Leveling the Playing Field - Fairness in AI Versus Human Game Benchmarks

Rodrigo Canaan
New York University

Christoph Salge
University of Hertfordshire

Leveling the Playing Field

Fairness in AI Versus Human Game Benchmarks

Rodrigo Canaan
Department of Computer Science and Engineering,
New York University
Brooklyn, NY
rodrigo.canaan@nyu.edu

Christoph Salge
School of Computer Science,
University of Hertfordshire
Hatfield, UK
ChristophSalge@gmail.com

Julian Togelius
Department of Computer Science and Engineering,
New York University
Brooklyn, NY
julian@togelius.com

Andy Nealen
School of Cinematic Arts,
University of Southern California
Los Angeles, CA
anealen@cinema.usc.edu

ABSTRACT

From the beginning if the history of AI, there has been interest in games as a platform of research. As the field developed, human-level competence in complex games became a target researchers worked to reach. Only relatively recently has this target been finally met for traditional tabletop games such as Backgammon, Chess and Go, which prompted a shift in current research focus to electronic games, which provide unique challenges. As is often the case with AI research, these results are liable to be exaggerated or misrepresented by either authors or third parties. The extent to which these games benchmark consist of "fair" competition between human and AI is also a matter of debate. In this paper, we review the statements made by authors and third parties in the general media and academic circle about these game benchmark results and suggest a taxonomy of dimensions to frame the debate of fairness in game contests between humans and machines. In particular, we argue that there is no completely fair way to compare human and AI performance on a game.

CCS CONCEPTS

• **Applied computing → Computer games**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence** .

KEYWORDS

Games, Game AI, AI Benchmarks, Fairness

ACM Reference Format:
Rodrigo Canaan, Christoph Salge, Julian Togelius, and Andy Nealen. 2019. Leveling the Playing Field: Fairness in AI Versus Human Game Benchmarks. In *Proceedings of Foundations of Digital Games (FDG 2019)*. ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.475/123_4

FDG 2019, August 2019, San Luis Obispo, California, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

Games have a long history of being used as Artificial Intelligence testbeds and benchmarks. The formalized representation, especially of board games, and the often explicit and clear reward structure makes them well suited for AI approaches. With recent advances in AI research there is also an emergent narrative that general game-play, the proficiency of playing a range of different games, is a necessary stepping stone towards Artificial General Intelligence (AGI) [40]. Setting this question aside, we want to focus on another related question: Are games a good way to test if an AI has human level intelligence?

This question is often brought up, given that humans are the only existing example of an AGI. Surpassing human level artificial intelligence seems to be a necessary step to reach AGI, and what better way to demonstrate this then to beat a human (or the best human) in a "fair" competition of intelligence, i.e. beat them in a game. Consequently, the media, and to some extend the scientific literature, often characterizes human-AI game competition as a way to determine if AIs has finally reached human level intelligence.

The central argument presented in this paper objects to this characterization. We argue that there is no, and that there can be no real "fair" comparison between an AI and a human that answer the question if an AI has human-level intelligence. To demonstrate the relevance of this argument, we will first give a short survey of how human AI game competitions are currently portrayed, both in the media and in scientific literature. We will then look at several existing examples of human AI game comparison and demonstrate while it is hard to establish a fair comparison for those specific examples. Based on these examples we will then make a more general argument, outlining how the range of different AIs is just too wide, to find any game that offers a fair comparison. As a corollary, we can say that for something to have human-level AI in a meaningful way it would have to be human, or very close to human, in both physical, mental and social embedding.

2 PORTRAYAL OF AI GAME BENCHMARK ACHIEVEMENTS

To set the stage for our discussion about comparing humans and AI benchmarks, we will look at important AI benchmarks in classic board games such as Backgammon, Chess an Go, as well as modern electronic games. We will inspect the claims that were made by the original authors of some systems that achieved success in game benchmarks, and how those results were subsequently discussed in the general media and follow-up academic papers. As an analysis tool, we will use previously published guidelines on how to write about AI research such as [3] and [26]. Our goal in this section is also to illustrate how game AI benchmarks are perceived by society, and what are the main concerns regarding the fairness of comparison between human and AI programs.

We note that the examples gathered in this sections are not meant to be taken as a statistically representative sample of all that has been written about each of these achievements.

2.1 TD Gammon

TD-Gammon [32], is a Backgammon-playing software developed by Gerald Tesauro at IBM using the temporal-difference learning, a reinforcement learning technique where a neural network learns through self-play by minimizing the difference in prediction of the outcome of the game between successful game states. Between 1991 and 1992, it played over a hundred games against some of the best players in the world across three different versions of the algorithm. The last version (TGD 2.1) came very close to parity with Bill Robertie, a former world champion, losing a 40-game series by a difference of a single point.

Tesauro highlights that observing the algorithm play has led to a change in how humans evaluate positions, especially in opening theory for the game. In particular, with some opening rolls, the system preferred "splitting" its back checkers rather than the more risky, but favored at the time option of "slotting" its 5-point. Since then, the splitting opening has been confirmed to be the superior choice by computer rollouts and is now the standard choice for the 2-1, 4-1 and 5-1 initial rolls.

When discussing applicability in other domains, Tesauro lists robot motor control and financial trading as potential applications while cautioning that the lack of a forward model and the scarcity of data might limit the success in these real world environments. Not much discussion of TD-Gammon's achievements was found in general media dating from the time of its release, but Woolsey, an analyst in Tesauro's paper [32] states that TD-Gammon's algorithm is "smart" and learns "pretty much the same way humans do", as opposed to "dumb" chess programs that merely calculate faster than humans.

2.2 Deep Blue

Deep Blue [4] is a computational system for playing chess, designed by a team at IBM led by Murray Campbell. It uses a combination of specialized hardware with software techniques, such as tree-search augmented heuristics for pruning and state evaluation crafted by human experts. It also uses a database of opening moves and endgame scenarios to select moves at the start of a match with little computational effort. It achieved enormous visibility in 1997 when it defeated the reigning champion Garry Kasparov in a six-game match with tournament with a score of $3\frac{1}{2} - 2\frac{1}{2}$. Kasparov had previously beat a former version of the algorithm in 1996 by $4 - 2$.

In their paper describing the system [4], the authors refrain from making speculative claims about the algorithm or its impact in the future of AI. However, the same cannot be said about the media. One article from the Weekly Standard, with the ominous title "Be Afraid" [37], first claimed that the system's "brute force" approach is "not artificial intelligence", but mere calculation.

From there, however, the article argues that in the second game, "Deep Blue won. Brilliantly. Creatively. Humanly" from a position that allegedly does not benefit as much from brute-force calculation. Then, they speculate that this amounts to passing a chess-specific Turing test and, if machines can pass this test, they might eventually pass the more general Turing test and grow beyond our control and understanding. Ultimately, they might become "creatures sharing our planet who not only imitate and surpass us in logic, who have perhaps even achieved consciousness and free will, but are utterly devoid of the kind of feelings and emotions that, literally, humanize human beings".

Other commentators, such as in this New York Times article [33] starts by characterizing both Deep Blue and the human brain as information processing machines, and in this view, the competition is "not a matter of man versus machine but machine versus machine". The main difference is that Kasparov and humans have feelings such as fear and regret, which help control the many activities that can be performed by a human. Deep Blue has none such feelings. However, they speculate that in the future, a potential machine (which they name "Deeper Blue") might be able to model its opponent and even have life goals outside of chess, such as achieving fame. A yet more advanced machine, ("Deepest Blue"), might also have a model of itself, which in their view, might count as consciousness. It might be vulnerable to psychological warfare, at which point humans would again stand a chance in a game of chess.

From these two article, we see how the success of AI chess (which, for humans, requires intelligence) quickly invites speculation about other features of human condition, such as creativity, emotions and consciousness, even though the underlying process of the AI is viewed by some as "mere calculation". We also see how super-human performance in one task automatically invites the thought that one day AI might achieve super-human performance in all tasks, which in some cases is painted as a scenario to be feared.

Another article, from the New York Times [34] focus on Kasparov's own reactions to the match, especially the last one, which Kasparov conceded after 19 moves claiming he had lost his fighting spirit and that he, as a human being, is afraid when faced with something he does not understand. Kasparov also said that the match should have been longer, as he needs time to rest, and that previous games by Deep Blue should be made available. This final remark might be justified by the fact that Deep Blue's "opening preparation was most extensive in those openings expected to arise in match play against Kasparov", although ultimately "none of the Kasparov-specific preparation arose in the 1997 match." [4]. In this article, we see a more nuanced discussion about whether the competition between Deep Blue and Kasparov was fair, in terms of psychology, fatigue and the availability of information. These are topics to which we will come back to in our own discussion about fairness.

2.3 Alpha Go, Alpha Go Zero and Alpha Zero

In 2016, AlphaGo [29], an agent developed by group of Google DeepMind researchers led by David Silver, became the first program to beat a human Go champion in a match against Lee Sedol, in which AlphaGo won by 4 – 1. The system uses a combination of Monte Carlo Tree Search with convolutional neural networks, which learned from professional human games and self play. In 2017, they announced a new version, AlphaGo Zero [30], which learned entirely from self play, with no human examples, and which was able to beat the previous AlphaGo version (AlphaGo Lee). Still in 2017, they announced AlphaZero, which uses a similar architecture (but different input representations and training) to beat other top engines in Go, Chess and Shogi.

The authors claim that the later versions of the system, (i.e., AlphaGo Zero and Alpha Zero) master the games without human help, or *Tabula Rasa*. These claims were scrutinized in a paper by Gary Marcus [16], who views the agent as an example of hybrid system. In particular, he points out the inability of the system to generalize to variations of the game without further training. The system is also unable to learn the paradigm of tree search or the rules of the game, which humans are capable of.

Similar to Deep Blue, AlphaGo and its successors also received wide media coverage. An article from Wired [38] states in its title that AlphaGo and Lee Sedol, together, "redefined the future", referring to two specific moves (which became famous as move 37 and move 78), the first by AlphaGo, the second by the human champion, which defied all expert opinions, and, indeed, were both evaluated by AlphaGo itself by having a probability of being played by a human close to one in ten million. An article by The Washington Post [38] also looks at move 37, and asks experts about its implications for creativity. One interviewee, Pedro Domingos, sees the move as creative, asking “if that’s not creative, then what is?”. Others, such as Jerry Kaplan, attribute the move to clever programming, not creativity of the software.

Another article, titled “Why is Elon Musk afraid of AlphaGo-Zero?” [39], describes the advancements from AlphaGo to AlphaGo Zero as an example of the AI becoming “smart and self-aware” and creating “its own AI which was as smart as itself if not smarter”. The article goes on to wonder about the risks of such an AI being able to generalize from data in other domains, such as in the defense and military industry, and potentially becoming “unsafe”. This is another example of extrapolating performance in a certain task (the game of Go) to performance in unrelated tasks, and raising fear based on this prospect.

2.4 Electronic games

Electronic games (or video games) offer additional challenges to AI researchers compared to traditional tabletop games. Due to a combination of almost continuous time scale (limited by the system’s frame rate) and potentially huge game state space and action space, electronic games are typically even more intractable by brute-force search than games as Go or Chess. As an example, an estimate by Ontañon et al, [18] quotes the state space of Starcraft as 10^{1685}, its branching factor as 10^{50} and its depth as 36000, whereas Go has corresponding values of roughly 10^{170}, 300 and 200. As such, a number of video game AI benchmarks have been proposed. While the use of video games as AI benchmarks goes back a long way, interest in these benchmarks has spiked since AlphaGo’s results of 2016, as Go, which was considered among the most challenging tabletop games, was finally beaten and new, harder challenges had to be explored.

Some of these benchmarks encourage the development of general techniques, that can be applied for a large number of domain problems, such as different games. That is the case of frameworks such as the Arcade Learning Environment (ALE) [2], where agents can be evaluated in one of hundreds of Atari 2600 games and the General Video Game AI Competition [23], where agents are evaluated in previously unseen arcade-like games.

Other examples benchmarks proposed for specific games are Vizdoom [11] (first person shooter), the Mario AI Benchmark [10] (platform game) and even benchmarks not focused on winning a game, but building a level for a platform game [28] or, inspired by the Turing test, playing in a way that is indiscernible from humans [9].

While all these benchmarks have garnered academic interest, none has arguably received as much general media coverage and player attention as AI challenges in the for Starcraft [36] after Google DeepMind and Blizzard, the game’s publisher, released an reinforcement learning for the game, and Valve’s Dota 2, where different versions of agents developed OpenAI went from defeating one of the best players in the world in a limited 1v1 version of the game in a showmatch in an official Valve tournament in 2017 [19], to defeating a team of 5 semi-professional players in the 99th percentile of skill in another showmatch in 2018 [21, 22] before eventually losing to professional players in a showmatch during The International 8 [20], the biggest Dota 2 event of 2018. The fact that both Starcraft and specially Dota 2 are popular eSports seems to have helped garner a lot of attention from the community of players as well.

Starcraft-playing agents are still unable to beat top human players, which has probably contributed to tone down the amount of speculation, but media outlets (and some researchers) seem to be betting on an AI victory in the near future [25].

For the remainder of this section, we will focus on Dota 2 AI media coverage, whose trajectory has been full of ups and downs and controversy.

A major point of debate has been the way the OpenAI agent visualizes and interacts with the game, as described in [21]. The high level features used by the agent in its observations allows it to "see" at any point in time, information such as the remaining health and attack value of all units in its view. A human would have to click on each unit, one by one, to view this information. Agents can also specify its actions at a high level by selecting ability, target, offset and even a time delay (from one to four frames). A human would have to make a combination of key presses and imprecise mouse movements to achieve the same effect.

An article on Motherboard [17] has described the advantages provided to the AI as “basically cheating”, summing up that, while humans have previously been disqualified from tournaments due to the use of illegal programmable macro actions, “Open AI Five plays like an entire team with programmable mice and telepathy”. The article also proposes that the agent should learn directly from visuals.

In a blog post [6], AI researcher Mike Cook, while ultimately having a positive view on the benchmark, also comments on the interface advantages, drawing attention to some highlights of the games where, even though the agents have a reaction speed of 200ms (in theory comparable to humans), they executed key actions such as interrupting a spell or coordinating powerful abilities in a way that is seemingly impossible for humans. Cook also warned about the potential of the AI to fall prey to techniques it has never encountered in its self play (such as the technique of pulling or unusual hero lineups) and that good performance in a few facets of the game (such as teamfighting) might give the illusion of greater overall competence in the game.

A final critique against OpenAI's agents came from the number of simplifications that had to be made to tackle a game as complex as Dota 2, such as playing with a reduced Hero pool, the innability to fight Roshan (a powerful NPC that tipically takes risky team-wide efforts to kill, but drops a valuable reward and is often the focus of game-deciding fights between teams) and the choice to have individual invulnerable couriers per player (as opposed to a vulnerable courier shared by the entire team). These demands can be seen in game forums such as [14, 24] and ultimately led to OpenAI's decision to drop most restrictions in preparation for the final matches at The International 8, which OpenAI lost [20].

3 DIMENSIONS OF FAIRNESS

We propose a (non-exhaustive) taxonomy of dimensions of fairness, each representing one possible interpretation of what is meant by asking whether a given competitive scenario between different game-playing systems (such as humans and AI agents) is fair.

- Perceptual fairness: do the two systems have the same input space, e.g. pixels from a screen? This is especially relevant in electronic games, where it might be the case that a human can only see the limited information on the screen, and needs to make specific actions to gather more information, such as scrolling the viewing window or clicking units to see their status. In contrast, an algorithm might have as input a structured list with the location and status of game objects. Even in cases where the input is ostensibly the same (such as an algorithm playing from pixels), is the image capturing apparatus equivalent to a human eye? As an example, the human retina has blind spots and lower peripheral resolution [31]. Does the computer suffer from the same limitations?
- Motoric fairness: do we have the same output space? The same actions to choose from? The same reaction speed? Does the computer have to actually press physically buttons or move pieces on a board? Does it have to deal with limited strength, speed and imprecision of human muscles?
- Historic fairness: have the human and the machine spent the same amount of time playing the game? Has it played the game under the same conditions? What about time spent playing closely related games?
- Knowledge fairness: have the agents had access to the same declarative knowledge (compiled by others) about the game? For example opening books, pre-trained neural networks, tables of state values?
- Compute fairness: do the agents have the same computational power? How should computational power be measured in this case? Some possible metrics are power consumption, number of neurons or storage capacity. These metrics have the shortcoming that, in the case of humans, it is not easy to determine how much of these resources are being used to play the game as opposed to other essential bodily functions. Another alternatives could be based on the number or depth of game states explored during tree search, which tends to heavily favor computers. Finally, the monetary cost of the total computation and the infrastructure required could also be considered, especially in the context of discussing the possibility that a machine might replace humans in a task after achieving "super-human performance", which is impractical if the artificial is several times more expensive than hiring a human to do the same task.
- Common-sense fairness: do the agents have the same knowledge about other things, such as may factor into the game? Have they gone to school? Do they know that dragons are typically dangerous and coins are typically desirable? Have they seen an American traffic sign? Do they know you can bribe police, or that walls with cracks in them are more likely to break when you bomb them? Do they understand instructions given in natural language by NPCs? With current technology, this is a dimension that tends to heavily favor humans.

Each of the dimensions is actually a continuum rather than a binary, and, as will be discussed below below, not all of them are equally relevant to all games. Regardless, it should be clear at this point that even determining what constitutes fairness at a game is a challenging task. We will argue that achieving complete fairness in all dimensions would require building a system that is essentially equivalent to a real, flesh-and-blood human.

4 A DISCUSSION OF FAIRNESS IN HUMAN-AI GAME BENCHMARKS

At a first glance, the issue of fair conditions in between a computer agent and a human seems more tractable in tabletop games, such as Backgammon, Chess and Go, than in electronic games. A major difference between the two domains seems to be that what we called Perceptual and Motoric fairness are less applicable, in a first glance, to tabletop games. As an example, AlphaGo required a human facilitator to input the current board state into the system, and to apply the move selected by the algorithm to the physical board. Modifying the system with a camera to read the board state and a robotic arm to move the pieces might be interesting Computer Vision and Robotic problems on their own, but it would be hard to argue that it would constitute a better Go player or that the competition with humans would be fairer.

Due to this significant difference, we divide discussion below between tabletop games and electronic games. Issues discussed for tabletop games in general also apply to electronic games, but the reverse is not true.

Another category that could be considered, but is not the focus of this work, is that of games involving direct interaction with the real world, such as the Robocup [13], where soccer is played between robots . These raise even more questions about fairness in the sensorimotor sense, as the physical speed, strength, weight and dimensions of the robots have to be constrained to a similar range that that of humans.

4.1 Fairness in Tabletop games

The first key issue affecting the fairness between human and artificial players in tabletop games are feelings such as fatigue, fear, anxiety, etc. In [34], Kasparov comments on the role these factors can play in a match. In [35], Ke Jie, another prominent Go player who has also lost to AlphaGo, stated that psychological factors are possibly “the weakest part of human beings”. It is a regular occurrence for sports commentators to also build a narrative around the mental factors going into an important match, especially one where a lot of pride or money is involved. The magnitude of the psychological effect is unclear from this brief study, but, to the degree in which it might change the outcomes, compensating for it is also not trivial. There is no straightforward way to account for these emotions in a computer simulation, and attempt to do so (e.g. by artificially injecting noise in the algorithm’s evaluation in situations of high stress) would defeat the purpose of building the best possible game-playing systems. These could all be considered examples of Motoric fairness, as the output of a human (its selected actions) are subject to limitations (in this case, psychological) that are not present in the computer system.

A second issue that can be raised is the use of look-up tables for specific points of a match, such as the opening and endgame, and the availability of information about a specific opponent in a match. These could be considered as considerations about Knowledge fairness. Look-up tables have been used in Deep Blue [4] and suggested as a potential improvement for TD-Gammon’s identified weakness in endgame situations. [32]. The use off similar resources in most competitive matches between humans is banned, but when playing versus a computer, should a human have access to the same tables that are available to the algorithm? Similarly, if an algorithm is capable of studying examples of human play in general (as is the case for the original AlphaGo [29] or even have some of its parameters or design decisions tuned to face a specific human player (as happened with Deep Blue [4]), wouldn’t it be fair for a human to review a large number of games by an artificial agent, receive a detailed summary of its preferred openings and strategies, perhaps even inspect the source code?

In the same vein, the use of a forward model to simulate future game states, as is done during tree search in Deep Blue [4] and the rollouts of AlphaGo [29] could be considered an issue of compute fairness . This could be compared to giving a human a set of extra boards and pieces with which to simulate potential lines off play during a match, which is again not allowed in competitive play. An important observation is that while it is possible for a human to fully simulate a game of Chess or Go in this way, it is harder to do the same for games that involve randomness and hidden information, and simulation becomes even harder for an unassisted human if the task involves continuous dimensions such as time and distance.

We would like to relate these issues to what is called System 1 thinking and System 2 thinking in dual-process theory [7]. System 1 thinking, often called intuition or heuristic thinking, has been related in the Reinforcement Learning context to the selection of actions without lookahead [1], such as using a look-up table for openings or a neural network pre-trained on a database of game states. System 2 thinking consists of conscious analytical reasoning, and has been related to Tree Search [1]. While both types of thinking in humans could be augmented for humans through the use of external tools such as notes on a piece of paper or extra boards for simulation of lines of play, an argument could be made following the Extended Mind [5] that whether such resources are internal or external to a system makes little difference when considering the system’s cognitive abilities. Taking this argument to the extreme, we could conceive of a situation where a complete artificial game-playing system is viewed as mere augmentation of a human’s cognitive abilities, leading to the absurd scenario of a "human versus AI" match where nonetheless all moves are selected by the same algorithm, one playing for itself, the other in the human’s stead.

In the opposite direction, we could attempt to reduce the computer’s advantage by restricting what kinds of techniques it is allowed to use, disallowing the ones that are viewed as inherently unfair. Ultimately, however, all of the aforementioned advantages could be reduced to following instructions on a paper, and it could be argued, similarly to the Chinese Room [27] thought experiment, that no AI achievements are ever to be considered as proof of mastery in a game.

A final, unrelated issue is the machine’s ability to generalize its learnings across different games or variations of the same game. According to Brooks [3], humans are prone to infer competence from performance. As humans, we might expect a system that performs as the best Go player in the world to be competent enough to play on a board of different dimensions, or play with a different goal (such as the intent to lose) or be at least a passable player in another similar game (such as chess). Marcus [16] points out that this is not the case with most existing techniques, and addressing this issue the motivations behind competition frameworks such as ALE [2], GVGAI [23] and GGP [8]. While this doesn’t strictly affect the fairness of competitions based on playing a single game, with a single ruleset, it is an important point to consider against the narrative that sees the success of AI in a new task as evidence that AGI is just around the corner.

4.2 Fairness in Electronic Games

The major difference between tabletop games and electronic games when it comes to perception of fairness seems to be rooted on the representation of the observation and action space, as well as reaction time, as discussed in [6, 17], or what we call perceptual and motoric fairness.

Regarding the observation space, a common paradigm to solve the issue is playing the game from pixels, rather than from higher level game features. This is the approach followed by Vizdoom [11] and ALE [2]. While the approach can be said to more closely emulate the way humans perceive video games, the comparison is not perfect. On one hand, favoring the AI, questions such as “is the difference between these two objects smaller than X?” are still much easier to answer accurately for an agent playing from pixels than for a human, and this perceptual advantage could benefit the agent when aiming an ability that affects area within a certain radius, for example. On the other hand, when a human sees pixels in the shape of a coin, a spider and fire, they can reasonably infer that the first object has to be collected, the second attacked and the third avoided, and such heuristic would work well for many games (what we call the common-sense dimmension of fairness). Embedding this representation and real-world knowledge in a visual AI system is an unsolved problem, which provides humans with an advantage that is not easy to surmount at the moment.

While objections to high level representations are valid, taken to the extreme, these objections would imply that no meaningful advancements could be made in video game-playing AI before the field of computer vision is essentially solved. This would be disappointing from a game AI perspective. After all, low level recognition of pixel patterns is not what immediately comes to mind we picture a human expertly playing a game. Results obtained on less structured or more general representations can fairly characterized to be more impressive, but the challenges involved in dealing with lower level representations don't necessarily capture what makes games such interesting AI problems in the first place.

For this reason, considerations about the input represetnation shouldn't be a barrier for game AI research, especially in environments where humans currently have the upper hand. We believe novel results using higher level representations are important, and further research that attempts to replicate these results while using less favorable or more general representations are also important and will likely naturally follow the initial results.

Similarly, the representation of the action space can take many forms, ranging from high level representations where an action is viewed as a tuple of [ability, target, offset] present in OpenAI's methodology [21], to simulating medium level user interface commands such as screen movements and unit highlighting in Starcraft [36] to directly simulating a virtual controller as in ALE [2]. The extreme position would be to insist on a robotic arm manipulating a physical controller or keyboard, which would again distract researchers from other legitimate game AI problems that can be tackled with higher level representations.

Excessively fast reaction speed is often cited as one of the factors that make an agent play in a perceived artificial fashion [12]. A popular solution, used by OpenAI [21] is to directly enforce a specific reaction time. Alternate solutions involve the "Sticky Action" and other methods discussed in [15] for the ALE environment. Interestingly, its original motivation was not to emulate human play, but to provide enough randomness to the otherwise deterministic ALE environment to force the agent to learn "closed loop policies" that react to a perceived game state, rather than potential "open loop policies" that merely memorize effective action sequences, but also works to avoid inhuman reaction speeds.

CONCLUSION

We have described in brief detail some of the most relevant game AI benchmark results in the past three decades, for both tabletop games (Backgammon, Chess and Go) and electronic games (specially Starcraft and Dota 2) and looked at some of the claims made by the authors of these game-playing systems and some third-party comments made by general media and research communities. From those, we conclude that there is a tendency to extrapolate from AI achievements in game Benchmarks to speculation about Artificial General Intelligence (AGI) scenarios where AI will eventually beat humans in all or most tasks. We have also seen examples of public concerns about the fairness of these benchmarks.

We have proposed a taxonomy of dimensions on which to evaluate fairness in a competition between two game-playing systems (such as a human and an AI agent). We provided examples of how these apply to tabletop games and electronic games, noting that there is greater focus on the Perceptual and Motoric dimensions of fairness for electronic games.

Ultimately, we argue that there are so many possible games, and so many possible architectures of game-playing agents, differing so widely in the dimensions of fairness, that it is impossible to infer human-level intelligence from success in any single game, and that a completely fair competition can only be achieved against an artificial system that is essentially equivalent to a flesh and blood human.

This conclusion should not serve to infer since complete fairness is impossible, accomplishments in game-playing AI are meaningless in general, or that benchmarks that score low in our fairness dimensions are not valuable. We observe that, usually, when a significant benchmark is reached, significant research often follows in order to make the system less reliant on human expertise, less reliant on specific hardware and impractical amounts of computing power, less reliant on extremely fast reaction speeds and more generalizable to similar, but different problems, leading to systems with fewer restrictions and wider applications.

REFERENCES

[1] Thomas Anthony, Zheng Tian, and David Barber. 2017. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems*. 5360–5370.

[2] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[3] Rodney Brooks. 2017. The Seven Deadly Sins of AI Predictions. https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

[4] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.

[5] Andy Clark and David Chalmers. 1998. The extended mind. *analysis* 58, 1 (1998), 7–19.

[6] Mike Cook. 2018. OpenAI Dota 2: Game Is Hard. http://www.gamesbyangelina.org/2018/08/openai-dota-2-game-is-hard/

[7] Jonathan St BT Evans. 1984. Heuristic and analytic processes in reasoning. *British Journal of Psychology* 75, 4 (1984), 451–468.

[8] Michael Genesereth, Nathaniel Love, and Barney Pell. 2005. General game playing: Overview of the AAAI competition. *AI magazine* 26, 2 (2005), 62.

[9] Philip Hingston. 2009. A turing test for computer game bots. *IEEE Transactions on Computational Intelligence and AI in Games* 1, 3 (2009), 169–186.

[10] Sergey Karakovskiy and Julian Togelius. 2012. The mario ai benchmark and competitions. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 1 (2012), 55–67.

[11] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. 2016. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*. IEEE, 1–8.

[12] Ahmed Khalifa, Aaron Isaksen, Julian Togelius, and Andy Nealen. 2016. Modifying MCTS for Human-Like General Video Game Playing.. In *IJCAI*. 2514–2520.

[13] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. 1997. Robocup: The robot world cup initiative. In *Proceedings of the first international conference on Autonomous agents*. ACM, 340–347.

[14] Team Liquid. 2018. OpenAI's Dota 2 bots vs. 5 top professionals in TI. https://www.liquiddota.com/forum/dota-2-general/534977-openais-dota-2-bots-vs-5-top-professionals-in-ti

[15] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. 2017. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *arXiv preprint arXiv:1709.06009* (2017).
[16] Gary Marcus. 2018. Innateness, AlphaZero, and Artificial Intelligence. *arXiv preprint arXiv:1801.05667* (2018).
[17] Motherboard. 2018. OpenAI Is Beating Humans at âĂŸDota 2âĂŹ Because ItâĂŹs Basically Cheating. https://motherboard.vice.com/en_us/article/gy3nvq/ai-beat-humans-at-dota-2
[18] Santiago Ontanón, Gabriel Synnaeve, Alberto Uriarte, Florian Richoux, David Churchill, and Mike Preuss. 2013. A survey of real-time strategy game AI research and competition in StarCraft. *IEEE Transactions on Computational Intelligence and AI in games* 5, 4 (2013), 293–311.
[19] OpenAI. 2017. Dota2. https://blog.openai.com/dota-2/
[20] OpenAI. 2018. The International 2018: Results. https://blog.openai.com/the-international-2018-results/
[21] OpenAI. 2018. OpenAI Five. https://blog.openai.com/openai-five/
[22] OpenAI. 2018. OpenAI Five Benchmark: Results. https://blog.openai.com/openai-five-benchmark-results/
[23] Diego Perez-Liebana, Spyridon Samothrakis, Julian Togelius, Simon M Lucas, and Tom Schaul. 2016. General video game ai: Competition, challenges and opportunities. In *Thirtieth AAAI Conference on Artificial Intelligence*. 4335–4337.
[24] Reddit. 2018. Team Human vs. OpenAI Five Match Discussions. https://www.reddit.com/r/DotA2/comments/94udao/team_human_vs_openai_five_match_discussions/
[25] MIT Technology Review. 2017. Humans Are Still Better Than AI at StarCraftâĂŤfor Now. https://www.technologyreview.com/s/609242/humans-are-still-better-than-ai-at-starcraftfor-now/
[26] Oscar Scwartz. 2018. 'The discourse is unhinged': how the media gets AI alarmingly wrong. https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong
[27] John Searle. 1999. The Chinese Room. (1999).
[28] Noor Shaker, Julian Togelius, Georgios N Yannakakis, Ben Weber, Tomoyuki Shimizu, Tomonori Hashiyama, Nathan Sorenson, Philippe Pasquier, Peter Mawhorter, Glen Takahashi, et al. 2011. The 2010 Mario AI championship: Level generation track. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 4 (2011), 332–347.
[29] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.
[30] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354.
[31] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. 2011. Peripheral vision and pattern recognition: A review. *Journal of vision* 11, 5 (2011), 13–13.
[32] Gerald Tesauro. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.
[33] The New York Times. 1997. Deep, Deeper, Deepest Blue. https://www.nytimes.com/1997/05/18/weekinreview/deep-deeper-deepest-blue.html
[34] The New York Times. 1997. Swift and Slashing, Computer Topples Kasparov. https://www.nytimes.com/1997/05/12/nyregion/swift-and-slashing-computer-topples-kasparov.html
[35] The New York Times. 2017. GoogleâĂŹs A.I. Program Rattles Chinese Go Master as It Wins Match. https://www.nytimes.com/2017/05/25/business/google-alphago-defeats-go-ke-jie-again.html
[36] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).
[37] WeeklyStandard. 1997. Be Afraid. https://www.weeklystandard.com/charles-krauthammer/be-afraid-9802
[38] Wired. 2016. IN TWO MOVES, ALPHAGO AND LEE SEDOL REDEFINED THE FUTURE. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/
[39] WorldAIShow. 2018. Why is Elon Musk afraid of AlphaGo-Zero? https://singapore.worldaishow.com/elon-musk-afraid-alphago-zero-ai/
[40] Georgios N. Yannakakis and Julian Togelius. 2018. *Artificial Intelligence and Games*. Springer. http://gameaibook.org.