# Political economy of superhuman AI

Mehmet S. Ismail*

Monday 6th March, 2023

**Abstract**

In this note, I study the institutions and game theoretic assumptions that would prevent the emergence of 'superhuman-level' arfiticial general intelligence, denoted by AI*. These assumptions are (i) the "Freedom of the Mind," (ii) open source "access" to AI*, and (iii) rationality of the representative human agent, who competes against AI*. I prove that under these three assumptions it is impossible that an AI* exists. This result gives rise to two immediate recommendations for public policy. First, 'cloning' digitally the human brain should be strictly regulated, and hypothetical AI*'s access to brain should be prohibited. Second, AI* research should be made widely, if not publicly, accessible. *JEL*: C70, C80

*Keywords*: Artificial intelligence, political economy, AGI, game theory, non-zero-sum games

## 1 Introduction

In 2015, over 150 artificial intelligence (AI) experts signed an open letter calling researchers from disciplines such as economics, law, and philosophy for doing future research on maximizing the societal benefit of AI.[1] The letter, which has now been signed by over 8000 people, comes with an accompanying paper by Russell, Dewey, and Tegmark (2015) who quote in part the following passage from Horvitz (2014).[2]

> ...we could one day lose control of AI systems via the rise of superintelligences that do not act in accordance with human wishes—and that such powerful systems would threaten humanity... Are such dystopic outcomes possible? If so, how might these situations arise? What are the paths to these feared outcomes? What might we do proactively to effectively address or lower the likelihood of such outcomes, and thus reduce these concerns? What kind of research would help us to better understand and to address concerns about the rise of a dangerous superintelligence or the occurrence of an "intelligence explosion"?

In this note, I study institutions and game theoretic assumptions that would prevent the emergence of superhuman-level arfiticial general intelligence. The first assumption is called "Freedom of the Mind,"

*Department of Political Economy, King's College London, London, UK. E-mail: mehmet.ismail@kcl.ac.uk

[1] https://futureoflife.org/2015/10/27/ai-open-letter/

[2] Whether human-level AI will be achieved or not has long been discussed; for a literature review, see, e.g., Everitt, Lea, and Hutter (2018), and the references therein. For potential economic and political harms of AI technologies in the short-term, see Acemoglu (2021).

which essentially prohibits 'cloning' digitally the human brain. The game theoretic aspect of this assumption is that AI* cannot take the representative (human) agent's strategy in a game as given. The second assumption is called 'Access' which gives the representative agent the permission to access AI*'s source code so as to take AI*'s strategy as given. The third assumption is 'Rationality' which means that the human agent chooses the strategy that maximizes their payoff, given the strategy of AI*. I prove that under these three assumptions it is impossible that an AI* exists. This result gives rise to two immediate policy recommendations. First, 'cloning' digitally the human brain should be strictly regulated, and potential AI*'s access to brain should be banned. Second, AI* research should be made widely, if not publicly, accessible.

# 2 The setup

## 2.1 Two-person (general-sum) perfect information games

Let $G = (N, X, I, u, S)$ 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 $i \in \{1, 2\}$, the set of pure actions $A_i = \bigcup_{x|I(x)=i} A_i(x)$ has finitely many elements.

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 $S'_i = \bigtimes_{x|I(x)=i} A_i(x)$ denote the set of all pure strategies of $i$, and $s' \in S' = \bigtimes_{i \in N} S'_i$ a pure strategy profile. A mixed strategy $s_i$ of player $i$ is a probability distribution over $S'_i$, and $S_i = \Delta(S'_i)$ is the set of all mixed strategies of player $i$. Let $s \in S$ 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 $u_i : S \to \mathbb{R}$. Let $s^*_i \in BR_i(s_j)$ denote a *best-response* of player $i$ to player $j$'s strategy $s_j$, i.e., $s^*_i \arg\max_{s'_i \in S_i} u_i(s'_i, s_j)$.

$G$ is two-player game played between a representative human agent, denoted by $H$, and a machine agent, denoted by $M$. 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.

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

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 $s \in S$ 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 $s'_i|x \in S_i|x$.

To define the nature of competition between the human agent and $M$, I introduce the following definition.

**Definition 3** (Repeated contest)**.** 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 $H$. Let $G^k_{1,2}$, $k \geq 1$, denote the *repeated contest* game in which each stage game consists of two games, $G_1$ and $G_2$, and each stage game is repeated $k$ times.

In simple words, the repeated contest between $H$ and $M$ is defined as the repeated game in which each stage game consists of two 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. For example, in the Chess World Championship, the players play an equal number of games with white pieces to account for any potential first-mover advantage. I formalize the concept of outperformance as follows.

**Definition 4** (Outperformance)**.** Let $G$ be a two-person perfect information game, $G^k_{1,2}$ the repeated contest, and $s$ be the players' strategy profile in $G^k_{1,2}$. Player $i \in \{H, M\}$ is said to *outperform* player $j \neq i$ if for any $k \in \{1, 2, ...\}$, $u_i(s) > u_j(s)$.

In plain words, player $i$ outperforms player $j$ in game $G$ if, no matter how many times the contest is repeated, player $i$'s expected payoff is strictly greater than player $j$'s.[3] However, the number of repetitions needed to determine the "better" player in practice 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

### 2.3.1 Superhuman machine

I define a 'superhuman' artificial intelligence, denoted by $M^*$, as an artificial general intelligence that is equipped with finite but significant computing power, and is able to take any game $G$ as an input and output a *solution*—i.e., a mixed strategy profile— based on its source code and computational 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. Updating its solution as the game proceeds is also possible, similar to chess engines.

Determining whether a machine is 'human-like' or 'superhuman' is a subjective matter that involves human judgments, such as the well-known Turing test (Turing, 1950). To define a superhuman machine, I first introduce a useful concept, namely the sample average of a two-player game played by a population of human agents.

**Definition 5** (Sample average)**.** Consider a population of human agents playing a two-player game $G$, and let $\{(u^1_1, u^1_2), (u^2_1, u^2_2), ..., (u^n_1, u^n_n)\}$ be the dataset of payoffs, where $(u^j_1, u^j_2)$ is the payoff received by player 1 and player 2 from the $j$th game of $G$. It is possible that each game is played by different players.

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

Then, the sample average is defined as follows:

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

The sample average $\mu(G)$ of a game $G$ is determined by the empirical average payoff received by a group of humans players who participate in playing the game. 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

1. there exists $G' \in \Gamma$ such that $M$ outperforms $H$ in $G'$,
2. for every $G \in \Gamma$, $M$ is not outperformed by $H$ in $G$, and
3. for every $G \in \Gamma$, there exists a strategy of human agent $s_H$ such that given machine's strategy $s_M$, $u_M(s_M, s_H) \geq \mu(G)$.

In simple terms, for an artificial intelligence to be classified as superhuman ($M^*$), it must outperform a human player ($H$) in some games and never be outperformed by $H$ in any game. Additionally, $M^*$ should be able to achieve at least the sample average payoff from every game. While these first two conditions would be sufficient for defining superhuman machine in zero-sum games, the third condition is necessary in non-zero-sum games in which cooperation is not only possible but also common.[4] Therefore, to avoid aggressive machine strategies that aim to minimize the human's payoff while also minimizing their own payoff in non-zero-sum games, I introduce the third condition. In general-sum games, a superhuman $M^*$ should not only outperform a rational agent $H$ but also perform well among a diverse population of humans. This leads to the first assumption of my paper.

**Assumption 1** (Superhuman Machine: **SHM**)**.** *Superhuman Machine (**SHM**) holds if $M$ is a superhuman machine, denoted by $M^*$.*

### 2.3.2 Machine Transparency

The assumption of Machine Transparency requires that the human agent $H$ has the permission to access the strategy of the superhuman machine $M^*$ and take it as given during the game. Formally, **MT** is stated as follows.

**Assumption 2** (Machine Transparency: **MT**)**.** *Machine Transparency (**MT**) is satisfied if for every game $G \in \Gamma$ and at any node $x$ in game $G$, $H$ takes $M^*$'s strategy $s_M \in S_M$ as given.*

In other words, $H$ is granted 'read' and 'copy' permissions in the sense that they can read the source code of $M^*$ and copy its strategy to give a response to this strategy. However, $H$ is not granted any

[4]For further discussion, see section 3.

other permissions and cannot necessarily modify the source code or the computing power of $M^*$. This assumption is crucial for analyzing the game-theoretic implications of the emergence of superhuman machine intelligence, as it ensures that $H$ plays the game with full knowledge of $M^*$'s strategy.

### 2.3.3 Rationality

In this note, I use the standard rationality assumption, which refers to the idea that the human player ($H$) is acting in a way that is consistent with their own self-interest.

**Assumption 3** (Rationality: **R**). *Player $i = H$ is rational if in every game $G \in \Gamma$, for every strategy $s_j$ of $M^*$, where $j \neq i$, $H$ chooses a strategy*

$$s_H^* \in \arg\max_{s_i' \in S_i} u_i(s_i', s_j). \tag{2.1}$$

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

In other words, $H$ chooses a strategy that maximizes their own expected utility given $M^*$'s strategy, which is feasible under the **MT** assumption.

### 2.3.4 Mental Privacy

The Mental Privacy assumption concerns the control of agent $H$ over their own mind. Specifically, it states that $H$ has exclusive control over their own brain and that $M^*$ cannot access or control it in any way. In the context of this paper, I assume that $M^*$ is not allowed to "digitally clone" $H$'s brain, which means that $M^*$ cannot program $H$'s brain in a way that would enable it to predict $H$'s decisions either deterministically or non-deterministically. This assumption ensures that $H$ has the freedom to choose their actions independently of any assumptions or predictions made by $M^*$, thus preserving the unpredictability of $H$'s strategy.

**Assumption 4** (Mental Privacy: **MP**). *Let $s_H' \in S_H$ be $M^*$'s prediction of $H$'s strategy in a game $G$. Mental Privacy (**MP**) holds if $H$ is free to choose a strategy $s_H \in S_H$ such that $s_H \neq s_H'$ regardless of $G$ in which $H$ has at least two pure strategies.*

The essence of the **MP** assumption is that if a human player has full control over their own brain, then $M^*$ cannot always predict their strategy. In other words, regardless of the code of $M^*$ regarding the strategy of $H$, $H$ can change their strategy in any way they wish. 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 assumed by $M^*$.

It is important to note that this is a mild assumption since a human player, who has access to the source code of $M^*$, can always change the strategy that $M^*$ assumes for them. Therefore, this assumption excludes the possibility that $M^*$ can control or replicate the brain of the human player. Additionally, Ismail (2022) shows that the 'mutual knowledge of rationality' and 'mutual knowledge of correct beliefs' do not hold in general in $n$-person games, including two-person games. This result means that it is impossible for both players to be rational and have correct predictions about the choices of the other player. However, since player $H$ has access to $M^*$, they can predict its strategy, making it impossible for $M^*$ to predict $H$'s choice if $M^*$ is rational.

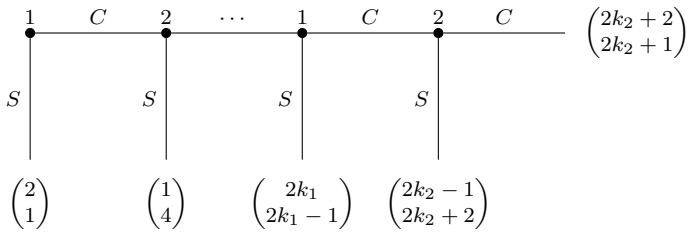


Figure 1: Payoff function of a linearly increasing-sum centipede game.

## 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 (1974) 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. For example, suppose that there are $m \geq 2$ (even) decision nodes and let $k_i \in \{1, 2, ..., \frac{m}{2}\}$ be the node such that player $i$ is active. Figure 1 illustrates the payoff structure of a linearly increasing-sum centipede game due to Aumann (1998).

There have been numerous experimental studies on the centipede game and its variations since the work of McKelvey and Palfrey (1992). These studies include, among others, Fey, McKelvey, and Palfrey (1996), Nagel and Tang (1998), Rubinstein (2007), Levitt, List, and Sadoff (2011), and Krockow, Colman, and Pulford (2016), 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. (2016), 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 (Krockow et al., 2016, 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:* $\mu(G) > 0.8m - 0.5$.

*Proof.* According to the meta study conducted by Krockow et al. (2016), 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 assumptions are called *consistent* if they do not lead to any logical contradiction. They are called *inconsistent* if they are not consistent.

The following theorem shows that the existence of a superhuman $M$ is impossible if the three main assumptions hold.

**Theorem 1** (Impossibility of $M^*$)**.** *The assumptions **MP**, **MT**, **R** and **SHM** are inconsistent.*

*Proof.* Assuming that **MP**, **MP**, and **R** hold and that $M^*$ exists, I will prove by contradiction that $H$ outperforms $M^*$ in an increasing-sum centipede game $G$.

To begin, let $s \in S$ be $M^*$'s solution in game $G$, defined by the payoff function in Figure 1. Suppose that $s_i|g \in BR_j(s_j|g)$ for every $i$ and every subgame $g$ of $G$, 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 is easy to see because 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 to the previous player 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.

Suppose $s_M(x_0)(S) > 0.75$, 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 **MT** assumption, $H$ takes the strategy $s_M$ of $M^*$ as given. Then, **R** implies that $H$ chooses $s^*_H \in \arg\max_{s'_i \in S_i} u_i(s'_i, s_M)$, i.e., $H$ best-responds to the strategy of $M^*$. Furthermore, the **MP** assumption implies that $H$'s strategy cannot be predicted by $M^*$, so $s_M \notin BR_M(s^*_H)$. 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 outcome is to choose S at the first node, which leads to a contradiction as shown above.

Therefore, $H$ outperforms $M^*$ in the repeated contest $G^k_{1,2}$ 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), $u_H(s^*_H, s_M) > u_M(s^*_H, s_M)$. This implies that $H$'s payoff must be strictly greater than $M^*$'s payoff in the repeated contest. The payoff function of $G$ ensures that the best-responding player receives a greater payoff than the other player unless player 1 chooses S with a high enough probability at the root of the game. 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 in seven main steps.

1. To reach a contradiction, suppose that **MP**, **MT**, **R**, and **SHM** all hold.
2. If $M^*$'s solution $s$ is an SPNE in the centipede game, then **SHM** must be violated due to Lemma 1.
3. Now suppose that $M^*$ stops at the first node with a high probability (but strictly below 1). But then it would be impossible for $M^*$ to receive the average sample payoff.
4. Therefore, $M^*$ must choose C at the root with a high enough probability to receive the average sample payoff.
5. Note that $H$ takes the strategy $s_M$ of $M^*$ as given by **MT**, $H$ best-responds to the strategy of $M^*$ by **R**, and $M^*$ cannot predict $H$'s strategy by **MP**.
6. 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 $M^*$.
7. Therefore, a contradiction is obtained. **MP**, **MT**, and **R** imply that **SHM** does not hold.

I next explore whether the four assumptions in Theorem 1 are *tight* in the sense that whether any three of the four assumptions are consistent.

**Proposition 1.** *The assumptions of Theorem 1 are tight.*

*Proof.* To prove this proposition, I drop each of the four assumptions **MP**, **MT**, **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 **MT**, **MP**, 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 Transparency**: Assuming that **MP** and **R** hold but **MT** 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, **MP**, **R**, and **SHM** are consistent.

**Rationality**: Assume that **MP** and **MT** 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, **MP**, **MT**, and **SHM** are consistent.

**Mental privacy**: Assuming that **MT** and **R** hold, but **MP** does not, $M^*$ might be able to program $H$'s brain and predict precisely what $H$ will choose and can best respond. However, Ismail (2022) shows that players cannot be both rational and predict the others' strategies correctly. 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 **MT**, **R**, and **SHM** are consistent. □

# 3 Discussion and conclusions

This paper examines the emergence of superhuman AI through a political economy perspective, considering institutional and societal factors that could impact its development. Using a game-theoretic framework to model strategic interactions between a human agent and a potential superhuman machine agent, I show that under certain assumptions, it is not possible for superhuman AI to consistently outperform humans in two-person games. This is a significant finding since many scholars have warned of the possible existential risks and moral dilemmas that superhuman AI could entail.

My analysis identifies four key assumptions underlying some of the arguments about the (dangers of) superhuman AI: Mental Privacy, Machine Transparency, Rationality, and Superhuman Machine. I show that these assumptions are inconsistent when taken together and "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 political economic context that can shape the development of superhuman AI.

Based on my findings, I propose two policy recommendations. The first is to regulate the cloning of human brains and restrict superhuman AI's access to human neural data. This recommendation is based on the Mental Privacy assumption, which suggests that digital cloning or copying of human brains should be prohibited to prevent the possibility of a superhuman AI predicting a human's strategy in any game. The second is to make superhuman AI research widely accessible for transparency and accountability based on the Machine Transparency assumption, which means that human agents should be able to access information about a superhuman AI's behavior.

It is worth noting that my analysis has some limitations. First, the game-theoretic framework I use does not fully account for the political economic implications of my assumptions. For instance, even if mental privacy laws are enacted, a superhuman machine may still be able to discover by itself what humans will choose in every game without accessing their brains. Second, the proof of my main theorem depends on constructing a counterexample using the centipede game. However, this counterexample is not a 'pathological' case, but rather an empirically validated example of a non-zero-sum game where humans can cooperate efficiently despite theoretical predictions. The proof could be generalized to other games where cooperation is crucial. Third, my analysis does not consider the possibility of multiple superhuman machines interacting with multiple human agents. This scenario may introduce new challenges for formalizing the cooperation and conflict between humans and the machines.

Despite its limitations, my analysis contributes to the ongoing debate about the emergence of superhuman AI by offering a formal game-theoretic framework for modeling potential strategic interactions between a human agent and a superhuman machine. By identifying the assumptions that underlie some of the existing arguments about the threats of superhuman AI, and showing their inconsistency when assumed together, this paper provides a new perspective on this complex issue.

# References

Acemoglu, D. (2021). Harms of AI. *Oxford Handbook of AI Governance*, forthcoming.

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

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.

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.

Horvitz, E. (2014). One hundred year study on artificial intelligence: reflections and framing. White paper, Stanford University, Stanford, CA (ai100.stanford.edu).

Ismail, M. S. (2022). Mutual knowledge of rationality and correct beliefs in $n$-person games: An impossibility theorem. *arXiv:2209.09847*.

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.

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.

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

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.

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

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

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

Russell, S., D. Dewey, and M. Tegmark (2015, Dec.). Research priorities for robust and beneficial artificial intelligence. *AI Magazine 36*(4), 105–114.

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

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