Levels of AGI: Operationalizing Progress on the Path to AGI

Meredith Ringel Morris[1], **Jascha Sohl-dickstein**[1], **Noah Fiedel**[1], **Tris Warkentin**[1], **Allan Dafoe**[1], **Aleksandra Faust**[1], **Clement Farabet**[1] **and Shane Legg**[1]
[1]Google DeepMind

We propose a framework for classifying the capabilities and behavior of Artificial General Intelligence (AGI) models and their precursors. This framework introduces levels of AGI performance, generality, and autonomy. It is our hope that this framework will be useful in an analogous way to the levels of autonomous driving, by providing a common language to compare models, assess risks, and measure progress along the path to AGI. To develop our framework, we analyze existing definitions of AGI, and distill six principles that a useful ontology for AGI should satisfy. These principles include focusing on capabilities rather than mechanisms; separately evaluating generality and performance; and defining stages along the path toward AGI, rather than focusing on the endpoint. With these principles in mind, we propose "Levels of AGI" based on depth (performance) and breadth (generality) of capabilities, and reflect on how current systems fit into this ontology. We discuss the challenging requirements for future benchmarks that quantify the behavior and capabilities of AGI models against these levels. Finally, we discuss how these levels of AGI interact with deployment considerations such as autonomy and risk, and emphasize the importance of carefully selecting Human-AI Interaction paradigms for responsible and safe deployment of highly capable AI systems.

Keywords: AI, AGI, Artificial General Intelligence, General AI, Human-Level AI, HLAI, ASI, frontier models, benchmarking, metrics, AI safety, AI risk, autonomous systems, Human-AI Interaction

Introduction

Artificial General Intelligence (AGI)[1] is an important and sometimes controversial concept in computing research, used to describe an AI system that is at least as capable as a human at most tasks. Given the rapid advancement of Machine Learning (ML) models, the concept of AGI has passed from being the subject of philosophical debate to one with near-term practical relevance. Some experts believe that "sparks" of AGI (Bubeck et al., 2023) are already present in the latest generation of large language models (LLMs); some predict AI will broadly outperform humans within about a decade (Bengio et al., 2023); some even assert that current LLMs *are* AGIs (Agüera y Arcas and Norvig, 2023). However, if you were to ask 100 AI experts to define what they mean by "AGI," you would likely get 100 related but different definitions.

The concept of AGI is important as it maps onto goals for, predictions about, and risks of AI:

Goals: Achieving human-level "intelligence" is an implicit or explicit north-star goal for many in our field, from the 1955 Dartmouth AI Conference (McCarthy et al., 1955) that kick-started the

[1] There is controversy over use of the term "AGI." Some communities favor "General AI" or "Human-Level AI" (Gruetzemacher and Paradice, 2019) as alternatives, or even simply "AI" as a term that now effectively encompasses AGI (or soon will, under optimistic predictions). However, AGI is a term of art used by both technologists and the general public, and is thus useful for clear communication. Similarly, for clarity we use commonly understood terms such as "Artificial Intelligence" and "Machine Learning," although we are sympathetic to critiques (Bigham, 2019) that these terms anthropomorphize computing systems.

Corresponding author(s): merrie@google.com

modern field of AI to some of today's leading AI research firms whose mission statements allude to concepts such as "ensure transformative AI helps people and society" (Anthropic, 2023a) or "ensure that artificial general intelligence benefits all of humanity" (OpenAI, 2023).

Predictions: The concept of AGI is related to a prediction about progress in AI, namely that it is toward greater generality, approaching and exceeding human generality. Additionally, AGI is typically intertwined with a notion of "emergent" properties (Wei et al., 2022), i.e. capabilities not explicitly anticipated by the developer. Such capabilities offer promise, perhaps including abilities that are complementary to typical human skills, enabling new types of interaction or novel industries. Such predictions about AGI's capabilities in turn predict likely societal impacts; AGI may have significant economic implications, i.e., reaching the necessary criteria for widespread labor substitution (Dell'Acqua et al., 2023; Ellingrud et al., 2023), as well as geo-political implications relating not only to the economic advantages AGI may confer, but also to military considerations (Kissinger et al., 2022).

Risks: Lastly, AGI is viewed by some as a concept for identifying the point when there are extreme risks (Bengio et al., 2023; Shevlane et al., 2023), as some speculate that AGI systems might be able to deceive and manipulate, accumulate resources, advance goals, behave agentically, outwit humans in broad domains, displace humans from key roles, and/or recursively self-improve.

In this paper, we argue that it is critical for the AI research community to explicitly reflect on what we mean by "AGI," and aspire to quantify attributes like the performance, generality, and autonomy of AI systems. Shared operationalizable definitions for these concepts will support: comparisons between models; risk assessments and mitigation strategies; clear criteria from policymakers and regulators; identifying goals, predictions, and risks for research and development; and the ability to understand and communicate where we are along the path to AGI.

Defining AGI: Case Studies

Many AI researchers and organizations have proposed definitions of AGI. In this section, we consider nine prominent examples, and reflect on their strengths and limitations. This analysis informs our subsequent introduction of a two-dimensional, leveled ontology of AGI.

Case Study 1: The Turing Test. The Turing Test (Turing, 1950) is perhaps the most well-known attempt to operationalize an AGI-like concept. Turing's "imitation game" was posited as a way to operationalize the question of whether machines could think, and asks a human to interactively distinguish whether text is produced by another human or by a machine. The test as originally framed is a thought experiment, and is the subject of many critiques (Wikipedia, 2023b); in practice, the test often highlights the ease of fooling people (Weizenbaum, 1966; Wikipedia, 2023a) rather than the "intelligence" of the machine. Given that modern LLMs pass some framings of the Turing Test, it seems clear that this criteria is insufficient for operationalizing or benchmarking AGI. We agree with Turing that whether a machine can "think," while an interesting philosophical and scientific question, seems orthogonal to the question of what the machine can do; the latter is much more straightforward to measure and more important for evaluating impacts. Therefore we propose that AGI should be defined in terms of *capabilities* rather than *processes*[2].

Case Study 2: Strong AI – Systems Possessing Consciousness. Philosopher John Searle mused, "according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the

[2] As research into mechanistic interpretability (Räuker et al., 2023) advances, it may enable process-oriented metrics. These may be relevant to future definitions of AGI.

right programs can be literally said to understand and have other cognitive states" (Searle, 1980). While strong AI might be one path to achieving AGI, there is no scientific consensus on methods for determining whether machines possess strong AI attributes such as consciousness (Butlin et al., 2023), making the process-oriented focus of this framing impractical.

Case Study 3: Analogies to the Human Brain. The original use of the term "artificial general intelligence" was in a 1997 article about military technologies by Mark Gubrud (Gubrud, 1997), which defined AGI as "AI systems that rival or surpass the human brain in complexity and speed, that can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed." This early definition emphasizes *processes* (rivaling the human brain in complexity) in addition to capabilities; while neural network architectures underlying modern ML systems are loosely inspired by the human brain, the success of transformer-based architectures (Vaswani et al., 2023) whose performance is not reliant on human-like learning suggests that strict brain-based processes and benchmarks are not inherently necessary for AGI.

Case Study 4: Human-Level Performance on Cognitive Tasks. Legg (Legg, 2008) and Goertzel (Goertzel, 2014) popularized the term AGI among computer scientists in 2001 (Legg, 2022), describing AGI as a machine that is able to do the cognitive tasks that people can typically do. This definition notably focuses on non-physical tasks (i.e., not requiring robotic embodiment as a precursor to AGI). Like many other definitions of AGI, this framing presents ambiguity around choices such as "what tasks?" and "which people?".

Case Study 5: Ability to Learn Tasks. In *The Technological Singularity* (Shanahan, 2015), Shanahan suggests that AGI is "Artificial intelligence that is not specialized to carry out specific tasks, but can learn to perform as broad a range of tasks as a human." An important property of this framing is its emphasis on the value of including metacognitive tasks (learning) among the requirements for achieving AGI.

Case Study 6: Economically Valuable Work. OpenAI's charter defines AGI as "highly autonomous systems that outperform humans at most economically valuable work" (OpenAI, 2018). This definition has strengths per the "capabilities, not processes" criteria, as it focuses on performance agnostic to underlying mechanisms; further, this definition offers a potential yardstick for measurement, i.e., economic value. A shortcoming of this definition is that it does not capture all of the criteria that may be part of "general intelligence." There are many tasks that are associated with intelligence that may not have a well-defined economic value (e.g., artistic creativity or emotional intelligence). Such properties may be indirectly accounted for in economic measures (e.g., artistic creativity might produce books or movies, emotional intelligence might relate to the ability to be a successful CEO), though whether economic value captures the full spectrum of "intelligence" remains unclear. Another challenge with a framing of AGI in terms of economic value is that this implies a need for *deployment* of AGI in order to realize that value, whereas a focus on capabilities might only require the *potential* for an AGI to execute a task. We may well have systems that are technically capable of performing economically important tasks but don't realize that economic value for varied reasons (legal, ethical, social, etc.).

Case Study 7: Flexible and General – The "Coffee Test" and Related Challenges. Marcus suggests that AGI is "shorthand for any intelligence (there might be many) that is flexible and general, with resourcefulness and reliability comparable to (or beyond) human intelligence" (Marcus, 2022b). This definition captures both *generality* and *performance* (via the inclusion of reliability); the mention of "flexibility" is noteworthy, since, like the Shanahan formulation, this suggests that *metacognitive* tasks such as the ability to learn new skills must be included in an AGI's set of capabilities in order to achieve sufficient generality. Further, Marcus operationalizes his definition by proposing five concrete

tasks (understanding a movie, understanding a novel, cooking in an arbitrary kitchen, writing a bug-free 10,000 line program, and converting natural language mathematical proofs into symbolic form) (Marcus, 2022a). Accompanying a definition with a benchmark is valuable; however, more work would be required to construct a sufficiently comprehensive benchmark. While we agree that failing some of these tasks indicates a system is *not* an AGI, it is unclear that passing them is sufficient for AGI status. In the Testing for AGI section, we further discuss the challenge in developing a set of tasks that is both necessary and sufficient for capturing the generality of AGI. We also note that one of Marcus' proposed tasks, "work as a competent cook in an arbitrary kitchen" (a variant of Steve Wozniak's "Coffee Test" (Wozniak, 2010)), requires robotic embodiment; this differs from other definitions that focus on non-physical tasks[3].

Case Study 8: Artificial Capable Intelligence. In *The Coming Wave*, Suleyman proposed the concept of "Artificial Capable Intelligence (ACI)" (Mustafa Suleyman and Michael Bhaskar, 2023) to refer to AI systems with sufficient performance and generality to accomplish complex, multi-step tasks in the open world. More specifically, Suleyman proposed an economically-based definition of ACI skill that he dubbed the "Modern Turing Test," in which an AI would be given $100,000 of capital and tasked with turning that into $1,000,000 over a period of several months. This framing is more narrow than OpenAI's definition of economically valuable work and has the additional downside of potentially introducing alignment risks (Kenton et al., 2021) by only targeting fiscal profit. However, a strength of Suleyman's concept is the focus on performing a complex, multi-step task that humans value. Construed more broadly than making a million dollars, ACI's emphasis on complex, real-world tasks is noteworthy, since such tasks may have more *ecological validity* than many current AI benchmarks; Marcus' aforementioned five tests of flexibility and generality (Marcus, 2022a) seem within the spirit of ACI, as well.

Case Study 9: SOTA LLMs as Generalists. Agüera y Arcas and Norvig (Agüera y Arcas and Norvig, 2023) suggested that state-of-the-art LLMs (e.g. mid-2023 deployments of GPT-4, Bard, Llama 2, and Claude) already *are* AGIs, arguing that *generality* is the key property of AGI, and that because language models can discuss a wide range of topics, execute a wide range of tasks, handle multimodal inputs and outputs, operate in multiple languages, and "learn" from zero-shot or few-shot examples, they have achieved sufficient generality. While we agree that generality is a crucial characteristic of AGI, we posit that it must also be paired with a measure of *performance* (i.e., if an LLM can write code or perform math, but is not reliably correct, then its generality is not yet sufficiently performant).

Defining AGI: Six Principles

Reflecting on these nine example formulations of AGI (or AGI-adjacent concepts), we identify properties and commonalities that we feel contribute to a clear, operationalizable definition of AGI. We argue that any definition of AGI should meet the following six criteria:

1. Focus on Capabilities, not Processes. The majority of definitions focus on what an AGI can accomplish, not on the mechanism by which it accomplishes tasks. This is important for identifying characteristics that are not necessarily a prerequisite for achieving AGI (but may nonetheless be interesting research topics). This focus on capabilities allows us to exclude the following from our requirements for AGI:

- Achieving AGI does not imply that systems *think* or *understand* in a human-like way (since this focuses on processes, not capabilities)

[3] Though robotics might also be implied by the OpenAI charter's focus on "economically valuable work," the fact that OpenAI shut down its robotics research division in 2021 (Wiggers, 2021) suggests this is not their intended interpretation.

- Achieving AGI does not imply that systems possess qualities such as *consciousness* (subjective awareness) (Butlin et al., 2023) or *sentience* (the ability to have feelings) (since these qualities not only have a process focus, but are not currently measurable by agreed-upon scientific methods)

2. Focus on Generality *and* Performance. All of the above definitions emphasize generality to varying degrees, but some exclude performance criteria. We argue that both generality and performance are key components of AGI. In the next section we introduce a leveled taxonomy that considers the interplay between these dimensions.

3. Focus on Cognitive and Metacognitive Tasks. Whether to require robotic embodiment (Roy et al., 2021) as a criterion for AGI is a matter of some debate. Most definitions focus on cognitive tasks, by which we mean non-physical tasks. Despite recent advances in robotics (Brohan et al., 2023), physical capabilities for AI systems seem to be lagging behind non-physical capabilities. It is possible that embodiment in the physical world is necessary for building the world knowledge to be successful on some cognitive tasks (Shanahan, 2010), or at least may be one path to success on some classes of cognitive tasks; if that turns out to be true then embodiment may be critical to some paths toward AGI. We suggest that the ability to perform physical tasks increases a system's generality, but should not be considered a necessary prerequisite to achieving AGI. On the other hand, metacognitive capabilities (such as the ability to learn new tasks or the ability to know when to ask for clarification or assistance from a human) are key prerequisites for systems to achieve generality.

4. Focus on Potential, not Deployment. Demonstrating that a system can perform a requisite set of tasks at a given level of performance should be sufficient for declaring the system to be an AGI; deployment of such a system in the open world should not be inherent in the definition of AGI. For instance, defining AGI in terms of reaching a certain level of labor substitution would require real-world deployment, whereas defining AGI in terms of being *capable* of substituting for labor would focus on potential. Requiring deployment as a condition of measuring AGI introduces non-technical hurdles such as legal and social considerations, as well as potential ethical and safety concerns.

5. Focus on Ecological Validity. Tasks that can be used to benchmark progress toward AGI are critical to operationalizing any proposed definition. While we discuss this further in the "Testing for AGI" section, we emphasize here the importance of choosing tasks that align with real-world (i.e., ecologically valid) tasks that people value (construing "value" broadly, not only as economic value but also social value, artistic value, etc.). This may mean eschewing traditional AI metrics that are easy to automate or quantify (Raji et al., 2021) but may not capture the skills that people would value in an AGI.

6. Focus on the Path to AGI, not a Single Endpoint. Much as the adoption of a standard set of Levels of Driving Automation (SAE International, 2021) allowed for clear discussions of policy and progress relating to autonomous vehicles, we posit there is value in defining "Levels of AGI." As we discuss in subsequent sections, we intend for each level of AGI to be associated with a clear set of metrics/benchmarks, as well as identified risks introduced at each level, and resultant changes to the Human-AI Interaction paradigm (Morris et al., 2023). This level-based approach to defining AGI supports the coexistence of many prominent formulations – for example, Aguera y Arcas & Norvig's definition (Agüera y Arcas and Norvig, 2023) would fall into the "Emerging AGI" category of our ontology, while OpenAI's threshold of labor replacement (OpenAI, 2018) better matches "Virtuoso AGI." Our "Competent AGI" level is probably the best catch-all for many existing definitions of AGI (e.g., the Legg (Legg, 2008), Shanahan (Shanahan, 2015), and Suleyman (Mustafa Suleyman and Michael Bhaskar, 2023) formulations). In the next section, we introduce a level-based ontology of AGI.

Levels of AGI

Performance (rows) x **Generality** (columns)	**Narrow** *clearly scoped task or set of tasks*	**General** *wide range of non-physical tasks, including metacognitive abilities like learning new skills*
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging *equal to or somewhat better than an unskilled human*	**Emerging Narrow AI** GOFAI[4]; simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	**Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023)
Level 2: Competent *at least 50th percentile of skilled adults*	**Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert *at least 90th percentile of skilled adults*	**Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	**Expert AGI** not yet achieved
Level 4: Virtuoso *at least 99th percentile of skilled adults*	**Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	**Virtuoso AGI** not yet achieved
Level 5: Superhuman *outperforms 100% of humans*	**Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	**Artificial Superintelligence (ASI)** not yet achieved

Table 1 | A leveled, matrixed approach toward classifying systems on the path to AGI based on depth (performance) and breadth (generality) of capabilities. Example systems in each cell are approximations based on current descriptions in the literature or experiences interacting with deployed systems. Unambiguous classification of AI systems will require a standardized benchmark of tasks, as we discuss in the Testing for AGI section.

In accordance with Principle 2 ("Focus on Generality and Performance") and Principle 6 ("Focus on the Path to AGI, not a Single Endpoint"), in Table 1 we introduce a matrixed leveling system that focuses on *performance* and *generality* as the two dimensions that are core to AGI:

- **Performance** refers to the *depth* of an AI system's capabilities, i.e., how it compares to human-level performance for a given task. Note that for all performance levels above "Emerging," percentiles are in reference to a sample of adults who possess the relevant skill (e.g., "Competent" or higher performance on a task such as English writing ability would only be measured against the set of adults who are literate and fluent in English).
- **Generality** refers to the *breadth* of an AI system's capabilities, i.e., the range of tasks for which an AI system reaches a target performance threshold.

This taxonomy specifies the minimum performance over most tasks needed to achieve a given rating – e.g., a Competent AGI must have performance at least at the 50th percentile for skilled adult humans on most cognitive tasks, but may have Expert, Virtuoso, or even Superhuman performance on a subset of tasks. As an example of how individual systems may straddle different points in our taxonomy, we posit that as of this writing in September 2023, frontier language models (e.g., ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama2 (Touvron et al., 2023), etc.) exhibit "Competent" performance levels for some tasks (e.g., short essay writing, simple coding), but are still at "Emerging" performance levels for most tasks (e.g., mathematical abilities, tasks involving factuality). Overall, current frontier language models would therefore be considered a Level 1 General AI ("Emerging AGI") until the performance level increases for a broader set of tasks (at which point the Level 2 General AI, "Competent AGI," criteria would be met). We suggest that documentation for frontier AI models, such as model cards (Mitchell et al., 2019), should detail this mixture of performance levels. This will help end-users, policymakers, and other stakeholders come to a shared, nuanced understanding of the likely uneven performance of systems progressing along the path to AGI.

The order in which stronger skills in specific cognitive areas are acquired may have serious implications for AI safety (e.g., acquiring strong knowledge of chemical engineering before acquiring strong ethical reasoning skills may be a dangerous combination). Note also that the rate of progression between levels of performance and/or generality may be nonlinear. Acquiring the capability to learn new skills may particularly accelerate progress toward the next level.

While this taxonomy rates systems according to their performance, systems that are *capable* of achieving a certain level of performance (e.g., against a given benchmark) may not match this level *in practice* when deployed. For instance, user interface limitations may reduce deployed performance. Consider the example of DALLE-2 (Ramesh et al., 2022), which we estimate as a Level 3 Narrow AI ("Expert Narrow AI") in our taxonomy. We estimate the "Expert" level of performance since DALLE-2 produces images of higher quality than most people are able to draw; however, the system has failure modes (e.g., drawing hands with incorrect numbers of digits, rendering nonsensical or illegible text) that prevent it from achieving a "Virtuoso" performance designation. While theoretically an "Expert" level system, *in practice* the system may only be "Competent," because prompting interfaces are too complex for most end-users to elicit optimal performance (as evidenced by the existence of marketplaces (e.g., (PromptBase)) in which skilled prompt engineers sell prompts). This observation emphasizes the importance of designing ecologically valid benchmarks (that would measure deployed rather than idealized performance) as well as the importance of considering how human-AI interaction paradigms interact with the notion of AGI (a topic we return to in the "Capabilities vs. Autonomy" Section).

The highest level in our matrix in terms of combined performance and generality is ASI (Artificial Superintelligence). We define "Superhuman" performance as outperforming 100% of humans. For instance, we posit that AlphaFold (Jumper et al., 2021; Varadi et al., 2021) is a Level 5 Narrow AI ("Superhuman Narrow AI") since it performs a single task (predicting a protein's 3D structure from an amino acid sequence) above the level of the world's top scientists. This definition means that Level 5 General AI ("ASI") systems will be able to do a wide range of tasks at a level that no

human can match. Additionally, this framing also implies that Superhuman systems may be able to perform an even broader generality of tasks than lower levels of AGI, since the ability to execute tasks that qualitiatively differ from existing human skills would by definition outperform all humans (who fundamentally cannot do such tasks). For example, non-human skills that an ASI might have could include capabilities such as neural interfaces (perhaps through mechanisms such as analyzing brain signals to decode thoughts (Tang et al., 2023)), oracular abilities (perhaps through mechanisms such as analyzing large volumes of data to make high-quality predictions), or the ability to communicate with animals (perhaps by mechanisms such as analyzing patterns in their vocalizations, brain waves, or body language).

Testing for AGI

Two of our six proposed principles for defining AGI (Principle 2: Generality *and* Performance; Principle 6: Focus on the Path to AGI) influenced our choice of a matrixed, leveled ontology for facilitating nuanced discussions of the breadth and depth of AI capabilities. Our remaining four principles (Principle 1: Capabilities, not Processes; Principle 3: Cognitive and Metacognitive Tasks; Principle 4: Potential, not Deployment; and Principle 5: Ecological Validity) relate to the issue of measurement.

While our *performance* dimension specifies one aspect of measurement (e.g., percentile ranges for task performance relative to particular subsets of people), our *generality* dimension leaves open important questions: What is the set of tasks that constitute the generality criteria? What proportion of such tasks must an AI system master to achieve a given level of generality in our schema? Are there some tasks that must always be performed to meet the criteria for certain generality levels, such as metacognitive tasks?

Operationalizing an AGI definition requires answering these questions, as well as developing specific diverse and challenging tasks. Because of the immense complexity of this process, as well as the importance of including a wide range of perspectives (including cross-organizational and multi-disciplinary viewpoints), we do not propose a benchmark in this paper. Instead, we work to clarify the ontology a benchmark should attempt to measure. We also discuss properties an AGI benchmark should possess.

Our intent is that an AGI benchmark would include a broad suite of cognitive and metacognitive tasks (per Principle 3), measuring diverse properties including (but not limited to) linguistic intelligence, mathematical and logical reasoning (Webb et al., 2023), spatial reasoning, interpersonal and intra-personal social intelligences, the ability to learn new skills and creativity. A benchmark *might* include tests covering psychometric categories proposed by theories of intelligence from psychology, neuroscience, cognitive science, and education; however, such "traditional" tests must first be evaluated for suitability for benchmarking computing systems, since many may lack ecological and construct validity in this context (Serapio-García et al., 2023).

One open question for benchmarking performance is whether to allow the use of tools, including potentially AI-powered tools, as an aid to human performance. This choice may ultimately be task dependent and should account for ecological validity in benchmark choice (per Principle 5). For example, in determining whether a self-driving car is sufficiently safe, benchmarking against a person driving without the benefit of any modern AI-assisted safety tools would not be the most informative comparison; since the relevant counterfactual involves some driver-assistance technology, we may prefer a comparison to that baseline.

While an AGI benchmark might draw from some existing AI benchmarks (Lynch, 2023) (e.g., HELM (Liang et al., 2023), BIG-bench (Srivastava et al., 2023)), we also envision the inclusion of

open-ended and/or interactive tasks that might require qualitative evaluation (Bubeck et al., 2023; Papakyriakopoulos et al., 2021; Yang et al., 2023). We suspect that these latter classes of complex, open-ended tasks, though difficult to benchmark, will have better ecological validity than traditional AI metrics, or than adapted traditional measures of human intelligence.

It is impossible to enumerate the full set of tasks achievable by a sufficiently general intelligence. As such, an AGI benchmark should be a *living* benchmark. Such a benchmark should therefore include a framework for generating and agreeing upon new tasks.

Determining that something is *not* an AGI at a given level simply requires identifying several[5] tasks that people can typically do but the system cannot adequately perform. Systems that pass the majority of the envisioned AGI benchmark at a particular performance level ("Emerging," "Competent," etc.), including new tasks added by the testers, can be assumed to have the associated level of generality for practical purposes (i.e., though in theory there could still be a test the AGI would fail, at some point unprobed failures are so specialized or atypical as to be practically irrelevant).

Developing an AGI benchmark will be a challenging and iterative process. It is nonetheless a valuable north-star goal for the AI research community. Measurement of complex concepts may be imperfect, but the act of measurement helps us crisply define our goals and provides an indicator of progress.

Risk in Context: Autonomy and Human-AI Interaction

Discussions of AGI often include discussion of risk, including "x-risk" – existential (for AI Safety, 2023) or other very extreme risks (Shevlane et al., 2023). A leveled approach to defining AGI enables a more nuanced discussion of how different combinations of performance and generality relate to different types of AI risk. While there is value in considering extreme risk scenarios, understanding AGI via our proposed ontology rather than as a single endpoint (per Principle 6) can help ensure that policymakers also identify and prioritize risks in the near-term and on the path to AGI.

Levels of AGI as a Framework for Risk Assessment

As we advance along our capability levels toward ASI, new risks are introduced, including misuse risks, alignment risks, and structural risks (Zwetsloot and Dafoe, 2019). For example, the “Expert AGI” level is likely to involve structural risks related to economic disruption and job displacement, as more and more industries reach the substitution threshold for machine intelligence in lieu of human labor. On the other hand, reaching “Expert AGI” likely alleviates some risks introduced by “Emerging AGI” and “Competent AGI,” such as the risk of incorrect task execution. The “Virtuoso AGI” and “ASI” levels are where many concerns relating to x-risk are most likely to emerge (e.g., an AI that can outperform its human operators on a broad range of tasks might deceive them to achieve a mis-specified goal, as in misalignment thought experiments (Christian, 2020)).

Systemic risks such as destabilization of international relations may be a concern if the rate of progression between levels outpaces regulation or diplomacy (e.g., the first nation to achieve ASI may have a substantial geopolitical/military advantage, creating complex structural risks). At levels below

[5] We hesitate to specify the precise number or percentage of tasks that a system must pass at a given level of performance in order to be declared a General AI at that Level (e.g., a rule such as "a system must pass at least 90% of an AGI benchmark at a given performance level to get that rating"). While we think this will be a very high percentage, it will probably not be 100%, since it seems clear that broad but imperfect generality is impactful (individual humans also lack consistent performance across all possible tasks, but remain generally intelligent). Determining what portion of benchmarking tasks at a given level demonstrate generality remains an open research question.

"Expert AGI" (e.g., "Emerging AGI," "Competent AGI," and all "Narrow" AI categories), risks likely stem more from human actions (e.g., risks of AI misuse, whether accidental, incidental, or malicious). A more complete analysis of risk profiles associated with each level is a critical step toward developing a taxonomy of AGI that can guide safety/ethics research and policymaking.

We acknowledge that whether an AGI benchmark should include tests for potentially dangerous capabilities (e.g., the ability to deceive, to persuade (Veerabadran et al., 2023), or to perform advanced biochemistry (Morris, 2023)) is controversial. We lean on the side of including such capabilities in benchmarking, since most such skills tend to be dual use (having valid applications to socially positive scenarios as well as nefarious ones). Dangerous capability benchmarking can be de-risked via Principle 4 (Potential, not Deployment) by ensuring benchmarks for any dangerous or dual-use tasks are appropriately sandboxed and not defined in terms of deployment. However, including such tests in a public benchmark may allow malicious actors to optimize for these abilities; understanding how to mitigate risks associated with benchmarking dual-use abilities remains an important area for research by AI safety, AI ethics, and AI governance experts.

Concurrent with this work, Anthropic released Version 1.0 of its Responsible Scaling Policy (RSP) (Anthropic, 2023b). This policy uses a levels-based approach (inspired by biosafety level standards) to define the level of risk associated with an AI system, identifying what dangerous capabilities may be associated with each AI Safety Level (ASL), and what containment or deployment measures should be taken at each level. Current SOTA generative AIs are classified as an ASL-2 risk. Including items matched to ASL capabilities in any AGI benchmark would connect points in our AGI taxonomy to specific risks and mitigations.

Capabilities vs. Autonomy

While capabilities provide prerequisites for AI risks, AI systems (including AGI systems) do not and will not operate in a vacuum. Rather, AI systems are deployed with particular interfaces and used to achieve particular tasks in specific scenarios. These contextual attributes (interface, task, scenario, end-user) have substantial bearing on risk profiles. AGI capabilities alone do not determine destiny with regards to risk, but must be considered in combination with contextual details.

Consider, for instance, the affordances of user interfaces for AGI systems. Increasing capabilities unlock new interaction paradigms, but *do not determine them*. Rather, system designers and end-users will settle on a mode of human-AI interaction (Morris et al., 2023) that balances a variety of considerations, including safety. We propose characterizing human-AI interaction paradigms with six **Levels of Autonomy**, described in Table 2.

These Levels of Autonomy are correlated with the Levels of AGI. Higher levels of autonomy are "unlocked" by AGI capability progression, though lower levels of autonomy may be desirable for particular tasks and contexts (including for safety reasons) even as we reach higher levels of AGI. Carefully considered choices around human-AI interaction are vital to safe and responsible deployment of frontier AI models.

We emphasize the importance of the "No AI" paradigm. There may be many situations where this is desirable, including for education, enjoyment, assessment, or safety reasons. For example, in the domain of self-driving vehicles, when Level 5 Self-Driving technology is widely available, there may be reasons for using a Level 0 (No Automation) vehicle. These include for instructing a new driver (education), for pleasure by driving enthusiasts (enjoyment), for driver's licensing exams (assessment), or in conditions where sensors cannot be relied upon such as technology failures or extreme weather events (safety). While Level 5 Self-Driving (SAE International, 2021) vehicles would likely be a Level

Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 0: No AI *human does every-thing*	Analogue approaches (e.g., sketching with pencil on paper) Non-AI digital workflows (e.g., typing in a text editor; drawing in a paint program)	No AI	n/a (status quo risks)
Autonomy Level 1: AI as a Tool *human fully controls task and uses AI to automate mundane sub-tasks*	Information-seeking with the aid of a search engine Revising writing with the aid of a grammar-checking program Reading a sign with a machine translation app	Possible: Emerging Narrow AI Likely: Competent Narrow AI	de-skilling (e.g., over-reliance) disruption of established industries
Autonomy Level 2: AI as a Consultant *AI takes on a substantive role, but only when invoked by a human*	Relying on a language model to summarize a set of documents Accelerating computer program-ming with a code-generating model Consuming most entertain-ment via a sophisticated recommender system	Possible: Competent Narrow AI Likely: Expert Narrow AI; Emerging AGI	over-trust radicalization targeted manipulation
Autonomy Level 3: AI as a Collaborator *co-equal human-AI collaboration; inter-active coordination of goals & tasks*	Training as a chess player through interactions with and analysis of a chess-playing AI Entertainment via social interactions with AI-generated personalities	Possible: Emerging AGI Likely: Expert Narrow AI; Competent AGI	anthropomorphization (e.g., parasocial relationships) rapid societal change
Autonomy Level 4: AI as an Expert *AI drives interaction; human provides guidance & feedback or performs subtasks*	Using an AI system to advance scientific discovery (e.g., protein-folding)	Possible: Virtuoso Narrow AI Likely: Expert AGI	societal-scale ennui mass labor displacement decline of human exceptionalism
Autonomy Level 5: AI as an Agent *fully autonomous AI*	Autonomous AI-powered personal assistants *(not yet unlocked)*	Likely: Virtuoso AGI; ASI	misalignment concentration of power

Table 2 | More capable AI systems unlock new human-AI interaction paradigms (including fully autonomous AI). The choice of appropriate autonomy level need not be the maximum achievable given the capabilities of the underlying model. One consideration in the choice of autonomy level are resulting risks. This table's examples illustrate the importance of carefully considering human-AI interaction design decisions.

5 Narrow AI (“Superhuman Narrow AI”) under our taxonomy[6], the same considerations regarding human vs. computer autonomy apply to AGIs. We may develop an AGI, but choose not to deploy it autonomously (or choose to deploy it with differentiated autonomy levels in distinct circumstances as dictated by contextual considerations).

Certain aspects of generality may be required to make particular interaction paradigms desirable. For example, the Autonomy Levels 3, 4, and 5 ("Collaborator," "Expert," and "Agent") may only work well if an AI system also demonstrates strong performance on certain metacognitive abilities (learning when to ask a human for help, theory of mind modeling, social-emotional skills). Implicit in our definition of Autonomy Level 5 ("AI as an Agent") is that such a fully autonomous AI can act in an aligned fashion without continuous human oversight, but knows when to consult humans (Shah et al., 2021). Interfaces that support human-AI alignment through better task specification, the bridging of process gulfs, and evaluation of outputs (Terry et al., 2023) are a vital area of research for ensuring that the field of human-computer interaction keeps pace with the challenges and opportunities of interacting with AGI systems.

Human-AI Interaction Paradigm as a Framework for Risk Assessment

Table 2 illustrates the interplay between AGI Level, Autonomy Level, and risk. Advances in model performance and generality unlock additional interaction paradigm choices (including potentially fully autonomous AI). These interaction paradigms in turn introduce new classes of risk. The interplay of model capabilities and interaction design will enable more nuanced risk assessments and responsible deployment decisions than considering model capabilities alone.

Table 2 also provides concrete examples of each of our six proposed Levels of Autonomy. For each level of autonomy, we indicate the corresponding levels of performance and generality that "unlock" that interaction paradigm (i.e., levels of AGI at which it is possible or likely for that paradigm to be successfully deployed and adopted).

Our predictions regarding "unlocking" levels tend to require higher levels of performance for Narrow than for General AI systems; for instance, we posit that the use of AI as a Consultant is likely with either an Expert Narrow AI or an Emerging AGI. This discrepancy reflects the fact that for General systems, capability development is likely to be uneven; for example, a Level 1 General AI ("Emerging AGI") is likely to have Level 2 or perhaps even Level 3 performance across some subset of tasks. Such unevenness of capability for General AIs may unlock higher autonomy levels for particular tasks that are aligned with their specific strengths.

Considering AGI systems in the context of use by people allows us to reflect on the interplay between advances in models and advances in human-AI interaction paradigms. The role of model building research can be seen as helping systems’ capabilities progress along the path to AGI in their performance and generality, such that an AI system’s abilities will overlap an increasingly large portion of human abilities. Conversely, the role of human-AI interaction research can be viewed as ensuring new AI systems are *usable* by and *useful* to people such that AI systems successfully extend people’s capabilities (i.e., "intelligence augmentation" (Brynjolfsson, 2022)).

[6] Fully autonomous vehicles might arguably be classified as Level 4 Narrow AI ("Virtuoso Narrow AI") per our taxonomy; however, we suspect that in practice autonomous vehicles may need to reach the Superhuman performance standard to achieve widespread social acceptance regarding perceptions of safety, illustrating the importance of contextual considerations.

Conclusion

Artificial General Intelligence (AGI) is a concept of both aspirational and practical consequences. In this paper, we analyzed nine prominent definitions of AGI, identifying strengths and weaknesses. Based on this analysis, we introduce six principles we believe are necessary for a clear, operationalizable definition of AGI: focusing on capabilities, not processes; focusing on generality *and* performance; focusing on cognitive and metacognitive (rather than physical) tasks; focusing on potential rather than deployment; focusing on ecological validity for benchmarking tasks; and focusing on the path toward AGI rather than a single endpoint.

With these principles in mind, we introduced our Levels of AGI ontology, which offers a more nuanced way to define our progress toward AGI by considering generality (either Narrow or General) in tandem with five levels of performance (Emerging, Competent, Expert, Virtuoso, and Superhuman). We reflected on how current AI systems and AGI definitions fit into this framing. Further, we discussed the implications of our principles for developing a living, ecologically valid AGI benchmark, and argue that such an endeavor (while sure to be challenging) is a vital one for our community to engage with.

Finally, we considered how our principles and ontology can reshape discussions around the risks associated with AGI. Notably, we observed that AGI is not necessarily synonymous with autonomy. We introduced Levels of Autonomy that are unlocked, but not determined by, progression through the Levels of AGI. We illustrated how considering AGI Level jointly with Autonomy Level can provide more nuanced insights into likely risks associated with AI systems, underscoring the importance of investing in human-AI interaction research in tandem with model improvements.

Acknowledgements

Thank you to the members of the Google DeepMind PAGI team for their support of this effort, and to Martin Wattenberg, Michael Terry, Geoffrey Irving, Murray Shanahan, Dileep George, and Blaise Agüera y Arcas for helpful discussions about this topic.

References

Blaise Agüera y Arcas and Peter Norvig. Artificial General Intelligence is Already Here. Noema, October 2023. URL https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

Amazon. Amazon Alexa. URL https://alexa.amazon.com/. accessed on October 20, 2023.

Rohan Anil, Andrew M. Dai, Orhan Firat, and et al. PaLM 2 Technical Report. *CoRR*, abs/2305.10403, 2023. doi: 10.48550/arXiv.2305.10403. URL https://arxiv.org/abs/2305.10403.

Anthropic. Company: Anthropic, 2023a. URL https://www.anthropic.com/company. Accessed October 12, 2023.

Anthropic. Anthropic's Responsible Scaling Policy, September 2023b. URL https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf. accessed on October 20, 2023.

Apple. Siri. URL https://www.apple.com/siri/. accessed on October 20, 2023.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter,

Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing AI Risks in an Era of Rapid Progress. *CoRR*, abs/2310.17688, 2023. doi: 10.48550/arXiv.2310.17688. URL https://arxiv.org/abs/2310.17688.

Jeffrey Bigham. The Coming AI Autumn. Blog Post, 2019. URL https://jeffreybigham.com/blog/2019/the-coming-ai-autumnn.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *CoRR*, abs/2307.15818, 2023. doi: 10.48550/arXiv.2307.15818. URL https://arxiv.org/abs/2307.15818.

Erik Brynjolfsson. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *CoRR*, abs/2201.04200, 2022. doi: 10.48550/arXiv.2201.04200. URL https://arxiv.org/abs/2201.04200.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. doi: 10.48550/arXiv.2303.12712. URL https://arxiv.org/abs/2303.12712.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *CoRR*, abs/2308.08708, 2023. doi: 10.48550/arXiv.2308.08708. URL https://arxiv.org/abs/2308.08708.

Murray Campbell, A. Joseph Hoane, and Feng-hsiung Hsu. Deep Blue. *Artif. Intell.*, 134(1–2):57–83, jan 2002. ISSN 0004-3702. doi: 10.1016/S0004-3702(01)00129-1. URL https://doi.org/10.1016/S0004-3702(01)00129-1.

Xi Chen, Xiao Wang, Soravit Changpinyo, and et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *CoRR*, abs/2209.06794, 2023. doi: 10.48550/arXiv.2209.06794. URL https://arxiv.org/abs/2209.06794.

Brian Christian. *The Alignment Problem*. W. W. Norton & Company, 2020.

Millon Madhur Das, Punyajoy Saha, and Mithun Das. Which One is More Toxic? Findings from Jigsaw Rate Severity of Toxic Comments. *CoRR*, abs/2206.13284, 2022. doi: 10.48550/arXiv.2206.13284. URL https://arxiv.org/abs/2206.13284.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the Jagged

Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Management Unit Working Paper Number 24-013*, September 2023.

Kweilin Ellingrud, Saurabh Sanghvi, Gurneet Singh Dandona, Anu Madgavkar, Michael Chui, Olivia White, and Paige Hasebe. Generative AI and the future of work in America. McKinsey Institute Global Report, July 2023. URL https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america.

Center for AI Safety. Statement on AI Risk, 2023. URL https://www.safe.ai/statement-on-ai-risk.

Ben Goertzel. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 01 2014. doi: 10.2478/jagi-2014-0001.

Google. Google Assistant, your own personal Google. URL https://assistant.google.com/. accessed on October 20, 2023.

Grammarly, 2023. URL https://www.grammarly.com/.

Ross Gruetzemacher and David B. Paradice. Alternative Techniques for Mapping Paths to HLAI. *CoRR*, abs/1905.00614, 2019. doi: 10.48550/arXiv.1905.00614. URL http://arxiv.org/abs/1905.00614.

Mark Gubrud. Nanotechnology and International Security. *Fifth Foresight Conference on Molecular Nanotechnology*, November 1997.

IBM. IBM Watson. URL https://www.ibm.com/watson. accessed on October 20, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of Language Agents. *CoRR*, abs/2103.14659, 2021. doi: 10.48550/arXiv.2103.14659. URL https://arxiv.org/abs/2103.14659.

Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher. *The Age of AI*. Back Bay Books, November 2022.

Shane Legg. Machine Super Intelligence. Doctoral Dissertation submitted to the Faculty of Informatics of the University of Lugano, June 2008.

Shane Legg. Twitter (now "X"), May 2022. URL https://twitter.com/ShaneLegg/status/1529483168134451201. Accessed on October 12, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, and et al. Holistic Evaluation of Language Models. *CoRR*, abs/2211.09110, 2023. doi: 10.48550/arXiv.2211.09110. URL https://arxiv.org/abs/2211.09110.

Shana Lynch. AI Benchmarks Hit Saturation. Stanford Human-Centered Artificial Intelligence Blog, April 2023. URL https://hai.stanford.edu/news/ai-benchmarks-hit-saturation.

Gary Marcus. Dear Elon Musk, here are five things you might want to consider about AGI. "Marcus on AI" Substack, May 2022a. URL https://garymarcus.substack.com/p/dear-elon-musk-here-are-five-things?s=r.

Gary Marcus. Twitter (now "X"), May 2022b. URL https://twitter.com/GaryMarcus/status/1529457162811936768. Accessed on October 12, 2023.

J. McCarthy, M.L. Minsky, N. Rochester, and C.E. Shannon. A Proposal for The Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth Workshop, 1955.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145%2F3287560.3287596.

Meredith Ringel Morris. Scientists' Perspectives on the Potential for Generative AI in their Fields. *CoRR*, abs/2304.01420, 2023. doi: 10.48550/arXiv.2304.01420. URL https://arxiv.org/abs/2304.01420.

Meredith Ringel Morris, Carrie J. Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. The Design Space of Generative Models. *CoRR*, abs/2304.10547, 2023. doi: 10.48550/arXiv.2304.10547. URL https://arxiv.org/abs/2304.10547.

Mustafa Suleyman and Michael Bhaskar. *The Coming Wave: Technology, Power, and the 21st Century's Greatest Dilemma*. Crown, September 2023.

OpenAI. OpenAI Charter, 2018. URL https://openai.com/charter. Accessed October 12, 2023.

OpenAI. OpenAI: About, 2023. URL https://openai.com/about. Accessed October 12, 2023.

OpenAI. GPT-4 Technical Report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

Orestis Papakyriakopoulos, Elizabeth Anne Watkins, Amy Winecoff, Klaudia Jaźwińska, and Tithi Chattopadhyay. Qualitative Analysis for Human Centered AI. *CoRR*, abs/2112.03784, 2021. doi: 10.48550/arXiv.2112.03784. URL https://arxiv.org/abs/2112.03784.

PromptBase. PromptBase: Prompt Marketplace. URL https://promptbase.com/. accessed on October 20, 2023.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark. *CoRR*, abs/2111.15366, 2021. doi: 10.48550/arXiv.2111.15366. URL https://arxiv.org/abs/2111.15366.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. April 2022. URL https://cdn.openai.com/papers/dall-e-2.pdf.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. *CoRR*, abs/2207.13243, 2023. doi: 10.48550/arXiv.2207.13243. URL https://arxiv.org/abs/2207.13243.

Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, Tomas Lozano-Perez, Vikash Mansinghka, Christopher Pal, Blake Richards, Dorsa Sadigh, Stefan Schaal, Gaurav Sukhatme, Denis Therien, Marc Toussaint, and Michiel Van de Panne. From Machine Learning to Robotics: Challenges and Opportunities for Embodied Intelligence. *CoRR*, abs/2110.15245, 2021. doi: 10.48550/arXiv.2110.15245. URL https://arxiv.org/abs/2110.15245.

SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, April 2021. URL https://www.sae.org/standards/content/j3016_202104. Accessed October 12, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR*, abs/2205.11487, 2022. doi: 10.48550/arXiv.2205.11487. URL https://arxiv.org/abs/2205.11487.

John R. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–424, 1980. doi: 10.1017/S0140525X00005756.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality Traits in Large Language Models. *CoRR*, abs/2307.00184, 2023. doi: 10.48550/arXiv.2307.00184. URL https://arxiv.org/abs/2307.00184.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of Assistance over Reward Learning, 2021. URL https://openreview.net/forum?id=DFIoGDZejIB.

Murray Shanahan. *Embodiment and the Inner Life*. Oxford University Press, 2010.

Murray Shanahan. *The Technological Singularity*. MIT Press, August 2015.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks. *CoRR*, abs/2305.15324, 2023. doi: 10.48550/arXiv.2305.15324. URL https://arxiv.org/abs/2305.15324.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529:484–489, 2016. doi: 10.1038/nature16961.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the Game of Go Without Human Knowledge. *Nature*, 550:354–359, 2017. doi: 10.1038/nature24270.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and

Go through Self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *CoRR*, abs/2206.04615, 2023. doi: 10.48550/arXiv.2206.04615. URL https://arxiv.org/abs/2206.04615.

Stockfish. Stockfish - Open Source Chess Engine, 2023. URL https://stockfishchess.org/.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic Reconstruction of Continuous Language from Non-invasive Brain Recordings. *Nature Neuroscience*, 26:858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. *CoRR*, abs/2311.00710, 2023. doi: 10.48550/arXiv.2311.00710. URL https://arxiv.org/abs/2311.00710.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

A.M. Turing. Computing Machinery and Intelligence. *Mind*, LIX:433–460, October 1950. URL https://doi.org/10.1093/mind/LIX.236.433.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Research*, 50:D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *CoRR*, abs/1706.03762, 2023. doi: 10.48550/arXiv.1706.03762. URL https://arxiv.org/abs/1706.03762.

V. Veerabadran, J. Goldman, S. Shankar, and et al. Subtle Adversarial Image Manipulations Influence Both Human and Machine Perception. *Nature Communications*, 14, 2023. doi: 10.1038/s41467-023-40499-0.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent Analogical Reasoning in Large Language Models. *Nature Human Behavior*, 7:1526–1541, 2023. URL https://doi.org/10.1038/s41562-023-01659-w.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *CoRR*, abs/2206.07682, 2022. doi: 10.48550/arXiv.2206.07682. URL https://arxiv.org/abs/2206.07682.

Joseph Weizenbaum. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM*, 9(1):36–45, jan 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL https://doi.org/10.1145/365153.365168.

Kyle Wiggers. OpenAI Disbands its Robotics Research Team. VentureBeat, July 2021. URL https://venturebeat.com/business/openai-disbands-its-robotics-research-team/.

Wikipedia. Eugene Goostman - Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Eugene_Goostman, 2023a. Accessed October 12, 2023.

Wikipedia. Turing Test: Weaknesses — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Turing_test, 2023b. Accessed October 12, 2023.

Terry Winograd. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *MIT AI Technical Reports*, 1971.

Steve Wozniak. Could a Computer Make a Cup of Coffee? Fast Company interview: https://www.youtube.com/watch?v=MowergwQR5Y, 2010.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *CoRR*, abs/2309.17421, 2023. doi: 10.48550/arXiv.2309.17421. URL https://arxiv.org/abs/2309.17421.

Remco Zwetsloot and Allan Dafoe. Thinking about Risks from AI: Accidents, Misuse and Structure. *Lawfare*, 11:2019, 2019. URL https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.