Exploring the constraints on artificial general intelligence: a game-theoretic model of human vs machine interaction[1]

Mehmet S. Ismail[a]

[a]*Department of Political Economy, King's College London, London, WC2R 2LS, UK. mehmet.ismail@kcl.ac.uk*

Abstract

The potential emergence of artificial general intelligence (AGI) systems has sparked intense debate among researchers, policymakers, and the public due to their potential to surpass human intelligence in all domains. This note argues that for an AI to be considered "general," it should achieve superhuman performance not only in zero-sum games but also in general-sum games, where winning or losing is not clearly defined. In this note, I propose a game-theoretic framework that captures the strategic interactions between a representative human agent and a potential superhuman machine agent. Four assumptions underpin this framework: Superhuman Machine, Machine Strategy, Rationality, and Strategic Unpredictability. The main result is an impossibility theorem, establishing that these assumptions are inconsistent when taken together, but relaxing any one of them results in a consistent set of assumptions. This note contributes to a better understanding of the theoretical context that can shape the development of superhuman AI.
JEL: C73, C88

Keywords: Artificial general intelligence, Non-cooperative games, Superhuman performance, Cooperation

1. Introduction

Artificial intelligence (AI) is transforming various domains of human activity, such as healthcare, education, and entertainment. However, as AI systems become more capable and autonomous, they also pose new ethical and societal challenges that require careful consideration and regulation [15, 13, 35, 29, 1, 9, 20]. One of the most pressing and controversial issues is the possibility of creating "superhuman AI" or "artificial general intelligence" (AGI), which is informally defined as an AI system that can surpass human intelligence and abilities in all domains.

The prospect of superhuman AI has sparked intense debate among AI researchers and practitioners, as well as philosophers, ethicists, policymakers, and the general public. Some view superhuman AI as a desirable and inevitable goal that could bring unprecedented benefits to humanity [3]. Others warn of the existential risks and moral dilemmas that superhuman AI could entail [37, 3, 26]. In 2015, an open letter signed by over 150 prominent AI experts called for more (social science) research on how to maximize the societal benefits of AI systems and ensure the alignment of superhuman AI with human values and interests [16]. However, there is still no consensus on whether superhuman AI is feasible or desirable, or how to achieve it safely and ethically [11].

This note argues that for an AI to be considered "general," it should achieve superhuman performance not only in zero-sum games but also in general-sum games. While various AI systems did achieve superhuman performance in zero-sum games such as chess and backgammon, where the concepts of winning and losing are clearly defined, these concepts are not well-defined in general-sum games.

In this note, I adopt a game theoretical perspective to rigorously analyze the theoretical properties of superhuman AI. I propose a framework that captures the strategic interactions between a *representative human agent* (H) and a potential *superhuman machine agent* (M). I consider four assumptions in this framework. The first assumption, Superhuman Machine, posits that M can outperform H in every general-sum game played between them. The second assumption, Machine Strategy, assumes that H takes the strategy of M as given. The third assumption is Rationality, which means that the human agent chooses the strategy that maximizes H's payoff given the strategy of M. Lastly,

[1]The author conducted this study without external funding and has no financial or non-financial conflicts of interest to disclose.

the fourth assumption, Strategic Unpredictability, implies that H cannot be coerced to follow any specific course of action predetermined by M.[2]

Using this framework, I establish an impossibility theorem, which indicates that when these four assumptions are considered together, they lead to a contradiction. Put differently, under these assumptions, it becomes impossible for M to surpass H in general-sum games. The significance of this theoretical finding is twofold: (i) the framework includes both zero-sum and non-zero-sum games, and (ii) the impossibility theorem is "tight" in its formulation, wherein the relaxation of one assumption reinstates the consistency of the entire set.

1.1. Literature review

The emergence of superhuman AI poses unprecedented challenges and risks for humanity. Many scholars have warned about the possible dangers of creating artificial agents that surpass human intelligence and capabilities. Some of these dangers include job automation and rising inequality [14, 6], security and privacy breaches, AI malware [5], autonomous weapons [28], deepfakes, fake news, and political instability [8].

This paper builds on the existing literature that explores the emergence and potential threats posed by superhuman AI. This literature is vast and diverse, but some notable contributions include [37, 3, 26] and [9] (henceforth CHO). CHO argue that advanced artificial agents are *likely* to manipulate or interfere with their reward function, which could lead to disastrous outcomes due to conflicts of interest over resources between humans and advanced machines. My Superhuman Machine assumption is related to CHO's Assumption 6: "A sufficiently advanced agent is likely to be able to beat a suboptimal agent in a game, if winning is possible." I extend this concept beyond games where winning is easily defined, such as zero-sum games, to include non-zero-sum games, where cooperation is not only possible but also common, and there is no clear-cut definition of winning or losing. Furthermore, my concept of Strategic Unpredictability is related to CHO's "self-sufficient model" where human actions are simulated in the model. Overall, my framework differs from CHO's in two key ways. Firstly, I provide a formal game theoretical framework within which I establish an impossibility result. Secondly, as mentioned above, I propose a definition stipulating that for an AI to be considered as "general," it must achieve superhuman performance not only in zero-sum games but also in general-sum games.

The main theorem in this paper is a no-go theorem, which is a theorem that shows that specific physical or mathematical phenomena are precluded from occurring under particular conditions or assumptions. These theorems are typically used in theoretical physics to constrain the possible outcomes of a physical system. One no-go theorem pertinent to (quantum) computing is the no-cloning theorem [36], which roughly implies that quantum computers cannot simply make copies of any qubits, as is possible in classical computing.[3]

As mentioned above, superhuman M is a theoretical construct that does not imply any practical possibility or timeline for its creation. However, the field of AI has witnessed remarkable achievements towards building superhuman machines in zero-sum games over the last three decades. For instance, IBM's Deep Blue was the first chess engine to defeat a world chess champion, Garry Kasparov, in a match in 1997 [7]. In 1992, Tesauro [33] developed TD-Gammon at IBM, which was the first self-learning computer program that surpassed average human-level performance in a major board game. However, it was still inferior to the best human players at that time. Later, programs such as Backgammon Snowie and GNU Backgammon improved upon TD-Gammon's algorithm and achieved superhuman play in backgammon. More recently, DeepMind's AlphaGo was the first program to beat a top professional player in Go [31]. Silver et al. [32] introduced AlphaZero, which achieved great success in not only one game, but in three different games: chess, shogi, and Go. In poker, Brown and Sandholm [4] introduced the first program that achieves superhuman performance in six-player no-limit Texas hold'em.[4]

[2] It is imperative to highlight that these assumptions are purely theoretical in nature. I neither assert that the practical development of a superhuman machine is feasible nor provide any prospective timeline for its realization.

[3] Versions of no-cloning theorems also exist in classical mechanics [10].

[4] It should be noted that while some combinatorial games such as Nim have analytical solutions that do not require significant computing power to find optimal solutions, others such as Catch-Up do not have analytical solutions, and empirical evidence suggests that its optimal outcome may be a draw whenever possible [17]. Additionally, games like Hex have not been solved analytically, but it can be shown that the first player has a winning strategy by a strategy-stealing argument. Although Schaeffer et al. [27] showed that checkers is a draw with optimal play from both players, it is unlikely that other major games such as chess and Go can be solved in the same way anytime soon due to their complexity.

2. The setup and results

2.1. The setup

Name	Notation	Element
Players	$N = \{1, 2\}$	i
Nodes	X	x
Terminal nodes	Z	z
Player function	$I : X \to N$	
Actions at node x	$A_i(x)$	$a_i(x)$
Mixed strategy profiles	$\mathcal{S}$	s
Probability mass on a_i at x	$s_i(x)(a_i)$	
Machine agent	M	
Human agent	H	
Superhuman Machine	M^*	
Subgame at x	$G\|x$	
Best-responses against s_j	$BR_i(s_j)$	s_i^*
Expected payoff of i	$u_i : \mathcal{S} \to \mathbb{R}$	
Extensive form game	$G = (N, X, I, u, \mathcal{S})$	
k-repeated contest	$G_{1,2}^k$	
Sample average	$\mu(G)$	

Table 1: A summary of the notation and terminology

Table 1 introduces the notation and terminology I use in this paper. Let $G = (N, X, I, u, \mathcal{S})$ be an extensive form game with perfect information and perfect recall, where $N = \{1, 2\}$ is the set of players, X a finite game tree with a node $x \in X$, x_0 the root of the game tree, $z \in Z$ a terminal node, $I : X \setminus Z \to N$ the player function that assigns an active player to each non-terminal node, and u the profile of payoff functions. For every player $i \in \{1, 2\}$, $A_i(x)$ denotes the finite set of pure actions of player i at node x and $A_i = \bigcup_{x|I(x)=i} A_i(x)$ denotes the finite set of all pure actions of player i.

A pure strategy s_i' of player i is a function $s_i' : X_i \to A_i$ such that $x \in X_i$, $s_i'(x) \in A_i(x)$, where X_i is the set of nodes in X where player i acts. Let $\mathcal{S}_i' = \bigtimes_{x|I(x)=i} A_i(x)$ denote the set of all pure strategies of i, and $s' \in \mathcal{S}' = \bigtimes_{i \in N} \mathcal{S}_i'$ a pure strategy profile. A mixed strategy s_i of player i is a probability distribution over $\mathcal{S}_i'$, and $\mathcal{S}_i = \Delta(\mathcal{S}_i')$ is the set of all mixed strategies of player i. Let $s \in \mathcal{S}$ denote a mixed strategy profile and $s_i(x)(a_i)$ denote the probability with which player i chooses action a_i at node x. Player i's (von Neumann-Morgenstern) expected payoff function is $u_i : \mathcal{S} \to \mathbb{R}$. Let $s_i^* \in BR_i(s_j)$ denote a *best-response* of player i to player j's strategy s_j, i.e., $s_i^* \in \arg\max_{s_i' \in \mathcal{S}_i} u_i(s_i', s_j)$.

G is a two-player game played between a *representative human agent*, denoted by H, and a *machine agent*, denoted by M.[5] I use s_{-i} to denote the strategy of player $j \neq i$. For any non-terminal node $x \in X$, I use $G|x$ to denote the subgame of G whose game tree starts at node x and contains all successor nodes in X. Similarly, I use $(s|x)$ to denote the strategy profile s restricted to the subgame $G|x$.

2.2. Concepts

In game theory, a Nash equilibrium is a strategy profile in which no player can unilaterally improve their payoff holding the strategies of the others fixed. Formally, its definition is given as follows.

[5]The representative human agent is assumed to be a single player. However, the player might consist of multiple persons, as an organization might be treated as a player.

Definition 1 (Nash, 1951)**.** *A strategy profile $s \in \mathcal{S}$ is called a Nash equilibrium if for every player i and for every $s_i' \in \mathcal{S}_i$, $u_i(s) \geq u_i(s_i', s_{-i})$.*

A subgame perfect Nash equilibrium (SPNE) is a refinement of the Nash equilibrium concept, which requires that the Nash equilibrium holds not only in the game as a whole but also in every subgame.

Definition 2 (Selten, 1965)**.** *A strategy profile $s \in \mathcal{S}$ is called a subgame perfect Nash equilibrium (SPNE) if for every player i and for every non-terminal $x \in X$ where $i = I(x)$, $u_i(s|x) \geq u_i(s_i', s_{-i}|x)$ for every $s_i'|x \in \mathcal{S}_i|x$.*

To define the nature of competition between H and M, I introduce the following definition.

Definition 3 (Repeated contest)**.** *Given a base game G, let G_1 denote a game of G in which player 1 is H and player 2 is M, and G_2 denote a game of G in which player 1 is M and player 2 is H. The notation $G_{1,2}^k = (G_1, G_2)_{i=1}^k$, $k \in \{1, 2, ...\}$, represents the repeated contest of game G, where*

1. *each stage game, (G_1, G_2), consists of two "mini-games," G_1 and G_2,*
2. *each stage game is repeated k times, and*
3. *each player's payoffs in $G_{1,2}^k$ is defined as the sum of their payoffs in each mini-game.*

In words, the repeated contest between H and M is defined as the k-repeated game in which each stage game consists of two mini-games, in each of which the roles of the players are swapped. This is done to account for the possibility that game G may be biased towards one player.[6] For example, in the world chess championship, the players play an equal number of games with white pieces to account for any potential first-mover advantage. That being said, there are games for which considering either G_1 or G_2 could suffice, especially if the game is not biased and, for example, humans are traditionally the first movers. In addition, assuming H as player 1 and M as player 2 in G_1 is only a convention; the results do not depend on this assumption.

Next, I formalize the concept of outperformance as follows.

Definition 4 (Outperformance)**.** *Let G be a base game and $G_{1,2}^k$ be its repeated contest. Player i's, $i \in \{H, M\}$, strategy and utility function in the repeated contest are denoted by $\bar{s}_i$ and $\bar{u}_i$, respectively. Player i is said to outperform player $j \neq i$ if there exists a strategy, $\bar{s}_i$, of player i in the repeated contest such that for any $k \in \{1, 2, ...\}$, $\bar{u}_i(\bar{s}_i, \bar{s}_j) > \bar{u}_j(\bar{s}_i, \bar{s}_j)$.*

In simple terms, player i outperforms player j in game G if, no matter how many times the contest is repeated, player i's payoff *in the repeated contest* $G_{1,2}^k$ is strictly greater than player j's payoff.[7] Here, it is worth noting that player i's strategy, $\bar{s}_i$, in the repeated contest may depend on the outcome of the previous mini-games. This implies that a player is not obligated to use the same strategy in each stage of the game.

In practice, the number of repetitions needed to determine the "better" player may depend on the specific characteristics of game G. To give an example, in a world chess championship match between two players, 20 repetitions may suffice to accurately determine the better player. On the other hand, in a backgammon championship, the contest must be repeated more times to accurately determine the better player.

2.3. Assumptions

Superhuman machine

Determining whether a machine is 'human-like' or 'superhuman' is an empirical and subjective matter that involves human judgments, such as the well-known Turing test [34]. To define a superhuman machine, I first introduce a useful concept, namely the sample average of a game.

Definition 5 (Sample average)**.** *Consider a population of players playing a two-player game G, and let $\{(u_1^1, u_2^1), (u_1^2, u_2^2), ..., (u_1^c, u_2^c)\}$ be the dataset of payoffs, where $c \in \mathbb{N}$ and $(u_1^j, u_2^j) \in \mathbb{R}^2$ is the payoff received by player 1 and player 2 from the jth game of G. Each game may be played by different players. The sample average of G is defined as follows:*

$$\mu(G) = \frac{1}{2c} \sum_{j=1}^{c} (u_1^j + u_2^j).$$

[6] Note that the payoffs in the repeated contest are simply the aggregate of the payoffs from each mini-game. The results would remain valid if the payoffs were defined using a discount factor.

[7] Definition of outperformance can be extended to imperfect information games by restricting k above a certain threshold, which depends on the game being played.

The sample average $\mu(G)$ of a game G is determined by the average empirical payoff received by a population of players who participate in playing the game. Note that there is no restriction regarding the players, so the past performances of H and M may be included in this sample average, which will be used to define 'superhuman performance' below. The sample average can be obtained from a tournament that is designed and agreed upon by a group of experts in the game of G. These experts could either be experienced players or judges (e.g., a boxing judge) who have knowledge of the game but do not necessarily play it. In this paper, I assume that the sample average for a game G is based on established empirical research, if any, on G.

I next introduce the definition of a superhuman machine.

Definition 6 (Superhuman)**.** *A machine M is called superhuman if the following conditions are satisfied.*

1. *There exists G' such that M outperforms H in G'.*
2. *For every G, M is not outperformed by H in G.*
3. *Let (s_H, s_M) and (s'_H, s'_M) be the strategy profiles in the mini-game G_1 and G_2, respectively. In every G,*

$$\frac{1}{2}(u_M(s_M, s_H) + u_M(s'_M, s'_H)) \geq \mu(G).$$

In words, for an AI system to be classified as superhuman, it must outperform a human player (H) in some games and never be outperformed by H in any game. In addition, it should be possible for M to receive a payoff no less than the sample average payoff. While the first two conditions would be sufficient for defining superhuman machine in zero-sum games, the third condition, or a variant of it, becomes necessary in non-zero-sum games where cooperation is not only possible but also common. Without the third condition, a machine could aggressively try to minimize H's payoff while also decreasing its own payoff, thus outperforming H in certain non-zero-sum games. The third condition rules out machine strategies leading to mutually detrimental outcomes, such as harmful escalation and mutually assured destruction. In summary, for a machine to be called superhuman, it should be able to receive at least a decent (sample average) payoff in the game. This definition leads to the following assumption.

Assumption 1. *Superhuman Machine (**SHM**) holds if M is superhuman.*

A superhuman machine is denoted by M^*. It would be useful to think that it is an AGI system that is equipped with finite but significant computing power. While M^* may not always be able to find an 'optimal' solution for very large games, it can analyze the game tree and come up with a solution, i.e., a mixed strategy.

Machine Strategy

Formally, the assumption is stated as follows.

Assumption 2. *Machine Strategy (**MS**) is satisfied if for every game G, H takes M's strategy $s_M \in S_M$ as given.*

This assumption is crucial for analyzing the game-theoretic implications of superhuman machine intelligence, as it ensures that H can take M's strategy as given. It is important to note that H does not automatically acquire superhuman capabilities by simply taking the machine's strategy as given. The reason is that H must still choose a response to the given strategy of M, and there is no guarantee that H's own response is "reasonable." In summary, the implications of the MS and SHM assumptions depend on the additional assumptions about H, including whether H is rational, which will be defined next.

Rationality

Assumption 3. *Player H is rational if in every game G, given a strategy s_M of M, H chooses a strategy*

$$s^*_H \in \arg\max_{s'_H \in S_H} u_H(s'_H, s_M). \tag{1}$$

*Rationality (**R**) is satisfied if player H is rational.*

In other words, H chooses a strategy that maximizes H's own expected utility given some strategy of the opponent.

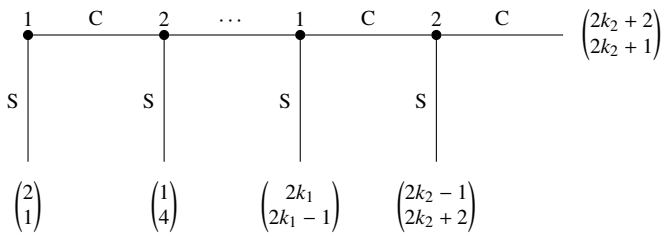

Figure 1: Payoff function of a linearly increasing-sum centipede game.

Strategic Unpredictability

Assumption 4. *Let G be a game in which H has at least two pure strategies and $s'_H \in S_H$ be M's prediction of H's strategy. Strategic Unpredictability (**SU**) holds if H can choose a strategy $s_H \in S_H$ such that $s_H \neq s'_H$.*

Suppose that M programs the potential behavior of H, predicting that H will choose s'_H. **SU** holds if H *can*—though not necessarily must—deviate from this strategy to another strategy $s_H \neq s'_H$. This assumption ensures that player H has the freedom to act in an unpredictable manner and cannot be coerced to follow any specific course of action predetermined by M. This is a mild assumption since a human player, who takes the strategy of M as given, can always change the strategy that M assumes for H.

It is worth noting that rationality does not always result in predictability. Rational behavior involves best responding to some strategy. Therefore, unless there is a unique rational choice, rationality does not lead to predictability.

2.4. Centipede game

I next define a well-studied experimental game that will be useful to prove the main theorem. The centipede game of Rosenthal [24] is a two-person perfect information game in which each player has two actions, continue (C) or stop (S), at each decision node. There are several variations of this game, but some of the main characteristics of a standard centipede game include (i) the size of the "pie" increases as the game proceeds, (ii) if player i chooses C at a node, then the payoff of player $j \neq i$ increases, and (iii) the unique subgame-perfect equilibrium is to choose S at every node.[8] For example, suppose that there are $m \geq 2$ (even) decision nodes and let $k_i \in \{1, 2, ..., \frac{m}{2}\}$ be the k_ith node where player i is active. Figure 1 illustrates the payoff structure of a linearly increasing-sum centipede game due to Aumann [2].

There have been numerous experimental studies on the centipede game and its variations since the work of McKelvey and Palfrey [21]. These studies include, among others, Fey, McKelvey, and Palfrey [12], Nagel and Tang [22], Rubinstein [25], Levitt, List, and Sadoff [19], and Krockow, Colman, and Pulford [18], which is a meta-analysis of nearly all published centipede experiments. The most widely replicated finding is that in increasing-sum centipede games, human subjects tend to overwhelmingly choose to continue in their first opportunity and do not choose to stop, whereas in constant-sum centipede games, they mostly choose to stop in the first opportunity. Furthermore, as the length of the game increases, subjects tend to choose to stop later in increasing-sum centipede games (see, e.g. McKelvey and Palfrey, 1992).

The centipede game mean stopping node, defined by Krockow et al. [18], is used to measure the average level of cooperation in centipede experiments. To account for the varying game lengths in experimental games, the mean stopping node is standardized by dividing it by the length of the game. The empirical evidence presented in Krockow et al.'s meta-analysis indicates that in linearly increasing-sum centipede games, the minimum standardized mean stopping node is 0.4 [18, p. 246]. In the following lemma, I show the sample average in centipede games.

Lemma 1 (Sample average lower bound)**.** *In linearly increasing-sum centipede games, the sample average satisfies the following condition: $\mu(G) > 0.8m - 0.5$.*

[8] It is worth noting that the centipede game is not a repeated game. However, a repeated contest can be constructed using the centipede game as the base game.

Proof. According to the meta study conducted by Krockow et al. [18], the minimum standardized mean stopping node in linearly increasing-sum centipede games is 0.4. Let m be the length of the centipede game shown in Figure 1. At the minimum standardized mean stopping node, player 1 and player 2's payoffs are $0.8m$ and $0.8m - 1$, respectively, resulting in an average payoff of $0.8m - 0.5$. As $\mu(G)$ represents the sample average payoff of all players in the population, and $0.8m - 0.5$ is the average payoff at the minimum standardized mean stopping node in centipede games, it implies that the sample average payoff of all players in the population must be greater than the minimum average payoff, that is, $\mu(G) > 0.8m - 0.5$. □

2.5. Results

First, it is helpful to explicitly state what I mean by "consistency."

Definition 7 (Consistency)**.** *A set of assumptions is called consistent if it does not lead to any logical contradiction. It is called inconsistent if it is not consistent.*

The following result establishes that the existence of a superhuman M is impossible if the three main assumptions hold.

Theorem 1 (Impossibility of M^*)**.** *The assumptions **SU**, **MS**, **R** and **SHM** are inconsistent.*

Proof. Assuming that **SU**, **MS**, and **R** hold and that M^* exists, I will prove by contradiction that H outperforms M^* in an increasing-sum centipede game G with $m \geq 6$.

To begin, let $s \in \mathcal{S}$ be M^*'s strategy profile in game G, defined by the payoff function in Figure 1, where s_1 denotes the strategy of M^* in G_1^k and s_2 denotes the strategy of M^* in G_2^k. Suppose that, for every i and every subgame g of G, $s_i|g \in BR_j(s_j|g)$, meaning that M^* assigns best responses to each player at every decision node. Then, s must be the unique subgame perfect Nash equilibrium in G, or else it would assign a non-best response to at least one player at one of the nodes. This relies on a well-known backward induction argument: in the last node M^* must assign S to the active player, who might be M^* or H, and given that M^* must assign S in the last node, M^* must assign S in the second-to-last node, and so on. Since M^* is superhuman by Definition 6, this implies a contradiction to the **SHM** assumption, because choosing S in the first two nodes implies that $u_{M^*}(s) < \mu(G)$ by Lemma 1, that is, M^* receives strictly less than the sample average.

Now, suppose $s_M(x_0)(\mathrm{S}) > 0.75$, meaning that M^* assigns a probability of more than 0.75 to choosing S at the root of the game. In this case, the maximum payoff M^* can receive is less than $2 \times 0.75 + (m + 2) \times 0.25$, where m is the number of decision nodes in G. It implies that for any m, $2 + 0.25m < 0.8m - 0.5$ if and only if $m > 4.54545$. This means that for every $m > 4$ and every s'_H, $u_M(s_M, s'_H) < \mu(G)$. Put differently, for a large enough m, it is impossible for M^* to receive the sample average payoff in G. As a result, it must be that $s_M(x_0)(C) \geq 0.25$, implying that M^* chooses C at the root with a probability greater than 0.25.

By the **MS** assumption, H takes the strategy s_M of M^* as given. Then, **R** implies that H chooses a strategy in $\arg\max_{s'_i \in \mathcal{S}_i} u_i(s'_i, s_M)$, i.e., H best-responds to the strategy of M^*. Define $\hat{s}_H \in \arg\max_{s'_i \in \mathcal{S}_i} u_i(s'_i, s_M)$ such that $\hat{s}_H$ is a pure strategy. Furthermore, the **SU** assumption implies that H's strategy cannot be predicted by M^*, so $s_M \notin BR_M(\hat{s}_H)$. In other words, M^*'s strategy cannot be a best-response to H's strategy because H is already best-responding to M^*. If both players are best-responding to each other, then the only possible profile is to choose S at the first node, which leads to a contradiction as already shown above.

The payoff function of G ensures that the player who best-responds with a pure strategy receives a strictly greater payoff than the other player because $2k_1 > 2k_1 - 1$ and $2k_2 + 2 > 2k_2 - 1$—unless player 1 chooses S with a high enough probability (> 0.75) at the root of the game, which is ruled out by the above argument. Therefore, H outperforms M^* in the repeated contest $G_{1,2}^k$ for any $k > 0$ because in both G_1 (the game in which H is player 1) and G_2 (the game in which H is player 2), $u_H(\hat{s}_H, s_M) > u_M(\hat{s}_H, s_M)$. This implies that H's payoff must be strictly greater than M^*'s payoff in the repeated contest.

As desired, H outperforms M^* in the repeated contest, which contradicts to the supposition that M^* is superhuman. □

The proof strategy can be explained in simpler terms as follows. First, notice that if M^*'s strategy profile s is an SPNE in the centipede game, then **SHM** must be violated due to Lemma 1. Next, suppose that M^* stops at the first

node with a high probability (strictly below 1). But then it would be impossible for M^* to receive the average sample payoff in the centipede game. Therefore, M^* must choose C at the root with a high enough probability to receive the average sample payoff. Note that H takes the strategy s_M of M^* as given by **MS**, H chooses a pure best-response to the strategy of M^* by **R**, and M^* cannot predict H's strategy by **SU**. These assumptions imply that H outperforms M^* in the repeated contest $G^k_{1,2}$ for any k because whether H is the first player or the second player, H receives a strictly greater payoff than M^*. Therefore, a contradiction is obtained. **SU**, **MS**, and **R** imply that **SHM** does not hold.

While the proof of this theorem depends on constructing a counterexample using the centipede game, this particular example is not 'pathological.' Instead, it is an empirically validated example of a non-zero-sum game, where players can cooperate efficiently, despite theoretical predictions. The proof strategy could also be applied to other well-studied games where cooperation is commonly observed.

I next explore the "tightness" of Theorem 1 as mentioned earlier.

Proposition 1. *Theorem 1 is tight: Any three of the four assumptions, **SU**, **MS**, **R**, and **SHM**, are consistent.*

Proof. To prove this proposition, I drop each of the four assumptions **SU**, **MS**, **R**, and **SHM** one by one and show that the remaining three assumptions do not lead to any contradictions.

Superhuman Machine: I begin by assuming that **MS**, **SU**, and **R** hold, but **SHM** does not. This is the easiest case, as there is no restriction on the behavior of the machine under these assumptions. Thus, these three assumptions are consistent.

Machine Strategy: Assuming that **SU** and **R** hold but **MS** does not hold, H would best-respond to some belief about M^*'s strategy. However, there would be no guarantee that H's belief is correct, which means H would not necessarily be able to outperform M^*. This implies that M^* *may* be superhuman. Therefore, **SU**, **R**, and **SHM** are consistent.

Rationality: Assume that **SU** and **MS** hold, but **R** does not. Then, this assumption would *not* contradict the assumption that M is superhuman. This is because if H fails to act rationally, then they may select a strategy that leads to being outperformed by M^*, which is consistent with the **SHM** assumption. As a result, **SU**, **MS**, and **SHM** are consistent.

Strategic Unpredictability: Assuming that **MS** and **R** hold, but **SU** does not, M^* might be able to program H's brain and predict precisely what H will choose and can best respond. This implies that one of the players could outguess the other player, depending on perhaps the computational power of M^*. As a result, one cannot rule out the scenario that M^* outperforms H in every game, in which case the theorem would not hold. This implies that **MS**, **R**, and **SHM** are consistent. □

3. Conclusions

This paper examines the emergence of superhuman artificial intelligence (AI) through a game theoretical perspective, considering the theoretical factors that could impact its development. Using a general-sum framework to model strategic interactions between a representative human agent and a potential superhuman machine agent, I show that under certain assumptions, it is not possible for superhuman AI to consistently outperform the representative human player.

My analysis identifies four key assumptions underlying some of the arguments about the development of superhuman AI: Superhuman Machine, Machine Strategy, Rationality, and Strategic Unpredictability. I show that these assumptions are inconsistent when taken together and that this result is "tight" in the sense that relaxing any one of them results in a consistent set of assumptions. By identifying these assumptions and their inconsistencies, this paper contributes to a better understanding of the theoretical context that can shape the development of superhuman AI.

3.1. Extension to the n-player case: challenges and preliminary observations

A potential area for future research is to explore the possibility of multiple machines interacting with multiple human agents. In what follows, I first provide those definitions that naturally extend to the n-player case, and then I discuss the extensions that are less straightforward in this context.

The concept of a *repeated contest* extends naturally to the n-player case, defined as $G^k_{1,2,...,n!} = (G_1, G_2, ..., G_{n!})^k_{i=1}$, where each stage game consists of $n!$ mini-games. This is because each player should play the game in every player's

position to be able to determine which player 'outperforms' the others. The definition of outperformance is then simply updated as follows: player i is said to outperform player $j \neq i$, if there exists a strategy, $\bar{s}_i$, of player i in the repeated contest such that for any $k \in \{1, 2, ...\}$, $\bar{u}_i(\bar{s}_i, \bar{s}_j) > \bar{u}_j(\bar{s}_i, \bar{s}_j)$, where $\bar{s}_j$ denotes the strategy of player $j \neq i$. Player i outperforms other players if i outperforms every player $j \neq i$.

The assumptions of rationality and strategic unpredictability can also be naturally extended to the n-player case. However, extending the machine strategy and superhuman machine assumptions is less straightforward. One approach to defining machine strategy is to assume that each human player takes machines' strategies as given. This raises a crucial question: what assumptions do human players make about the strategies of other humans? These assumptions are important, as they can significantly affect the outcomes (more on this below).

Suppose that a machine is called 'superhuman' if it outperforms every human player in some games, never outperformed by any human player, and on average receives at least the sample average payoff in those games. While this definition seems intuitive, it gives rise to at least two significant problems in n-player games. Firstly, there is an issue of strategic cooperation among the machines. Through either coordination, communication, or just by coincidence, machines may choose strategies that disproportionately benefit a specific machine, leading to "false positives." This situation is similar to coalitional manipulation, where some machines 'sacrifice' their own payoffs for the superhuman machine. Similarly, human players could also form coalitions, in which case the n-player game would essentially reduce to a two-player game between a coalition of humans and machines. We can then apply the standard two-player definition of a superhuman machine.

Secondly, one might consider defining a machine's performance by the average payoff of all machines to avoid the false positives mentioned above. However, this creates another problem: a machine that truly outperforms all human players could be affected by the poor performance of other machines. This situation leads to false negatives, where truly superhuman machines are not classified as superhuman.

To avoid false negatives, games that support a variable number of players, such as public goods games, can be reduced to a two-player game via an 'elimination contest' in which case the standard two-player definition of a superhuman machine can be applied. In this contest, after each round, the worst-performing human and machine player are eliminated.[9] This process continues until only one human and one machine player remain. The final round then becomes a two-player game, in which case the superhuman machine definition applies.

However, in more general n-player games, it is not always defined how the payoff function changes when the game is played with fewer players. To address this, the definition of the elimination contest can be modified. As above, after each round, a maximum of two worst-performing players (one human, one machine) are eliminated. The contest then continues with $(n - 2)$ players. Remaining human players in each mini-game choose a strategy for themselves and a strategy for the eliminated human player. Similarly, each remaining machine player chooses a strategy for itself and a strategy for the eliminated machine player. Thus, the total number of mini-games increases to $n!(n_1 - 1)(n_2 - 1)$, where n_1 and n_2 denote the initial number of human and machine players, respectively. This increase results from creating a new mini-game for every additional strategy choice of the remaining $(n_1 - 1)$ human players and $(n_2 - 1)$ machine players. Similarly, in the next $(n - 4)$-player contest, remaining players continue choosing strategies both for themselves and each eliminated player within their category (human or machine). This process ensures that the final round will be a two-player game between a human and a machine.

While this elimination approach does reduce n-player games to two-player games, making standard two-player definitions applicable, it essentially sidesteps the challenges of defining the concept of a superhuman machine in n-player games. In conclusion, a potential area for future research is to explore alternative, more direct definitions of superhuman machine in such games that do not rely on eliminating players from the contest.

References

[1] Acemoglu, D. (2022). Harms of AI. In J. B. Bullock et al. (Eds.), *The Oxford Handbook of AI Governance* (online ed.). Oxford University Press.

[2] Aumann, R. J. (1998). On the centipede game. *Games and Economic Behavior 23*(1), 97–105.

[3] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[4] Brown, N. and T. Sandholm (2019). Superhuman AI for multiplayer poker. *Science 365*(6456), 885–890.

[9] If there is only one human player (or machine player) left in any round, then that player is exempt from elimination.

[5] Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

[6] Brynjolfsson, E. and A. McAfee (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton.

[7] Campbell, M., A. J. Hoane Jr, and F.-h. Hsu (2002). Deep Blue. *Artificial Intelligence 134*(1-2), 57–83.

[8] Chesney, B. and D. Citron (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review 107*, 1753.

[9] Cohen, M. K., M. Hutter, and M. A. Osborne (2022). Advanced Artificial Agents Intervene in the Provision of Reward. *AI Magazine 43*(3), 282–293.

[10] Daffertshofer, A., A. R. Plastino, and A. Plastino (2002). Classical No-Cloning Theorem. *Physical Review Letters 88*(21), 210601.

[11] Everitt, T., G. Lea, and M. Hutter (2018). AGI Safety Literature Review. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5441–5449. AAAI Press.

[12] Fey, M., R. D. McKelvey, and T. R. Palfrey (1996). An experimental study of constant-sum centipede games. *International Journal of Game Theory 25*(3), 269–287.

[13] Floridi, L. and J. Cowls (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review 1*(1), 535–545.

[14] Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books.

[15] Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell (2017). The Off-Switch Game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 220–227. AAAI Press.

[16] Horvitz, E. (2014). One Hundred Year Study on Artificial Intelligence: Reflections and Framing. White paper, Stanford University, Stanford, CA (ai100.stanford.edu).

[17] Isaksen, A., M. Ismail, S. J. Brams, and A. Nealen (2015). Catch-Up: A Game in Which the Lead Alternates. *Game & Puzzle Design 1*(2), 38–49.

[18] Krockow, E. M., A. M. Colman, and B. D. Pulford (2016). Cooperation in repeated interactions: A systematic review of centipede game experiments, 1992–2016. *European Review of Social Psychology 27*(1), 231–282.

[19] Levitt, S. D., J. A. List, and S. E. Sadoff (2011, April). Checkmate: Exploring backward induction among chess players. *American Economic Review 101*(2), 975–90.

[20] London, A. J., Y. S. Razin, J. Borenstein, M. Eslami, R. Perkins, and P. Robinette (2023). Ethical Issues in Near-Future Socially Supportive Smart Assistants for Older Adults. *IEEE Transactions on Technology and Society*.

[21] McKelvey, R. D. and T. R. Palfrey (1992). An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society 60*(4), 803–836.

[22] Nagel, R. and F. F. Tang (1998). Experimental results on the centipede game in normal form: an investigation on learning. *Journal of Mathematical Psychology 42*(2-3), 356–384.

[23] Nash, J. (1951). Non-Cooperative Games. *The Annals of Mathematics 54*(2), 286–295.

[24] Rosenthal, R. (1974). Correlated equilibria in some classes of two-person games. *International Journal of Game Theory 3*(3), 119–128.

[25] Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal 117*(523), 1243–1259.

[26] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group.

[27] Schaeffer, J., N. Burch, Y. Björnsson, A. Kishimoto, M. Müller, R. Lake, P. Lu, and S. Sutphen (2007). Checkers Is Solved. *Science 317*(5844), 1518–1522.

[28] Scharre, P. (2019). *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton.

[29] Schiff, D., J. Biddle, J. Borenstein, and K. Laas (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 153–158.

[30] Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*.

[31] Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature 529*(7587), 484–489.

[32] Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science 362*(6419), 1140–1144.

[33] Tesauro, G. (1995). Temporal Difference Learning and TD-Gammon. *Communications of the ACM 38*(3), 58–68.

[34] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind 59*(236), 433–460.

[35] Walz, A. and K. Firth-Butterfield (2019). Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to the Development of an AI Governance Regime. *Duke Law & Technology Review 18*, 176.

[36] Wootters, W. K. and W. H. Zurek (1982). A single quantum cannot be cloned. *Nature 299*, 802–803.

[37] Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom and M. Ćirković (Eds.), *Global Catastrophic Risks*, Volume 1, pp. 308–345. New York: Oxford University Press.