Intelligence 104 (2024) 101832

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Defining intelligence: Bridging the gap between human and artificial perspectives

Gilles E. Gignac[a,*], Eva T. Szodorai[b]

[a] School of Psychological Science, The University of Western Australia, M053, 35 Stirling Highway, Perth, WA 6009, Australia
[b] Curtin University, Australia

ARTICLE INFO

Keywords:
Human intelligence
Artificial intelligence
General intelligence
Artificial general intelligence
Nomenclature
AI metrics

ABSTRACT

Achieving a widely accepted definition of human intelligence has been challenging, a situation mirrored by the diverse definitions of artificial intelligence in computer science. By critically examining published definitions, highlighting both consistencies and inconsistencies, this paper proposes a refined nomenclature that harmonizes conceptualizations across the two disciplines. Abstract and operational definitions for human and artificial intelligence are proposed that emphasize maximal capacity for completing novel goals successfully through respective perceptual-cognitive and computational processes. Additionally, support for considering intelligence, both human and artificial, as consistent with a multidimensional model of capabilities is provided. The implications of current practices in artificial intelligence training and testing are also described, as they can be expected to lead to artificial achievement or expertise rather than artificial intelligence. Paralleling psychometrics, 'AI metrics' is suggested as a needed computer science discipline that acknowledges the importance of test reliability and validity, as well as standardized measurement procedures in artificial system evaluations. Drawing parallels with human general intelligence, artificial general intelligence (AGI) is described as a reflection of the shared variance in artificial system performances. We conclude that current evidence more greatly supports the observation of artificial achievement and expertise over artificial intelligence. However, interdisciplinary collaborations, based on common understandings of the nature of intelligence, as well as sound measurement practices, could facilitate scientific innovations that help bridge the gap between artificial and human-like intelligence.

1. Introduction

Human intelligence ranks as one of psychology's oldest and most vigorously debated dimensions (Deary, 2020; Jensen, 1998). Even attempts to achieve a commonly agreed-upon definition of human intelligence have proven difficult (Bartholomew, 2004; Sternberg & Detterman, 1986). The area of computer science has also produced numerous definitions of artificial intelligence (Legg & Hutter, 2007a; Monett & Lewis, 2018). Disagreements and inconsistencies in conceptualisations of constructs can be expected to hinder progress in a scientific field, if they lead to fragmented research efforts, impede the development of a unified theoretical framework, and create barriers to effective communication and collaboration among researchers (Flake & Fried, 2020; Kuhn, 1962). In the following, we review some published definitions of intelligence in the areas of psychology and computer science (i.e., human and artificial) with two aims: (1) highlight inconsistencies; and (2) suggest a common nomenclature that may facilitate scientific progress across both fields. With agreed upon conceptualisations and definitions of words and terms such as 'intelligence', 'achievement', 'expertise', and 'general intelligence', the fields of psychology and computer science could achieve enhanced precision in research, foster more meaningful interdisciplinary dialogues, and potentially pave the way for scientific innovations.

2. Constructs: psychological and computational

Though disagreements abound on what human intelligence is, there is broad agreement that human intelligence is a psychological construct (Johnson, 2013; Plomin, 2018; Sternberg, 2012). A psychological construct is an abstract, unobservable, hypothetical entity inferred from postulated thoughts and observable behaviours, representing patterns of psychologically related phenomena (Cronbach & Meehl, 1955; Sijtsma,

[*] Corresponding author at: School of Psychological Science, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia.
E-mail address: gilles.gignac@uwa.edu.au (G.E. Gignac).

https://doi.org/10.1016/j.intell.2024.101832
Received 6 February 2024; Received in revised form 9 March 2024; Accepted 1 April 2024
Available online 8 April 2024
0160-2896/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2006). In plain language, a psychological construct is a concept developed to describe a specific aspect of the mind or behaviour that is not directly observable, but is inferred from patterns in thoughts, feelings, and actions. In addition to intelligence, examples of well-established psychological constructs include anxiety (characterized by feelings of tension, worried thoughts, and physical changes associated with arousal of the autonomic nervous system; Reiss, 1997), self-esteem (which involves one's overall subjective emotional evaluation of their own worth; Pyszczynski et al., 2004), and motivation (the process that initiates, guides, and sustains goal-directed behaviours; Touré-Tillery & Fishbach, 2014). Constructs are recognised across many scientific disciplines, including physics (e.g., energy, which refers to an unobservable property possessed by all matter and systems, capable of transformation from one form to another; Papadouris & Constantinou, 2012); chemistry (e.g., chemical bond, which refers to the forces holding atoms together in molecules; Shahbazian & Zahedi, 2006), and biology (e.g., fitness, an organism's ability to survive and reproduce in its environment; Grafen, 2015).

Artificial intelligence is not a psychological construct, as it does not originate from the same underlying human cognitive or emotional processes. Instead, artificial intelligence may be considered a *computational construct*, as it is inferred from the outcomes of simulated aspects of human thought and decision-making, which are facilitated by data processing, machine learning techniques, and algorithmic principles (Prasad et al., 2023; Schoser, 2023). Additionally, artificial intelligence has evolved through computer science and engineering advancements (Kumar et al., 2023), marked by human-initiated intervention, intellectual effort, and purposeful innovation. By comparison, human intelligence has evolved primarily through natural selection, marked by organic adaptation and neurological optimization (Gabora & Russon, 2011).

Constructs are essential tools in psychological research and theory, as they help conceptualize and organize complex psychological phenomena in a way that allows for systematic investigation, prediction, and explanation.[1] In practice, psychological constructs are inferred from responses to various stimuli and performance on tasks (Strauss & Smith, 2009). To measure psychological constructs effectively, it is essential to have clear and specific definitions of those constructs (Messick, 1981; Slaney & Racin, 2013). Arguably, these principles should also apply to computational constructs, in order to help advance the field scientifically.

Given the fundamental role constructs play in structuring our understanding of complex phenomena, we propose abstract and operational definitions for both human and artificial intelligence in the following section. Our definitions are not only grounded in their respective domains and established frameworks, but also reflect an appreciable degree of conceptual coherence to facilitate interdisciplinary scientific discourse.

3. What is human intelligence?

Since its inception as a psychological attribute more than a century ago, many definitions of human intelligence have been proposed. In fact, when Sternberg and Detterman (1986) surveyed two dozen experts in the field of intelligence, two dozen different definitions were provided. Though definitions of human intelligence tend to differ in precise terms, there is a convergence around certain core ideas. Correspondingly, a total of 52 professors with expertise in intelligence signed an editorial that defined intelligence as "the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience" (Gottfredson, 1997, p. 13). This definition has since been recognised by additional experts in the field (e.g., Colom, 2020; Deary et al., 2006; Halpern, 2014). Furthermore, expert definitions generally align with implicit theories of human intelligence held by specialists in other fields and laypeople alike (Sternberg, 1985; Sternberg et al., 1981).

Despite the above, there is reason to be dissatisfied with commonly written and endorsed definitions of intelligence, such as that included in Gottfredson (1997), as well as other similar definitions in other sources (e.g., Humphreys, 1984; Wechsler, 1944).[2] Specifically, such definitions are essentially a list of examples of subdimensions of intelligence, rather than representing an encompassing concept that recognises the large number of cognitive abilities that have been recognised over the years (Carroll, 1993; Schneider & McGrew, 2018). Arguably, an acceptable definition of intelligence needs to strike a delicate balance between being both sufficiently abstract and sufficiently detailed, in order to facilitate a theoretical understanding of what does and what does not constitute intelligence.

Drawing upon Gignac (2018, p. 440), we define human intelligence as a human's "maximal capacity to achieve a novel goal successfully using perceptual-cognitive [processes]." There are three important characteristics to this definition. First, when a person's intelligence is considered, it is in the context of their maximal capacity to solve novel problems, not a person's typically manifested intelligent behaviour. For instance, while some people may exhibit high intelligence levels on formal tests, they might not consistently apply this capacity in daily activities due to varying motivational factors (e.g., need for cognition; see von Stumm & Ackerman, 2013). Correspondingly, the correlation between overall intelligence and typical intellectual engagement is only approximately 0.45 (Chamorro-Premuzic et al., 2006).

Secondly, the essence of human intelligence is closely tied to its application in novel contexts (Davidson & Downing, 2000; Raaheim & Brun, 1985). This entails solving problems that a person has not previously encountered, rather than those with which they are already familiar. The concept of novelty in intelligence, and its distinction from academic achievement and expertise, is explored in greater detail further below. To foreshadow one of our key conclusions, we will provide evidence to suggest that current AI systems may be suggested to have demonstrated artificial achievement, and perhaps artificial expertise in some cases, whereas there is far less evidence for artificial intelligence.

Thirdly, human intelligence is underpinned by perceptual-cognitive functions (Thomson, 1919), which, at a basic level, encompass a range of mental processes, including attention, visual perception, auditory perception, and sensory integration (i.e., multiple modalities). In a wide array of contexts, one or more basic perceptual-cognitive processes would be required to identify and process relevant information, thereby enabling effective interaction with the environment. The capacity to process sensory inputs is necessary for the manifestation of human cognitive abilities, including subdimensions such as memory span, for example, as discussed further below.

Though our recommended abstract definition of human intelligence may help elucidate its conceptual nature, it lacks concreteness to be sufficiently useful to guide the development of corresponding psychometric measures of intelligence. Echoing Humphreys (1984, p. 22): "A scientist has not merely a right but a duty to define concepts in a way compatible with measurement operations and with the data resulting

[1] It is acknowledged that there are contrasting perspectives on the precise nature and usefulness of constructs (Borsboom, 2023).

[2] Humphreys (1984, p. 243) defined intelligence as the "the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at anyone period of time." Wechsler (1943, p. 3) defined intelligence as "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment." Sternberg (2011, p. 55) defined intelligence as "the capacity to learn from experience, using metacognitive processes to enhance learning, and the ability to adapt to the surrounding environment, which may require different adaptations within different social and cultural contexts."

from those operations." Consequently, drawing upon Gignac (2018, p. 440), we define human intelligence operationally "as [a person's] maximal capacity to complete a novel standardized task with veridical scoring using perceptual-cognitive [processes]." A novel standardized task implies a task (or test) for which the examinee has had no exposure, ideally not even awareness of the type of questions that will be asked during the testing session, so as to reduce the chances of preparation. Standardized implies a test for which there are clear instructions and procedures that are followed for all examinees (Sireci, 2020). As the primary purpose of a psychological test is to compare the performance of people (Cronbach, 1960), it is essential that all examinees have the same opportunities to perform, a situation facilitated by following procedures in the same manner for all.

Finally, veridical scoring implies that the assessment and interpretation of responses are based on objective, verifiable criteria for which there is wide agreement. This approach helps ensure that the test scores are free from subjective biases and/or inconsistent grading standards, factors that may be expected to impact test score reliability and validity adversely.[3] For example, a vocabulary test that included the word 'ambiguous' might use a multiple-choice format where examinees select the correct definition from a set of options (e.g., "Having more than one possible meaning or interpretation."). The scoring is veridical as it relies on predetermined, correct answers recognised and agreed upon by language experts. This ensures that each examinee's understanding of the word is assessed against a consistent, objective standard. All valid tests of intelligence, including matrix reasoning tests, memory span tests, and quantitative reasoning tests, for example, include veridical scoring.

4. What is artificial intelligence (AI)?

Like human intelligence, many definitions of artificial intelligence have been proposed, as documented in comprehensive reviews (Legg & Hutter, 2007a; Monett & Lewis, 2018). Artificial intelligence is perhaps most commonly defined as "the ability of machines to perform tasks that typically require human intelligence" (e.g., Minsky, 1961; Prasad et al., 2020; Schoser, 2023). As such a definition does not define human intelligence, it is circular and lacks specificity. Furthermore, such a definition may arguably be more accurately considered a goal of AI, not a definition of AI.

In addition to the limited definition of artificial intelligence above, there are four definitions that have emerged within the literature that, to some degree, intersect the fields of psychology and computer science: Goertzel (2010), Chollet (2019), Wang (2022), and Legg and Hutter (2007b). Next, we present and evaluate each definition, taking into consideration the desirability to have definitions of human and artificial intelligence that are complementary.

First, Goertzel (2010); Goertzel & Yu, 2014) defined artificial intelligence as a system's ability to recognise patterns quantifiable through the observable development of actions or responses while achieving complex goals in complex environments. Goertzel's reference to the ability to recognise patterns is consistent with human intelligence definitions, particularly in the context of fluid intelligence (Hayes et al., 2015) and logical-mathematical intelligence (Gardner & Hatch, 1989). However, Goertzel's reference to 'achieving complex goals' falls short by not adequately differentiating between novel and non-novel goals. This distinction is crucial, as we elaborate further below, particularly when considering the difference between achievement or expertise and intelligence, which inherently involves the ability to deal with novel challenges.

Second, Chollet (2019, p. 27) defined the intelligence of a system as "a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty." At the core of Chollet (2019) definition is learning (skill acquisition), however, as we document further below, learning is only one of many subdimensions of human intelligence, indicating a need for a broader conceptualisation. Reference to generalisability in the Chollet (2019) definition is an important one, as it helps distinguish intelligence from achievement and expertise, as we describe further below. However, generalisability is not necessarily easy to identify in every context. Therefore, it may be more precise to state that intelligence is manifested when entities successfully complete tasks that are novel to them, as unpractised challenges are fundamental to the valid measurement of human intelligence (Davidson & Downing, 2000; Raaheim & Brun, 1985).

Next, Wang (2022, p. 35) defined intelligence as "the ability of an information processing system to adapt to its environment while working with insufficient knowledge and resources." The inclusion of the concept of adaptability in Wang (2022) definition is consistent with many abstract definitions of human intelligence (McIntosh et al., 2005; Pintner, 1923; Sternberg, 2011). However, the capacity to adapt to the environment may be excessively broad, given the number of different types of factors that can lead to adaptation. The term 'insufficient knowledge' in Wang (2022) definition conveys the notion of novelty, in the sense that the system was not specifically trained on the problem, which is a good contextualisation of a definition of intelligence, as achievement and expertise are not intelligence, as we detail further below.

Finally, in a paper devoted to defining machine intelligence, Legg and Hutter (2007b, p. 402) defined intelligence as "an agent's ability to achieve goals in a wide range of environments", which is a definition that has core similarities to the Gignac (2018) definition we endorsed. However, there are some important differences. In particular, the definition does not make clear that the goals must be novel, an essential criterion, as we noted above. Another characteristic of Legg & Hutter (2007b) definition is that it makes reference to "wide ranging environments", which Legg and Hutter (2007a, 2007b) suggest to imply performance across diverse situations, tasks, and problems, i.e., generalisability. Though our favoured definition of human intelligence aligns with this viewpoint, specifying 'novel goals' instead of 'wide-ranging environments' is more precise. Additionally, Legg and Hutter (2007b) definition does not explicitly acknowledge that intelligence should be regarded as an entity's maximal capacity. Finally, Legg and Hutter's (2007b) definition does not specify the mechanisms by which intelligent behaviour is underpinned, an important limitation with respect to differentiating between computational functionality and cognitive processes that underpin human intelligence.

In light of the above, and considering the need to balance coherence and distinctiveness in psychology and computer science disciplines, we propose defining artificial intelligence abstractly as the maximal capacity of an artificial system to successfully achieve a novel goal through computational algorithms.[4] Our abstract definition of AI is identical to the definition of human intelligence we outlined above, with two exceptions. First, we replaced 'human' with 'artificial system' to reflect the fundamental distinction between organic, human cognitive processes versus synthetic, computer-based operations inherent in AI systems. Secondly, novel goals are specified to be achieved through the use of computational algorithms, not perceptual-cognitive processes. A

[3] Test score reliability refers to the consistency (or precision) of test scores (Traub & Rowley, 1980). Practically, it determines the level of confidence that can be attributed to a case's test score, such as a score of 100 on an IQ test. Validity pertains to whether a score can justifiably be interpreted to indicate a particular attribute (e.g., human intelligence; Messick, 1989). Reliability is a necessary but not sufficient condition for validity.

[4] In AI, a "goal" typically refers to a target state or outcome that the system is programmed to achieve (Russel & Norvig, 2016). These goals are set by human programmers and are not autonomously generated by the AI itself. For instance, a chess-playing AI has the 'goal' of winning the game, but this goal is a programmed objective, not a product of the AI's own volition or desire.

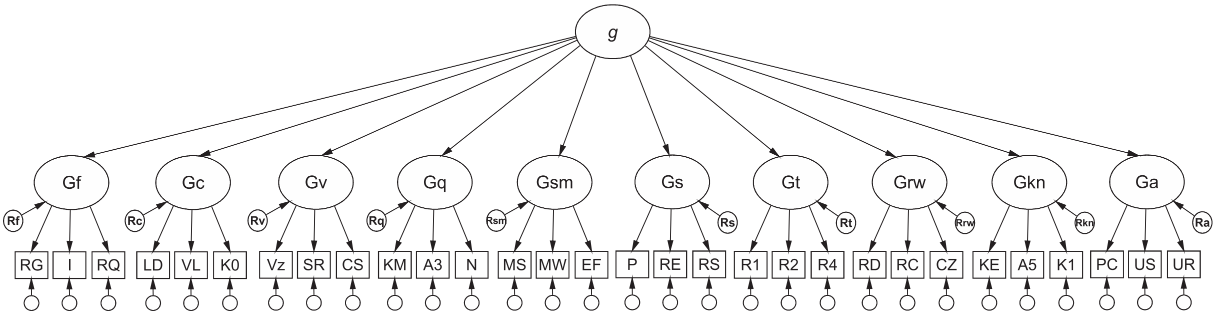

Fig. 1. Example Higher-Order Model of Intelligence.
Note. Circles represent latent dimensions; squares represent observe variables (i.e., test scores); *Gf* = fluid reasoning abilities (*Gf*); *Gc* = comprehension knowledge abilities; *Gv* = visual intelligence abilities; *Gq* = quantitative knowledge; *Gsm* = Short-term memory; *Gs* = cognitive processing speed; *Gt* = Decision and reaction speed; *Grw* = reading and writing; *Gkn* = domain specific knowledge; *Ga* = Auditory processing; RG = general sequential reasoning; I = induction; RQ = quantitative reasoning; LD = language development; VL = lexical knowledge; K0 = general (verbal) information; Vz = visualization; SR = spatial relations; CS = closure speed; KM = mathematical knowledge; A3 = mathematical achievement, N = numeracy; MS = memory span; MW = working memory; EF = learning efficiency; P = perceptual speed; RE = reaction time; RS = reading speed; R1 = simple reaction time; R2 = choice reaction time; R4 = semantic processing speed; RD = reading decoding; RC = reading comprehension; CZ = close speed; KE = general science information; A5 = geography achievement; K1 = general science information; PC = phonetic coding; US = speech sound discrimination; UR = resistance to auditory stimulus distortion; (see Flanagan & Dixon, 2013; Schneider & McGrew, 2018).

computational algorithm encompasses any set of rules or procedures used by a computer to solve a problem or accomplish a task (Cormen, 2013). Computational algorithms in AI can range from simple, rule-based instructions to complex processes like those found in machine learning and deep learning (Raj, 2019). These algorithms can involve pattern recognition, data processing, decision-making, and learning from data, for example.

Consistent with the operational definition of human intelligence we established above, we propose that artificial intelligence be defined operationally as an artificial system's maximal capacity to complete a novel standardized task with veridical scoring using computational algorithms. The only difference between the human operational definition and the artificial definition is reference to an artificial system and computational algorithms.

5. A note on AI metrics

Given our operational definition of artificial intelligence, which makes reference to standardized tasks, we note that *AI metrics* is an emerging discipline focusing on AI systems' performance measurement and evaluation which parallels psychometrics in psychology (Goertzel, 2014; Welty et al., 2019). In AI metrics, 'dataset' refers to collections of problems for AI systems to solve, similar to 'test banks' in psychology, which consist of questions for evaluating human behaviour. Cronbach (1960, p. 21) defined a psychological test as "a systematic procedure for comparing the behaviour of two or more people." Consequently, in AI metrics, a test may be defined as a systematic procedure for assessing and comparing artificial entities' capabilities across various tasks or domains. Typically, AI test questions or problems have answers or optimal solutions that can be clearly identified, aligning with the concept of veridical scoring in intelligence measurement.

A number of AI system capability tests have been published. For example, the HumanEval test (or dataset) is comprised of programming challenges that have been crowdsourced from a diverse group of contributors (Lu et al., 2010; Siddiq et al., 2023). Each challenge within the test is specifically crafted to assess an AI model's proficiency in generating programming code. Other examples include the AI2 Reasoning Challenge (ARC; Clark et al., 2018), a benchmark dataset specifically designed for question answering in the domain of science, TruthfulQA (Lin et al., 2021), a dataset aimed at measuring the truthfulness of AI models' responses, particularly in situations where misconceptions or popular false beliefs are involved, and HellaSwag (Zellers et al., 2019), a dataset created to challenge AI models in commonsense natural language inference, using strategically crafted scenarios that are simple for humans but difficult for AI systems. The AI metric tests listed above are similar to how an IQ test battery comprises various cognitive subtests, each aimed at evaluating different aspects of human cognitive ability. Awareness of AI capability tests (and terminology) is useful for intelligence researchers, as it not only illuminates the methodologies used in AI evaluation but also underscores the potential for interdisciplinary research that leverages insights from both fields.

Next, to help substantiate our definitions of intelligence, both human and artificial, we will expand upon earlier statements, emphasizing that intelligence is distinct from achievement, expertise, and adaptation. These distinctions are important when considering whether AI systems have, thus far, attained artificial achievement or expertise in certain domains, rather than intelligence.

6. Intelligence is not achievement or expertise

In simple terms, it has been suggested that intelligence is what one does when one does not know what to do.[5] This definition highlights the importance of novelty to the entity when encountering and solving intellectual problems, a crucial component of valid intelligence testing in humans (Davidson & Downing, 2000; Gignac, 2018; Jensen, 1998). That is, it is imperative that participants are not made available prior specific knowledge of test items, effective goal management strategies, or practice with similar problems, in order to ensure the validity of the test as a measure of cognitive ability, rather than a reflection of learned responses or familiarity.[6] As we show next, the assumption of novelty is violated in the context of the demonstration of achievement and expertise.

Intelligence is a broad construct that facilitates the potential to achieve across multiple domains (Gottfredson, 2002). By comparison, achievement is a realization of this potential to varying degrees within a specific domain through instruction and/or practice (Preckel et al., 2020). Though achievements may be recognised across a broad range of domains (e.g., academic, professional, sporting, etc.), academic achievement is perhaps the most commonly considered type of achievement by individual differences researchers. Academic achievement refers to the level of success a person has attained in educational

[5] While the popular quote "Intelligence is what you use when you don't know what to do" is often attributed to Jean Piaget, a definitive source for this statement has not yet been identified.

[6] "…if people had extensive practice or instruction on Raven problems, the goal management would become routine, thereby making the problems easier" (Carpenter et al., 1990, p. 428).

settings, typically measured through assessments including tests.[7] Correspondingly, achievement tests are specifically designed to evaluate a person's knowledge and skills in a particular area or subject, reflecting what they have intentionally learned or been trained in.

Two relatively well-known and comprehensive tests of academic achievement include the Programme for International Student Assessment (PISA; Turner & Adams, 2007), which assesses the reading, math, and science skills of 15-year-old students worldwide, and the National Assessment of Educational Progress (NAEP; Jones, 1996), which assesses the proficiency of American students in various subjects, such as mathematics, reading, science, and writing, at different grade levels.[8] In the US, two narrower tests of academic achievement include the Bar Examination (Merritt & Cornett, 2020), which is a rigorous assessment that law graduates are required to pass in order to be licensed to practice law in a given jurisdiction, and the United States Medical Licensing Examination (USMLE; Johnson, 2006), which a candidate's competence to apply medical knowledge, concepts, and principles necessary for effective patient care.

Though intelligence test scores have been found to associate positively and appreciably with academic achievement scores, they are not the same constructs (less than 50% shared variance; Lozano-Blasco et al., 2022; Pokropek et al., 2022). Furthermore, academic achievement tests are not intelligence tests, because people completing achievement tests would be expected to have studied specific content to help them answer questions or solve problems based on that knowledge, thus contravening the definitional expectation of novelty associated with our endorsed definitions of intelligence.

The recognition of achievement as conceptually distinct from intelligence is important for at least two reasons: (1) intelligence is known to predict more substantially a wide array of outcome variables in life than any particular achievement (Gottfredson, 2002)[9]; and (2) artificial intelligence systems are often trained on the content included in the tests used to measure their capacity, which, therefore, cannot reflect the system's intelligence, i.e., ability to demonstrate novel problem-solving ability. We provide evidence for statement (2) in a subsection further below. To foreshadow again, we will conclude that current AI systems demonstrate achievement, and perhaps expertise, whereas there is questionable evidence for the demonstration of artificial intelligence, as defined above.

Whereas achievement reflects the realization of one's potential to varying extents within a specific domain, expertise may be defined as the mastery of a comprehensive and structured set of knowledge acquired through extensive practice and experience, leading to an exceptionally high level of performance (Chi, 2006; Ericsson, 2006). There are instances where individuals have significantly improved their scores on cognitive tasks through extensive practice. For example, Chase and Ericsson (1982) described two college students (DD and SF) who increased their digit memory spans to 68 and 82 after approximately 250 h of practice. DD increased their digit span to 106 digits after an additional 350 h of practice (Ericsson & Staszewski, 1989). The typical adult has a digit span of approximately seven digits (Gignac & Weiss, 2015). Evidence for notable increases in cognitive performance after extensive practice in other areas, in the absence of any obvious benefits to other cognitive domains, have been documented, including memory span for chess positions (Gobet & Simon, 1998; Smith et al., 2021), spatial mapping ability among London taxi drivers (Woollett et al., 2009), and calculation prodigies (Jensen, 1990).

Importantly, improvements in cognitive performance in one area through training does not yield increases in performance in other cognitive abilities (Gobet & Sala, 2023; Sala & Gobet, 2019). Often, so limited are even near-transfer benefits that Norris et al. (2019) reported that training digit span failed to enhance letter span. Even when digit span was trained visually, the improvements failed to extend to the same digit task presented auditorily, highlighting the task-specific nature of cognitive training benefits.

We note that in computer science, some artificial systems are recognised as demonstrating expertise. For example, the mixture-of-experts (MoE) model (Nguyen & Chamroukhi, 2018). In MoE models, each 'expert' is a neural network trained on a specific aspect of a larger task. These networks, through their unique architecture and training, become highly proficient in their specialized domains, collectively contributing to the AI system's broader capabilities. This approach mirrors how human experts develop deep knowledge in specific areas through focused practice and experience. While AI systems differ in the degree to which they have been trained (see Lake et al., 2014, for an example of one-shot learning[10]), it may be contended that the development of artificial achievement and/or expertise, as opposed to artificial intelligence, plays a central role in enhancing a typical AI model's ability to execute human-like cognitive tasks.[11]

7. Intelligence is not adaptation

Human intelligence has been defined conceptually as a person's capacity to adapt to the environment successfully (Neisser et al., 1996; Sternberg, 2011). The principle of adaptive capacity is also central to several conceptualizations and definitions of artificial intelligence. For example, in an article devoted to defining artificial intelligence, Wang (2019, p. 17) endorsed the following definition of intelligence: "the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources." Russell and Norvig (2010) also emphasized the role of adaptation in intelligent agents, noting that these agents must be able to operate autonomously and adjust their behaviour based on changes in their environment. At first glance, these conceptualizations suggest a common ground between human and artificial intelligence, positing adaptability as a key facet of intelligent behaviour, indicative of an entity's ability not just to react, but also to learn and evolve within varied environmental contexts.

However, there are reasons not to consider adaptation as a defining characteristic of intelligence. First, evidence for adaptation to an environment is arguably too broad a concept to reflect intelligence. For

[7] By comparison, academic attainment refers to the level of success a person has attained in educational settings, as represented by educational awards (e.g., certificates, diplomas, degrees, etc.). Intelligence predicts both academic achievement and attainment positively (Keage et al., 2016; Lozano-Blasco et al., 2022).

[8] The Woodcock-Johnson IV Tests of Achievement (WJ IV ACH; Schrank, Mather, & McGrew, 2014) would be better described as an IQ test, not a test of achievement, because the subtests include test item content that people may not have been specifically trained upon.

[9] We note that while achievement is not intelligence, there is evidence that human intelligence may develop, to some degree, through formal education (Ritchie & Tucker-Drob, 2018), though the effects may be largely restricted to certain segments of the population (Peng & Kievit, 2020). Further complicating matters, testing co-development hypotheses statistically in an unambiguous manner is challenging (Curran & Hancock, 2021; Lüdtke & Robitzsch, 2022).

[10] Lake et al. (2014) applied a Hierarchical Hidden Markov model (HHMM) to classify (experiment 1) and generate (experiment 2) new (single-trial) speech sounds based on the Japanese language. Such tasks would require phonetic coding (PC) and speech sound discrimination (US) stratum I abilities within the Cattell-Horn-Carroll model of intelligence (McGrew, 2009).

[11] Several AI models, such as Grounded Language Acquisition models (Rasheed & Amin, 2016; Vong et al., 2024), eschew extensive text pre-training. Instead, by learning language through interaction and observation of the physical world, these models mimic human cognitive processes more closely, potentially advancing the realization of artificial intelligence. Additionally, PoseGPT (Feng et al., 2023) may have the capacity for limited reasoning in a narrow context (3D human body poses). These AI system developments do not undermine the primary goal of this paper, i.e., to develop a nomenclature for intelligence research across the psychology and computer science disciplines.

example, the human skin's ability to tan in response to sun exposure is an adaptive biological process driven by environmental factors, rather than a manifestation of cognitive ability. A more behaviourally relevant example is bird migration, a successful adaptation to the environment, but which is partly instinctual (i.e., genetically pre-programmed; Sweta et al., 2019), rather than the result of commonly recognised dimensions of cognitive ability. Secondly, successful adaptation is somewhat subjective and nuanced. For example, many human cognitive biases, such as confirmation bias, are generally considered to be adaptive tendencies; however, it is well-documented that these biases frequently lead to poor decision-making (Croskerry et al., 2013). Ultimately, the capacity to adapt to an environment should not be considered a defining characteristic of intelligence, as successful adaptation can be the result of primarily non-cognitive and/or instinctive characteristics, as well as unclear in some cases.

With human and artificial intelligence defined, we next define and describe general intelligence: a term used extensively in both psychology and computer science, though, as we show, with typically different understandings.

8. What is general intelligence (*g*)?

In psychology, general intelligence is a theoretical construct postulated to account for the empirical observation that test scores from a diverse collection of intelligence tests tend to correlate with each other positively (Jensen, 1998). In practical terms, people who tend to perform relatively well on verbal tasks also tend to perform relatively well on spatial tasks, memory span tasks, quantitative tasks, etc. Laypeople tend not to appreciate the degree of correspondence in performance across cognitive abilities (Rammstedt & Rammsayer, 2000). The relatively consistent order of people's performance across a wide variety of tasks (and modalities) yields a 'positive manifold': loosely speaking, a pattern of widespread positive correlations among different cognitive abilities (mean $r \approx 0.45$ to 0.50; Detterman & Daniel, 1989; Walker et al., 2023).

In human psychology, general intelligence (symbolized as '*g*') is a proposed construct to represent the empirical observation that individual differences in cognitive abilities correlate with each other positively, yielding a general factor when factor analysed (i.e., positive factor loadings from all tests; Jensen, 1998). Based on a factor solution derived from a factor analysis of an inter-ability correlation matrix, general factor scores can be derived for each case in the dataset. Such scores may be considered to represent 'psychometric *g*' (Jensen & Weng, 1994,).[12]

Some have theorized that mental energy (Spearman, 1927) or sustained concentration (Lykken, 2005) may be the mechanism by which *g* arises. In a more developed theory of intelligence that recognises *g*, Anderson (1992, 2001) contended that intelligence can be seen through the lens of different types of cognitive capacities, such as verbal and spatial, which at their core are uncorrelated because they are served by dedicated processing modules designed for specific tasks. However, the empirical observation that these abilities are positively correlated in individuals suggests a common underlying factor. This commonality, according to Anderson, is due to a shared information processing mechanism that underpins both types of primary cognitive abilities (verbal and spatial). Anderson emphasized that individual differences in the speed of this basic processing mechanism plays a crucial role in binding these diverse abilities together, contributing to the manifestation of *g*.

Jensen (2006) also considered processing speed, as measured by reaction time for example, may play a significant role in the emergence of *g*, because it reflects the basic efficiency and rapidity with which the brain can process information, perform cognitive tasks, and respond to stimuli. Faster processing speeds are thought to facilitate more efficient learning, problem-solving, and decision-making, which are key components of general intelligence.[13] Furthermore, Jensen (1998) asserted that it was "inescapable" (p. 249) that there must be generality of neural function to mediate positively correlated individual differences in cognitive abilities. For example, number of neurons, neural efficiency, or neural conduction velocity.

Despite the above, Jensen (1998) asserted that it is a misunderstanding to consider *g* a psychological process at all. Psychological processes can be identified and examined with a single person. For example, Ebbinghaus uncovered fundamental truths about learning and memory through self-experimentation, conducting studies where he was the sole subject (Postman, 1968). By contrast, general intelligence cannot be identified or examined, in the absence of individual differences data, as it is the variation between people that is core to its observation. Correspondingly, Jensen (1998, p. 74) stated that: "*g* may be thought of as a distillate of the common source [shared variance] of individual differences in all mental tests, completely stripped of their distinctive features of information content, skill, strategy, and the like."

In contrast to Jensen (1998) focus on variation between people, common variance in cognitive abilities can be examined from a within-person framework. A within-person approach can be insightful, as it offers an alternative approach to evaluate the plausibility of a common dimension that may impact performance across a number of different cognitive capacities. In an empirical investigation, Schmiedek et al. (2020) administered nine tests of cognitive ability (three each for working memory, processing speed, and episodic memory) to a sample of 101 participants who completed the tests on 100 occasions across a six month period. As a within-person investigation, Schmiedek et al. (2020) were specifically interested in the slight changes in cognitive performance across testing occasions, independently of the overall trend of increasing test scores that would be expected from repeated testing. Though the strength of the covariance between abilities was reduced in magnitude when examined from a within-person perspective, in comparison to the between-person data, there was nonetheless notable positively shared variance, especially between the working memory capacity and episodic memory latent variables. Though Schmiedek et al.'s (2020) sample was relatively small, and their measures lacked sufficient diversity to measure general intelligence respectably, the notable positive covariance between working memory and episodic memory within individuals reinforces the idea that a common cognitive foundation may underpin diverse abilities, i.e., affirming the plausibility of *g* as substantive psychological dimension.

In contrast to how general intelligence is typically conceptualised in psychology, artificial general intelligence definitions tend to reflect two perspectives: (1) functional-equivalence; and (2) capability-based. Below, we review the two categories of AGI conceptualisations, and we note how and why they should not be considered appropriate definitions of AGI.[14]

With respect to the functional-equivalence perspective, the term AGI is commonly defined as a quantitative level of artificial intelligence, specifically, a human level of intelligence (Amazon Web Services, 2024; Demasi et al., 2010; McLean et al., 2023; Obaid, 2023; Rayhan et al., 2023). Though the observation of AI at the level of typical human

[12] Many recognise the distinction between *g*, a theoretical construct advanced to account for the positive manifold, and psychometric *g*, a statistical representation of general intelligence typically derived from factor analyses conducted upon a correlation matrix of diverse subtests of cognitive abilities (Jensen, 1998).

[13] Some computer scientists consider efficiency a fundamental characteristic of artificial intelligence (Wang, 2006).

[14] Our review of definitions of AGI is not exhaustive. For example, Ozkural (2022) suggested that artificial general intelligence may be considered a mechanism capable of effectively performing operator induction, i.e., synthesizing data, generating hypotheses, and making predictions in a computationally efficient manner.

intelligence would be a remarkable technical accomplishment, defining AGI in such a way is inconsistent with descriptions of human general intelligence, as described above. Furthermore, human general intelligence, or psychometric *g*, is observed across all levels of human ability (Breit et al., 2022; Detterman & Daniel, 1989). Similarly, general intelligence factors have been identified in various species, including dogs (Arden & Adams, 2016), deer (Pastrana et al., 2022), and orangutans (Damerius et al., 2019). This cross-species observation suggests that *g* transcends mere cognitive complexity, highlighting its universal relevance across different levels of intelligence. Therefore, considering existing theory and empirical evidence on general intelligence, AGI may be expected to be observed across essentially the whole spectrum of AI system performance. Such a hypothesis could be tested empirically by administering a series of AI system performance tests to a diversity of AI systems, as we discuss further below.

With respect to the capability-based perspective, AGI is conceptualised as an AI system's ability to perceive, possess knowledge, understand, learn, and function either partially or completely autonomously in a variety of environments and tasks (Chollet, 2019; Huang, 2017; Maruyama, 2020; Mindt & Montemayor, 2020; Morris et al., 2023). As described above, such definitions are essentially consistent with definitions of intelligence (e.g., Gottfredson, 1997), not specifically general intelligence. Consequently, capability-based definitions are arguably not useful to distinguish AGI from AI. Furthermore, these conceptualisations of AGI do not recognise that the observation of AGI, like human general intelligence, arises from a) individual differences in AI system performance; and b) the observation of positive correlations between task performance across AI systems.

Like human intelligence, there are appreciable individual differences in AI model performance across various tasks (DeRose et al., 2020; Kumari, 2023). Consequently, drawing on the human general intelligence literature, artificial general intelligence may be defined as a theoretical construct representing the shared variance in AI systems' performance, demonstrated through their positively inter-correlated capabilities across a diverse range of AI metric tasks and multiple modalities (e.g., verbal and spatial). To our knowledge, our definition of AGI introduces a novel perspective that enhances alignment between the fields of psychology and computer science. As we discuss further below, there is some preliminary empirical evidence supportive of AGI as conceptualised here. Next, we describe empirically testable models of intelligence that both include and exclude general factors of intelligence.

9. Models of intelligence and *g*

Arguably, the most commonly recognised model of intelligence is the Cattell-Horn-Carroll model (CHC; Schneider & McGrew, 2018), a comprehensive framework that integrates a wide range of cognitive abilities. The CHC model categorises abilities across three strata, each representing a different level, or breadth, of cognitive functioning. At stratum I are abilities that are narrow in nature, representing specific cognitive tasks and processes. Examples include induction (I), reading comprehension (RC), spatial relations (SR), and working memory (MW).

Stratum II is the intermediate level and consists of relatively broad cognitive abilities in comparison to those abilities of Stratum I. Stratum II abilities arise because of relatively highly-correlated clusters of narrow stratum I abilities. Schneider and McGrew (2018) review listed a total of 17 stratum II abilities. In addition to the relatively well-known fluid reasoning factor (*Gf*), there are four factors that represent acquired-knowledge abilities, including comprehension–knowledge (*Gc*), domain-specific knowledge (*Gkn*), reading and writing (*Gw*), and quantitative knowledge (*Gq*). There are five specific sensory abilities, including visual (*Gv*), auditory (*Ga*), olfactory (*Go*), tactile (*Gh*), and kinesthetic (*Gk*). There are three memory factors, including working memory capacity (*Gwm*), learning efficiency (*Gl*), and retrieval fluency (*Gr*). There are also three speed-relevant abilities, including reaction/decision time (*Gt*), processing speed (*Gs*), and psychomotor speed (*Gps*). Finally, there is a psychomotor ability factor (*Gp*). The correlations between dimensions at stratum II are notably strong, typically in the $r \approx$ 0.60 to 0.65 region (Bryan & Mayer, 2020). That is, people who have high reasoning ability (*Gf*) also tend to have higher levels of comprehension-knowledge (*Gc*), for example.

Finally, stratum III is the top level, representing general intelligence or '*g*'. It may be considered a representation of overall cognitive ability (Carroll, 2003). To date, there are two approaches to the representation of general intelligence: (1) *g* as superordinate factor; and (2) *g* as breadth factor (Beaujean, 2015; Gignac, 2008). A visual representation of a higher-order version of the CHC model is depicted in Fig. 1. It can be seen that *g* is at the top with arrows leading to the stratum II abilities. Theoretically, the fact that the arrows lead from *g* to the stratum II abilities implies that *g* is the cause of the inter-correlations between the stratum II cognitive ability dimensions, even though some researchers do not believe that *g* is a cognitive process, as discussed above.

An alternative representation of a hierarchical model is the bifactor model, where *g* is a first-order factor like the stratum II dimensions, however, the *g* factor is associated with much greater breadth than the stratum II dimensions, and the stratum II dimensions are nested within the *g* factor (see Fig. 2). In the bifactor model of intelligence, *g* is considered to be a more direct cause of the inter-correlations between stratum I abilities/tasks, and the stratum II factors are all orthogonal to each other (and *g*). Though there is some empirical evidence in favour of a bifactor representation of human cognitive abilities (Cucina & Byle, 2017), there is as yet no comprehensive evidence supporting either the higher-order (superordinate) or breadth (bifactor) representation of *g*. Across a wide variety of IQ test batteries and samples, the general factor of intelligence is typically observed to account for 35 to 50% of the total variance in cognitive ability test performance (Canivez & Watkins, 2010; Chang et al., 2014; Dombrowski et al., 2018).

It would be misleading to suggest that there is consensus on the empirical and theoretical plausibility of *g* at all. Some prefer to consider a correlated factor model of intelligence, whereby there is a large number of inter-associations between the stratum II dimensions, as opposed to an overarching general factor (e.g., Horn, 1989).[15] An even more substantially disaggregated model of intelligence is a network model whereby only the inter-associations between the stratum I abilities are specified. According to network models of intelligence, cognitive abilities are seen more as a web of interconnected skills and processes, rather than being dominated or driven by a general factor or even group factors (van der Maas et al., 2017). In this view, intelligence is conceptualised as a dynamic system where various narrow cognitive abilities interact and influence each other in complex ways. The process overlap theory of intelligence is consistent with such a view (Kovacs & Conway, 2016). There is some psychometric (McGrew et al., 2023) and cognitive neuroscience evidence (Luppi et al., 2022) supportive of a network model conceptualisation of individual differences in intelligence.

A visual representation of a network model of intelligence is presented in Fig. 3. The circles represent narrow dimensions of ability. For example, VL represents lexical knowledge and RQ represents quantitative reasoning. Lines between circles represent shared variance, i.e., correlations. Furthermore, larger correlations are represented by progressively thicker lines. In network models, three or more nodes that are relatively highly intercorrelated with each other are often shown to have the same colour. In Fig. 3, there are three communities of narrow abilities (i.e., Gf, *Gc*, and *Gv*).

As a general statement, AI system research tends not to consider

[15] John B. Carroll consistently recognised the plausibility of a stratum III *g* factor, whereas other CHC theorists are more equivocal about the *g* factor, recommending researchers/practitioners to ignore it, if they prefer a correlated-factor model perspective (Schneider & McGrew, 2012).

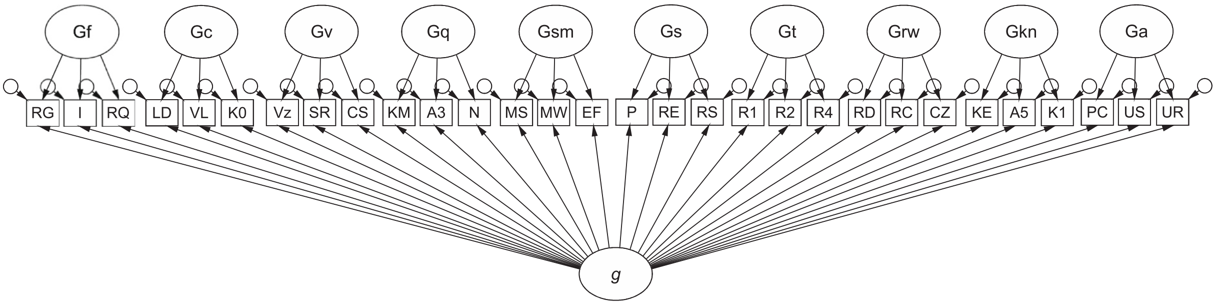

Fig. 2. Example Bifactor Model of Intelligence.
Note. Circles represent latent dimensions; squares represent observe variables (i.e., test scores); see Fig. 1 note for acronym spellings.

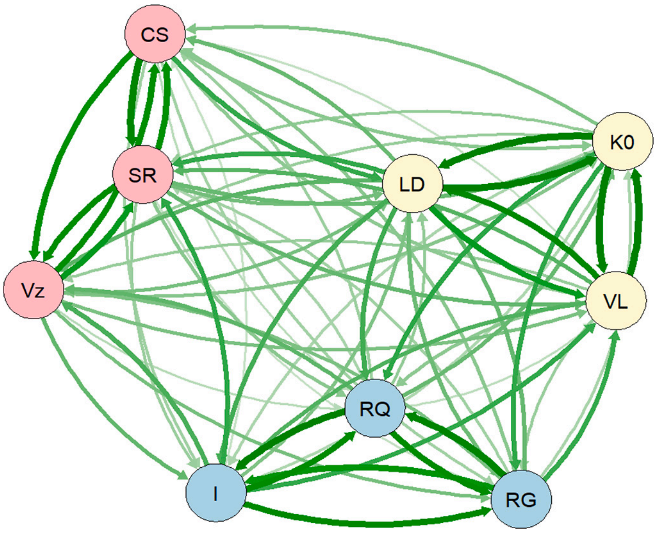

Fig. 3. Example Network Model of Intelligence Based on Nine CHC Stratum I Abilities.
Note. CS = closure speed; SR = spatial relations; Vz = visualization; I = induction; RQ = quantitative reasoning; RG = general sequential reasoning; LD = language development; K0 = general (verbal) information; VL = lexical knowledge; mauve coloured circles represent fluid reasoning abilities (*Gf*); yellow coloured circles represent comprehension knowledge abilities (*Gc*); blue coloured circles represent visual intelligence abilities (*Gv*).

taxonomies or models of ability as commonly as human intelligence researchers do. Exceptions include the Soar and LIDA architectures (or models) to represent AI system abilities. Ichise (2016) found that while the Soar and LIDA models shared some dimensional overlap with the CHC model of intelligence neither Soar nor LIDA were nearly as comprehensive. Consequently, it may be advantageous for computer scientists to adopt the CHC model, or a variation thereof, in AI system research, as it provides a comprehensive framework to evaluate and possibly develop AI systems with capabilities that mirror human cognitive abilities. It is noteworthy that many of the numerous AI system benchmark tests can, theoretically, be categorised within the diverse strata of the CHC model of intelligence. For example, Hellaswag (Zellers et al., 2019), which contains items relevant to commonsense reasoning to predict the most plausible continuation of a scenario, may be considered a measure of *Gf*, and Winogrande (Sakaguchi et al., 2021), which contains items primarily relevant to reading comprehension, may be classified as a measure of *Grw*.

10. Multidimensionality of intelligence

Although the concept of general intelligence remains controversial, there is a broad agreement among scholars that human intelligence is multi-dimensional (Neisser et al., 1996; Schneider & Newman, 2015). This is an important consideration, as some have argued that artificial intelligence represents essentially a single capacity. For example, the capacity to learn (e.g., Chollet, 2019) or to adapt to the environment (e.g., Wang, 2022). However, in psychology, the general factor of intelligence is derived from a diverse array of cognitive abilities (Jensen, 1998). Furthermore, explicitly recognising the multidimensionality of intelligence, whether alongside a general factor or not, is important, as it facilitates a more plausible depiction of how cognitively complex phenomena likely arise, including the manifestation of complex cognitive (or artificial) functions.

First, consider inspection time, which represents the minimum time required for the presentation of a stimulus on a screen for a person to detect a target (e.g., with 90% accuracy). In the typical inspection time measurement paradigm, the specific task is for a person to identify which of one of two lines presented vertically on a screen is the longest (Nettelbeck et al., 1996; Nettelbeck & Lally, 1976). Thus, the task relies upon essentially no prior knowledge, nor does it depend upon the demonstration of learning. A typical healthy adult inspection time is approximately 45 milliseconds, with appreciable individual differences (SD = 19; Crawford et al., 1998). Many studies have demonstrated a correlation between shorter inspection times and higher levels of more complex cognitive abilities, with a correlation coefficient of approximately −0.50 (Grudnik & Kranzler, 2001). Individual differences research suggests that inspection time is mostly associated with processing speed (*Gs*) and visual intelligence (*Gv*), with some unique effect associated with general intelligence (Crawford et al., 1998; O'Connor & Burns, 2003).

The ability to quickly perceive and interpret basic visual information is undeniably crucial for intelligent behaviour, as demonstrably evident in activities like human and AI-assisted motor vehicle driving, where rapid visual processing is essential (Roenker et al., 2003; Zhao, Zhao, et al., 2023; Zhao, Zhou, et al., 2023). Correspondingly, human intelligence has been found to be a positive predictor of driving ability in both driver simulated and non-simulated environments (Anderson et al., 2005; Ledger et al., 2019; Smith & Kirkham, 1982). Importantly, while processing speed is a recognised key factor in human intelligence (Jensen, 2006; Wechsler, 2008), the AI literature defining intelligence and AGI seldom addresses speed of information processing in a manner analogous to it treatment in human intelligence research, highlighting a key conceptual difference between these disciplines. Given the prominence with which processing speed is often considered a central feature that differentiates computer systems (Wang, 2020), it is interesting to speculate that variability in AI system efficiency may play a role in the possible observation of an artificial general intelligence factor.

Further support for conceptualizing intelligence as a construct broader than any single ability is found in cascading models of cognitive abilities. These models represent a hierarchical structure where foundational cognitive processes underpin more complex abilities. Empirical estimates (e.g., beta-weights) from well-fitting cascading models can shed light on the emergence of cognitive abilities, illustrating how basic intellectual processes gradually underlie more sophisticated forms of intelligence.

As an example, Fry and Hale (1996) measured processing speed ability, working memory capacity, and reasoning ability in a sample of children, adolescents, and young adults aged 7 to 19 years. They reported path analytic evidence in favour of a cascading model of abilities that lead from processing speed to working memory capacity, and working memory capacity to reasoning ability, as depicted in Fig. 4. Therefore, reasoning ability, a complex cognitive function, is partially grounded in relatively simpler processes like working memory and processing speed. Exploring whether AI systems exhibit a comparable cascading model of abilities could offer critical insights into their emergent dynamics. Such an understanding could inform the design of AI architectures that emulate the hierarchical progression of human cognitive development, thereby potentially improving their proficiency in complex tasks through a structured foundation in fundamental processes.

Additionally, consider face processing abilities, dimensions known to associate positively with *g* (Walker et al., 2023), and a significant current focus of AI system research (Hupont et al., 2022). Walker et al. (2023) provided both theoretical and empirical evidence for individual differences in human face processing abilities as consistent with simpler processes leading to more complex processes. Specifically, Walker et al. (2023) measured face detection ability (the ability to detect a face within a visual scene), face perception ability (the ability to distinguish faces within a group), face memory ability (the ability to recall a face) and face emotion expression recognition ability (the ability to correctly identify emotional expressions). Based on the correlations between corresponding latent variables, Walker et al. (2023) found that a cascading model of face processing abilities, leading from face detection to face perception to face memory to face emotional recognition ability, was consistent with the data (see Fig. 5). The findings by Walker et al. (2023) underscore the hierarchical nature of face processing abilities, suggesting that advancements in AI research on face recognition could benefit from adopting a similar cascading model. This approach might enhance AI's capability in complex face-related tasks by mirroring the stepwise development from basic detection to nuanced emotional recognition observed in humans.

In practice, cascading models of AI could be tested empirically, if there are positive inter-correlations between AI system performances (i.e., correlations between AI benchmark test scores). Beyond potentially establishing AI as multi-dimensional in nature, the empirical establishment of AI cascading models of abilities may foster a more integrated understanding of AI, one that may mirror the dynamic complexity of human intelligence to some extent. However, it is noteworthy that some large language models (LLMs) can execute high level language processing in some contexts (e.g., generate coherent narratives), but are surprisingly weak at executing other tasks humans find relatively easy (e.g., implicature; see Ruis et al., 2022), suggesting that LLMs may exhibit artificial achievement (i.e., independent skills acquired through specific training) rather than artificial intelligence.

In summary, definitions of AI (or AGI) that focus exclusively on a single dimension of ability (e.g., learning) risk oversimplifying its fundamental nature: an oversight that can foster misconceptions about AI's complexity and potentially impede the development of AI systems.

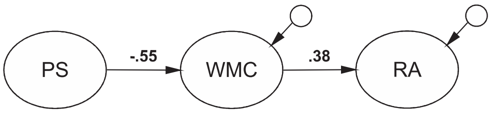

Fig. 4. Cascading Model of Cognitive Abilities.
Note. Adapted from Fry & Hale, 1996; PS = processing speed; WMC = working memory capacity; RA = reasoning ability; both coefficients were statistically significant, $p < .05$.

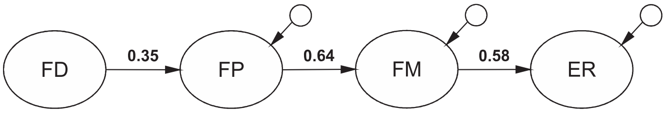

Fig. 5. Cascading Model of Cognitive Abilities: Face Processing.
Note. Adapted from Waller et al., 2023; FD = Face Detection; FP = Face Perception; FM = Face Memory; ER = Expression Recognition; coefficients in bold were statistically significant, $p < .05$.

11. Learning and intelligence

A substantial amount of AI system development is based on machine learning (Merkhofe et al., 2023; Singh et al., 2022), a fact that may motivate conceptualisations of artificial intelligence as the capacity to learn (Chollet, 2019). An exclusive focus on learning ability as the defining characteristic of artificial intelligence would be at odds with contemporary models of human intelligence, which acknowledge numerous subdimensions of cognitive abilities, only one of which is learning (Schneider & McGrew, 2018). Correspondingly, our definitions of intelligence and general intelligence do not specifically reference learning. Because machine learning is a central component in AI system development, and learning is a long researched construct in psychology, including a recognised indicator of intelligence (Schneider & McGrew, 2018), we offer some discussion on the commonalities and distinctions between human and artificial learning.

Like intelligence, learning is a construct: it is not observed directly, but inferred from the observation of inter-related behaviours. Borrowing from Jensen (1989, p. 40), human learning may be defined as a demonstrable change in the probability or intensity of a specific behaviour or behaviour potential, underpinned by neurological processes and cognitive strategies in response to various stimuli. This change excludes factors unrelated to learning, such as instinct or physical maturation. By comparison, AI learning may be defined as a demonstrable change in the probability or intensity of a specific response or decision-making potential in an artificial system, underpinned by computational algorithms and data. This change excludes factors unrelated to learning, such as programming updates or hardware modifications. Our complementary definitions of learning emphasize the role of the probability of responses in both human and AI domains, while also accounting for the distinct nature of learning in each domain.

Within the CHC model of intelligence, learning represents only a relatively small facet of the model. Specifically, according to Schneider and McGrew (2018), learning is represented by a relatively small stratum II ability known as a learning efficiency (*Gl*),[16] a dimension that represents "how much time and effort is needed to store new information in secondary memory [e.g., long-term memory]" (p. 97). Associative memory is considered an indicator (stratum I ability) of learning efficiency (*Gl*). A commonly used test of associative memory consists of face-name pairings (Rubiño & Andrés, 2018). In face-name pairings test, participants first view a series of face-name pairs and then, in the test phase, they are shown the faces again and asked to recall the associated names. This assesses their ability to form and retrieve associations, reflecting their learning efficiency in encoding and storing associative information in long-term memory. As the visual processing capacities of AI systems develop, their associative memory capacity could be measured with the validated face-name pairing test.

In addition to associative memory, meaningful memory is considered an indicator of *Gl*. Meaningful memory refers to the ability to remember information that is significant and conceptually rich, as opposed to rote memorization of arbitrary or unrelated facts. A psychometrically

[16] The better known stratum II *Glr* group-factor once comprised both learning efficiency and learning retrieval, but underwent a formal separation in Schneider and McGrew (2018).

established measure of meaningful memory is the Story Recall subtest within the Woodcock-Johnson IV (Schrank, Mather, & McGrew, 2014; Schrank, McGrew, & Mather, 2014). In the Story Recall test, participants are presented with a prerecorded short prose story, typically one to three paragraphs in length. They are then tasked with recalling and recounting the story in their own words. This free recall is assessed twice: once immediately following the presentation of the story and again after a 40-min delay, to evaluate both immediate and delayed meaningful memory recall capabilities. It should be acknowledged that the Story Recall subtest has been reclassified several times over time, suggesting a certainly level of instability in the conceptualisation and measurement of learning within contemporary human intelligence research. Nonetheless, given their sophisticated language processing capabilities, current LLMs could undertake the validated Story Recall test, with many expected to exhibit strong performance in both immediate and delayed recall tasks.

Research indicates that higher general intelligence enhances learning outcomes, with more intelligent individuals showing better responses to training (Vaci et al., 2019). Furthermore, having more prior knowledge (*Gk*) also improves learning on new tasks (Thurn et al., 2022). This research underscores the position that learning should be considered only one aspect of intelligence, one that arises from a complex interplay between other dimensions of cognitive abilities. It also raises an important question for the area of AI – can the same sorts of facilitation effects be observed?

A remarkable instantiation of human learning is the acquisition of language, a capacity that develops rapidly from infancy (Bergelson & Swingley, 2012). Furthermore, typically developing children acquire complex linguistic structures with minimal instruction (Rice, 1989; Tomasello, 2003). Stated more generally, humans excel in ‘one-shot learning,’ where they form concepts and generalize from minimal, sometimes singular, examples (e.g., Xu & Tenenbaum, 2007) - a stark contrast to AI’s need for extensive data and iterative training to achieve somewhat comparable concept formation (Zhao, Zhao, et al., 2023, Zhao, Zhou, et al., 2023; but see Lake et al., 2014). We note recent work on the Child’s View for Contrastive Learning model (CVCL) - a deep neural network for grounded word learning from slices of one child’s egocentric experience - demonstrates that many word-referent mappings (the connections between words and their real-world objects or concepts) present in a child’s everyday experience are learnable through relatively generic learning mechanisms from developmentally realistic data streams (Vong et al., 2024). Though certainly not consistent with one-shot learning, the CVCL model’s ability to acquire word-referent mappings, generalize to new visual referents, and align visual and linguistic conceptual systems partially reflects the innate human capability for rapid, efficient learning from relatively few examples.

We conclude this section by acknowledging that the degree to which learning ability is typically measured by well-established human intelligence batteries is rather limited, arguably due to the practical limitation of time. A psychologist may devote a maximum of 90 min to test a person’s intelligence comprehensively (e.g., WAIS-IV; Wechsler, 2008). Consequently, more sophisticated approaches to the measurement of individual differences in learning capacity is not feasible. By comparison, the measurement of individual AI system differences in the capacity to learn may be more feasible. Next, we discuss the learning effects derived from repeated exposures to intelligence test stimuli, as such occurrences are common in AI system development and assessment, which can be expected to invalidate the assessments as a test of artificial intelligence.

12. Impact of training on valid intelligence testing

The Advanced Progressive Matrices (APM; Raven et al., 1998a), considered one of the best measures of human fluid reasoning (Gignac, 2015), is a test that consists of 36 questions, each composed of a sequence of visual elements arranged according to abstract rules. People must first identify and encode the relevant visual features of the matrices, then induce the underlying rules governing the patterns, and finally apply those rules to generate a solution to determine the element that logically follows in the series. Carpenter et al. (1990) identified five analytic rules underlying the cognitive processes for solving APM items, distinguishing two simple, less predictive rules (constant in a row and quantitative pairwise progression) from three more complex, predictive ones (figure addition or subtraction, distribution of three values, and distribution of two values).[17] Matzen et al. (1994) found that the difficulty of each matrix reasoning question depends on the number and complexity of these rules.

Across several human studies, it has been found that repeated exposure to the APM leads to increases in test scores, but not fluid intelligence. For example, Bors and Vigneau (2003) administered the APM (36 questions) to participants on three occasions (≈ 45 day intervals). On average, people improved their performance by two additional questions answered correctly on occasion two and occasion three. Lozano and Revuelta (2020) expanded on Bors and Vigneau (2003) by examining if repeated APM test exposures enhance performance through the implicit learning of matrix reasoning rules. No improvement in ability from such repetitions was found. Instead, they found that improvers may achieve better familiarity with the test format and/or perceptual features. Based on an eye-tracking study, Hayes et al. (2015) found that improvements in APM performance through repeated testing were largely due to test-taking strategy refinement (based on changes in eye-fixation patterns), rather than increases in matrix reasoning ability. The research on the APM is consistent with the long established view that repeated exposure to test items can compromise the validity of IQ test scores (Cane & Heim, 1950; LeGagnoux et al., 1990).

We previously emphasized the importance of task novelty in intelligence testing to ensure scores reflect intelligence rather than achievement (Davidson & Downing, 2000; Raaheim & Brun, 1985). Practicing tests, or repeated exposure to similar test items, introduces factors related to achievement and expertise, thus distorting the scores as pure measures of intelligence. Consequently, many standardized intelligence tests (e.g., Wechsler scales; Wechsler, 2008; ICAR; Condon & Revelle, 2014) are regulated with respect to access to the test materials to help preserve their validity. By contrast, AI systems are often trained on similar or the same test items designed to assess their performance.

Consider Małkiński and Mańdziuk (2022) review of AI system methods for completing matrix reasoning problems. They reported that models are specifically trained on Raven’s type matrix questions. Furthermore, they stated that the Standard Progressive Matrices (Raven et al., 1998b) 60 items is insufficient for the purposes of training an AI system (p. 3). The practice of training AI systems on test items appears normative in the field of AI (Dahmen et al., 2021), which raises serious questions about intelligence demonstration when the AI systems solve the corresponding test item problems. Discerning if an AI system demonstrates intelligence or just mirrors achievement through extensive training is crucial, since intelligence, more than mere training, predicts success across diverse real-world contexts (Gottfredson, 2002).

Relatedly, in the development of large language models (LLMs), data leakage is a notable challenge. This issue arises when training data unintentionally influences the validation set, potentially skewing performance on benchmark tests (Bussola et al., 2021; Hannun et al., 2021; Linjordet & Balog, 2020; Qian et al., 2022). While efforts can be employed to help mitigate data leakage, in practice, it is difficult to avoid entirely (Lyu et al., 2021). Thus, while LLMs might not be explicitly trained on test-specific data, the inadvertent overlap between training and test datasets can compromise the unambiguous evaluation of their capabilities.

[17] Embretson (2002) later proposed an additional three relatively minor factors that contributed to APM performance, all three of which were perceptual in nature (i.e., not reliant upon induction or abstraction).

As another relevant example, the programming of goal management into AI systems may also limit the possibility of observing true artificial intelligence, as goal management is a crucial aspect of abstract problem-solving (Carpenter et al., 1990).[18] In valid human intelligence testing, the person tested is not provided with information pertinent to effective goal management. Consider the Tower of Hanoi (TOH) task which involves moving increasingly larger blocks from one peg to another (peg 1 to peg 3), without placing a larger block on a smaller one. Test items typically include between three to seven blocks. The TOH is considered primarily a measure of *Gs* and *Gf* within the CHC model of intelligence (Emick & Welsh, 2005; Jewsbury et al., 2016; Zook et al., 2004). Participants are not informed on how to strategize their approach or develop specific strategies for solving the puzzles efficiently. Instead, individuals must rely on their own cognitive abilities to devise a plan and adjust their strategy as needed, reflecting a more authentic assessment of their intelligence.

By contrast, the Optimal Ordered Problem Solver (OOPS; Schmidhuber, 2004), an AI system capable of solving Tower of Hanoi problems successfully, includes preprogrammed algorithms that define how to approach, organize, and solve the problems, significantly impacting the degree of novelty associated with the task – a defining characteristic of fluid intelligence (Carpenter et al., 1990). Thus, the programming dictated the system's goal-setting and problem-solving strategies, in contrast with human intelligence testing where goal management and strategy development are self-generated and managed. Given the OOPS' programmed instructions, it arguably demonstrated a capacity closer to achievement rather than intelligence.

We end this section with one final note about standardization and testing. Recall that standardized tests must be administered in the same manner across all occasions and cases, in order to interpret the test scores validly. In theory, there should be no problem with administering AI system tests such as HumanEval in a completely standardized manner, consistent with our operational definition of artificial intelligence. However, there is evidence that AI metrics tests are not always administered in a standardized manner, with developers choosing different environments and parameters for evaluation, which can significantly alter the results (Fortis, 2023; Kinsella, 2023; Vedula et al., 2022). Such variability introduces test biases and inconsistencies, making it difficult to directly compare performance outcomes across different AI systems or versions. Ensuring the reliability and validity of AI test scores necessitates strict adherence to standardized protocols, including uniform test sets and evaluation metrics, which is crucial for accurately assessing AI systems' capabilities, limitations, as well as investigating the possibility of artificial general intelligence.[19]

13. Testing for artificial general intelligence (AGI)

In practical terms, one approach to testing the possibility of a general factor of artificial intelligence is to submit a series of wide ranging tests to a large number of different AI systems.[20] That is, each AI system would undergo a comprehensive battery of tests, with scores recorded and analysed for inter-test correlations. Positive correlations among test scores would suggest the presence of an AGI factor. Further support for an AGI factor would be observed if a factor analysis of the test results revealed a single factor with positive loadings across all tests.

To date, two (unpublished) empirical investigations have reported correlations between test performance across AI systems.

Burnell et al. (2023) estimated the correlations between 27 tasks (Holistic Evaluation of Language Models; a.k.a., HELM benchmark) completed by 29 language models, including Anthropic-LM v4-s3, Cohere Command beta, GPT-3-davinci, and OPT. Burnell et al. (2023) reported a positive manifold with a mean inter-task correlation of 0.56, consistent with what is typically observed with human intelligence tests (Detterman & Daniel, 1989; Walker et al., 2023). Furthermore, an examination of a parallel analysis uncovered three dimensions. Based on a factor analysis, the three factors were labelled comprehension (33% variance explained), reasoning (31% variance explained), and language modeling (17% variance explained). Furthermore, the three factors inter-correlated positively (mean $r = 0.39$), suggesting the presence of a artificial general intelligence factor.

In addition to Burnell et al. (2023), Ilić (2023) reported a factor analysis that identified a single statistically significant and meaningful factor, based on a larger investigation of 1232 language models that completed 22 tasks. All 22 tasks loaded positively onto the factor which accounted for 85% of the variance in language model performance. Thus, the AGI factor identified by Ilić (2023) was stronger than Burnell et al. (2023), as well as stronger than that typically observed for human intelligence (Detterman & Daniel, 1989; Walker et al., 2023). Interestingly, the correlation between AGI factor scores and the size of the language model (i.e., parameter counts) was only 0.49, suggesting that model size does not correspond to a proportionate increase in AGI ability. Thus, other characteristics, such as model architecture, training data diversity, or optimization strategies, might play significant roles in the manifestation of AGI.

We note that, to date, no spatial AI factors have been identified, even though, in addition to LLMs, there are spatial models designed to process and interpret complex spatial data across various domains (e.g., Convolutional Neural Networks, Gu et al., 2018). Additionally, the Large Language and Vision Assistant (LLaVA; Liu, Li, et al., 2023), an end-to-end trained multimodal model that combines a vision encoder with an LLM for general-purpose visual (and language) processing, could potential solve basic visual intelligence test problems (e.g., Identical Pictures Test, Ekstrom et al. (1976); Mooney Face Detection Task, Verhallen & Mollon, 2016). Thus, in theory, a spatial artificial intelligence factor may be identified in future work.

In light of Burnell et al. (2023) and Ilić (2023), it may be suggested that there is tentative empirical evidence for the presence of AGI, or 'AI metric *g*' more precisely. However, it may be better described as artificial general achievement, as there are serious questions relevant to whether AI systems have in fact demonstrated intelligence, as discussed above. Additionally, in a manner similar to question marks over the plausibility of psychometric *g* (Ceci, 1990; Detterman, 1982), we knowledge that the empirical observation of AI metric *g* does not fully substantiate the construct of AGI, as multidimensional artificial intelligence may be better represented as a network model, as per human intelligence (McGrew et al., 2023).

Given the unsettled debates surrounding the interpretation of the positive manifold and the corresponding general factor of human intelligence (e.g., Gignac, 2016; van der Maas & Kan, 2016), the prospects emerging from the observation of a positive manifold for AI system performance are intriguing. Specifically, there may be unique opportunities to empirically test theories of general intelligence through experimental manipulation in ways that are not feasible with human subjects. For example, by systematically varying the processing speed and efficiency of AI systems, researchers can directly observe the effects on the strength and structure of the correlations between abilities, offering insights into whether these correlations—and by extension, a general intelligence factor—arise from underlying information

[18] Carpenter et al. (1990, p. 428): "a key component of analytic intelligence is goal management, the process of spawning subgoals from goals, and then tracking the ensuing successful and unsuccessful pursuits of the subgoals on the path to satisfying higher level goals." And "The use or organization of goals is a strategic level of thought, possibly involving metacognition or requiring reflection."

[19] The field of AI metrics is only nascent, with no apparent awareness of formal measurement properties including reliability and validity, even among the most recently published tests (e.g., PromptBench; MarkTechPost, 2023).

[20] An AI system is considered to encompass a wide range of artificial intelligence applications, from LLMs to visual recognition systems, and even more integrated systems that combine multiple types of AI.

processing mechanisms (as per Anderson, 1992, 2001). Over time, these types of experiments could narrow the divide between intelligence theories and actual data, possibly confirming the significance of central processing in both AI and human intelligence, for example. Ultimately, the ability of AI systems to complete a diversity of tests akin to human IQ assessments may not only illuminate AI's cognitive framework when analysed with methods well-established in human psychology, but also possibly deepen our understanding of human intelligence.

14. Memory span and intelligence

Memory span, defined as the maximum number of items a person can recall after a single exposure, usually within a brief period of about two seconds (Baddeley, 1990), is recognised as a key cognitive ability within models of human intelligence (Gignac, 2018). Notably, working memory, which involves the capacity to hold and manipulate information mentally, has been shown to be closely associated with fluid reasoning, sharing approximately 50% of its variance (Kane et al., 2005). Furthermore, some research posits that working memory and general intelligence might be nearly indistinguishable, or isomorphic (Colom et al., 2004). While the precise magnitude of the influence of memory span on problem-solving ability, which serve as critical markers of intelligence, continues to be debated, compelling evidence suggests that this effect is appreciable and likely causal (Hagemann et al., 2023).

It is interesting to observe that the AI literature on the nature of intelligence rarely ever recognises the potential role of memory span, given that an LLM's context window may be considered conceptually similar to human short-term memory. Context windows limit the number of text tokens LLMs can reference at once for generating responses or analysing input. This window is crucial for maintaining consistency and coherence in conversations or tasks, as it determines the extent of prior information the model can utilize. Thus, the similarities between context windows and human memory span include the limited capacity to hold information, focusing on the most recent or immediately relevant data for current tasks, and the mechanism of forgetting older information as new data comes in. Interestingly, LLM's appear to manifest primacy and recency effects in a manner similar to that observed in humans (Liu, Lin, et al., 2023). AI systems other than LLMs have mechanisms similar to context windows, such as the memory cells in Long Short-Term Memory (LSTM) networks used in sequence modeling (Yu et al., 2019), the field of view in Convolutional Neural Networks for image processing (Samy et al., 2018), and the state perception in Reinforcement Learning agents (Sheng et al., 2022).

In light of the above, it is noteworthy that Burnell et al. (2023) did not identify a factor relevant to memory span. This could be because none of the AI benchmark tests directly measure a dimension akin to memory span, an arguably noteworthy limitation, in our view. Given the variability in memory retrieval capacities across AI systems,[21] it is reasonable to postulate the plausibility of an AI memory span factor. Additionally, such a factor might be relatively clearly characterized as an aspect of intelligence; or, at least, it would not be obviously considered a dimension of achievement. If an AI system memory span dimension is observed to correlate positively with other AI system performance dimensions, it could support the plausibility of our novel conceptualisation of AGI, which parallels human general intelligence. A similar argument could be made for AI system processing speed, given the important role processing speed has played in the human intelligence literature (Gignac, 2018). Further research to address these possibilities is encouraged.

15. AGI and predictive validity

Human intelligence, encompassing both the broad measure of psychometric *g* and specific cognitive abilities, is a vital psychological construct, valued not merely for indicating performance on a range of intelligence tests but more so for its ability to predict key social outcomes. Research has consistently shown that general intelligence is a strong predictor of critical factors such as academic success (Pokropek et al., 2022), educational and occupational attainment (Salgado et al., 2003; Strenze, 2007), income and financial stability (Shaffer, 2020; Zagorsky, 2007), as well as essential aspects of personal safety and health, including risk assessment in daily life (e.g., avoiding fatal accidents; O'Toole, 1990) and overall longevity (Gottfredson & Deary, 2004). This extensive predictive ability, a hallmark of criterion validity, underscores the significance of psychometric *g*: it is not just a measure of how people perform on intelligence tests, but a robust indicator of future behaviour, outcomes, and achievements in diverse life domains.

Ideally, to fully validate the constructs of AI and AGI, it would be necessary to demonstrate their predictive validity through a set of complex, socially valuable criteria tailored to their unique functions. These criteria could include their efficacy in driving technological and scientific advancements, boosting human productivity, and improving overall quality of life. Admittedly, identifying suitable predictive validity criteria in the AI context presents a significant challenge, given the tremendously diverse nature of AI applications. Nonetheless, further research is encouraged in AI metrics to develop and refine these criteria, ensuring they justifiably reflect the impact and utility of AI systems in diverse real-world scenarios.

16. Conclusion

AI systems have demonstrated their capability to solve cognitive ability test problems, primarily through guided training (e.g., Zhuo & Kankanhalli, 2020) or programmed approaches to transform problems into algorithmically solvable formats (e.g., Schmidhuber, 2004). While remarkable, it is debatable whether these accomplishments signify intelligence, given that the capabilities of most current AI systems are limited to specific programming and/or training data, without the necessary demonstration of novel problem-solving ability characteristic of human intelligence (Davidson & Downing, 2000; Raaheim & Brun, 1985). Consequently, many AI systems might be more aptly recognised as having the capacity to exhibit artificial achievement or artificial expertise. Despite not reaching the threshold of artificial intelligence, artificial achievement and expertise systems should, nonetheless, be regarded as remarkable scientific accomplishments, ones that can be anticipated to impact many aspects of society in significant ways. Furthermore, with clear and coherent conceptualisations and definitions of achievement, expertise, intelligence, and general intelligence adopted by the fields of psychology and computer science, greater collaborations and insights may be facilitated, which may ultimately help bridge the gap between artificial and human-like intelligence.

CRediT authorship contribution statement

Gilles E. Gignac: Writing – review & editing, Writing – original draft, Conceptualization. **Eva T. Szodorai:** Investigation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the first author used chatGPT in order to improve readability of some passages. After using chatGPT, the first author reviewed and edited the content as needed and takes full

[21] As of March 8th, 2024, the context window sizes for GPT-3, GPT-3.5-turbo, and GPT-4 Turbo are 2048 tokens, 16,385 tokens, and 128,000 tokens respectively (OpenAI, 2023, November 6; OpenAI, 2024). Anthropic's Claude 2.1 surpasses these, with a capacity to process 200,000 tokens or roughly 150,000 words, equivalent to about 500 pages of text (Anthropic, 2023, November 21). While these specifications may evolve, it underscores the individual differences in context windows among different LLMs.

responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

No data was used for the research described in the article.

References

Amazon Web Services. 2024 (n.d.). What is AGI? Retrieved January 20th, 2024 from https://aws.amazon.com/what-is/artificial-general-intelligence/.

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Oxford, UK: Blackwell.

Anderson, M. (2001). Conceptions of intelligence. *Journal of Child Psychology and Psychiatry, 42*(3), 287–298.

Anderson, S. W., Rizzo, M., Shi, Q., Uc, E. Y., & Dawson, J. D. (2005, June). Cognitive abilities related to driving performance in a simulator and crashing on the road. In *, Vol. 3, No. 2005. Driving assessment conference*. University of Iowa.

Anthropic. (2023, November 21). Introducing Claude 2.1. https://www.anthropic.com/index/claude-2-1.

Arden, R., & Adams, M. J. (2016). A general intelligence factor in dogs. *Intelligence, 55*, 79–85.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Hillsdale, NJ: Erlbaum.

Bartholomew, D. J. (2004). *Measuring intelligence: Facts and fallacies*. Cambridge University Press.

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence, 3*(4), 121–136.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's advanced progressive matrices. *Learning and Individual Differences, 13*(4), 291–312.

Borsboom, D. (2023). Psychological constructs as organizing principles. In L. A. van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on contemporary psychometrics*. https://doi.org/10.1007/978-3-031-10370-4_5. Methodology of educational measurement and assessment.

Breit, M., Brunner, M., Molenaar, D., & Preckel, F. (2022). Differentiation hypotheses of intelligence: A systematic review of the empirical evidence and an agenda for future research. *Psychological Bulletin, 148*(7–8), 518–554.

Bryan, V. M., & Mayer, J. D. (2020). A meta-analysis of the correlations among broad intelligences: Understanding their relations. *Intelligence, 81*, Article 101469.

Burnell, R., Hao, H., Conway, A. R., & Orallo, J. H. (2023). *Revealing the structure of language model capabilities*. arXiv preprint arXiv:2306.10062.

Bussola, N., Marcolini, A., Maggio, V., Jurman, G., & Furlanello, C. (2021). AI slipping on tiles: Data leakage in digital pathology. In *Pattern recognition. ICPR international workshops and challenges: Virtual event, January 10–15, 2021, proceedings, part I* (pp. 167–182). Springer International Publishing.

Cane, V. R., & Heim, A. W. (1950). The effects of repeated retesting: III. Further experiments and general conclusions. *Quarterly Journal of Experimental Psychology, 2* (4), 182–197.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment, 22*(4), 827–836.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review, 97*, 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 5–21). Pergamon.

Ceci, S. J. (1990). *On intelligence...More or less: A bioecological treatise on intellectual development*. Prentice-Hall.

Chamorro-Premuzic, T., Furnham, A., & Ackerman, P. L. (2006). Ability and personality correlates of general knowledge. *Personality and Individual Differences, 41*(3), 419–429.

Chang, M., Paulson, S. E., Finch, W. H., Mcintosh, D. E., & Rothlisberg, B. A. (2014). Joint confirmatory factor analysis of the woodcock-Johnson tests of cognitive abilities, and the Stanford-Binet intelligence scales, with a preschool population. *Psychology in the Schools, 51*(1), 32–57.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *Vol. 16. The psychology of learning and motivation* (pp. 1–58). Academic Press.

Chi, M. T. H. (2006). Two approaches to the study of experts characteristics. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21–30). Cambridge, MA: Cambridge University Press.

Chollet, F. (2019). *On the measure of intelligence*. arXiv preprint arXiv:1911.01547.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). *Think you have solved question answering? Try ARC, the AI2 reasoning challenge*. arXiv preprint arXiv:1803.05457.

Colom, R. (2020). Intellectual abilities. In A. Gallagher, C. Bulteau, D. Cohen, & J. Michaud (Eds.), *Handbook of clinical neurology: Neurocognitive development - normative development* (pp. 109–120). Elsevier.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*, 277–296.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64.

Cormen, T. H. (2013). *Algorithms unlocked*. MIT Press.

Crawford, J. R., Deary, I. J., Allan, K. M., & Gustafsson, J. E. (1998). Evaluating competing models of the relationship between inspection time and psychometric intelligence. *Intelligence, 26*(1), 27–42.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality and Safety, 22*(Suppl. 2), 58–64.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence, 5*(3), 27.

Curran, P. J., & Hancock, G. R. (2021). The challenge of modeling co-developmental processes over time. *Child Development Perspectives, 15*(2), 67–75.

Dahmen, U., Osterloh, T., & Roßmann, J. (2021, December). Generation of virtual test scenarios for training and validation of ai-based systems. In *In 2021 IEEE international conference on Progress in informatics and computing (PIC)* (pp. 64–71). IEEE.

Damerius, L. A., Burkart, J. M., van Noordwijk, M. A., Haun, D. B., Kosonen, Z. K., Galdikas, B. M., & van Schaik, C. P. (2019). General cognitive abilities in orangutans (pongo abelii and Pongo pygmaeus). *Intelligence, 74*, 3–11.

Davidson, J. E., & Downing, C. L. (2000). Contemporary models of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 34–52). New York, NY: Cambridge University Press.

Deary, I. J. (2020). *Intelligence: A very short introduction* (2nd ed.). Oxford University Press.

Deary, I. J., Spinath, F. M., & Bates, T. C. (2006). Genetics of intelligence. *European Journal of Human Genetics, 14*(6), 690–700.

Demasi, P., Szwarcfiter, J. L., & Cruz, A. J. (2010, June). A theoretical framework to formalize AGI-Hard problems. In *3d conference on artificial general Intelligence (AGI-2010)* (pp. 64–65). Atlantis Press.

DeRose, J. F., Wang, J., & Berger, M. (2020). Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1160–1170.

Detterman, D. K. (1982). Does "g" exist? *Intelligence, 6*, 99–108.

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence, 13*(4), 349–359.

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018). Hierarchical exploratory factor analyses of the Woodcock-Johnson IV Full Test Battery: Implications for CHC application in school psychology. *School Psychology Quarterly, 33*(2), 235–250.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Kit of factor-referenced cognitive tests. *ETS Research and Development.*

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. Irvine, & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 219–250). Mahwah, NJ: Erlbaum.

Emick, J., & Welsh, M. (2005). Association between formal operational thought and executive function as measured by the Tower of Hanoi-Revised. *Learning and Individual Differences, 15*(3), 177–188.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge, MA: Cambridge University Press.

Ericsson, K. A., & Staszewski, J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In D. Klahr, & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 235–267). Hillsdale, NJ: Erlbaum.

Feng, Y., Lin, J., Dwivedi, S. K., Sun, Y., Patel, P., & Black, M. J. (2023). *PoseGPT: Chatting about 3D human pose*. arXiv preprint arXiv:2311.18836.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465.

Flanagan, D. P., & Dixon, S. G. (2013). The Cattell-Horn-Carroll theory of cognitive abilities. In C. Reynolds, K. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education: A reference for the education of children, adolescents, and adults with disabilities and other exceptional individuals* (4th ed.). Wiley.

Fortis, S. (2023, December 7). *Is Google's Gemini smarter than OpenAI's Chatgpt? Community sleuths find out*. Cointelegraph. https://cointelegraph.com/news/is-google-s-gemini-really-smarter-than-openai-s-gpt-4-community-sleuths-find-out.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7*, 237–241.

Gabora, L., & Russon, A. (2011). The evolution of human intelligence. In R. Sternberg, & S. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 328–350). Cambridge University Press.

Gardner, H., & Hatch, T. (1989). Multiple intelligences go to school: Educational implications of the Theory of multiple intelligences. *Educational Researcher, 18*(8), 4–10.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science, 50*(1), 21–43.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of *g. Intelligence, 52*, 71–79.

Gignac, G. E. (2016). On the evaluation of competing theories: A reply to van der Maas and Kan. *Intelligence, 57*, 84–86.

Gignac, G. E. (2018). Conceptualizing and measuring intelligence. In V. Zeigler-Hill, & T. Shackelford (Eds.), *Vol. 1. The SAGE handbook of personality and individual differences* (pp. 439–464). Sage.

Gignac, G. E., & Weiss, L. G. (2015). Digit span is (mostly) related linearly to general intelligence: Every extra bit of span counts. *Psychological Assessment, 27*(4), 1312–1323.

Gobet, F., & Sala, G. (2023). Cognitive training: A field in search of a phenomenon. *Perspectives on Psychological Science, 18*(1), 125–141.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*(3), 225–255.

Goertzel, B. (2010). Toward a formal characterization of real-world general intelligence. In *Proceedings of the 3d conference on artificial general Intelligence (2010)* (pp. 74–79). Atlantis Press.

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1–46.

Goertzel, B., & Yu, G. (2014, July). From here to AGI: A roadmap to the realization of human-level artificial general intelligence. In *2014 international joint conference on neural networks (IJCNN)* (pp. 1525–1533). IEEE.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23.

Gottfredson, L. S. (2002). G: Highly general and highly practical. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 331–380). Erlbaum.

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*(1), 1–4.

Grafen, A. (2015). Biological fitness and the fundamental theorem of natural selection. *The American Naturalist, 186*(1), 1–14.

Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence, 29*(6), 523–535.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., … Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377.

Hagemann, D., Ihmels, M., Bast, N., Neubauer, A. B., Schankin, A., & Schubert, A. L. (2023). Fluid intelligence is (much) more than working memory capacity: An experimental analysis. *Journal of Intelligence, 11*(4), 70.

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Taylor & Francis.

Hannun, A., Guo, C., & van der Maaten, L. (2021, December). Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in artificial Intelligence* (pp. 760–770). PMLR.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1–14.

Horn, J. L. (1989). Models of intelligence. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 29–73). University of Illinois Press.

Huang, T. J. (2017). Imitating the brain with neurocomputer a "new" way towards artificial general intelligence. *International Journal of Automation and Computing, 14* (5), 520–531.

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds, & R. T. Brown (Eds.), *Perspectives on Bias in mental testing* (pp. 221–247). Springer.

Hupont, I., Tolan, S., Gunes, H., & Gómez, E. (2022). The landscape of facial processing applications in the context of the European AI act and the development of trustworthy systems. *Scientific Reports, 12*(1), 10688.

Ichise, R. (2016). An analysis of the CHC model for comparing cognitive architectures. *Procedia Computer Science, 88*, 239–244.

Ilić, D. (2023). *Unveiling the general intelligence factor in language models: A psychometric approach.* arXiv preprint arXiv:2310.11616.

Jensen, A. R. (1989). The relationship between learning and intelligence. *Learning and Individual Differences, 1*(1), 37–62.

Jensen, A. R. (1990). Speed of information processing in a calculating prodigy. *Intelligence, 14*(3), 259–274.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Jensen, A. R., & Weng, L. J. (1994). What is a good g? *Intelligence, 18*(3), 231–258.

Jewsbury, P. A., Bowden, S. C., & Strauss, M. E. (2016). Integrating the switching, inhibition, and updating model of executive function with the Cattell—Horn—Carroll model. *Journal of Experimental Psychology: General, 145*(2), 220–245.

Johnson, D. A. (2006). The United States Medical Licensing Examination (USMLE): Maintaining the integrity of the examination process. *Journal of Medical Regulation, 92*(3), 16–19.

Johnson, W. (2013). Whither intelligence research? *Journal of Intelligence, 1*(1), 25–35.

Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25*(7), 15–22.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66–71.

Keage, H. A., Muniz, G., Kurylowicz, L., Van Hooff, M., Clark, L., Searle, A. K., … McFarlane, A. (2016). Age 7 intelligence and paternal education appear best predictors of educational attainment: The Port Pirie cohort study. *Australian Journal of Psychology, 68*(1), 61–69.

Kinsella, B. (2023, December 21). *CMU study shows Gemini Pro Trails GPT-3.5 and GPT-4 in performance benchmarks. Is Gemini Pro as robust as Google says?* Synthedia. https://synthedia.substack.com/p/cmu-study-shows-gemini-pro-trails.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*(3), 151–177.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kumar, A. V. S., Sharma, P., Irawati, I. D., Chandrashekar, D. V., Musirin, I. B., Abdullah, H. M. A., & Rao, M. L. (2023). Artificial intelligence in computer science: An overview of current trends and future directions. In S. Suman Rajest, B. Singh, A. J. Obaid, R. Regin, & K. Chinnusamy (Eds.), *Advances in artificial and human intelligence in the modern era* (pp. 43–60). IGI Global.

Kumari, P. (2023, October 4). *Comparing language models through parameters vs. real-life experiments*. Labellerr. https://www.labellerr.com/blog/comparing-language-models-through-parameters-vs-real-life-experiments/.

Lake, B., Lee, C. Y., Glass, J., & Tenenbaum, J. (2014). One-shot learning of generative speech concepts. In *, vol. 36, No. 36. Proceedings of the annual meeting of the cognitive science society*.

Ledger, S., Bennett, J. M., Chekaluk, E., & Batchelor, J. (2019). Cognitive function and driving: Important for young and old alike. *Transportation Research Part F: Traffic Psychology and Behaviour, 60*, 262–273.

LeGagnoux, G., Michael, W. B., Hocevar, D., & Maxwell, V. (1990). Retest effects on standardized structure-of-intellect ability measures for a sample of elementary school children. *Educational and Psychological Measurement, 50*(3), 475–492.

Legg, S., & Hutter, M. (2007a). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications, 157*, 17.

Legg, S., & Hutter, M. (2007b). Universal Intelligence: A definition of MachineIntelligence. *Minds and Machines, 17*, 391–444.

Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring how models mimic human falsehoods.* arXiv preprint arXiv:2109.07958.

Linjordet, T., & Balog, K. (2020, September). Sanitizing synthetic training data generation for question answering over knowledge graphs. In *Proceedings of the 2020 ACM SIGIR on international conference on theory of information retrieval* (pp. 121–128).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual instruction tuning.* arXiv preprint arXiv: 2304.08485.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). *Lost in the middle: How language models use long contexts*. arXiv preprint arXiv: 2307.03172.

Lozano, J. H., & Revuelta, J. (2020). Investigating operation-specific learning effects in the Raven's advanced progressive matrices: A linear logistic test modeling approach. *Intelligence, 82*, Article 101468.

Lozano-Blasco, R., Quílez-Robres, A., Usán, P., Salavera, C., & Casanovas-López, R. (2022). Types of intelligence and academic performance: A systematic review and meta-analysis. *Journal of Intelligence, 10*(4), 123.

Lu, Y., Xu, D., Wang, L., Hartley, R., & Li, H. (2010, July). Illumination invariant sequential filtering human tracking. In *, Vol. 4. In 2010 international conference on machine learning and cybernetics* (pp. 2133–2138). IEEE.

Lüdtke, O., & Robitzsch, A. (2022). A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(6), 888–907.

Luppi, A. I., Mediano, P. A., Rosas, F. E., Holland, N., Fryer, T. D., O'Brien, J. T., … Stamatakis, E. A. (2022). A synergistic core for human brain evolution and cognition. *Nature Neuroscience, 25*(6), 771–782.

Lykken, D. T. (2005). Mental energy. *Intelligence, 33*(4), 331–335.

Lyu, Y., Li, H., Sayagh, M., Jiang, Z. M., & Hassan, A. E. (2021). An empirical study of the impact of data splitting decisions on the performance of AIOps solutions. *ACM Transactions on Software Engineering and Methodology (TOSEM), 30*(4), 1–38.

van der Maas, H. L., Kan, K. J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence, 5*(2), 16.

Małkiński, M., & Mańdziuk, J. (2022). *Deep learning methods for abstract visual reasoning: A survey on Raven's progressive matrices*. arXiv preprint arXiv:2201.12382.

MarkTechPost. (2023, December 23). *Microsoft researchers introduce PromptBench: A PyTorch-based Python package for evaluation of Large Language Models (LLMs)*. MarkTechPost. https://www.marktechpost.com/2023/12/23/microsoft-researchers-introduce-promptbench-a-pytorch-based-python-package-for-evaluation-of-large-language-models-llms/.

Maruyama, Y. (2020). The conditions of artificial general intelligence: Logic, autonomy, resilience, integrity, morality, emotion, embodiment, and embeddedness. In *Artificial general intelligence: 13th international conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, proceedings 13* (pp. 242–251). Springer International Publishing.

Matzen, L. B. V., Van der Molen, M. W., & Dudink, A. C. (1994). Error analysis of Raven test performance. *Personality and Individual Differences, 16*(3), 433–445.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10.

McGrew, K. S., Schneider, W. J., Decker, S. L., & Bulut, O. (2023). A psychometric network analysis of CHC intelligence measures: Implications for research, theory, and interpretation of broad CHC scores "Beyond g". *Journal of Intelligence, 11*(1), 19.

McIntosh, D. E., Dixon, F. A., & Pierson, E. E. (2005). Use of intelligence tests in the identification of giftedness. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 504–520). Guilford Press.

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence, 35*(5), 649–663.

Merkhofe, E., Chaudhari, D., Anderson, H. S., Manville, K., Wong, L., & Gante, J. (2023). *Machine learning model attribution challenge*. arXiv preprint arXiv:2302.06716.

Merritt, D. J., & Cornett, L. (2020). Building a better bar: The twelve building blocks of minimum competence. *AccessLex Institute Research Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3793580.

Messick, S. (1981). Constructs and their vicissitudes. *Psychological Bulletin, 89*, 575–588.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Mindt, G., & Montemayor, C. (2020). A roadmap for artificial general intelligence: intelligence, knowledge, and consciousness. *Mind and Matter, 18*(1), 9–37.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Monett, D., & Lewis, C. W. (2018). Getting clarity by defining artificial intelligence—A survey. In *Philosophy and theory of artificial intelligence 2017* (pp. 212–214). Springer International Publishing.

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., … Legg, S. (2023). *Levels of AGI: Operationalizing Progress on the Path to AGI*. arXiv preprint arXiv:2311.02462.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101.

Nettelbeck, T., & Lally, M. (1976). Inspection time and measured intelligence. *British Journal of Psychology, 67*(1), 17–22.

Nettelbeck, T., Rabbitt, P. M. A., Wilson, C., & Batt, R. (1996). Uncoupling learning from initial recall: The relationship between speed and memory deficits in old age. *British Journal of Psychology, 87*(4), 593–607.

Nguyen, H. D., & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), Article e1246.

Norris, D. G., Hall, J., & Gathercole, S. E. (2019). Can short-term memory be trained? *Memory & Cognition, 47*, 1012–1023.

Obaid, O. I. (2023). From machine learning to artificial general intelligence: A roadmap and implications. *Mesopotamian Journal of Big Data, 2023*, 81–91.

O'Connor, T. A., & Burns, N. R. (2003). Inspection time and general speed of processing. *Personality and Individual Differences, 35*(3), 713–724.

OpenAI. (2023). New models and developer products announced at DevDay. https://openai.com/blog/new-models-and-developer-products-announced-at-devday.

OpenAI. (2024). *Models overview*. OpenAI (n.d.) https://platform.openai.com/docs/models/overview.

O'Toole, B. J. (1990). Intelligence and behavior and motor vehicle accident mortality. *Accident Analysis and Prevention, 22*, 211–221.

Ozkural, E. (2022). Measures of intelligence, perception and intelligent agents. In B. Goertzel, M. Ikle, & A. Potapov (Eds.), *Artificial general intelligence: 14th international conference, AGI 2021* (pp. 174–183). Springer.

Papadouris, N., & Constantinou, C. P. (2012). Middle school students using energy analysis in diverse phenomena. *Review of Science, Mathematics and ICT Education, 6*(1), 73–87.

Pastrana, C. I., Gonzalez, F. J. N., Inostroza, M. G. P., Arbulu, A. A., Bermejo, J. V. D., & Aguilera, M. J. R. (2022). Study of variability of cognitive performance in captive fallow deer (Dama dama) through *g* and *c* factors. *Journal of Veterinary Behavior, 47*, 70–85.

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives, 14*(1), 15–20.

Pintner, R. (1923). *Intelligence testing*. Holt, Rinehart, & Winston.

Plomin, R. (2018). *Blueprint: How DNA makes us who we are*. MIT Press.

Pokropek, A., Marks, G. N., & Borgonovi, F. (2022). How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities? *Journal of Educational Psychology, 114*(5), 1121–1135.

Postman, L. (1968). Hermann Ebbinghaus. *American Psychologist, 23*(3), 149–157.

Prasad, A., Kumar, A. S., Sharma, P., Irawati, I. D., Chandrashekar, D. V., Musirin, I. B., & Abdullah, H. M. A. (2023). Artificial Intelligence in Computer Science: An Overview of Current Trends and Future Directions. *Advances in Artificial and Human Intelligence in the Modern Era*, 43–60.

Prasad, A., Kumar, A. S., Sharma, P., Irawati, I. D., Chandrashekar, D. V., Musirin, I. B., … Worrell, F. C. (2020). Talent development in achievement domains: A psychological framework for within-and cross-domain research. *Perspectives on Psychological Science, 15*(3), 691–722.

Preckel, F., Golle, J., Grabner, R., Jarvin, L., Kozbelt, A., Müllensiefen, D., … Worrell, F. C. (2020). Talent development in achievement domains: A psychological framework for within-and cross-domain research. *Perspectives on Psychological Science, 15*(3), 691–722.

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., & Schimel, J. (2004). Why do people need self-esteem? A theoretical and empirical review. *Psychological Bulletin, 130*(3), 435–468.

Qian, H., Wang, B., Ma, P., Peng, L., Gao, S., & Song, Y. (2022, November). Managing dataset shift by adversarial validation for credit scoring. In *Pacific rim international conference on artificial intelligence* (pp. 477–488). Cham: Springer Nature Switzerland.

Raaheim, K., & Brun, W. (1985). Task novelty and intelligence. *Scandinavian Journal of Psychology, 26*, 35–41.

Raj, D. J. S. (2019). A comprehensive survey on the computational intelligence techniques and its applications. *Journal of IoT in Social, Mobile, Analytics, and Cloud, 1*(3), 147–159.

Rammstedt, B., & Rammsayer, T. H. (2000). Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences, 29*(5), 869–880.

Rasheed, N., & Amin, S. H. (2016). Developmental and evolutionary lexicon acquisition in cognitive agents/robots with grounding principle. *Computational Intelligence and Neuroscience, 8571265*. https://doi.org/10.1155/2016/8571265

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Raven manual section 4: Advanced progressive matrices*. Oxford Psychologists Press.

Raven, J. C., Raven, J. E., & Court, J. H. (1998b). *Progressive matrices*. Oxford Psychologists Press.

Rayhan, A., Rayhan, R., & Rayhan, S. (2023). *Artificial general Intelligence: Roadmap to achieving human-level capabilities*.

Reiss, S. (1997). Trait anxiety: It's not what you think it is. *Journal of Anxiety Disorders, 11*(2), 201–214.

Rice, M. L. (1989). Children's language acquisition. *American Psychologist, 44*, 149–156.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*(8), 1358–1369.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors, 45*(2), 218–233.

Rubiño, J., & Andrés, P. (2018). The face-name associative memory test as a tool for early diagnosis of Alzheimer's disease. *Frontiers in Psychology, 9*, 1464.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). *Large language models are not zero-shot communicators*. arXiv preprint arXiv:2210.14986.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach* (3rd ed.). Pearson.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM, 64*(9), 99–106.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*(1), 9–20.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology, 88*, 1068–1081.

Samy, M., Amer, K., Eissa, K., Shaker, M., & ElHelw, M. (2018). Nu-net: Deep residual wide field of view convolutional neural network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 267–271).

Schmidhuber, J. (2004). Optimal ordered problem solver. *Machine Learning, 54*, 211–254.

Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2020). Within-person structures of daily cognitive performance differ from between-person structures of cognitive abilities. *PeerJ, 8*, e9290.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed., pp. 99–144). Guilford Press.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan, & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). The Guilford Press.

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review, 25*(1), 12–27.

Schoser, B. (2023). Framing artificial intelligence to neuromuscular disorders. *Current Opinion in Neurology, 36*(5), 424–426.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV tests of achievement*. Rolling Meadows, IL: Riverside.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV tests of cognitive abilities. *Journal of Psychoeducational Assessment, 33*(4), 381–390.

Shaffer, J. A. (2020). Forethought and intelligence: How conscientiousness, future planning, and general mental ability predict net worth. *Personality and Individual Differences, 159*, Article 109853.

Shahbazian, S., & Zahedi, M. (2006). The role of observables and non-observables in chemistry: A critique of chemical language. *Foundations of Chemistry, 8*(1), 37–52.

Sheng, J., Wang, X., Jin, B., Yan, J., Li, W., Chang, T. H., … Zha, H. (2022). Learning structured communication for multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems, 36*(2), 50.

Siddiq, M. L., Santos, J., Tanvir, R. H., Ulfat, N., Rifat, F. A., & Lopes, V. C. (2023). *Exploring the effectiveness of large language models in generating unit tests*. arXiv preprint arXiv:2305.00418.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika, 71*, 451–455.

Singh, B., Kumar, R., & Singh, V. P. (2022). Reinforcement learning in robotic applications: A comprehensive survey. *Artificial Intelligence Review*, 1–46.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105.

Slaney, K. L., & Racin, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*, 4–12.

Smith, D. I., & Kirkham, R. W. (1982). Relationship between intelligence and driving record. *Accident Analysis & Prevention, 14*(6), 439–442.

Smith, E. T., Bartlett, J. C., Krawczyk, D. C., & Basak, C. (2021). Are the advantages of chess expertise on visuo-spatial working-memory capacity domain specific or domain general? *Memory & Cognition, 49*(8), 1600–1616.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49*(3), 607.

Sternberg, R. J. (2011). Intelligence and giftedness. In R. J. Sternberg, L. Jarvin, & E. L. Grigorenko (Eds.), *Explorations in giftedness* (pp. 54–81). Cambridge University Press.

Sternberg, R. J. (2012). Intelligence in its cultural context. In M. Gelfand, C.-Y. Chiu, & Y.-Y. Hong (Eds.)*, Vol. 2. Advances in cultures and psychology* (pp. 205–248). New York, NY: Oxford University Press.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*(1), 37–55.

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence?* Contemporary viewpoints on its nature and definition: Ablex Publishing.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401–426.

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139*(4), 841–869.

Sweta, A., Pratap, D., & Haldar, C. (2019). Stunning facts of bird migration: Mini-review. *Journal of Endocrinology and Reproduction, 23*(1), 44–47.

Thomson, G. A. (1919). On the cause of hierarchical order among correlation coefficients. *Proceedings of the Royal Society A, 95*, 400–408.

Thurn, C., Nussbaumer, D., Schumacher, R., & Stern, E. (2022). The role of prior knowledge and intelligence in gaining from a training on proportional reasoning. *Journal of Intelligence, 10*(2), 31.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass, 8*(7), 328–341.

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4*(4), 517–545.

Turner, R., & Adams, R. J. (2007). The programme for international student assessment: An overview. *Journal of Applied Measurement, 8*(3), 237–248.

Vaci, N., Edelsbrunner, P., Stern, E., Neubauer, A., Bilalić, M., & Grabner, R. H. (2019). The joint influence of intelligence and practice on skill development throughout the life span. *Proceedings of the National Academy of Sciences, 116*(37), 18363–18369.

Vedula, S. S., Ghazi, A., Collins, J. W., Pugh, C., Stefanidis, D., Meireles, O., … Sachdeva, A. K. (2022). Artificial intelligence methods and artificial intelligence-enabled metrics for surgical education: A multidisciplinary consensus. *Journal of the American College of Surgeons, 234*(6), 1181–1192.

Verhallen, R. J., & Mollon, J. D. (2016). A new Mooney test. *Behavior Research Methods, 48*(4), 1546–1559.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504–511.

Walker, D. L., Palermo, R., Callis, Z., & Gignac, G. E. (2023). The association between intelligence and face processing abilities: A conceptual and meta-analytic review. *Intelligence, 96*, Article 101718.

Wang, L. (2020, October). Analysis of factors affecting computer data processing speed. In *, Vol. 1648, No. 2. Journal of physics: Conference series* (p. 022136). IOP Publishing.

Wang, P. (2006). *Rigid flexibility: The logic of intelligence. 2006*. Springer.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37.

Wang, P. (2022, April). Intelligence: From definition to design. In *International workshop on self-supervised learning* (pp. 35–47). PMLR.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Williams & Wilkins.

Wechsler, D. (2008). *Wechsler adult intelligence scale—Fourth edition: Technical and interpretive manual*. Pearson Assessment.

Welty, C., Paritosh, P., & Aroyo, L. (2019). *Metrology for AI: From benchmarks to instruments*. arXiv preprint arXiv:1911.01875.

Woollett, K., Spiers, H. J., & Maguire, E. A. (2009). Talent in the taxi: A model system for exploring expertise. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1522), 1407–1416.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245.

Yu, N., Weber, C., & Hu, X. (2019). Learning sparse hidden states in long short-term memory. In *Artificial neural networks and machine learning–ICANN 2019: Deep learning: 28th international conference on artificial neural networks, Munich, Germany, September 17–19, 2019, proceedings, part II 28* (pp. 288–298). Springer International Publishing.

Zagorsky, J. L. (2007). Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence, 35*(5), 489–501.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *HellaSwag: Can a machine really finish your sentence?*. arXiv preprint arXiv:1905.07830.

Zhao, J., Zhao, W., Deng, B., Wang, Z., Zhang, F., Zheng, W., … Burke, A. F. (2023). Autonomous driving system: A comprehensive survey. *Expert Systems with Applications, 122836*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J. R. (2023). *A survey of large language models*. arXiv preprint arXiv:2303.18223.

Zhuo, T., & Kankanhalli, M. (2020). *Solving Raven's progressive matrices with neural networks*. arXiv preprint arXiv:2002.01646.

Zook, N. A., Davalos, D. B., DeLosh, E. L., & Davis, H. P. (2004). Working memory, inhibition, and fluid intelligence as predictors of performance on Tower of Hanoi and London tasks. *Brain and Cognition, 56*(3), 286–292.

van der Maas, H. L., & Kan, K. J. (2016). Comment on "Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence" by Gilles E. Gignac (2016). *Intelligence, 57*, 81–83.