact that it is not clear how this result can be applied to solve real-world problems or simulate human abilities, e.g., common sense, and also exhibit a form of generalized AI. An evaluation based uniquely on the capacity of playing games, even when enriched by the capability of learning complex rules without any human supervision, is bound to be unsatisfactory. As the internet has dramatically changed the cultural habits and social interaction of users, who continuously exchange information with intelligent agents, it is quite natural to consider cooperation as the next step in AI software evaluation. Although this concept has already been explored in the scientific literature in the fields of economics and mathematics, its consideration in AI is relatively recent and generally covers the study of cooperation between agents. This paper focuses on more complex problems involving heterogeneity (specifically, the cooperation between humans and software agents, or even robots), which are investigated by taking into account ethical issues occurring during attempts to achieve a common goal shared by both parties, with a possible result of either conflict or stalemate. The contribution of this research consists in identifying those factors (trust, autonomy, and cooperative learning) on which to base ethical guidelines in agent software programming, making cooperation a more suitable benchmark for AI applications.

Keywords: AI; evaluation; game; cooperation; trust; autonomy

Citation: Barbierato, E.; Zamponi, M.E. Shifting Perspectives on AI Evaluation: The Increasing Role of Ethics in Cooperation. *AI* **2022**, *3*, 331–352. https://doi.org/10.3390/ai3020021

Academic Editors: Pablo Rivas, Gissella Bejarano and Javier Orduz

Received: 7 March 2022
Accepted: 13 April 2022
Published: 19 April 2022

1. Introduction

This introduction sums up the most notable attempts to define intelligence and the methodological frameworks to measure it (see Table 1).

The definition of intelligence is controversial, as it has been debated from different perspectives without necessarily reaching an agreement. In the Middle Ages, clerics and theologists studied different qualities of thought, such as memory, abstraction, and imagination (https://thepsychologist.bps.org.uk/volume-29/november-2016/looking-back-medieval-mind, accessed on 6 March 2022), in an attempt to define the human mind. Memory covered a different role, as it allowed theologists to conceive complex interpretations of the holy texts: not just a static device, but an enormous, sophisticated machine able to recreate instantaneously places in time. The capacity to recollect memories required the development of sophisticated techniques such as the Ars Magna (a method invented by the Catalan Ramon Llull (1232–1316) and described mostly by pictures and schemas, meant to be used as a mechanical process to pursue the truth in every field of

knowledge. Llull's works had a deep influence on philosophers such as Leibniz and, more generally, on artificial intelligence. See for example Fidora et al. [1]).

Table 1. A summary of methodological frameworks defining intelligence.

Person/Group/Society	Proposed Definition of Intelligence	Time
Ramon Llull	*Ars Magna*, intelligence emerges from learning based on mnemonic skills	1305–1308
Renè Descartes	Mind and body are separated. Intelligence in animals is unlikely	1641
Enlightenment	Rational thought (i.e., Newton's Principia Mathematica)	1685
American Psychological Association	Intelligence can emerge in different ways, and it should be measured according to different techniques	1892
Alan Turing	Intelligence as a mechanical process. The imitation game	1950
Dartmouth manifesto	Mind as software executed in the brain	1956
Edward Feigenbaum	Capability to perform inferences on a set of rules and a knowledge base	1965
Jobin et al.	Rational behavior associated with moral principles	2019

While according to Middle Ages thinkers the human mind was embodied, the French philosopher Descartes, a thinker of the Modern Era, thought the opposite, separating the spiritual from the material essence. Furthermore, in the XVII century, the Enlightenment promoted intelligence as a means to bring light, in the darkness of ignorance, and maturity, as a sign of independence of thought and courage (I. Kant, "What is Enlightenment?", 1784). In his book *The Critique of Pure Reason*, Kant discussed some of the principles of theories of mind establishing de facto the foundations of psychology and cognitive sciences. The American Psychological Association provides an interesting definition of intelligence (https://www.apa.org/topics/intelligence, (accessed on 6 March 2022) "Intelligence refers to intellectual functioning. Intelligence quotients, or IQ tests, compare your performance with other people your age who take the same test. These tests don't measure all kinds of intelligence, however. For example, such tests can't identify differences in social intelligence, the expertise people bring to their interactions with others. There are also generational differences in the population as a whole. Better nutrition, more education and other factors have resulted in IQ improvements for each generation."), which is focused on one side on its measurement through tests, and on the other on admitting that a quantitative assessment is per se not possible: such a contradiction is however intellectually stimulating, as it motivates researchers to deepen studies around human intelligence.

Artificial intelligence (AI) stems from a group of American researchers. The Dartmouth manifesto (http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf, accessed on 6 March 2022), written by J. McCarthy, M. L. Minsky, N. Rochester, and C.E. Shannon proposed a list of objectives aiming to prove that software could solve problems of different natures, simulating the way the human brain works. In a few years, the task proved to be much more complex than initially thought, although a second issue emerged, consisting of the difficulty of assessing a procedure to prove that the re-enacting of a cognitive task was fulfilled or not. For example, one of the points in Dartmouth concerned the role of casualty in creativity: evaluating if this task has been achieved is, in itself, another task, even more complex than the previous one. The problem of evaluating an AI software requires in the first place the definition of a set of metrics, which are required to meet a few criteria: (i) non-ambiguity (i.e., they must be interpreted in one way only); (ii) transparency (they must be clearly identifiable within an AI software architecture); (iii) sharing (they must be

validated and accepted by the AI community); and (iv) commensurability (they must be comparable to those used to measure human intelligence).

Historically, one of the ways to assess the capacity of an AI software to mimic the human mind was by testing its capacity to play games—mainly structured into deterministic rules—against humans. However, other means of evaluation were possible, for example, chatbots, i.e., software capable of performing natural language processing and establishing a conversation with a human; problem solvers, i.e., applications able to separate the data from the logic programming and—in principle—working on any kind of problem, assuming that it can be described according to well-formulated formula; finally, software used to exhibit some kind of creativity, such as the generation of metaphors, the composition of poetry (sometimes in the form of limericks (https://artificial-intelligence.leeds.ac.uk/limericks-submitted-to-the-microsoft-limerick-competition-at-aaai-20/ (accessed on 6 March 2022) is an example of submitted composition to the Microsoft Research ran a limerick competition at their stand at AAAI-20 in New York, February 2020.)) or even puns.

However, these are not the only tests proposed to evaluate potential artificial general intelligence. The most famous one is the Turing Test [2], in which the ability of a machine to be (mistakenly) recognized as human is the assessment criterion. As this is determined by human judges, the evaluation could be considered biased by subjectivity. Additionally, Steve Wozniak, the Apple co-founder, has hypothesized that the ability of a computer to go into an unknown kitchen and figure out how to make a cup of coffee could be an indicator of general intelligence as well. Other benchmarks relate to the capability of achieving a college degree [3] or learning and performing jobs ordinarily done by humans [4].

The marketplace turned its attention to AI in the 1980s, when the market for both AI-related software and hardware grew up to USD 425 million (http://world-information.org/wio/infostructure/100437611663/100438659445, accessed on 6 March 2022). Machine vision (a branch of AI) only, rose in value to USD 80 million. In more recent times, Datamin, a company developing AI software, established the mark of USD 392 million, and Pony.ai, working on self-driving units, assessed its growth up to USD 102 million (https://towardsdatascience.com/the-secrets-of-successful-ai-startups-whos-making-money-in-ai-part-ii-207fea92a8d5, accessed on 6 March 2022). However, AI commercialization and value in terms of trust and reliability collapsed during two time periods (1974–1980 and 1987–1993). In 1973, the Lighthill report ("Artificial Intelligence: A General Survey" James Lighthill: in Artificial Intelligence: a paper symposium, Science Research Council) reported that "In no part of the field have the discoveries made so far produced the major impact that was then promised". One of the most promising AI paradigms, the *expert system*, resulted in a type of software that was very difficult and expensive to maintain. In 1981, the Fifth Generation Computer program was launched in Japan with a budget of USD 850 million without reaching any of the objectives targeted.

In the light of the successes and the failures of AI, the necessity emerges to establish those criteria that can guide the development of AI software in terms of commercialization. The ability to play games or even learn the rules of a game without supervision might not be fully relevant in this context, favoring instead paradigms based on cooperation. Besides modeling social interactions with closer detail and realism with respect to the ability to play games, cooperation raises interesting aspects concerning trust, autonomy, and possible conflicts between the involved parties when they share the same goal. As these topics inevitably involve moral concepts and ethical guidelines, we argue that AI software programming must include ethics when cooperation is involved.

In general, ethics in AI is a topic that has attracted a number of researchers offering an interdisciplinary perspective: starting with a philosophical standpoint, where it is agreed that a machine should denote moral principles in its behavior. For example, according to Kušić [5], there is a difference between humans exploiting technology and the ethics of machines. While the former considers AI as a neutral tool, the latter suggests that a moral agent possesses a moral value shifting between rights and responsibilities. Jobin et al. [6] reviewed 84 articles regarding ethical guidelines concerning AI, highlighting eleven

principles, including privacy and fairness (another interesting topic concerns sustainability, a concept that can be declined in different ways, from the study of techniques to reduce the environmental cost due to complex computer hardware (or even data storage), to the claim of Halsband regarding intergenerational equity ([7]), which takes into account the bond of preserving what is considered of paramount importance by future generations). Privacy has attracted significant attention in recent years due to the increasing circulation of large amounts of information regarding the individual due to social media and invasive forms of advertising. More specific concerns such as the theft of data over the internet require a strong cybersecurity framework, as pointed out by Khosravy et al. [8,9] (both the papers refer to a model inversion attack (MIA), a technique exploited by a malicious user meant to acquire the dataset used to train an AI model, for example, to attack a facial recognition system). In contrast, the concept of fairness concerns scenarios where, for example, an AI model has processed data about facial recognition, loan requests, or criminal records, providing openly biased classifications towards specific ethnic and gender groups.

The original contribution of this work consists of (i) revisiting AI's capability to play and even learn the rules of games (such as chess, checkers, and Atari video games) against humans; (ii) assessing the effectiveness of this approach as a means of evaluating AI software in terms of commercialization; (iii) characterizing the advantages and shortcomings of deploying agents acting in cooperation (instead of competing) with humans; (iv) establishing a more robust benchmark for ethical AI based on the definition of three pillars (trust, autonomy, and cooperative learning).

The article is organized as follows: Section 2 presents an overview of related work with respect to the evaluation of intelligence and AI; Section 3 discusses the literature about games as a benchmark for AI; Section 4 focuses on AI evaluation through cooperation, as a paradigm contraposed to gaming; Section 5 presents several thought experiments centered on the cooperation between humans and software agents, followed by definitions of the ethical factors related to this form of interaction. Finally, Section 6 provides the conclusions and future work.

2. Definition of the Problem

The evaluation of intelligence is a difficult problem, originating from the complexity that derives from providing a unique definition. The problem can be partially loosened by considering a restricted (and simplified) domain of AI, nevertheless, this choice results in a variety of perspectives, as shown in Table 2.

Table 2. Examples of AI evaluation criteria.

Origin	Proposed Evaluation Criteria
Dartmouth School	Not clearly defined
Meta-Rationality	Hard to achieve because of the incommensurability of the different disciplines
IEEE Spectrum	Review classic measures (such as accuracy) in the light of baselines
Forbes	Data connectors, flexibility, ease-of-use, and ethics vs bias
GoDataDriven	Analytical capability (data, people, and technology) and business adoption (executive support, funding and implementation)
Kang and C. T. Haas	Qualitative criteria
Ish et al.	Accuracy of the classification and average time spent on the datapoint

Meta-Rationality (https://metarationality.com/artificial-intelligence-progress, accessed on 6 March 2022) suggests an interesting metaphor for AI, which is regarded as a *wolpertinger*, an imaginary mammal with parts of other animals. Such is AI, which conveys—from an epistemological perspective—many theoretical aspects derived from mathematics, science, engineering, and other disciplines, making it very difficult to measure the different

evaluation criteria owned by each discipline. Because of its multidisciplinary character, this problem cannot be put aside. The author highlights the enthusiasm in the last century around the so-called good old-fashioned AI (GOFAI), a collection of paradigms that proved to be wrong for many reasons, ultimately because of the lack of a coherent assessment framework. However, the situation does not seem to have changed, since the classical epistemological path consisting of (i) hypothesis formulation and (ii) validation by experiment does not seem to be necessarily followed by AI. Another issue consists of the impossibility of explaining the results produced by an AI application (a typical issue of neural networks, which is often considered a black box). On the other hand, the classic evaluation criteria used in mathematics (i.e., asymptotic convergence of a method) are not really applicable in AI.

Special awareness needs to be paid to hypes, as denoted by IEEE Spectrum (https://spectrum.ieee.org/learn-the-red-flags-of-overhyped-ai-claims, accessed on 6 March 2022), especially when a mix of buzzwords (such as AI, blockchain and Internet of Things are mixed up) are poorly defined. Some of the AI claims can be verified by checking the data used to train the algorithm against the presence of bias, which is a more complex issue. Furthermore, the statement that a machine learning model achieves a very high accuracy—such as 99%—is hardly meaningful, unless a baseline for comparison is provided. A more interesting drawback emerges when ethics are considered, as it is very difficult to adjust the contraposition between the ethical principles followed by an AI model and the ethics guidelines that a company should follow, the latter being driven by financial motives.

According to Forbes (https://www.forbes.com/sites/tomtaulli/2021/07/09/how-to-evaluate-ai-software/?sh=7fe6927b7322, accessed on 6 March 2022) AI software evaluation can be a challenging task. As any process of this sort requires a set of carefully chosen criteria to assess the performance of the competitors, AI can be measured according to (i) data connectors, i.e., the capacity to connect to a data source, (ii) flexibility, i.e., the capacity to avoid one unique focus and respond to different needs, implying the capacity of an AI software to choose the best formalism for the job; (iii) ease-of-use, i.e., a non-tech person must be able to use the software without any a-priori scientific or technical capability and finally (iv) ethical-AI, the capability of the software to produce not-biased classifications in terms of gender, ethnical groups or other sensitive data attributes.

Furthermore, in a whitepaper recently published by GoDataDriven (https://cdn2.hubspot.net/hubfs/697348/GDD%20General%20files/WP-AI-Maturity-Journeys/AI%20Maturity%20Journey-Whitepaper-GoDataDriven.pdf, accessed on 6 March 2022), the authors evaluate the maturity of AI applications like a process composed of analytical capability and business adoption. The first pillar can be decomposed into three sublevels, notably data, people, and technology: the example proposed to clarify this perspective concerns the capability of considering and implementing a data lake (A non-structured, highly scalable, indexed, and cataloged storage repository, queriable by different programming languages contraposed—for example—to a data warehouse, which is structured and follows the relational database organization, where data can be queried by SQL only) (data level), data scientists (people level), and finally data analytics applications (technology level). The second pillar is focused on how much AI is embedded within a professional organization and is articulated into three levels: executive support, funding, and implementation. The first motivator relates to the presence of a leader, driven by the will of pushing the organization toward AI; the second defines the financial support necessary to fund an AI enterprise, and the last regards the necessity to involve the business organization in any AI-related activity. The authors claim that the passage from an immature AI application to a more mature one is composed of different steps, such as (i) initialization, typical of a company that is not necessarily fully aware of the capabilities offered by machine learning algorithms—often, the attention is focused on a single AI application and on the produced results; (ii) continuous experimentation, where the company has recognized the value of AI applications and needs to consolidate the knowledge acquired so far: this goal is achieved typically by recruiting skilled staff, by setting working standards and procedures; (iii) enterprise

empowerment, where the company has acquired a vision and takes decisions based on the intelligent analysis of data and finally, (iv) AI democratization, the stage where AI is firmly embedded in the company philosophy and organization, which extends the capability of learning from data, providing a suitable infrastructure that grants access to AI applications to the business.

However, the evaluation of AI software may depend on the specific application domain. For example, Kang and C. T. Haas [10] discuss the evaluation with regard to tools that can automate conformance checking—a time-consuming process if done manually—in the engineering and construction sector. The authors consider different commercial AI tools within different case studies, ranging from natural language processing (NLP) to image search tools. However, the analysis taken seems to be qualitative rather than quantitative.

According to Ish et al. [11], the performance of an AI software can be evaluated as the capacity of a trained algorithm to classify new data, and by the average time the system spends when processing a specific datapoint. The second metric could be used to define the requirements of the AI software. Specifically, typical classifiers in machine learning (able to provide a prediction in the contraposed fashion of yes or no) use four types of measures, namely (i) true positive rate or recall, (ii) true negative rate or specificity, (iii) positive predictive value or precision, and finally (iv) negative predictive value. An interesting parameter is provided by the accuracy, defined as the sum of the true positive and true negative values. The authors also considered the cost calculated by taking into account (i) the benefit of a true positive, (ii) the cost of a false positive, (iii) the benefit of a true negative, and (iv) the cost of a true negative. Each quantity is multiplied, respectively, by the corresponding percentage and summed up. The authors point out that a classifier produces a score that is compared against a threshold, which ultimately should be regarded as a metric too, as it decides on the final response.

3. Evaluating AI through Games

A plausible objection to the evaluation of AI software through quantitative measures is based on the fact that human intelligence cannot be measured entirely by numerical methods. Instead, more complex tests can be considered, such as techniques aiming to analyze non-verbal intelligence, abstract thought, or even creativity. The latter plays a significant role in games for at least two reasons: firstly, with regard to the ability to learn the rules (in a supervised way or autonomously), and secondly, with concern to the capability of planning a strategy to defeat the opponent.

From a historical perspective (Figure 1), the relationship between games and artificial intelligence started with three theoretical premises. The first one can be attributed to Charles Babbage [12]. As he and Ada Lovelace were working on the Analytical Engine [13], he had the intuition of eventually testing the intellectual skills of the machine through its capability to play games, such as chess. Despite the Analytical Engine not going past the prototype stage, the conceptualization of games as a testing environment for an artificial intelligence remained applicable. The second one can be attributed to Alan Turing. He was the first to theorize a computer program able to learn and master chess, thus speculating on the idea that games could be a means to develop artificial intelligence itself. Finally, Claude Shannon suggested that games, chess in particular, could be intended as a "wedge in attacking other problems of a similar nature and of greater significance" [14], such as performing mathematical operations, translating one language to another, and orchestrating a melody. From this perspective, games became also relevant as a research environment for more practical problems. One could notice that chess was considered the epitome of intellectual games; in fact, it was an activity associated with prominent, brilliant individuals. Thus, the ability to play the game was considered an indicator of remarkable intelligence. However, the first game-playing program was not related to chess, as the one developed by A.S. Douglas as a part of his doctoral dissertation at Cambridge was devised to play Tic-Tac-Toe.

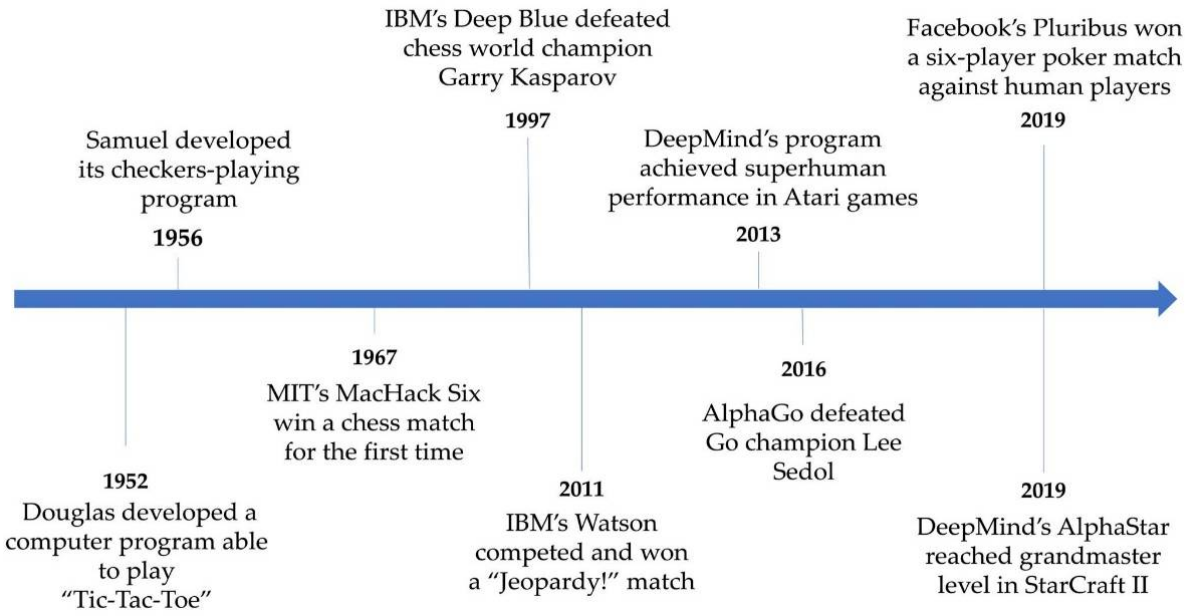

Figure 1. A brief history of game-playing AI.

In 1956, Arthur Samuel, one of Shannon's fellow attendees at the notorious Dartmouth Conference and IBM researcher, developed a checkers-playing program. The research carried out by Samuel on his checkers-playing program ended with the realization of the program itself, which was only able to play checkers [15]. However, a solution to a more functional problem was not produced from the effort. Even though Samuel's effort focused on checkers, the research community adopted chess as the main research environment in the following years. Chess provided some important advantages over the other board games [16]. The extensive body of historical and theoretical literature helped the researchers in accessing reliable data to develop their programs, and it also served as a benchmark to validate the performances against past recorded matches. The popularity of correspondence chess, consisting of unknown opponents competing against each other, smoothed the introduction of a computer opponent. Plus, the chess community had developed a ranking system to appraise the relative, overall strength of each player. In this way, it was possible to check the improvements of the computer programs in relation to a reliable criterion. Even if at first the human chess tournaments served only as a model to shape the computer one, by the late sixties, computers were already competing in conventional tournaments. In 1967, the MacHack Six, developed at MIT, became the first computer program to defeat a human player in an official competition [17]. However, this was just the start of a long history of achievements in the chess-playing domain of artificial intelligence, which culminated with the defeat of chess champion Garry Kasparov. With this event, one paradigm has fully emerged for the evaluation of artificial intelligence, which is the defeat of the human opponent. Therefore, the first computer program to win over a chess grandmaster was purposely devised with such intent. John McCarthy has critiqued this new direction of focusing on building just stronger performers, saying that he believed that for a company it would have been easier "to justify work on computer chess as a means of getting publicity than as a research tool" (McCarthy, 1997). Deep Blue could be indeed considered more successful in re-establishing IBM's brand image after a few years of decline than in providing original contributions in the field of artificial intelligence. In fact, the "human against artificial intelligence" established itself as a powerful narrative to astonish the public [18]. As a matter of fact, IBM used this strategy again when deciding to take on the development of a computer program able to compete in "Jeopardy!". IBM's Watson, similar to Deep Blue, was incredibly successful in catching the public's attention, but far less in producing a significant advancement in the research area. However, importantly,

Watson can be considered the only achievement in the game-playing domain that was actually applied in real-life applications. Although, the capabilities displayed during the demonstration match were not thoroughly present.

DeepMind, following the blueprint of IBM, was able to make a name for itself with the development of a computer program able to reach superhuman performance in various Atari games. The accomplishment was influential in cementing the theory that games could be a way to develop and test a more general type of artificial intelligence. A program able to master more than one game at a time could indicate the development of a broader set of skills. Nonetheless, DeepMind's consecutive attainment, a computer program able to play Go, exhibited only the abilities necessary to play one, single game. The program named AlphaGo [19] won in a public match over Go champion Lee Sedol. Similar to Deep Blue, it was retired right after the event, but three other updated versions were developed in the following years. Whereas AlphaGo Zero has been only able to play Go, AlphaZero and MuZero were more versatile, mastering more than one board game at times. DeepMind has experienced some complications in replicating the same outstanding results obtained in the research field into real-life solutions. Hence, the company has begun to make use of video games as a new research environment. Video games indeed seem to offer similar challenges to the real world. DeepMind has extensively aimed its attention at the development of AlphaStar [20], a computer program able to play StarCraft II. In this case, the computer program was not able to consistently beat human players, as in its past game-playing efforts. At the same time, the shift toward video games has been prevalent in the research community. This trend has better emphasized the shortcomings of developing computer programs to match against human players. While board games only imply the use of reasoning capabilities, video games also require physical skills. Thus, the disparity between human and artificial players has become more noticeable. Considering AlphaStar, the program had to be limited to view only part of the game's map at a time, and the actions per minute were reduced to match the restrictions of a human player. However, AlphaStar still displayed an exceptional, constant efficiency in making choices, which is generally uncommon in humans [21]. Hence, there are some undeniable, structural differences between humans and artificial intelligence, which question whether the adversarial comparison between the two is a suitable method for appraisal. Figure 1 recaps the history of game-playing AI.

Are Games Effective Benchmarks?

Benchmarks in the research field of artificial intelligence have long been associated with performances on a defined, narrow task. Chollet [22] has traced back this tendency to one of the pioneers in artificial intelligence, Marvin Minsky. As the evolutionary theory had inspired cognitive psychologists to intend intelligence as a result of special-purpose adaptations, Minsky favored the idea of also evaluating artificial intelligence on specific tasks. In fact, akin to Minsky's assumption, the major accomplishments in the domain of artificial intelligence have been in narrow, precise endeavors, such as chess or image recognition. Hence, four approaches have been used to assess the capability to complete a single task: human review, white-box analysis, peer confrontation, and benchmarks. Benchmarks have the advantages of providing a fixed test set, which is the same for every attempt, and of being scalable and flexible enough to cover a wide range of possible tasks. However, benchmarks tend to focus on a single metric (e.g., winning a chess match) that encourages shortcuts to reach the final goal. On this matter, three frameworks based on games and used to assess more general skills of artificial intelligence are going to be considered.

The first one is the General Game Playing Competition, which has been held since 2005. The concept of the competition was based on rethinking the game-playing domain in the aftermath of Deep Blue's performance. The intention was to move away from research on narrow tasks and develop programs able to display more adaptive capabilities. The competition's goal is to aim researchers at developing programs able to play more

than one game without human intervention. The distinguishable characteristic of this framework is that the programs are tested on unseen variants of board games, such that the algorithms cannot be designed to tailor winning strategies in advance. In this way, special-purpose skills cannot be embedded in the programs. The evaluation of the artificial participants is divided into two phases: a qualification round and a runoff competition. In the qualification round, the programs engage in various games, ranging from single player to multiplayer. In this phase, the assessment is based on consistent legal play, the ability to attain winning positions, and playing time [23]. The best ones participate in the runoff round, where artificial participants are pitted against each other in a series of games of increasing complexity. Hence, the final evaluation is based on a peer comparison, from which the overall winner is selected.

The second framework can be considered an outgrowth of the previous one, specifically designed for video games. In fact, the General Video Game AI Competition has a similar scope, that is, the creation of an artificial intelligence able to play a broad, and in principle unlimited, set of games. Additionally, the challenge is that games on which the game-playing programs are tested are unknown a priori. In this case, the evaluation is done through a ranking system: over all the attempts at each game, the number of victories, the total sum of points, and the total time spent playing are the three criteria that determine the relative ranking. After all the entrants are ranked in each game, points are awarded according to each ranking position achieved. Hence, the program with the higher number of points is declared the winner.

The third one, that is the Arcade Learning Environment, differently from the past frameworks, does not provide a set of unseen games for testing the programs. The framework can be used to interface with the famous Atari console games. The set of available games is wide enough to provide different sets for training and testing. The main evaluation is based on the score obtained by the programs in each game. However, as it is more troublesome to compare performances in different games, a few techniques need to be used to solve the problem, such as the normalization of each score obtained in each game.

In all these frameworks, the metrics are a proxy for a general set of skills of programs and facilitate the comparison between different programs. However, they are strictly related to the game environments. This prevents a deployment in the real world, and it can be considered indeed the major drawback of this type of benchmark. Additionally, it emphasizes how games, even though they have been a primary driver for research in artificial intelligence, have been less useful in producing techniques that can be applied in a real-world application. One exemplification of this shortcoming is the scarcity of real-life solutions backed by reinforcement learning. In recent years there has been a resurgence of the technique to develop game-playing programs, especially in the view of DeepMind's various accomplishments in Atari games, Go, chess, shogi, and StarCraft II. Yet the remarkable performances have been limited to game environments.

As a representative case, the basic research on artificial intelligence through chess illustrates how the objective has been progressively sidetracked to favor the development of stronger performers. This is why Deep Blue has been considered the dead-end to research in the chess domain. Arguably, Deep Blue did not display any sign of possessing general skills as the applicability of the chess supercomputer was limited to chess, and it was specifically built to defeat the chess champion, Garry Kasparov. The playing strategy was based on a brute force approach, and Deep Blue had been able to search up to 330 million positions per second during the match with Kasparov in 1997 [24]. The hardware requirements needed to top Deep Blue could not be easily replicated by other researchers at the time. The brute force techniques were principally suitable for playing chess, lacking the potential to be a solution in more practical domains. Hence, chess lost the initial appeal that has driven the researchers for years, as Deep Blue was the proof that the capability of beating a world chess champion was not an indicator of artificial general intelligence. It has to be accounted for that research in chess was not very successful outside their own domain, even before Deep Blue. Chess can be considered a two-player zero-sum game, given that a

player wins if and only if the other loses. In this scenario, one player will try to predict the opponent's move and play the best response to it, whereas the other player anticipates this strategy and plays a move that guards against any adverse outcome. Chess is also a game of perfect information, which can be solved by backward induction—in reference to Kuhn's and Zermelo's Theorems. Theoretically, as every finite game of perfect information has a backward induction solution, it is possible for players to predict the outcome of a game and settle on strategies that constitute a Nash equilibrium. In chess, the ability of human players to do backward induction is very limited, especially in comparison with computers. This can be considered the main reason for the immediate successes of chess-playing programs in the early days of artificial intelligence. Even so, the ability of computers is also still far from solving the game entirely and various techniques are used to limit the exploration to promising branches. However, almost all real-life interactions cannot be described as two-player zero-sum games, with few exceptions such as a transaction between a buyer and a seller, making the results obtained in the chess domain intrinsically not generalizable. Considering this, the premise of chess as a steppingstone to developing practical solutions has shown to not have concrete support. This suggests a deeper consideration of whether games fail to be an appropriate research domain for real-life solutions. The argument used for chess applies to all the two-player zero-sum games with perfect information; hence, it can be extended to games such as checkers or *Go*. As a matter of fact, efforts in both games have been unsuccessful in producing reproducible results outside the research domain. However, does the argument stand still for other types of games?

Backgammon is a two-player zero-sum game with incomplete information, as it possesses a stochastic component. In games with incomplete information, the players lack some relevant information about their opponent, but they are aware of all the moves previously taken. Specifically, in backgammon, the board is visible to both players, but given that the possible moves are determined by dice rolling, the strategies are more concealed. One of the most famous cases of a computer program able to play backgammon is TD-Gammon, which was developed by Gerald Tesauro in the early nineties. The accomplishment of the program was the ability to learn how to play backgammon by solely playing against itself. TD-Gammon has sparked an unprecedented interest in the research community to reproduce the results obtained in backgammon outside the game domain [25]. In two-player games with incomplete information, it is still possible to find a Nash equilibrium, given that all the possible outcomes of the current state of the game are known. The stochastic component adds a significant uncertainty in predicting the opponent's next moves, differently from deterministic games. In the latest version of the program, TD-Gammon's decisions were based on evaluating all the possible combinations of dice rolls and consequent moves for at least one turn. The program would pick the move with the higher probability to lead to a positional advantage in the next one. Whereas TD-Gammon was an important accomplishment in the game domain, the main limitation relates to the fact that all outcomes are rarely known in real-world situations. Moreover, as a zero-sum game, backgammon is not representative of the majority of real-world interactions.

Poker is a board game that can be played by more than two players, and it is also a game of imperfect information, given the partial observability of the opponents' cards. Games of imperfect information can still be solved by finding a Nash equilibrium in a subgame, or in other words, in a given turn [26]. However, it is more difficult to do so in games with more than two players. Pluribus [27] is a joint research effort from Facebook and Carnegie Mellon University. The program was able to defeat professional human players in the most played poker format in the world, which is six-player no-limit Texas Hold'em poker. In games of imperfect information, the behavior displayed by the players acts as a substitute for the missing clues. In fact, the moves selected by the players signal their intentions. In poker, bluffing and other strategies are commonly adopted by the players to disguise their real strategies. Pluribus was able to account for the possible shifts in strategies by the players at each hand, instead of assuming that a single strategy would be played by each single player from that point on. This was necessary to avoid the

situation that the program would set on a continuation of the same strategy when making a decision, as it would have been easy for the opponents to exploit such weakness. So, at each hand, the gameplay of Pluribus would remain unpredictable. Even though more real-life situations could be represented through games with imperfect information, in theory, the results are still bound to a simplified environment. All the types of games that have been presented are finite, but in reality, most interactions have an ongoing nature. Additionally, they are structurally competitive, restraining the applicability outside basic research. The video game domain falls under the same reasoning. Taking into consideration Open AI's effort in this domain [28], Dota2 is a multiplayer online battle area video game, in which two teams compete against each other. It can be classified as a zero-sum game with imperfect information. Even if there are additional challenges compared to board games, mainly related to the larger action space, Dota2 is a more uncomplicated environment than the real world. The advantage of games has indeed been the ability to provide a simplified but still challenging environment. Without a doubt, the game domain has been fundamental for the advancement in artificial intelligence. However, outside basic research, games seem to not offer an appropriate testbed for real-world situations. In fact, games-based benchmarks are likely to overlook some real-world challenges, such as the lack of exhaustive data for effective training or the burden of safety considerations. Moreover, there is not enough attention given to developing a system explainable to somebody not familiar with the technicality of artificial intelligence, which could prevent an eventual adoption of a real-life solution. The game domain further fails to account for ethical considerations [29]. Hence, a follow-through experiment from a game environment to the real world becomes more troublesome, as more effort is required toward working on these additional issues. Table 3 provides a summary of this discussion. However, a new direction in the game domain has been undertaken recently, with a specific focus on the interactions between artificial intelligence and humans, thus moving away from the adversarial perspective that has dominated this field of research. An interesting experiment has also shown how the self-play technique that has been widely used in the research domain of games may produce systems that do not work well when paired to play with humans. Additionally, on the same lines, the need for subjective metrics to evaluate the cooperation among artificial intelligence and humans when paired together in teams has been proposed [30]. The emerging argument is that some necessary factors for the deployment and the adoption of artificial intelligence, such as trust or interpretability from the human counterparts, should not be neglected in the research phase. As both efforts were made through the use of games, Overcooked and Hanabi, respectively, it follows that the games can still represent a resourceful environment, but a new approach is needed to overcome the limitations presented in this section. Additionally, the focus on cooperation between artificial intelligence and humans emerges as an important matter that needs additional consideration.

Table 3. Summary of game environments and their limitations.

Type of Theoretical Game	Example of the Game	Solution	Limitations to A Real-World Application
Perfect two-player zero-sum game	Chess (IBM's Deep Blue)	Predict the outcome by backward induction and settle on the Nash equilibrium strategy	Most real-life interactions cannot be described as two-player zero-sum games, except one between buyers and sellers
Two-player zero-sum game with incomplete information (stochastic component)	Backgammon (Tesauro's TD-Gammon)	Find a Nash equilibrium, given that the probabilities of the different outcomes can be computed	In real-life, all outcomes of uncertainty are rarely known, and, consequently, it is not possible to account for each one of them
Multiplayer games with imperfect information	Poker (Facebook's Pluribus)	A subgame Nash equilibrium can be found for each turn	The results are bound to simplified turn-based interactions
Real-time multiplayer games with imperfect information	Dota2 (OpenAI's OpenAI Five)	Even if the choice of the action happens simultaneously, it is still possible to find subgame Nash equilibriums	A more realistic environment with a high branching factor does not entail considering real-world challenges

4. Evaluating AI through Cooperation

Aside from the capacity to play a game and even learn its rules (for example, by using reinforcement learning, such as in AlphaGo), an AI system can be evaluated also by the ability to cooperate with other agents, artificial or humans.

The problem of cooperation has been extensively studied in cognitive sciences as the attention revolves around the concepts of interference, a term borrowed from physics to denote both opposition (conflict between goals) and agreement (pursuing a common goal) and benevolence (the attempt to make other agents' functioning easier). Millot and Lemoine [31] identify two types of cooperation, called, respectively, horizontal and vertical. The former is heterarchical and deploys the agents at the same level, assuming that the corresponding tasks are independently defined; the latter is hierarchical as one agent has decision power over the others. Furthermore, both formalisms exploit techniques oriented to managing interference and facilitating the accomplishment of other agents' goals. When agents possess similar know-how and they can be replicated, then cooperation can be augmentative.

March [32] highlights the pervasive signs of the widespread forms of collaboration between AI agents and humans in different areas, such as decision-making procedures, assuming that deception is not considered, i.e., humans are aware that their counterparts within the considered interaction are agents. Quite interestingly, the author stresses the fact that some cooperation models include the presence of obstinate human subjects reluctant to negotiate their share. Some other types of interactions might lead to the manipulation of the subject's belief, exactly as it happens with the well-known prisoner's dilemma: one of the takeaways of this work is that cooperation must rely on trust to be effective.

The industry is the typical scenario where humans and robots are working together, the latter executing a type of work considered either repetitive or dangerous by a human being, preserving however two important shared objectives: production and human safety. Cesta et al. [33] argue that robustness is an important account in the human–robot cooperation (HRC) paradigm, including the ability to change a robot's speed or trajectory, but more notably, its functional aspect, such as changing its tasks-coordination algorithm. These methods do perform well with temporal uncertainties, i.e., deviations from the scheduled trajectories, although unforeseen scenarios (i.e., for which no previous historical data used to train machine learning algorithms exist) are obviously problematic. The authors depict the so-called collaborative tasks, for example, collaborative assembly, which consists of assisting the user to assemble an object made of different components by applying planning algorithms, which are able to solve scheduling problems. Another interesting topic regards the action undertaken by a robot in order to hand an object over to a human, which implies a recognition of what the human counterpart is expecting. Problems of this sort are usually solved by K-Nearest Neighbors and, more generally, re-enforcement algorithms.

Physical cooperation between a human and a robot is described by Mörtl et al. [34], where the common task consists of moving a bulky object—a goal involving both geometrical and dynamical challenges, i.e., constraints related to the environment, shared awareness of the shape and the dynamics of the grasped object. The authors define a mathematical model by accounting for all the actions taken by the agent in terms of negotiation and actuation redundancy, producing as output the desired trajectory of the object. Quite interestingly, not only the success of the cooperation is measured by quantitative measures, such as the time taken to complete the task and the interaction between the force deployed by the agents, but also from a subjective perspective, interviewing the participants and asking about their efforts, frustration, physical and mental demands.

According to Guo and Yang [35], trust between a human and an autonomous agent stabilizes over repeated interactions, but the negative experiences have a greater influence than the positive ones in the long run. Chong et al. [36] have analyzed the contest of AI-assisted decision-making in which humans are responsible for the final decision. The study

confirms the adverse effect of poor performances on the human counterpart, as humans tend to negatively judge themselves for relying on the agent. Bender et al. [37] argue that the perceived adaptiveness of the autonomous agent is strongly connected with trust; on the other hand, a system dysfunction, while accomplishing a task, negatively impacts confidence. Hence, it has emerged that the ability of an agent to avoid making errors is fundamental to strengthening cooperation. Hanoch et al. [38] have argued that autonomous agents could encourage risk-taking behavior in humans. In the experiment conducted, participants were required to press the spacebar on a computer keyboard to inflate a balloon displayed on the screen before it could blow up. When encouragement came from a robot sitting next to the participants, a more risk-inclined behavior was observed. Participants tended to rely less on the experience accumulated in the previous trials to follow the instructions coming from the robot, especially if they have led to a positive outcome in the past. However, excessive trust can lead as well to negative consequences. From the insights collected by Compagna et al. [39] at the Fabrication Laboratory "MTI-engAge" at the Technical University of Berlin, the routinization of the cooperation between humans and agents in the workplace could pose a safety risk. The reduction of the cognitive tasks when humans work with a predictable agent could endanger the trust built over time. The lack of attention could result in an important safety issue when humans and autonomous agents physically work in the same environment. Errors could become fatal in routine tasks or could lead to a high level of stress in human workers, as they could be considered at fault. With a higher degree of delegation in decision-making, humans may feel less responsible for the performance of an autonomous agent. However, it is important to consider whether increasing the autonomy of the agent may shrink that of the human counterparts. Formosa [40] has debated that the relationship between autonomous agents and humans should not be seen as a zero-sum game; thus, incrementing the autonomy of an agent may be beneficial in cooperation. However, this is not always true, and there is the possibility that an autonomous agent could be programmed to nudge the human counterpart into conforming to their behavior, even to one desired by an external party. Karpus et al. [41] have focused on the opposite issue, which is whether humans could take advantage of an autonomous agent. Whereas humans are equally likely to cooperate with another human and an AI agent, they are more inclined to take advantage of the latter one. The explanation for such difference is the belief that the autonomous agent is programmed to cooperate; hence, humans feel less guilty about betraying a counterpart that does not reciprocally sacrifice any interests. One way to prevent this exploitative behavior from humans could be linked to anthropomorphizing the embedment in which the agent is presented. Kulms and Kopp [42] have demonstrated that anthropomorphisms enhance the perceived trustworthiness of an autonomous agent. However, humans tend to behaviorally develop the same level of trust if a continued competent performance is carried on, whether the agent exhibits human-like cues or not. Maehigashi et al. [43] argue that there is a significant difference in how humans react to errors made by AI systems and robots. While trust decline is greater for a mistake made by AI software, humans seem to be more forgiving of anthropomorphic robots. The reason for such difference is that humans are keener to assume error possibilities in an agent that somewhat resembles themselves. In fact, while humans acknowledge the possibility of errors for fellow humans, they tend to not do so for agents that do not show any human-like traits. Thus, this bias could be the source of the misattribution of errors investigated by Chong et al. [ibid.]. However, a concern surrounding anthropomorphism is that humans may cultivate a social relationship with an entity that is not truly social [44]. The question is whether such deception can be considered ethical, even if it can further the cooperation between an autonomous agent and the human counterpart. A significant scenario is the one where the trust developed by the human counterpart is exploited. On this matter, Dafoe et al. [45] argue that one of the main challenges in human–AI cooperation is the involvement of humans themselves. As the issue of privacy can become an ethical constraint, data collection may be limited. Especially if data is collected by a third party, such as corporations or employers in the case

of work environments. Even though this could prevent a malevolent use of a non-sentient, autonomous agent, it can also reduce the necessary data to provide appropriate training to the agent itself. Figure 2 illustrate an approximate rule of thumb for using trust as a proxy for evaluation.

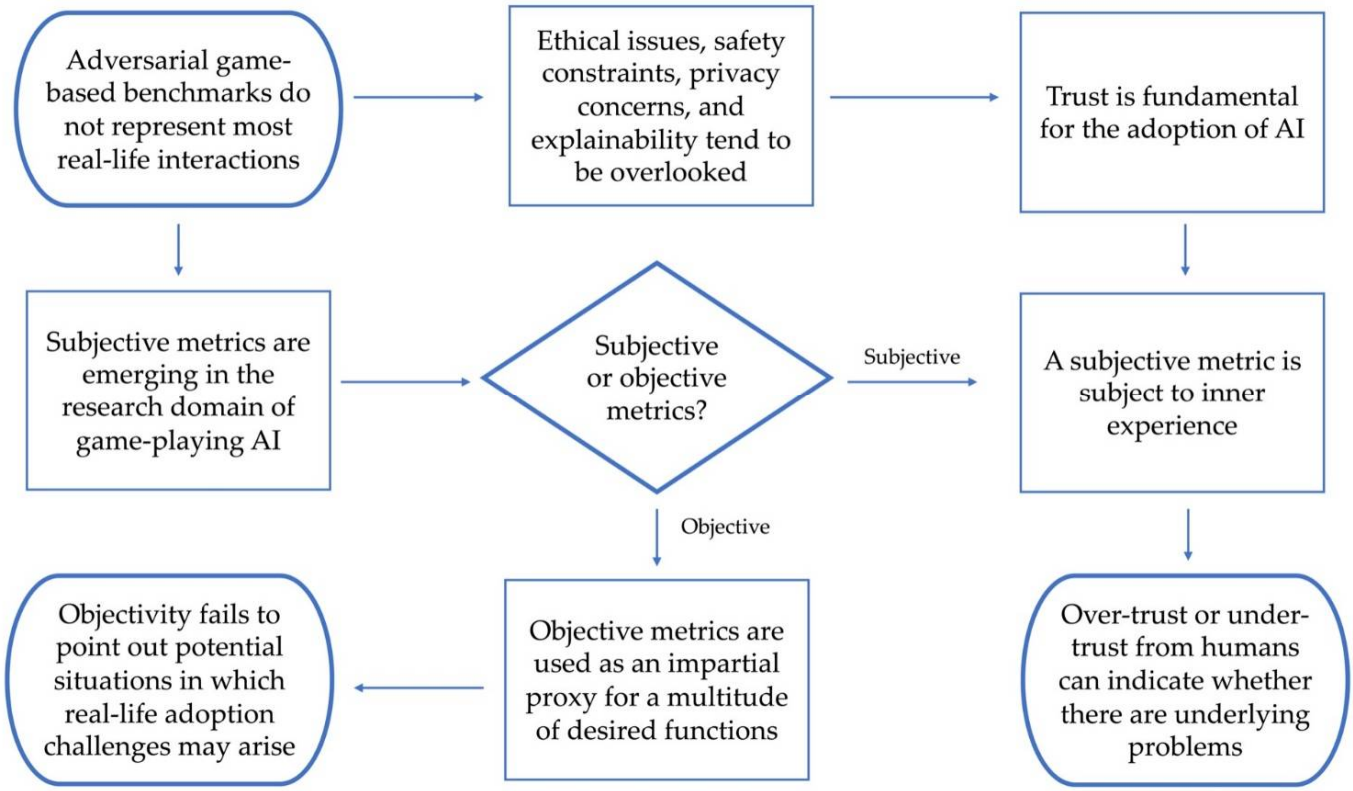

Figure 2. Flowchart on the use of trust as a proxy for evaluating an AI agent.

5. Ethical Concerns in Human–Machine Cooperation

The emergence of ethical concerns in scenarios based on cooperation is particularly evident in hybrid situations, i.e., where interaction occurs between a human and an agent (As per the definition provided by Russell and Norvig ([46]) an agent is "Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators". This concept can be augmented by including the notion of *rationality*, so that a *rational agent* is "An agent that acts so as to maximize the expected value of a performance measure based on past experience and knowledge."). This section presents three case studies (see Table 4) based on thought experiments, which are discussed in the light of a set of key factors, defined as follows.

Trust is a belief based on reliability; autonomy is the capacity to make decisions in conditions of uncertainty; finally, cooperative learning [46] is a term borrowed from educational studies (contrasted with forms of competitive learning), where the student's learning process is rewarded by other members of the group gaining new skills and achieving goals. An interesting, similar approach can be found in cooperative machine learning [47], which consists of a modification of the usual ML algorithm life cycle: the algorithm is trained on partially labeled data and is used to predict novel observations. A manual process revises all the labels with low evidence: the same data are used again to re-train the model.

Table 4. Thought experiments between concerns and realizations.

Scenario	Concerns	Realizations
Soccer match	Cooperative Learning within a humans/robots scenario is hard to achieve because of violation of expectations (i.e., robots would play rationally, humans could play on an emotional basis)	RoboCup, though players are not humans yet
Battlefield	Lack of ethical programming could lead to casualties for both parties. Logical contradictions might occur when ethical programming is applied, due to the difficulty of differentiating human hostiles from artificial hostiles (i.e., tanks, planes)	Boston Dynamics, though some side problems (such as noise) need to be resolved
Restaurant	Cooperative learning relies on the confidence that humans will not be replaceable. It is necessary that teaming up with a robot should not cause frustration and not be a source of problems inside. Free riding could be likely from the human counterpart.	Robot waiters have been introduced in various locations around the world such as at Pizza Hut (https://www.businessinsider.com/pepper-robot-to-start-work-at-pizza-hut-2016-5?op=1&r=US&IR=T, accessed on 6 March 2022), Denny's (https://thetakeout.com/dennys-robot-waiter-automation-server-replace-humans-la-1848132060, accessed on 6 March 2022), or a local restaurant in The Netherlands (https://www.theverge.com/2020/5/31/21276318/restaurant-netherlands-robot-waiters-social-distancing-pandemic, accessed on 6 March 2022).

5.1. The S Scenario: A Soccer Match

A soccer team H is composed of some human players, which are challenged to compete against a team R, including an equivalent number of robots (regarded as agents). It is assumed that the members of H and R have equivalent physical skills (speed, resistance to fatigue, etc.). Furthermore, typical player roles and play styles, such as goalkeeper, central midfielder, and so forth would be replicated in R. With respect to the latter, trust is denoted by the knowledge that players of the same team will coordinate their actions to maximize the number of goals scored against the competitors and those kept by the goalkeeper. Cooperative learning is described by (i) taking all the actions needed to score a goal (attack), (ii) not committing foul play, as it would result in penalties; and (iii) preventing H from scoring (defense). Autonomy reflects the capacity of an agent to plan a strategy providing the highest probability to score a goal in conditions of uncertainty (i.e., the outcome of every action, such as crossing, dribbling, and so forth, is stochastic).

Discussion. It is not guaranteed that H would play a fair game because of social and psychological factors (such as anger, frustration, and pressure from the coach or the supporters), though *R* has been programmed to play in such a manner. Since both H and R are homogenous, no conflicts would arise inside the teams as the main objective—winning the game—is shared.

To review cooperative learning, a slightly different scenario S' is considered, where each one of the X and Y teams is composed of a mix of humans and robots. Trust should be regarded in two ways, i.e., from the perspective of humans and robots. In the first case, a human player *h* would trust a robot *r* of the same team to be programmed to win the game and would pass the ball indifferently to a human or robotic member of the same team only in the case he or she was educated to do so for a long time (i.e., from a young age). Cultural bias or simply a conservative attitude would push the player to pass the ball to other humans instead of robots in circumstances that require them to take a decision in a matter of seconds. Vice-versa, a robotic player would take any decision rationally, planning a strategy and rescheduling continuously the best actions serving the purpose of scoring a goal. As its human counterpart playing in the same team might behave emotionally or even irrationally, being not motivated to play well, after a certain period of time a robot could consider passing the ball not to human players anymore. Furthermore, while all the robotic

players are identical, no original playing styles or attitudes could emerge during a game. In absence of strong psychological references (i.e., a robot particularly good at performing a particular action), human players would prefer to play with other humans, compromising the objective of winning the game. Another issue about Cooperative learning occurs in a situation where a robot would estimate the probability to win the game to be very low (or lower than a fixed threshold)—well in advance before the end of the game: it would not make sense for the robotic team to carry on playing until the end. Under S this would not be a problem, though it would be in S′, where humans playing would notice that suddenly all the robots within the same team would not stop playing one by one. From a material perspective, information processing could be subject to significant delays, and adaptability (i.e., understanding that the performance of a human player might change during the game) might be anyway very difficult to achieve.

Finally, autonomy should not be impacted in either S or S′, apart from hardware malfunctioning of a robot when, for example, visual aids (i.e., embedded cameras combined with artificial vision) have stopped working. In this case, a robot would depend on a human providing direction about the game and it should be replaced.

5.2. The B Scenario: A Battlefield

In a battlefield scenario, human soldiers HS and robot soldiers RS fight together against a common enemy. Though HS and RS fight side by side, they do it on a plane of inequality, as the latter are considered expendable and much stronger (better senses, higher precision when shooting, resistance to fatigue, better armor, etc.) than the former. Furthermore, robots would precede humans in open fire fights; though robots are programmed to be autonomous, they would take orders from human troopers when needed. It is expected that HS would fully trust RS because of the way they have been programmed. However, no humans would feel obliged to protect or save a robot, being undistinguishable and replicable. For this reason, a robot could trust only others of its kind (i.e., when a robot is injured and incapacitated, it would be reasonable to program its peers to rescue him instead of leaving it to the enemy's forces). With regard to cooperative learning, the obvious goal, shared by both the parties, would consist in drawing a strategy leading to the victory against a common enemy considering as a constraint the lowest number of casualties among HS. Depending on the cost of building and training a member of HS, there could also be a motivation to preserve its integrity, if not on the whole, enough to reuse its parts. With concern to autonomy, it is expected that the components of RS would be programmed to react to uncertainty and adapt its objectives according to unforeseen changes in the scenario (i.e., the worst case, where the entire HS has been obliterated, the strength of the enemy has been underestimated and so forth), although the final goal would be preserved. However, their autonomy could be overruled by the orders of humans.

As a robot in a war zone might injure or even cause the death of a human being in a shooting, its programming should include moral and ethical principles to prevent cruelty and violation of human rights. In literature, robotics ethics often refers to the well-known Asimov's laws of robotics (Isaac Asimov, *The Complete Robots*, Harper Voyager books, 2018), which could not be used as a whole in this case as they would inhibit any offensive action from RS. At first sight, the laws might be modified, for example stating that no harm—physical or psychological—should be caused to an unharmed human, belonging to HS or to an enemy who has surrendered or is not capable of offense. However, it has to be noticed that some situations might lead to a stall, for example where one or more HS soldiers were ordered to take a mission highly risky or even suicidal: the order would be overruled by a robot (Asimov's stories about robots are often based on logical contradictions emerging from the robotic laws. Note that the latter have sometimes been reviewed and modified in recent years: Van Dang et al. [45] designed an agent oriented to home service, proposing a variation of Asimov's idea as follows: "Modified first law: A robot may not serve harmful food to a human being") in order to protect their lives. Furthermore, human enemies might be indistinguishable from HS, therefore causing RS to refuse to fight. Alternatively, RS

could be programmed to attack only means of offense, such as tanks, planes, or cannons, and not enemy soldiers. This would imply disguising some of the truth (i.e., both tanks or planes are ultimately driven by humans) and potentially undermine even more Trust, not to mention the fact that the enemy might use civilians or hostages as human shields.

It can be argued that in a different war scenario B', where HS would be ruthless (potentially even criminals) and ethical programming of RS would be omitted, the consequences of this choice might seriously harm HS, in case RS would judge the former expendable to defeat the enemy.

5.3. The R Scenario: A Restaurant

In a restaurant called RH, a staff composed of human and robot waiters is tasked with optimizing customer service. It is assumed that robots will not automate the job completely, such that human waiters are still an important resource inside RH. Cooperative learning is based on (i) the common goal that should bring together the robot and the human waiters, (ii) a less tedious job should serve as an incentive for the inclusion of the robot waiters in the workplace, (iii) an increased organization among the staff should facilitate the functioning of the robots. Autonomy is related to the supposition that robot waiters will have capabilities on par with their human counterparts in ten years. Robot waiters have already been successfully employed inside restaurants, but their role is limited to bringing the food to the right table and going straight back to the kitchen to get ready to take on the next order. Human waiters are needed to set up the orders and indicate the table to the robots. Thus, it is not that far-fetched that robots will eventually be capable of arranging an order for the right table by themselves. This does not imply that human waiters will be obsolete in this scenario. Instead, their presence would be essential for optimal service, especially in unexpected situations, where robots may lack the intuition to solve a problem with clients.

Trust is a direct consequence of the necessity that human waiters should not feel threatened by the robots in any way. Neither their job should be at stake because of eventual replacement, nor should their performance be negatively affected by the robots. This implies that the robots cannot act unpredictably. A lack of this condition may compromise and disincentivize cooperative learning, as human waiters may start a coalition against the robots. In a staff composed of more than one robot, it is unavoidable that the robots should be prevented from teaming up together and leaving out the human waiters. This will have a similar effect to the previous situation, putting at hazard the willingness to cooperate.

What emerges from this scenario are two important premises for the cooperative learning of human and robot waiters. First, teaming up with a robot should not cause frustration in the human counterparts. It is important that the robots have room to act autonomously, but some constraints are necessary. The robots should be reliable in serving clients and should not be a source of problems inside RH. In fact, the risk is that a robot waiter may become the scapegoat for disputes and issues in the workplace. This further implies that when problems may arise in the software, the latter should be developed to allow for quick intervention by the rest of RH's staff. In this way, the regular working schedule is not compromised by the malfunctioning of the robots. Second, robots should be put in the position of functioning optimally. The work environment should be organized to allow the robots to work efficiently. A lack of staff management may lead to robots overworking, and human waiters free riding on their effort. In this case, robots may be put under stress that may be detrimental to their hardware. Assuming that robots will rely on rechargeable batteries, their power should not be exhausted during their working schedule because of overwork. As this forfeit from robot waiters is out of their control, the responsibility should be on the staff.

5.4. Discussion

The objective of this research is to propose three pillars of a future ethical framework to evaluate AI agents cooperating with humans across different scenarios (Table 5). The

outcome of this work is described in terms of trust, autonomy, and cooperative learning by generalizing some of the considerations that emerged during the study of S, B and R thought experiments. Firstly, trust should be regarded as connected to explainability and reliability. Explainability prescribes that all the actions taken by an agent must be understandable, i.e., it must always be possible to generate a minimal, finite causal chain. In contrast, Reliability requires that an agent is able to perform its task without interruption: this requirement is primarily related to the agent's hardware and is usually achieved by duplicating its critical hardware components. Secondly, autonomy serves a two-fold purpose: on the one hand, it states the responsibility of the actions taken by an agent to pursue a common target, and on the other hand, it defines the awareness that every decision drawn determines an impact on neighboring agents and humans. Finally, cooperative learning has proven to be successful in educational areas (especially with children) and has been adopted in machine learning, applying algorithms based on reinforcement learning.

Table 5. The three pillars of AI evaluation in an ethical framework.

Factor	Depends on
Trust	Explainability and Reliability
Autonomy	Responsibility, ethical behavior social awareness
Cooperative Learning	Reinforcement learning, possible contradictions resulting in lack of action

However, some difficulties might emerge in hybrid contexts such as those described so far. For example, irrational (or not ethical) human behavior might decrease the level of trust of an agent, resulting in a lack of action or competition and therefore conflict in achieving a shared goal. It follows that a hybrid scenario (e.g., such as S, B, or R) would be unstable when humans violate one of the three pillars.

With respect to real-world scenarios, a few case studies taken from both scientific literature and the industry can be considered precursors. RoboCup (https://www.robocup.org/, accessed on 6 March 2022) stems from a work [48] of Professor A. Mackworth published in 1992 and the Workshop on Grand Challenges in Artificial Intelligence held in Tokyo the same year, where the idea of implementing the game of soccer on robots was discussed and assessed from a technological and social impact point of view. The resulting project was called "Robot World Cup" (RoboCup); the first soccer game with real robots took place in November 1996, in Osaka. RoboCup has raised the objectives considerably when the organizers claimed that their main objective is to have a game played against human opponents by 2050 (https://www.bbc.com/news/business-58662246, accessed on 6 March 2022). Robot competitions (where well-known AI paradigms, such as Reinforcement Learning are used, see Martins et al. [49]) are not new in benchmarking AI capabilities—see Anderson et al. [50] where the authors point out that a common issue in these types of competitions consists of considering mostly the goal of winning against an opponent, instead of the scientific theory leading to advancements. Furthermore, some aspects of RoboCup are not realistic—starting from the flat idealized surface or the fact that the roboplayer, though having a humanoid shape, is very small in size. HuroCup (https://firaworldcup.org/leagues/fira-sports/hurocup/, accessed on 6 March 2022) is another competition oriented to humanoid robotics, articulated in a series of events aiming at evaluating intelligence at large by considering a single robot performing different challenges; however, aspects such as cooperative learning and trust are outside of the scope.

The origin of military robots (https://eda.europa.eu/webzine/issue14/cover-story/robotics-in-defence, accessed on 6 March 2022) (armed, unmanned devices used in warfare) is rather old, starting from World War II: as per the most recent developments in the field, drones and autonomous fighter jets have been successfully used in war scenarios. The software deployed is essentially based on deep learning algorithms using object identification, together with knowledge bases in order to learn (https://www.technologyreview.com/20

19/11/06/132036/the-us-army-is-creating-robots-that-can-follow-ordersand-ask-if-they-dont-understand/, accessed on 6 March 2022) the best strategy to achieve a goal. The French army is reported to have tested reconnaissance robotic dogs developed by Boston Dynamics (https://www.businessinsider.com/boston-dynamics-spot-robot-french-military-combat-out-of-battery-2021-4/, accessed on 6 March 2022), proving that the use of these kinds of units can save soldiers' lives. Robot service dogs are currently deployed by the 325th Security Forces Squadron at Tyndall US Air Force (USAF) base (https://www.af.mil/News/Article-Display/Article/2551037/robot-dogs-arrive-at-tyndall-afb/#:~:text=The%20first%20official%20semi%2Dautonomous,of%20protection%20to%20the%20base, accessed on 6 March 2022) to check and detect route anomalies. Although *Ghost Robotics* has equipped its unit with remotely operated guns, in some cases the usage of canine-like robots has proven not suitable (https://www.scientificamerican.com/article/the-nypds-robot-dog-was-a-really-bad-idea-heres-what-went-wrong/, accessed on 6 March 2022), as people were either frightened or threatened. Even the US military has admitted that *Big Dog* robots, developed by *Boston Dynamics*, are too noisy to be used on the battlefield with human soldiers.

From a different perspective, AI software can be used to assist human activities in situations where humans are subjected to stress due to the high amount of stimuli received from the environment, for example in a space mission: as part of the next mission, NASA will officially test whether AI could be beneficial to future astronauts. The effort has been in partnership with Amazon, and a voice-activated *Alexa* speaker will be integrated into *Artemis I*'s inaugural mission (According to a communication from Nasa's website available at: https://www.nasa.gov/feature/callisto-technology-demonstration-to-fly-aboard-orion-for-artemis-i, accessed on 6 March 2022). The technology will support only a limited range of capabilities, such as providing the traveling speed or switching on the lights on demand. For the *Artemis I* mission, humans will interact with the voice assistant from the control center. However, in the long term, the plan is to develop more functionalities that would allow the technology to have more control over the spacecraft while astronauts are on board [51]. For now, this situation cannot be categorized as a form of cooperation, given the lack of autonomy of the technology under scrutiny. Yet, it can be considered an early phase in the development of such a situation into space missions.

With events caused by the COVID-19 pandemic, there has been an acceleration in the robotics applications inside the workplace. Service robots, in particular, experienced a boost in popularity in the hospitality sector [52]. An example of service robots is the one previously mentioned of robot waiters: their introduction was supported mainly by the labor shortage and the need for social distancing. However, the limited level of autonomy has prevented the complete automatization of the job [53]. The integration of robots with restaurant staff can be considered an early form of cooperation. For instance, the waiter robots are autonomously able to carry the food from the kitchen to the right table. This should free the human waiters to focus more on the clients, instead of rushing from one table to the other. Hence, an approximate, common shared goal is noticeable, focusing on enhancing the customer service (As in the case of Sergio's restaurant where robot waiters were used to take the heavy lifting from human waiters, in order to reduce overwork. The article from which the case is extracted is available at: https://miami.cbslocal.com/2021/05/20/sergios-restaurant-astro-robot/, accessed on 6 March 2022).

6. Conclusions and Future Work

From a historical perspective, games have emerged as one of the most popular benchmarks in the field of artificial intelligence. Major milestones have been achieved by engaging in the game-playing domain, such as IBM's Deep Blue or DeepMind's AlphaGo. However, such accomplishments tend to lack a subsequent real-life application. This can be considered a significant shortcoming in using games to develop, test, and benchmark artificial intelligence. The competitive structures of the games that have been prevalently used as a research environment fail to represent real-life interactions and challenges. For

instance, the lack of exhaustive datasets, the need for safety constraints and explainability, or the ethical issues that may arise in developing real-life applications are overlooked by the characteristic metrics (e.g., overall score or win) of the game-playing domain. However, a new trend is emerging in this area. A few influential pieces of research have been carried out to understand how the potential cooperation between humans and artificial intelligence could play out through game environments. This is part of an important, relevant research direction that specifically focuses on the forthcoming collaborative relationship between artificial intelligence and humans in everyday situations. Hence, considering the necessity of smoothing the development and the adoption of artificial intelligence into real-life scenarios, the question is whether games can hold any interest in today's research efforts. To find an answer a comparison between the ability to play games and the ability to cooperate successfully with other parties will be reviewed.

Games have a clear advantage, which is the objective metrics offered by the score or the win over an opponent. The main reason for the success of games in the research area is the possibility of comparing the different progress with impartiality. Cooperation, studied either in a real-life scenario or in games, can mainly be only subjectively assessed by the individuals directly involved in the experiments. Trust has emerged as the proxy for appraising the progress in cooperative interactions, especially when humans are involved. However, trust is based on inner experiences; thus, the measure may be considered biased. On the other hand, it better encapsulated the issues that may arise in real-life applications. A lack of trust may indicate problems with the comprehensibility of the choices, or safety concerns that humans have encountered during experimentation.

At this point, considerations need to be made regarding whether games offer an adequate research environment to study cooperative interactions involving humans and artificial intelligence.

Just the purpose of researching cooperation leaves out all the popular games that have an adversarial structure, such as chess, backgammon, or poker. Only games that entail a cooperative mode offer an appropriate domain; hence, the choice is limited to a handful of games with that characteristic, such as Hanabi, Minecraft, or sports games, such as soccer. In this case, the crucial issue is the over-simplification of real-life interactions and environments. This can lead to neglecting the effects of the misattributions of successes and failures, the anthropomorphic traits, and the routinization of the long-term interplay between humans and artificial intelligence. Some barriers to cooperation could be more visible in environments resembling a workplace, such as the issue of free-riding and over-trust in artificial intelligence. Nevertheless, board and video games offer an accessible and cost-efficient platform that can further research in the area of cooperation. Games also have the benefit of presenting a vast amount of data to train and evaluate algorithms, whereas real-life scenarios may be scarce in this regard. Hence, the game-playing domain can still be a fruitful research environment, if a more comprehensive perspective is adopted that goes beyond winning over humans or achieving the highest score.

Future work will analyze the thought experiments from a numerical perspective, providing a quantitative characterization of trust, autonomy, and cooperative learning.

Author Contributions: Both the authors have equally contributed to the manuscript in terms of conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Catholic University of the Sacred Heart, with grant D.3.2 "Ethic and scientific impacts of AI applications".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fidora, A.; Sierra, C.; Institut d'Investigació en Intelligència Artificial (Eds.) *Ramon Llull: From the Ars Magna to Artificial Intelligence*; Artificial Intelligence Research Institute: Barcelona, Spain, 2011; Volume IIIA.
2. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *LIX*, 433–460. [CrossRef]
3. Goertzel, B. What Counts as a Conscious Thinking Machine? New Scientist. Available online: https://www.newscientist.com/article/mg21528813-600-what-counts-as-a-conscious-thinking-machine/ (accessed on 3 January 2022).
4. Nilsson, N.J. Human-Level Artificial Intelligence? Be Serious! *AI Mag.* **2005**, *26*, 68.
5. Kušić, M.; Nurkić, P. Artificial morality: Making of the artificial moral agents. *Belgrade Philos. Annu.* **2019**, *32*, 27–49. [CrossRef]
6. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
7. Halsband, A. Sustainable AI and Intergenerational Justice. *Sustainability* **2022**, *14*, 3922. [CrossRef]
8. Khosravy, M.; Nakamura, K.; Hirose, Y.; Nitta, N.; Babaguchi, N. Model Inversion Attack by Integration of Deep Generative Models: Privacy-Sensitive Face Generation from a Face Recognition System. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 357–372. [CrossRef]
9. Khosravy, M.; Nakamura, K.; Hirose, Y.; Nitta, N.; Babaguchi, N. Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System. *KSII Trans. Internet Inf. Syst.* **2021**, *15*, 1100–1118. [CrossRef]
10. Kang, S.; Haas, C.T. Evaluating artificial intelligence tools for automated practice conformance checking. *ISARC Proc. Int. Symp. Autom. Robot. Constr.* **2018**, *35*, 1–8.
11. Ish, D.; Ettinger, J.; Ferris, C. *Evaluating the Effectiveness of Artificial Intelligence Systems in Intelligence Analysis*; RAND Corp.: Santa Monica, CA, USA, 2021. Available online: https://www.rand.org/pubs/research_reports/RRA464-1.html (accessed on 6 March 2022).
12. Babbage, C.; Campbell-Kelly, M. *Passages from the Life of a Philosopher*; Rutgers University Press: New Brunswick, NJ, USA; IEEE Press: Piscataway, NJ, USA, 1994.
13. Bromley, A.G. Charles Babbage's Analytical Engine, 1838. *IEEE Ann. Hist. Comput.* **1998**, *20*, 29–45. [CrossRef]
14. Shannon, C.E. Programming a Computer for Playing Chess. In *Computer Chess Compendium*; Levy, D., Ed.; Springer: New York, NY, USA, 1988; pp. 2–13. [CrossRef]
15. Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [CrossRef]
16. Ensmenger, N. Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Soc. Stud. Sci.* **2011**, *42*, 5–30. [CrossRef] [PubMed]
17. Bory, P. Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Converg. Int. J. Res. New Media Technol.* **2017**, *25*, 627–642. [CrossRef]
18. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with Deep Reinforcement Learning. *arXiv* **2013**, arXiv:1312.5602.
19. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609. [CrossRef] [PubMed]
20. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354. [CrossRef] [PubMed]
21. Madan, C. Considerations for Comparing Video Game AI Agents with Humans. *Challenges* **2020**, *11*, 18. [CrossRef]
22. Chollet, F. On the Measure of Intelligence. *arXiv* **2019**, arXiv:1911.01547.
23. Perez-Liebana, D.; Samothrakis, S.; Togelius, J.; Schaul, T.; Lucas, S.; Couetoux, A.; Lee, J.; Lim, C.-U.; Thompson, T. The 2014 General Video Game Playing Competition. *IEEE Trans. Comput. Intell. AI Games* **2015**, *8*, 229–243. [CrossRef]
24. Campbell, M.; Hoane, A.J., Jr.; Hsu, F.-H. Deep Blue. *Artif. Intell.* **2002**, *134*, 57–83. [CrossRef]
25. Tesauro, G. Programming backgammon using self-teaching neural nets. *Artif. Intell.* **2002**, *134*, 181–199. [CrossRef]
26. Dutta, P.K. *Strategies and Games: Theory and Practice*; MIT Press: Cambridge, MA, USA, 1999.
27. Brown, N.; Sandholm, T. Superhuman AI for multiplayer poker. *Science* **2019**, *365*, 885–890. [CrossRef] [PubMed]
28. Brockman, G.; Chan, B.; Cheung, V.; Debiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; Hesse, C. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv* **2019**, arXiv:1912.06680.
29. Farisco, M.; Evers, K.; Salles, A. Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence. *Sci. Eng. Ethics* **2020**, *26*, 2413–2425. [CrossRef] [PubMed]
30. Siu, H.C.; Pena, J.D.; Chang, K.C.; Chen, E.; Zhou, Y.; Lopez, V.J.; Palko, K.; Allen, R.E. Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. *arXiv* **2021**, arXiv:2107.07630.
31. Millot, P.; Lemoine, M. An attempt for generic concepts toward human-machine cooperation. In *SMC'98 Conference Proceedings, Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218), San Diego, CA, USA, 14 October 1998*; IEEE: Piscataway, NJ, USA, 1998; Volume 1, pp. 1044–1049. [CrossRef]
32. March, C. Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *J. Econ. Psychol.* **2021**, *87*, 102426. [CrossRef]
33. Cesta, A.; Orlandini, A.; Umbrico, A. Fostering Robust Human-Robot Collaboration through AI Task Planning. *Procedia CIRP* **2018**, *72*, 1045–1050. [CrossRef]

34. Mörtl, A.; Lawitzky, M.; Kucukyilmaz, A.; Sezgin, M.; Basdogan, C.; Hirche, S. The role of roles: Physical cooperation between humans and robots. *Int. J. Robot. Res.* **2012**, *31*, 1656–1674. [CrossRef]
35. Guo, Y.; Yang, X.J. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *Int. J. Soc. Robot.* **2020**, *13*, 1899–1909. [CrossRef]
36. Chong, L.; Zhang, G.; Goucher-Lambert, K.; Kotovsky, K.; Cagan, J. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Comput. Hum. Behav.* **2021**, *127*, 107018. [CrossRef]
37. Bender, N.; Faramawy, S.E.; Kraus, J.M.; Baumann, M. The role of successful human-robot interaction on trust—Findings of an experiment with an autonomous cooperative robot. *arXiv* **2021**, arXiv:2104.06863.
38. Hanoch, Y.; Arvizzigno, F.; García, D.H.; Denham, S.; Belpaeme, T.; Gummerum, M. The Robot Made Me Do It: Human–Robot Interaction and Risk-Taking Behavior. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 337–342. [CrossRef]
39. Compagna, D.; Weidemann, A.; Marquardt, M.; Graf, P. Sociological and Biological Insights on How to Prevent the Reduction in Cognitive Activity that Stems from Robots Assuming Workloads in Human–Robot Cooperation. *Societies* **2016**, *6*, 29. [CrossRef]
40. Formosa, P. Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy. *Minds Mach.* **2021**, *31*, 595–616. [CrossRef]
41. Karpus, J.; Krüger, A.; Verba, J.T.; Bahrami, B.; Deroy, O. Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* **2021**, *24*, 102679. [CrossRef]
42. Kulms, P.; Kopp, S. More Human-Likeness, More Trust? In Proceedings of the Mensch und Computer 2019, Hamburg, Germany, 8 September 2019; pp. 31–42. [CrossRef]
43. Maehigashi, A.; Tsumura, T.; Yamada, S. Comparison of human trust in an AI system, a human, and a social robot as a task partner. *arXiv* **2022**, arXiv:2202.01077.
44. Salles, A.; Evers, K.; Farisco, M. Anthropomorphism in AI. *AJOB Neurosci.* **2020**, *11*, 88–95. [CrossRef] [PubMed]
45. Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; Graepel, T. Cooperative AI: Machines must learn to find common ground. *Nature* **2021**, *593*, 33–36. [CrossRef] [PubMed]
46. Gillies, R.M. Cooperative Learning: Review of Research and Practice. *Aust. J. Teach. Educ.* **2016**, *41*, 39–54. [CrossRef]
47. Dong, M.; Sun, Z. On human machine cooperative learning control. In Proceedings of the 2003 IEEE International Symposium on Intelligent Control ISIC-03, Houston, TX, USA, 8 October 2003; pp. 81–86. [CrossRef]
48. Mackworth, A.K. On Seeing Robots. In *Computer Vision: Systems, Theory and Applications*; World Scientific: Singapore, 1993; pp. 1–13. [CrossRef]
49. Martins, F.B.; Machado, M.G.; Bassani, H.F.; Braga, P.H.M.; Barros, E.S. rSoccer: A Framework for Studying Reinforcement Learning in Small and Very Small Size Robot Soccer. *arXiv* **2022**, arXiv:2106.12895.
50. Anderson, J.; Baltes, J.; Cheng, C.T. Robotics competitions as benchmarks for AI research. *Knowl. Eng. Rev.* **2011**, *26*, 11–17. [CrossRef]
51. Grush, L. Amazon's Alexa and Cisco's Webex Are Heading to Deep Space on NASA's Upcoming Moon Mission. *The Verge*. 5 January 2022. Available online: https://www.theverge.com/2022/1/5/22866746/nasa-artemis-i-amazon-alexa-cisco-webex-lockheed-martin-orion (accessed on 14 January 2022).
52. Zhang, Y. A Big-Data Analysis of Public Perceptions of Robotic Services Amid COVID-19. *Adv. Hosp. Tour. Res.* **2021**, *9*, 234–242. [CrossRef]
53. Garcia-Haro, J.M.; Oña, E.D.; Hernandez-Vicen, J.; Martinez, S.; Balaguer, C. Service Robots in Catering Applications: A Review and Future Challenges. *Electronics* **2020**, *10*, 47. [CrossRef]