When Are We Done with Games?

Niels Justesen
IT University of Copenhagen
Copenhagen, Denmark
noju@itu.dk

Michael S. Debus
IT University of Copenhagen
Copenhagen, Denmark
msde@itu.dk

Sebastian Risi
IT University of Copenhagen
Copenhagen, Denmark
sebr@itu.dk

***Abstract*—From an early point, games have been promoted as important challenges within the research field of Artificial Intelligence (AI). Recent developments in machine learning have allowed a few AI systems to win against top professionals in even the most challenging video games, including Dota 2 and StarCraft. It thus may seem that AI has now achieved all of the long-standing goals that were set forth by the research community. In this paper, we introduce a black box approach that provides a pragmatic way of evaluating the fairness of AI vs. human competitions, by only considering motoric and perceptual fairness on the competitors' side. Additionally, we introduce the notion of extrinsic and intrinsic factors of a game competition and apply these to discuss and compare the competitions in relation to human vs. human competitions. We conclude that Dota 2 and StarCraft II are not yet mastered by AI as they so far only have been able to win against top professionals in limited competition structures in restricted variants of the games.**

I. Introduction

Games provide a suite of diverse and challenging tasks, helping to benchmark progress in the field of Artificial Intelligence (AI). Performing well in renowned games such as Chess, Go, Dota 2 (Valve Corporation, 2013), and StarCraft II (Blizzard Entertainment, 2010) are important milestones in AI; due to their immense computational complexities, playing them well seemed to require some form of human-level intelligence. These games are, after all, designed to challenge the human mind on a highly competitive level. However, due to the increase in computational resources and recent developments in machine learning, deep neural networks now seem to achieve human-level control in these games, from self-play alone or with the help of human demonstrations. With DeepMind's 'AlphaGo' beating the world's best Go player Lee Sedol, OpenAI's 'Dota Five' beating a team of professionals in a restricted game of Dota 2, and DeepMind's 'AlphaStar' beating the professional StarCraft II player Grzegorz 'MaNa' Komincz on one selected map, it may seem that AI has now achieved all of these long-standing goals that were set forth by the research community. So where does this leave us? Are we, as AI researchers, done with games?

This paper provides a discussion of designing and evaluating fairness in human vs. AI game competitions. Ultimately, we argue that a claim of superiority of AIs over humans is unfounded, until AIs compete with and beat humans in competitions that are structurally the same as common human vs. human competitions. These competitions are, after all, designed to erase particular elements of unfairness within the game, the players, or their environments.

We take a black-box approach that ignores some dimensions of fairness such as learning speed and prior knowledge, focusing only on perceptual and motoric fairness. Additionally, we introduce the notions of game *extrinsic* factors, such as the competition format and rules, and game *intrinsic* factors, such as different mechanical systems and configurations within one game. We apply these terms to critically review the aforementioned AI achievements and observe that game extrinsic factors are rarely discussed in this context, and that game intrinsic factors are significantly limited in AI vs. human competitions in digital games.

Claimed AI achievements in games were also reviewed by Canaan et al. [14], focusing on how researchers and the media have portrayed the achieved results. Additionally, they proposed six dimensions of fairness in human vs. AI game competitions: 1) Perceptual: do the two competitors use the same input space? 2) Motoric: do the two competitors use the same output space? 3) Historic: did the two systems spend the same amount of time on training? 4) Knowledge: do the two systems have the same access to declarative knowledge about the game? 5) Compute: do the two systems have the same computational power? 6) Common-sense: do the agents have the same knowledge about other things? Based on their six dimensions of fairness, they concluded that "a completely fair competition can only be achieved against an artificial system that is essentially equivalent to a flesh and blood human". We will return to their evaluation in a later paragraph.

Our main critique of current evaluation procedures is twofold. AI superiority in games cannot be claimed without carefully treating the game extrinsic and intrinsic factors, such as the competition's structure and rules, and the game's configurations. First of all, it is necessary that AIs compete in a game's extrinsic tournament structures, as already employed in human vs. human competitions. Thus, to formulate a proper competition between humans and AI systems, we argue that we must first study the extrinsic game factors that are in play when humans are competing and then formulate an experimental setup that imitates them, without limiting the game's intrinsic variables, such as maps, races, heroes, etc., as is currently be done.

Through the discussion of two areas (the competitors perceptual and motoric abilities; and the game's extrinsic and intrinsic factors) we hope to show that, so far, no fair compe-

978-1-7281-1884-0/19/$31.00

tition between AIs and human has occurred in Dota 2 and StarCraft II. We argue that, if these factors are accounted for in the future, and we ignore the competitors individual characteristics of knowledge acquisition (considering them as black boxes), we can construct a competition that is capable of producing a fair evaluation of the competitors' *output*. This output can then form the basis for future discussions on AI vs. human intelligence. For instance, if the AI wins in our hypothetically fair competition, does this mean it is more intelligent? If the human wins, what are the areas in which the AI has to improve? Are there still other factors that we did not observe and account for? In this manner, our current approach does not claim ultimate correctness, but constitutes a step forward in the area of human vs. AI competition, by critically evaluating the current state and proposing areas of consideration for future competitions. Thus, our approach is different from Canaan et al.'s [14] in that they claim the impossibility of a fair competition due to differences in the competitors, which we simply treat as black boxes and focus on the competition instead.

While this paper focuses on claims of super-human performance in games, there are other claims made in this context that are worth discussing; for example, whether a system has learned from *tabula rasa* [29]. Due to our black-box approach, these questions will not be discussed in this paper.

In particular, we aim to address the game AI community with this paper and its discussion of how to create a competition that enables us to claim AI superiority. The implications for players, communities, game designers and game studies researchers outside of AI are outside the scope of this paper. Furthermore, the scope of this paper is limited to only considering AI systems in the role of playing a game competitively. It is, however, important to not neglect that AI/CI has many other roles for games such as procedural content generation, player modeling, and data mining [50]. When we ask the question *are we done with games?*, we are thus only concerned with the traditional branch of AI in games where the goal is to create a system that plays competitively.

II. Games for AI

Since the birth of the research field of AI in 1956, games have been a vital test bed to compare algorithms and measure the progress in the field. Several games have through the years been promoted by researchers as key challenges for AI with the ultimate goal of defeating the best human players. Initially, traditional board games stood the test, such as Checkers, Backgammon, and Chess. Shannon wrote in his seminal paper on computational Chess, that "chess is generally considered to require 'thinking' for skillful play; a solution of this problem will force us either to admit the possibility of a mechanized thinking or to further restrict our concept of 'thinking'" [40] and similarly Herbert Simon wrote "if one could devise a successful chess machine one would seem to have penetrated the core of human intellectual endeavor" [42]. Today, it is undisputed that computers play these games at a super-human level. After Chess, Go was promoted as the next grand AI challenge due to its significantly higher complexity [9, 11, 22, 28, 30]. Demis Hassabis, the CEO of the company DeepMind that eventually developed the first computer system to beat a human professional in Go, has described Go in an interview as the "Mount Everest" for AI scientists[1].

Video games, on the other hand, present a new array of challenges for AI, which has led researchers to promote the Atari 2600 arcade video games [10], Dota 2 [20], and StarCraft [12, 32, 45, 47] as even harder challenges. Researchers have considered StarCraft to be the hardest games for computers to play [16, 51], which ultimately suggests that this game is the final grand challenge for AI in games before tackling real-world problems.

III. The Blackbox Approach

In this paper, we propose a pragmatic way of evaluating AI against human intelligence in game competitions. To be able to do this, some obvious differences have to be pointed out and disregarded. First of all, for the purpose of comparability, we consider AI systems as black boxes. Their training, knowledge, common sense and idiosyncratic function will not be considered. Treating an AI as a black box thus disregards four of the six dimensions of fairness introduced in [14]. This leaves us with perceptual and motoric fairness as they deal with the system's interaction with the game. In fact, human vs. human competitions also take this approach: we ignore how contestants have prepared for the competition and their IQ scores are not considered important. Only in few cases (e.g. weight in boxing or gender in physical sports) is it deemed necessary to impose further restrictions and limitations. In the case of electronic sports (eSports), gender segregation is a topic of an ongoing debate, which to cover would exceed the scope of this paper. However, to explicate the 'obvious differences' that will be disregarded, a discussion of types of potential superintelligences by Nick Bostrom is useful. While his position on the emergence of superintelligences and its consequences are arguable, we deem his threefold distinction as useful here, to describe and understand ways in which machines are simply and obviously different from humans.

In the book *Superintelligence* [8], Bostrom distinguished three different forms of possible superintelligences [8, p.63] that we will use to explicate the aforementioned differences. *Speed intelligence* describes a superintelligence that "[...] can do all that a human intellect can do, but much faster" [8, p. 64]. AI systems can usually speed up matches and play them much faster than a human player can. It is thus an obvious difference that AIs can train *faster* than humans if the number of games over time (not the learning outcome) is the measure of speed. *Collective intelligence* describes "a system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system" [8, p. 65]. In the context of learning, many machine learning algorithms fit this description, as neural networks usually are trained in

[1] http://www.taipeitimes.com/News/front/archives/2016/03/14/2003641528

several instances in parallel to ultimately combine the gathered information into one system – an option that a human player does not have. In fact, AlphaGo, Dota Five, and AlphaStar all relied heavily on these advantages. AlphaGo played 1,280,000 games against itself using 50 GPUs [41], Dota Five played around 180 years of real-time Dota per day on 256 GPUs and 128,000 CPU cores [36], and for AlphaStar, a league of agents each agent played up to 200 years of real-time StarCraft, each using 16 TPUs [48].

Finally, Bostrom describes *quality superintelligence* as one system that is "[...] at least as fast as a human mind and vastly qualitatively smarter" [8, p. 68]. Simply put, this resembles the difference between two humans with different IQs; or more theoretically, the encounter with an alien race that "thinks on a different level", incomprehensible to us. This quality super intelligence is one potential interpretation of the outcomes of a fair competition between AIs and humans.

Especially when it comes to learning and training, the AIs have advantages over humans in what Bostrom called speed and collective intelligence. During training, developers can speed up games to a degree which exceed human capabilities. In addition to that, one version of a system can play hundreds of games simultaneously and gather the gained information afterward. This is a strength of artificial intelligence that we should embrace and not handicap. This factor is thus ignored by our black box approach.

As discussed here, AI systems and humans are different in many regards, and we agree with Canaan et al.'s [14] conclusion that a final, holistic comparison of the two is nonsensical. However, in the current paper, we want to discuss what happens if we exclude the obvious differences (as partially done in human vs. human competitions as well). Our interest lies in an evaluation of whether, or to what extent, the competitions between AI systems and humans were actually carried out on equal grounds. In other words, we exclude the characteristics of the participants, to evaluate the characteristics of the competition.

To be able to discuss an AI vs. human competition within games, we need a prototypical example of a human vs. human competition and a rudimentary distinction of how such competitions regulate game extrinsic and intrinsic factors. The purpose of this is to establish certain terms that enable an in-depth discussion and comparison of AI vs. human competitions and their idiosyncratic structures.

IV. A Prototypical Human Competition

For the following discussion, some definitions of terms are necessary. First of all, the questions *what a game is* and what should be considered parts of games have been considered in many publications [5, 6, 13, 25, 43, 44] as well as some heated debates [1, 21, 26, 31]. The authors accept that ultimately defining games might be an impossible task that bears normative and discriminatory potential. However, for the current purpose of structuring our argument and observations, we will develop a makeshift model of game intrinsic and extrinsic factors. This model will be based on previous research on the ontology of games [3], and metagames [15, 18].

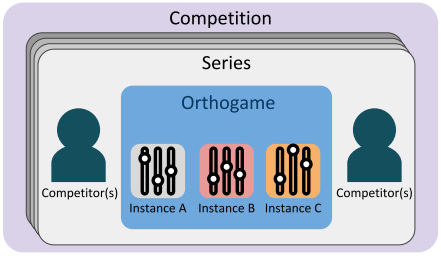

Fig. 1: A prototypical human vs. human competition consisting of one or more series between two teams or individuals. Each series consists of multiple game instances of the same orthogame.

Aarseth and Calleja's cybermedia model [3] constitutes a descriptive model that covers games but intentionally also other phenomena. They describe games as a player's perspective on a cybermedia object, which consists of a *materiality*, a *sign system*, and a *mechanical system*. Especially in the case of the mechanical system, it is possible that one cybermedia object contains several systems. Their example is *World of Warcraft* (Blizzard Entertainment, 2004), which contains multiple mechanical systems, such as questing, raiding, PvP arena, and PvP battlegrounds.

To avoid the confusion that the term *game* brings with it and to avoid the processual perspective on games that Aarseth and Calleja take[2], we will use Carter et al.'s term "orthogame" [15] instead of Aarseth and Calleja's "cybermedia objects" [3]. The orthogame describes "[...] what players collectively consider to be the 'right and correct game'" [15]. We understand the orthogame as the digital artifact installed on a computer or physical artifact as used for play (including its rules). This explicitly excludes the player from the object itself. It is important to note that especially digital orthogames have various different 'starting configurations', i.e. maps or chosen races. These starting configurations determine individual *game instances*: subsets of the orthogame with one particular configuration. A *series* occurs between two teams or individuals across multiple instances of the orthogame, i.e. a best-of-five series. Finally, all of these concepts are encompassed by an "added metagame" [18, p. 5]. Added metagames are structures regulating leagues, ladders, tournaments, competitions, etc.

We will now put these terms into work in an example of a StarCraft II instance. It must be noted that it is a prototypical example and a more detailed model could be drawn (as discussed at the end of this section). However, the developed terms will still be applicable in those cases, even though the structure could be expanded. Figure 1 illustrates this prototypical version of a human vs. human competition.

In our example, two players are competing for the world championship in StarCraft II. Over the course of the last months, they both proceeded through an *added metagame*: a ladder, as well as a KO system in the finals, which are

[2]One of the unsolved problems regarding games is the question whether they are objects of processes [2] The ontological commitment of a process perspective onto games is that the player constitutes an element of the game itself; a perspective that the authors do not share in the current endeavor.

held at a physical location. Now they face off in the grand final. The grand final is constituted by a best-of-five series. This means that the players face each other at least three times, playing the same orthogame (StarCraft II), but different instances of it. These instances (circle, triangle, and diamond shapes) are usually determined by one of two processes. On the one hand, there exists a "material metagame" [18, p. 5], which encompasses drafting armies or heroes in some games. In our StarCraft II example, there is a map selection procedure before a series, where players can veto maps that will be removed from the map pool. On the other hand, the particular configurations can be regulated by the added metagame beforehand, such as 1 vs. 1 competitions in Dota 2. These limit the orthogame to a particular player composition (two players) and a spatial layout (middle lane only). Thus, the example actually constitutes a combination of both processes, through the limitation of a map pool in StarCraft II (added metagame restriction), and the subsequent selection of maps from the pool by the players (material metagame process).

We can expand the model and include the whole added metagame of the world championships by adding more series to Figure 1, as indicated by the additional 'series' frame. These additional series, in turn, consist of (potentially) differently configured instances of the orthogame, played by different players. Another possible constellation is an added metagame between the same players but within different orthogames. However, we further argue that the function of added metagames is to regulate game extrinsic and intrinsic factors within the added metagame, with the purpose of balancing and fairness. We will elaborate on these concepts in the following sections.

V. Game Extrinsic and Intrinsic Factors

To reiterate the previous section, we can split a competition into three general areas: the added metagame, series within the added metagame and instances of an orthogame played within the series. We argue that the function of the added metagame is to regulate both, the format of the series (extrinsic factors), as well as the particular configurations of the orthogame (intrinsic factors). The extrinsic and intrinsic factors of regulation will be exemplified in the following sections.

A. Fairness: Game Extrinsic Factors

Added metagames generally regulate two game extrinsic areas in a competition: the structure or format of the competition and the competitors participating in it. An example of an added metagame are ladder system, as implemented in many contemporary online multiplayer games, such as League of Legends (Riot Games, 2009), Heroes of the Storm (Blizzard Entertainment, 2015) or StarCraft II. In the world championship of soccer, a mini-league system, in which teams play in discrete groups, and a KO round, in which teams eliminate each other in best-of-one series, are combined. Other sports require the finalists of a competition to face each other in several instances, for example, a best-of-five-series. The aim of this is to minimize arbitrary factors that could predetermine a victory as much as possible. In Chess, for example, the white player starts the game with a slight advantage. To balance white's advantage, an extrinsic system (the series) is employed, which guarantees that both players will start with white. Another potential reason for the implementation of series is to negate factors such as day to day performance, weather, or home turf advantage. In other words, these regulations aim to make sure the best player wins and not a player that was lucky to gain a game extrinsic advantage.

Yet in other sports, as well as eSports, it is common to divide participants into groups, depending on their physical attributes. In boxing, for example, the added metagame divides participants into groups by weight. In nearly every sport it is also common to distinguish between youth- and adult-leagues. The intention here is not only to make the individual instances more interesting for the spectators but also to prevent injuries. A similar, but a more controversial division is related to the participants' gender (e.g. [46]). This division is also employed in eSports, where its usefulness is a topic of ongoing debate. We will discuss the necessity of limiting 'physical' capabilities of participants in AI vs. human competitions in the section "Critique of AI Achievements in Games."

B. Fairness: Game Intrinsic Factors

As discussed earlier, contemporary digital game artifacts contain multiple mechanical systems. Aarseth and Calleja [3] mention raiding, questing and player vs. player (PvP) as examples in World of Warcraft. Different models identify different (types of) elements of games for different purposes (e.g. [4, 7, 49]). Elverdam and Aarseth [19], for example, identify dimensions of games for the purpose of classification. These include the players' relation, virtual and physical space, struggle (including the game's goals), and more. The individual categories are less important than the observation that contemporary video game can rarely be described in only one manner. While it is easy to argue that Chess particularly requires two opposed players, it is impossible to make the same general statement about the digital artifact StarCraft II, which contains the potential player compositions of single player, two player, multiplayer, and multiteam (see[19]).

Another game intrinsic factor is different maps (e.g. in StarCraft II), which would be comparable to different boards in Chess or Go. These maps influence the individual strategies, but also the available player numbers. While StarCraft II also allows playing three different races, it is common for a player to stick to one race throughout their 'career'. In Dota 2, however, players need to be able to play a range of different heroes, as each series includes the drafting and banning of heroes from a pool of (as of writing) around 115 heroes. Selecting heroes can be considered a "material metagame" [18] to determine the configuration of the game instance. Added metagames can also regulate clear orthogame intrinsic content, such as available items inside of Dota 2. In some cases, items that heroes use to increase their strength of abilities were banned by leagues, as they were considered overpowered or were simply bugged.

When comparing AI and human intelligence in games, the here discussed multifacetedness of games must be considered and both participants' capabilities of engaging with the *artifact as a whole* needs to be examined. This insight leads us to the next section, in which we examine particular games and competitions within them, using the here developed distinction of game extrinsic and intrinsic factors, as well as the prototypical human vs. human added metagame.

VI. Critique of AI Achievements in Games

In this section we will review a few selected milestone achievements of AI in games, focusing on particular systems and their comparisons to human professionals. We will discuss the fairness of these competitions and whether it is valid to claim that the AI systems are super-human using the black box approach and the coherence of game extrinsic and intrinsic factors. The goal is to identify to what extent the presented milestones adhere to or deviate from the prototypical human vs. human added metagame. Any deviations, we argue, indicates an unfairness in the competition (for either side) and a claim of superiority must be treated carefully in these cases.

A. AlphaGo

AlphaGo, developed by DeepMind, is the first Go-playing system to win against professional Go players without handicap on a full-sized board. The first of these games were against the 2-dan European Go champion Fan Hui in 2015, where AlphaGo won 5-0 [41]. The 9-dan 18-time world champion Lee Sedol was the next target for AlphaGo, and a five-game match was scheduled in 2016 with a one million dollar prize and following the official rules: Chinese ruleset, a 7.5-point komi, and two-hour time limit for each player. Prior to the match, the Fan Hui games were published allowing Lee Sedol to prepare against AlphaGo.[3] AlphaGo won 4-1 with a loss in game four, showing a weakness in the system that might be exploitable. A new version of AlphaGo called 'Master' was anonymously registered to the 'Tygem' and 'FoxGo' Go servers, playing a total of 50 game instances, with a shorter time limit than usual game instances, against professional and top players, winning all of them[4]. AlphaGo's last game instances were at the Future of Go Summit, where AlphaGo won aginst several top players including the highest ranked player Ke Jie and a team of five human players, without losing a single game. After this event, AlphaGo was retired[5].

We argue that AlphaGo ultimately competed fairly in non-restricted matches against numerous top professional players both online and in settings similar to human competitions both in terms of game intrinsic and extrinsic factors.

[3]https://web.archive.org/web/20160214135238/https://gogameguru.com/an-younggils-pro-go-videos-alphago-vs-fan-hui-game-2/

[4]https://www.nature.com/news/google-reveals-secret-test-of-ai-bot-to-beat-top-go-players-1.21253

[5]https://deepmind.com/research/alphago/alphago-vs-alphago-self-play-games/

B. OpenAI Five

In 2016, the company OpenAI decided to pursue the challenge of beating human professionals in the multiplayer online battle arena (MOBA) game Dota 2 [35]. Dota 2 is a fast-paced real-time game, has partially observable states, high-dimensional observation and action spaces, and has long time horizons [36]. Normally in Dota 2, two teams of five players play against each other while there also is a one vs. one (1v1) variant. In 2017, OpenAI developed a bot capable of playing the 1v1 version that beat the former professional player 'Blitz' 3-0, the professional players 'Pajkatt' 2-1, 'CC&C' 3-0, the top 1v1 player 'Sumail' 6-0, and 'Dendi' 2-0. The standard (or most popular) variant of Dota 2 is played in teams of five players. However, because there exist serious 1v1 Dota competitions, the added metagame does adhere to at least an existing version of the added human vs. human metagame, mimicking a termination tournament with five players. The bot was updated by the developers between each series [34]; possibly bugs were fixed and control parameters were tuned. 'Sumail' also played against the previous version of the bot and won this time 2-1.

It can be argued that altering the bot in-between game instances is a violation of the black-box approach, as it effectively becomes a new system. However, human players usually have the ability to discuss strategies with a coach in-between game instances, so perhaps an AI should also be allowed to be influenced by a 'coach'. In any case, it depends on whether the developers that are modifying it, are considered part of the entity that is competing, which we would argue, they are not. Human modification of the AI system should thus not take place within a series of individual game instances, or even whole added metagames. Because both positions (for and against human intervention) are arguable, it appears necessary to develop an explicit regulation in this area for future human vs. AI competitions.

These series were played under standard 1v1 tournament rules. The bot had direct access to the features of the game artifact from the API, instead of being presented to the visual representation of the game artifact. The bot could only access the same information that would have been available to a human player but it was structured differently. For instance, humans have to infer the position of heroes and thus estimate the distances between units, which are important for ranged attacks, while the bot can access the exact positions and thus calculate the exact distances, instantly. This arguably goes against perceptual fairness because the input space should be the same. Here, the input space includes the same information for both the AI system and the human, but by perceiving the game state in a different way than humans, the AI system might have an advantage or disadvantage. The bot had access to the same actions as humans players and they were performed at similar frequencies but with a quicker reaction time of 80ms [33]. The reaction time was, however, reduced in later competitions.

After the 1v1 win, OpenAI let the bot play thousands of

games against various players, where several exploits were found to overcome the bot [34]. This setup mimics a human ladder where we would expect experienced human players not to have trivial exploits. The AI, however, would quickly descend the ladder due to these discovered exploits.

In 2018, a newer version of the bot named *OpenAI Five* was able to beat a team consisting of 99.95th percentile players 'Blitz', 'Cap', 'Fogged', 'Merlini', and 'MoonMeander' (some are former professionals) in a restricted version of the 5v5 game [33]. This series was named the *OpenAI Five Benchmark*. Some of the restrictions include a fixed hero pool of 18 heroes (instead of 117) resulting in 11 million possible game instances, no summons/illusions, and no Scan. The reaction time was increased from 80ms to 200ms in an attempt to match that of humans. The bot won the first two game instances where it did the hero drafting itself and lost the third game where the audience did the draft. The restricted hero pool is a significant limitation of the game intrinsic factors, effectively reducing the possible game instances to a much smaller subset than are usual in human vs. human competitions.

'OpenAI Five' later played two show series against the two teams of top professionals 'paiN Gaming' and 'Big God' and lost.[6] In 2019, 'OpenAI Five' won a best-of-three series 2-0 (the OpenAI Five Finals) against the Dota team OG, which consists of top-professional players. In this series, the hero pool was further restricted to just 17 heroes [37]. Playing against just one team mimics the final series of a tournament but not an entire tournament, ignoring the game extrinsic factors of complete added metagame structures.

After the win against OG, the *OpenAI Five Arena* allowed anyone to play against OpenAI Five. These games had the same restrictions as earlier. OpenAI Five won 7,215 games and lost 42 (99.4% win rate) against a total of 15,019 players[7]. One team, mainly consisting of the players 'ainodehna', 'backtoashes', 'CANYGODXXX', '.tv/juniorclanwar', and 'gazeezy', was able to reach a ten game winning streak.

The OpenAI Arena is basically an extensive ladder setup cohering to game extrinsic factors. A ladder challenges the bot to be robust to many different strategies and playing styles. The fact that it won almost every game but one team was able to beat it repeatedly is interesting. If this result was due to a trivial exploit, then most teams, knowing about the exploit, would be able to beat it; for a human opponent this would not be the case. However, the bot won 99.4% of the games in an extrinsically fair setup, which we would not expect even from human world champions. The criteria for being the best Dota 2 team on a ladder is not to have a 100% win rate, and we thus should not impose that expectation on the bot.

C. AlphaStar

StarCraft was, along with other real-time strategy games, proposed as a new challenge for AI in 2003 [12] with a renewed interest by research teams at Facebook in 2016 [45] and DeepMind in 2017 [47]. In 2019, DeepMind played their bot *AlphaStar* against the two top professional players Dario 'TLO' Wünsch and Grzegorz 'MaNa' Komincz and won both series 5-0 [48]. All games in these two series were Protoss vs. Protoss on the standard medium-sized map *CatalystLE*.

It was claimed that these series adhered to professional match conditions [48], while this is in fact not the case. Tournaments never use just a single map for a whole series but instead a predefined map pool, thus the competition did not adhere to the game extrinsic factors. Additionally, in professional tournaments, players face multiple players controlling any of the three races, and not just Protoss.

After the match 'MaNa' also mentioned that he made a few mistakes because they played an earlier version of StarCraft II than the one he was used to; he also did not warm up, which he would usually do [27, 19:50]. The actions per minute (APM) count of AlphaStar was around 280, which is lower than professional players, and with a reaction time of 350ms on average. AlphaStar had only access to visual information from the game, similarly, but not exactly identical, to the screen pixels presented to human players [47]. This is arguably a violation of the game intrinsic factors, similarly to OpenAI Five, since AlphaStar has a different input space than human players have. It is, however, a weak violation since AlphaStar's representation of the game state has the same information, while it is just structured differently.

Importantly, AlphaStar was not restricted to the limited view of the camera, which a human player has to control manually. As DeepMind puts it: "it could observe the attributes of its own and its opponents visible units on the map directly, without having to move the camera - effectively playing with a zoomed out view of the game." [48]. 'MaNa' expressed this as being "very unfair" [27, 1:17:15]. This is, however, a clear advantage of AlphaStar on both the levels of perceptional capabilities and motoric necessities. Furthermore, it could even be argued that this alters the "perspective dimension" [19] of StarCraft II from vagrant to omnipresent, which is arguably an alteration of the orthogame itself.

Later, the professional player 'MaNa' played against a prototype of AlphaStar that controlled the camera as well, in a single-game series and won. He found a weakness in AlphaStar during a Warp Prism harassment with Immortals, continuously warping in units, picking them up, and escaping. Whether this weakness was due to the camera control or if it was a critical exploit of AlphaStar is not known. It may, however, seem that he won the last game because it was played at a later date than the others and he had time to prepare against its style. Specifically, he said that "We ('TLO' and 'MaNa') noticed that the agent sticks to the basic units a lot. It's very confident in its micro, and it should be, it's great micro, but it doesn't really transition out of it." [27, 1:19:15]. 'MaNa' said his new plan was to "... defeat AlphaStar with simply better unit composition rather than unit control" [27, 1:20:20]. We notice here, that in the two 5-0 wins against 'TLO' and 'MaNa', they did not have a chance to scrutinize any recorded games played by AlphaStar, which professional

[6]https://liquipedia.net/dota2/The_International/2018/OpenAI_Showmatches
[7]https://arena.openai.com/#/results

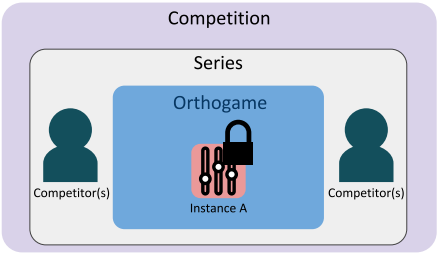

Fig. 2: A typical AI vs. human competition consisting of one series between two teams or individuals. The series often consist of identical game instances, or a limited set of game instances, of the same orthogame.

player typically can do before important human vs. human series. In contrast, the developers of AlphaStar picked and knew the opponent in advance. 'MaNa' said, commenting on his first series against AlphaStar: "I was completely in the dark ... I don't know what to expect. If you are a StarCraft player you are familiar with people you are playing on the ladder ... you know what their styles are." [27, 18:40]. Compared to a prototypical human vs. human competition, this is an unusual setup of the competition, as professional players know each others' play styles before playing. To observe the problem from another angle: human vs. human added metagames never keep their participants secret from their participants, as was the case in 'MaNa' vs. AlphaStar.

VII. Conclusions

We introduced a black box approach that can be used when designing and evaluating human vs. AI game competitions as well as the notions of game extrinsic and intrinsic factors. We applied these to discuss the fairness of recent AI achievements of AlphaGo, Dota Five, and AlphaStar. It appears that the added metagame's role in an AI vs. human competition has a different focus than in the human vs. human competitions. The added metagame in a human vs. human competition regulates mostly the bigger structure of extrinsic factors, such as a sequence of series, number of instances in a series and groupings due to a physical difference between the competitors. The added metagame's role in the AI vs. human competitions, however, is focused on the regulation of game intrinsic factors. This means, in competitions between AIs and humans in digital games, the orthogame is so far always limited to either one particular configuration or a very small amount of possible configurations, compared to human vs. human competitions. A visualization of this setup is shown in Figure 2. Dota Five, for example, is capable of playing only 17 out of (approximately) 115 heroes of Dota. Thus, the orthogame (Dota 2) in the competition between Dota Five and humans had to be limited to these 17 heroes.

Dota Five had direct access to game state variables while humans must infer positions, attack ranges, and health from a raw visual representation, which ultimately leads to 'educated guesses' more than factual knowledge. AlphaStar, in fact, used a visual representation but a different one than what is presented to humans. Furthermore, Dota Five and AlphaStar are incapable of 'misclicking', which is the act of giving a command unintended by the player. These two factors constitute imbalances in the perceptual and motoric capabilities of the competitors, which must be accounted for in the future. To have a fair competition, the AIs must be handicapped through their interaction with the game to imitate how humans are interacting with it, i.e. if humans have imprecise and slow means of interacting with the game it should be the same for the AI system. It can be argued that this is something that naturally occurs in human vs. human competition, as every human participant is implicitly restricted to human capabilities. Thus, the addition of a handicap to the AI simply constitutes an explicit correction through game extrinsic factors.

We thus conclude that we are not done with games. The games proposed as ultimate AI challenges, Dota 2 and StarCraft II, are not yet mastered by AI. As we identified, so far AI vs. human competitions are different in that (1) the AIs do not compete in a tournament structure, but are simply matched with the best (available) human player and (2) they limit the orthogame in particular ways, such as range of maps, heroes or races. To be able to claim that 'We are done with games' the AI has to engage in a fair competition with humans that is constituted by the same external factors, such as several matches against the same opponent, as well as different opponents. Additionally, it should not limit game internal factors, such as only allowing certain playable heroes or maps. Only then, the claim that AI is superior to humans in games is justified, given that StarCraft II and DotA 2 are and remain the most complex games to beat.

We are, however, not neglecting the significant progress made towards achieving the goals, as no system before Dota Five and AlphaStar could win against professionals in any competition in these games. The fact that game extrinsic factors have been largely ignored in human vs. AI competitions, might indicate where to focus our future efforts. Adapting to a variety of possibly unseen game instances as well as different players are important open challenges; they deal with problems of generality [17, 23, 52], transfer learning [39], and life-long learning [38], which are all active areas of focus in machine learning research [24].

An ultimate goal that would demonstrate that an AI system can fully master a game, beyond the extrinsic factors of human vs. human competitions, would be to allow anyone to play against it over a long period of time. This setup would be similar to OpenAI Arena, without restricting any intrinsic game factors. This goal was to some degree achieved by AlphaGo when it played on the Tygem and FoxGo Go servers without losing, and without restricting the intrinsic factors.

VIII. Acknowledgements

This research has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (Grant Agreement No [695528] MSG: Making Sense of Games).

References

[1] E. Aarseth. Ludology. In M. J. P. Wolf, editor, *The Routledge Companion to Video Game Studies*. Routledge, 2014.

[2] E. Aarseth. Ontology. In M. J. P. Wolf, editor, *The Routledge companion to video game studies*, pages 510–518. Routledge, 2014.

[3] E. Aarseth and G. Calleja. The word game: The ontology of an undefinable object. In *Foundations of Digital Games Conference*, 2015.

[4] E. Aarseth, S. M. Smedstad, and L. Sunnanå. A multidimensional typology of games. In *DiGRA Conference*, 2003.

[5] J. Arjoranta. Game definitions: A wittgensteinian approach. *Game Studies: the international journal of computer game research*, 14, 2014.

[6] E. M. Avedon and B. Sutton-Smith. *The study of games*. John Wiley & Sons, 1971.

[7] S. Bjork and J. Holopainen. *Patterns in game design*. Charles River Media, 2004.

[8] N. Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

[9] B. Bouzy and T. Cazenave. Computer go: an ai oriented survey. *Artificial Intelligence*, 132(1):39–103, 2001.

[10] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[11] J. Burmeister and J. Wiles. The challenge of go as a domain for ai research: a comparison between go and chess. In *Proceedings of Third Australian and New Zealand Conference on Intelligent Information Systems. ANZIIS-95*, pages 181–186. IEEE, 1995.

[12] M. Buro. Real-time strategy games: A new ai research challenge. In *IJCAI*, volume 2003, pages 1534–1535, 2003.

[13] R. Caillois. *Man, play, and games*. University of Illinois Press, 2001.

[14] R. Canaan, C. Salge, J. Togelius, and A. Nealen. Leveling the playing field-fairness in ai versus human game benchmarks. *arXiv preprint arXiv:1903.07008*, 2019.

[15] M. Carter, M. Gibbs, and M. Harrop. Metagames, paragames and orthogames: A new vocabulary. In *Proceedings of the international conference on the foundations of digital games*, pages 11–17. ACM, 2012.

[16] M. Čertickỳ and D. Churchill. The current state of starcraft ai competitions and bots. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017.

[17] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning. *arXiv preprint arXiv:1812.02341*, 2018.

[18] M. S. Debus. Metagames: on the ontology of games outside of games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, page 18. ACM, 2017.

[19] C. Elverdam and E. Aarseth. Game classification and game design: Construction through critical analysis. *Games and Culture*, 2(1):3–22, 2007.

[20] J. M. F. Fernandez and T. Mahlmann. The dota 2 bot competition. *IEEE Transactions on Games*, 2018.

[21] G. Frasca. Ludologists love stories, too: notes from a debate that never took place. In *DiGRA conference*, 2003.

[22] S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesvári, and O. Teytaud. The grand challenge of computer go: Monte carlo tree search and extensions. *Communications of the ACM*, 55(3):106–113, 2012.

[23] N. Justesen, R. R. Torrado, P. Bontrager, A. Khalifa, J. Togelius, and S. Risi. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

[24] N. Justesen, P. Bontrager, J. Togelius, and S. Risi. Deep learning for video game playing. *IEEE Transactions on Games*, 2019.

[25] J. Juul. *Half-real: Video games between real rules and fictional worlds*. MIT press, 2011.

[26] B. Keogh. Across worlds and bodies: Criticism in the age of video games. *Journal of Games Criticism*, 1(1):1–26, 2014.

[27] G. M. Komincz. Deepmind starcraft 2 demonstration - mana's personal experience. URL https://www.youtube.com/watch?v=zgIFoepzhIo.

[28] K. L. Kroeker. A new benchmark for artificial intelligence. *Commun. ACM*, 54(8):13–15, 2011.

[29] G. Marcus. Innateness, alphazero, and artificial intelligence. *arXiv preprint arXiv:1801.05667*, 2018.

[30] M. Müller. Computer go. *Artificial Intelligence*, 134(1-2):145–179, 2002.

[31] J. H. Murray. The last word on ludology v narratology in game studies. In *International DiGRA Conference*, 2005.

[32] S. Ontanón, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. A survey of real-time strategy game ai research and competition in starcraft. *IEEE Transactions on Computational Intelligence and AI in games*, 5(4):293–311, 2013.

[33] OpenAI. Openai five benchmark. https://openai.com/blog/openai-five-benchmark/, .

[34] OpenAI. More on dota 2. https://openai.com/blog/more-on-dota-2/, .

[35] OpenAI. Openai five. https://openai.com/five/, .

[36] OpenAI. Openai five. https://blog.openai.com/openai-five/, .

[37] OpenAI. How to train your openai five. https://openai.com/blog/how-to-train-your-openai-five/, .

[38] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

[39] E. Parisotto, J. L. Ba, and R. Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*, 2015.

[40] C. E. Shannon. Programming a computer for playing chess. In *Computer chess compendium*, pages 2–13. Springer, 1988.

[41] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[42] H. Simon and W. Chase. Skill in chess. In *Computer chess compendium*, pages 175–188. Springer, 1988.

[43] J. Stenros. The game definition game: A review. *Games and Culture*, 12(6):499–520, 2017.

[44] B. Suits. *The Grasshopper-: Games, Life and Utopia*. Broadview Press, 2014.

[45] G. Synnaeve, N. Nardelli, A. Auvolat, S. Chintala, T. Lacroix, Z. Lin, F. Richoux, and N. Usunier. Torchcraft: a library for machine learning research on real-time strategy games. *arXiv preprint arXiv:1611.00625*, 2016.

[46] S. Teetzel. On transgendered athletes, fairness and doping: An international challenge. *Sport in Society*, 9(2):227–251, 2006.

[47] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[48] O. e. a. Vinyals. Alphastar: Mastering the real-time strategy game starcraft ii. URL https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

[49] M. J. Wolf. Inventing space: Toward a taxonomy of on-and off-screen space in video games. *Film Quarterly (ARCHIVE)*, 51(1):11, 1997.

[50] G. N. Yannakakis. Game ai revisited. In *Proceedings of the 9th conference on Computing Frontiers*, pages 285–292. ACM, 2012.

[51] G. N. Yannakakis and J. Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

[52] C. Zhang, O. Vinyals, R. Munos, and S. Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.