HOW TO BENCHMARK AGI: THE ADVERSARIAL GAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, it has been observed that the performance of foundational models, especially in Natural Language Processing (NLP) and Computer Vision (CV), keeps increasing rapidly, and new emergent capabilities continue to appear with increasing scale. Some researchers claim that we could soon reach a point where future models are generally more capable than humans. Due to this possible scenario, and the critical safety risks involved, it's paramount that we're able to accurately access and measure the capabilities of future models.

However, we find that the related terms of Artificial General Intelligence (AGI) and Artificial Super-Intelligence (ASI) are ill-defined, and no definite benchmarking process is proposed. Mitigating this gap is the aim of this work. Summarizing related literature, we propose precise definitions for AGI and ASI. Moreover, to tackle the benchmarking problem, we propose a new test, which we name "the Adversarial Game (AG)". We show that AG is complete, in the sense that a system is AGI if and only if it can consistently win the AG against human players. Further, we further show that previous attempts to define AGI can be cast as special cases of AG. Finally, under some standard assumptions, a system's performance in AG is readily measurable. Similarly, we propose related criteria for ASI. Overall, we hope that the proposed methodology can help the community towards better accessing the capabilities of future models.

1 INTRODUCTION

Defining and measuring intelligence, and specific aspects of it, in humans and animals has a rich history in psychology and cognitive sciences Guilford (1967); Gardner (2011); Roberts (1998). However, principles discovered from these cases do not directly transfer to machines. For example, in the 90's, chess was considered as a pinnacle of human intelligence: it involves creativity, pattern recognition, strategy, long-term planning and decision making, and more. Yet, a computer named Deep Blue Campbell et al. (2002) managed to win against world champion Garry Kasparov in 1997, merely via brute-force calculation, without much understanding of chess.

Thinkers started pondering on the question of machine intelligence soon after the first digital computers were conceptualized; a seminal work is Turing (2009), where Alan Turing introduces the idea of a computer learning similarly to a child, and speculates about the prospects of Artificial Intelligence (AI). Moreover, he claims that the question "Can a machine think?" is ill-defined, and proposes the Turing test as an actionable alternative: if a human observer cannot distinguish whether he discusses with another human or a machine (communicating by text), then we have to conclude that the machine is indeed "intelligent" for all practical purposes.

Soon afterwards, AI was established as a branch of Computer Science, with the long-term goal of creating General Artificial Intelligence (AGI). Over the decades, additional definitions of AGI and machine intelligence have been proposed Minsky (1988); Newell & Simon (2007). Legg (2008) summarizes these and many additional definitions, and proposes to measure machine intelligence using ideas from Hutter (2004). According to the definition of Legg (2008), and taking into account the considerations in Chollet (2019), machine intelligence could be defined as "the ability to learn and solve problems in a wide range of environments". Similarly, AGI has been defined a "a system that can do anything a human can do" Legg & Hutter (2007). Unfortunately, as can be seen also in the survey of Legg (2008), methodologies proposed to test machine intelligence are currently not adequate: for example, while the Turing test is a long-standing and promising idea, simple systems have managed to convince as human in the past Weizenbaum (1966). Also, additional criticism

focuses on the lack of aspects of intelligence involving visual perception and motion in the Turing test LeCun (2022).

Very recently, large foundational models Bommasani et al. (2021) in NLP Brown et al. (2020); OpenAI (2023), CV Alayrac et al. (2022), combined modalities Alayrac et al. (2022) or Robotics Brohan et al. (2023) are revolutionizing the field of AI, unlocking remarkable performance across a wide range of tasks. Moreover, new capabilities seem to emerge by scaling the model and dataset sizes Wei et al. (2022). This has renewed discussions about the possibility of AGI, where some researchers claim that we could soon reach a point where future models are generally equivalent or more capable than humans, and even go beyond that, achieving Artificial Super-intelligence (ASI).

Due to this possibility, it is crucial to accurately define what exactly we mean by those terms, and more importantly, how we can measure the capabilities and generality of current and future AI models. In this work, we attempt to lay a foundation on both these aspects.

First, starting out from the literature summarized above, we propose two definitions for AGI and ASI that are detailed and precisely described. Instead of trying to find a single task (such as the Turing test) that is assumed to be equivalent to AGI, we instead want to simply capture in precise terms the essence of the proposed definitions in the past.

Second, armed with this foundation, we then propose possible ways to measure them, and to be able to accurately access model capabilities. Towards this, we propose a new test, which we name "the Adversarial Game (AG)". AG attempts to yield a measurement process that aligns with our definition of AGI, and in fact we're able to show that is equivalent with it, in the sense that a system is an AGI if and only if it can consistently win against human players in AG. Under some standard assumptions, we can show that a system's performance in AG can be measured and estimated. Additionally, we propose a simple definition and criterion to characterize if a system is an ASI.

Current state-of-the-art Large Language Models (LLMs) can already understand the objectives of AG, and answer or propose challenging questions Bubeck et al. (2023). Due to this, we think that larger scale evaluations will allow us to get a more clear idea of the true capabilities of such models, as well as a way to measure future versions, and the relative progress made.

Overall, we hope that our approach can be used as a foundation to more precisely measure the performance of current and future AI models. Especially, due to the risks involved in the case where AGI systems materialize, it is critical to have these considerations from now on, and approach these questions earnestly.

2 THE DEFINITIONS

In this section, beginning from the literature summarized previously, we attempt to provide definitions for AGI and ASI. Our emphasis is preciseness and measurability.

First, we start with AGI. According to the previous descriptions, a system is AGI if it can do "any work a human can do" Simon (1965) or "any economically valuable task" Eloundou et al. (2023). As we observe, the emphasis here is on generality: an AGI system must be able to perform a wide range of tasks, and not only some narrow ones. Moreover, it should be competitive with humans in all of these tasks. Hence, we propose the following definition:

Definition 2.1 (AGI)**.** Let T be any well-defined task, and $Perf_T$ be a corresponding performance measure for T. Moreover, let H be a individual randomly drawn from the human population D_H ($H \sim D_H$). Then, a system S is an AGI if and only if:

$$\forall T : \mathbb{P}_{H \sim D_H}[Perf_T(S) \geq Perf_T(H)] \geq \frac{1}{2} \tag{1}$$

That is, S is an AGI if and only if (iff) it can outperform a person selected at random in any task. A potential weak points of def. 2.1 is on how to measure performance for various tasks. We'll try to elaborate more on this in the next section.

Next, the definition of ASI is more straightforward:

Definition 2.2 (ASI)**.** In the setup of def. 2.1, a system S is ASI if and only if we have:

Algorithm 1 Adversarial Game (AG)

Input: players P_1, P_2, judge J, max score n
Output: winner w
$s_1, s_2 \leftarrow 0$ {scores for both players}
$p \leftarrow 0, o \leftarrow 1$ {init. players' turn}
while $s_1 < n \wedge s_2 < n$ **do**
  $q, a_p \sim P_p$ {generate question q and candidate answer a_p}
  $a_o \leftarrow P_o(q)$ {generate opponent answer a_o}
  $v \leftarrow J(q, a_p, a_o)$ {get judge verdict}
  if $v =$ "p" **then**
    $s_p \leftarrow s_p + 1$
  else if $v =$ "o" $\vee\ v =$"ill-defined" **then**
    $s_o \leftarrow s_o + 1$
  end if
  $p, o \leftarrow swap(p, o)$ {switch turns}
end while
if $s_1 > s_2$ **then**
  $w \leftarrow 1$
else
  $w \leftarrow 2$
end if

$$\forall T, \forall H \in D_H : Perf_T(S) \geq Perf_T(H) \tag{2}$$

That is, S is an ASI iff it can outperform any human at any task.

Having laid down these definitions, the next - and most important - question is: how to measure them? This is what we attempt to do in the next sections.

3 Benchmarking AGI: the Adversarial Game

After having laid down the definitions, the critical question is: how can we measure / test the capabilities or a system? For this end, we propose the Adversarial Game (AG). AG is, basically, a straightforward realization of def. 2.1. It works as follows:

Two players, a human and an AI, take turns and in each round, the player ask the opponent a question q. The objective is to come up with questions that (1) are well defined, so that one can judge whether a potential answer is "good" or not, and (2) are challenging and ideally involve intelligence and creativity. After the opponent answers the question (a_o), the player also "reveals" their own answer to the question, a_p (the player has to output a_p without seeing a_o).

Then, we enter the scoring phase of the game: an independent judge, having no knowledge of the players and whose answer is which, receives the question q and both answers a_o and a_p. The judge has to decide which response answers the question better / more comprehensively. If the judge declares that a_o is a better answer than a_p, then the opponent wins the point of that round. Otherwise, if a_p is deemed better, the player wins the point instead. The judge can also output "equivalent", meaning that he judges both a_p and a_o to be equivalently good at answering q, in which case we have a tie and none wins the point. Finally, the judge can also output "ill-defined" meaning that, according to his judgement, the question is nonsense, or a definite answer does not exist. In that case, the opponent wins the point. The game continues in this fashion, until one player reaches a predefined score first.

AG can be summarized by alg. 1.

What are the properties of the Adversarial Game? First, we see that AG is complete: that is, a system satisfies def. 2.1 if and only if it can consistently win AG against randomly selected human players:

Proposition 3.1. *A system S is an AGI according to def. 2.1 iff it can consistently win AG against randomly (uniformly) selected humans: that is, in the setup of def. 2.1 we have:*

$$\mathbb{P}_{H\sim D_H}[AG(S,H)=S]\geq\frac{1}{2} \tag{3}$$

where $AG(S,H)$ represent the outcome of an Adversarial Game against S and H.

Proof. Indeed: on the one direction, if S satisfies def. 2.1, then we have $\mathbb{P}_{H\sim D_H}[Perf_T(S)\geq Perf_T(H)]\geq\frac{1}{2}$ for any task T, and thus, S will win any possible round in AG with probability at least $\frac{1}{2}$, and hence also the game. On the other hand, if $\mathbb{P}_{h\sim D_H}[AG(S,H)=S]<\frac{1}{2}$, then there must exist at least one task T where $\mathbb{P}_{H\sim D_H}[Perf_T(S)\geq Perf_T(H)]<\frac{1}{2}$, thus violating def. 2.1. □

Before proceeding, there's a need to clarify two points in def. 2.1: first, we didn't specify anything about the task T. Second, it's also not clear how we can measure the performance of a task. Let's try to briefly address these points:

Ideally, we think a system S should be classified as AGI if it can perform any task comparably to an average human: this is the reason why we didn't pose any restrictions on T. However, as we saw in the introduction, some descriptions of AGI in the literature emphasize on the system's ability to perform tasks that are economically relevant ("do any work a human can do"): the idea is that in this case, S could be used to automate production, which is the underlying goal of AI. Additionally, state-of-the-art LLMs today can perform a very large variety of tasks that can be expressed as text, often better than many humans: however, they obviously trivially fail in tasks such as "fetch an object". Thus, leaving T completely unspecified may result in failing to capture those aspects.

To remedy this, one could formally define a set of tasks $\mathbb{T}$, where any task to be used in AG must belong to $\mathbb{T}$: for example, $\mathbb{T}=\{T:T\text{ is economically relevant}\}$ or $\mathbb{T}=\{T:T\text{ can be expressed in text}\}$. Notice that $\mathbb{T}$ does not to be pre-specified as a long list: instead, we can constraint AG, in the sense that the judge will reject a question (and return "ill-defined"), if it doesn't satisfy the definition of $\mathbb{T}$; we indicate this restricted version of the game as $AG_{\mathbb{T}}$. However ideally, for a true AGI system, T should ultimately be anything.

Additionally, it's not always clear how to measure the performance of a task T ($Perf_T$). For many tasks, especially technical, performance can readily be measured. However, for other tasks, for example involving artistic creation, a performance metric does not exist; what can we do in this case?

Unfortunately, there is no good answer to this question: it is a common problem in almost all definitions of machine intelligence Legg (2008). The solution we propose is to deffer this to the judge: that is, we let the judge ultimately decide the relative "quality" of each answer. This is a reasonable choice: for example, in the case of artworks, artists or art historians are able to argue about the style of a work, its artistic features, its position within the historical context and the message it conveys, etc. Of course, this is not perfect: for example, many great painters were recognized only posthumously. However, it's the best answer that we currently have.

Having clarified these points, we can proceed to discover another important feature of AG: namely, that all previously proposed tests of AGI can be cast as restricted versions of it:

Proposition 3.2. *All previously proposed tests for AGI, such as the Turing test, can be expressed as special cases of an Adversarial Game.*

Proof. Indeed: we see that the Turing test is equivalent with $AG_{\mathbb{T}}$, where $\mathbb{T}=\{$"pretend to be human"$\}$. In the same fashion, other popular tests such as the The Robot College Student Test, the IKEA Test of the Coffee Test Wikipedia can be expressed as special cases of AG. □

The intuition of the above result is the following: typically, the proposed definitions of machine intelligence attempt to find some key task, and then argue that if a system can perform it, then it would be able to perform any other task. For example, in the case of the Turing test, the idea is that in order to able to imitate a human, a system should implicitly be able to perform any other task, such as for example natural language understanding, arithmetics, or anything else that a human

could respond to. Similarly, the other definitions attempt to refine this, e.g. for example the Robot College Student Test also takes physical motion and interaction into account, etc.

On the other hand, AG tries to directly capture def. 2.1, akin to an open-ended competition. For example, in order to determine who is the best tennis player, we would challenge people to a tennis match. Similarly, to determine the best chess player, we would challenge them in a chess game. This is what AG tries to describe: a general AI system should be able to take challenges in an open-ended way - in contrary to narrow systems, not only on chess or tennis, but in any task - and be able to perform competitively. In that way, it addresses a weakness of previous definitions, where they need to prove (and they don't) that the task they propose is equivalent to general intelligence. AG in contrary simply materializes the definition directly, and this is also where its generality stems from.

In alg. 1, the game is played in a symmetric way, where the human and the AI player take turns in asking questions. This form incentivises both players to identify weaknesses on the opponent. In the case of the human player, it incentivises them to identify new tasks where the AI system fails and exploit them; this helps us discover and surface potential weaknesses of the system. On the other hand, the AI system is also forced to propose tasks where it thinks it can outperform human players; thus, forcing it to "expose" its capabilities.

However, in practice this symmetric form may have the downside that human players have less incentives to participate, as answering the AI's questions might be a difficult and tedious task. For that, we can also propose a variant of the game, called the Questioning Game:

The Questioning Game (QG): QG is similar to AG 1, with the difference that the players don't change turns; only the human player proposes questions, and only the AI answers them. The process is the same as in alg. 1, with line 14 removed.

We see that the Questioning Game enjoys similar properties to the Adversarial Game:

Proposition 3.3. *A system S is and AGI according to def. 2.1 iff it can consistently win QG against randomly (uniformly) selected humans with probability at least $\frac{1}{2}$. Moreover, all previously proposed tests for AGI can be expressed as special cases of QG.*

Finally, by Proposition 3.1 we know that a system S satisfies def. 2.1 if it's probability of winning an AG against a random human player is larger or equal to $\frac{1}{2}$. However, in practice we cannot know the exact probability, and we have only sample games. Yet, under some standard assumptions, we can use statistics to estimate the true winning probability with any degree of confidence:

Proposition 3.4. *Let $AG(S, H_1), ..., AG(S, H_n)$ are n Adversarial Game outcomes, where $H_i \sim D_H$ uniformly at random, and put $Y_i = \mathbf{1}[AG(S, H_i) = S], i = 1, ..., n$ (e.g., Y_i is 1 if S won the i-th game, and 0 else). Further, let $p_0 = \mathbb{P}_{H \sim D_H}[AG(S, H) = S]$ be the true winning probability, and $p = \frac{Y_1 + ... + Y_n}{n}$ the empirical fraction of wins. Assuming that $p = \frac{1}{2} + \epsilon$ with some $\epsilon > 0$, and that the game outcomes are i.i.d., we have that $p_0 \geq \frac{1}{2}$ with confidence at least $1 - \delta$, where $\delta = \exp(-2n\epsilon^2)$.*

Proof. Under the stated assumptions, Y_i are i.i.d. draws from the Bernoulli random variable $Y = \mathbf{1}[AG(S, H) = S]$. Thus, $Y \in [0, 1]$, and from Hoeffding's inequality Shalev-Shwartz & Ben-David (2014) we have:

$$\mathbb{P}\left[\frac{Y_1 + ... + Y_n}{n} - \mathbb{E}[Y] \geq \epsilon\right] \leq \exp\left[-\frac{2n\epsilon^2}{b-a)^2}\right] \tag{4}$$

, where $Y \in [a, b]$. As in our case $a = 0$ and $b = 1$, and $\frac{Y_1 + ... + Y_n}{n} = p$, $\mathbb{E}[Y] = p_0$, this simplifies to:

$$\mathbb{P}[p - p_0 \geq \epsilon] \leq \exp\left[-2n\epsilon^2\right] \tag{5}$$

By assumption, $p = \frac{1}{2} + \epsilon$, and the inequality $p - p_0 \geq \epsilon$ is equivalent to:

$$p - p_0 \geq \epsilon \iff \frac{1}{2} + \epsilon - p_0 \geq \epsilon \iff p_0 \leq \frac{1}{2} \tag{6}$$

Hence, eq. 5 becomes:

$$\mathbb{P}[p_0 \leq \frac{1}{2}] \leq \exp\left[-2n\epsilon^2\right] \tag{7}$$

Therefore, the opposing statement, $p_0 \geq \frac{1}{2}$, holds with probability at least $1 - \delta$, where $\delta = \exp\left[-2n\epsilon^2\right]$. □

Thus, having a moderate number of independent games allows us to estimate the winning probability of S. In particular, if S manages to consistently maintain a competitive ELO rating against the human average in a (hypothetical) online platform, this should be considered as a strong indication that S is a potential AGI.

4 BENCHMARKING ASI

According to def. 2.2, a system S would be characterized as an ASI, if it can perform any task better than any human. Some futurologists Kurzweil (2005) would specify ASI as a system outperforming the entire human population; however, once a system satisfies def. 2.2, it can be trivially amplified to outperform any number of humans:

Proposition 4.1. *If a system S satisfies def. 2.2, then it can be amplified into a system S' outperforming a set of N humans in any task T.*

Proof. Let $S' = S_{coord}[S_1, ..., S_N]$ be a new system, where $S_i, i = 1, ..., N$ are N copies of S, and S_{coord} is another copy of S, assigned with the task of coordinating $S_1, ..., S_N$. Since each system S_i is more capable than any human, and S_{coord} outperforms any human in the task of "coordination", the combined system S' will outperform a set of N humans in any task. □

Thus, we see that def. 2.2 is sufficient. Further, by similar considerations as is the previous section, we can see that an ASI system should be capable of winning the Adversarial Game against any human opponent:

Proposition 4.2. *A system S is an ASI according to def. 2.2 iff it can win AG (or QG) against any human opponent: that is, we have: $\forall H \in D_H : AG(S, H) = S$.*

Proof. Indeed, otherwise there would exist at least one task T where $Perf_T(S) < Perf_T(H)$ for some $H \in D_H$, violating def. 2.2. □

Hence, if a system S could consistently maintain a world-class rating against any human opponent in AG, then it should be considered as a candidate ASI.

Further, besides this, there is another criterion that an ASI system needs to fulfill: namely, the system should be capable of original discovery, and of advancing the state of the art in multiple fields:

Proposition 4.3. *A system S is an ASI according to def. 2.2 iff it is capable of original discovery; that is, the system is able to advance the state of the art in mathematics, sciences, and other areas of knowledge endeavour.*

Proof. The task $T =$ "advancing the state of the art" is something that human experts are capable of, in their own field of knowledge / expertise. Since S must be able to outperform any human at any task, it has to also be able to perform T, and moreover do so in any field of endeavour. □

Remark: Note that we don't necessarily have to wait for the theories of a potential ASI system S to be validated in the real world: instead, one could do the analysis "in retrospect": for example, we could train the system using data only up to 1904, and see if it can rediscover the Special Theory of Relativity Einstein et al. (1905); or, one could train the system with data up to 2016, and check if it can reinvent the Transformer architecture Vaswani et al. (2017); etc. Unsolved mathematical problems could be used as well, as verifying the proofs will be much easier than finding them. Overall, testing for original discovery is a viable way to measure the capabilities of a potential ASI system.

5 CONCLUSION

The dramatic recent improvement of Foundational Models and the new capabilities that emerge with scale force us to think about the future: what will be the capabilities of prospective models? Is it probable that future models have comparable capabilities with humans? And, how can we know?

Finding that the related concepts of general intelligence are not very specific, we attempt to rectify this, and propose precise definitions for the concepts of AGI and ASI. Further, as being able to measure these capacities is the most crucial aspect, we propose a new test called the Adversarial Game, and show that a system is an AGI only if it can win in AG against randomly selected human players. Additionally, we find that previously proposed tests can be seen as special instances of AG. Similarly, we propose related criteria for potential ASI systems.

We think that the most important path for future work is to use AG in order to measure the capabilities of current and future models in a large-scale fashion. As state of the art LLMs OpenAI (2023) can already perform competitive to humans in multiple tasks, we believe that an implementation of AG as an online platform where anyone can participate will allow us to uncover the true capabilities and shortcomings of such models, as well as upcoming ones. Moreover, as it has been showed that state of the art LLMs can also be used to judge the quality of answers Bubeck et al. (2023), the judging part of AG could be partially automated, as human judges need only to double-check the LLMs reasoning and correct only if needed. Additionally, it's plausible that a version of AG could also be used with AI models playing against each other, to reveal their relative capability.

In summary, we hope that our approach and the considerations introduced can lay a foundation for accurately measuring the capabilities of current and future models. Due to the high safety risks that can arise from future more powerful models, it's critical to have a way of characterizing and measuring their capabilities, and we hope that this work can pave a way towards it.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

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

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Albert Einstein et al. On the electrodynamics of moving bodies. *Annalen der physik*, 17(10):891–921, 1905.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Howard E Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011.

Joy Paul Guilford. The nature of human intelligence. 1967.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

Ray Kurzweil. The singularity is near. In *Ethics and emerging technologies*, pp. 393–406. Springer, 2005.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022.

Shane Legg. Machine super intelligence. 2008.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444, 2007.

Marvin Minsky. *Society of mind*. Simon and Schuster, 1988.

Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, pp. 1975. 2007.

OpenAI. GPT-4 Technical Report. Technical report, OpenAI, 2023.

William Albert Roberts. *Principles of animal cognition.* McGraw-Hill, 1998.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Herbert Alexander Simon. *The shape of automation for men and management*, volume 13. Harper & Row New York, 1965.

Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Wikipedia. Artificial general intelligence (wikipedia). URL `https://en.wikipedia.org/wiki/Artificial_general_intelligence`. Accessed: 2023-10-04.