How much like us do we want AIs to be?

Considering both intelligence and psychology when thinking about "human-like" AI

Eric Dietrich
Philosophy Department, Binghamton University,
Binghamton, New York, 13902, USA
dietrich@binghamton.edu

Chris Fields
23 Rue des Lavandières,
Caunes Minervois, 11160, FRANCE
ORCID: 0000-0002-4812-0744
fieldsres@gmail.com

John P. Sullins
Philosophy Department, Sonoma State University,
Rohnert Park, California, 94928, USA
john.sullins@sonoma.edu

Bram Van Heuveln
Cognitive Science Department, Rensselaer Polytechnic Institute,
Troy, New York, 12180, USA
heuveb@rpi.edu

Robin Zebrowski*
Cognitive Science Department, Beloit College,
Beloit, Wisconsin, 53511, USA
zebrowsr@beloit.edu

*Corresponding author: Please send all correspondence to zebrowsr@beloit.edu

Abstract:

Replicating or exceeding human intelligence, not just in particular domains but in general, has always been a major goal of Artificial Intelligence (AI). We argue here that "human intelligence" is not only ill-defined, but often conflated with broader aspects of human psychology. Standard arguments for replicating it are morally unacceptable. We then suggest a reframing: that the proper goal of AI is not to replicate humans, but to complement them by creating diverse intelligences capable of collaborating with humans. This goal renders issues of theory of mind, empathy, and caring, or community engagement, central to AI. It also challenges AI to better understand the circumstances in which human intelligence, including human moral intelligence, fails.

Keywords: Artificial General Intelligence, Cooperative AI, Deep learning systems, Ethics, Human psychology, Human variation, Theory of mind

Statements and Declarations: The authors declare no conflicts of interest relevant to the current work. This work received no external funding.

1. Introduction: Human-like intelligence as a goal of AI

While some have argued that it is not a major or even a real goal of AI research (e.g. Whitby, 2003), replicating or exceeding human-level (or "human-like") intelligence in artificial systems has always been both explicitly-stated and highly publicized as AI's primary objective. Turing's (1950) imitation game is, after all, about imitating a human. The title of Newell and Simon's report on the General Problem Solver (GPS, 1961) reads, "GPS, A program that simulates human thought." Feigenbaum and Feldman (1963) chose the title *Computers and Thought* at a time when the only recognized exemplars of thought, at least in academia, were humans. In their Turing Award lecture, Newell and Simon (1976) explicitly characterize the "empirical research" of AI as understanding human intelligence by replicating it. The goal of the CYC project (Lenat, Prakash and Shepherd, 1986) is, similarly, to replicate human common-sense reasoning. Nilsson (2006) is perhaps the most explicit, characterizing the goal of AI as building machines that can do every job humans are paid to do. Prominent projects including the ACT-R model (Anderson, 1993), SNePS (Shapiro, 2000), and the Soar cognitive architecture (Laird, 2012) were not just intended to take us further down the road toward building a human-level intelligence, but were explicitly characterized as actually replicating at least some aspects of human-level intelligence. These were, moreover, some of the most important, visible, and well-funded projects in the history of AI. So, building a human-level intelligence has, as a matter of historical record, been a serious and perfectly explicit goal of AI from the start.[1] It is also a current goal, despite the arguments of detractors like Whitby and the efforts of some prominent AI researchers, e.g. Brooks (1991), to set alternative goals.

From Lucas (1961) to Penrose (1989) and beyond, the "strong AI" goal of replicating human-like intelligence, or perhaps to exceed it by creating

artificial general intelligence (AGI) (Goertzel, 2014), has drawn the ire of philosophers, scientists in other disciplines, and much of the public (see Dietrich et al., 2021 for an extended discussion). Part of the furor has always been about replicating human consciousness, not just human intelligence (HI) as a set of abstracted problem-solving capabilities. The relationship between consciousness and intelligence, and in particular, whether intelligence requires consciousness, remains highly controversial. While we will refrain from discussing this issue in detail, we reject *a priori* claims that AIs either must be or (much more commonly) cannot be conscious in favor of a position that acknowledges the moral hazard of AIs possibly turning out to be conscious, either now or at some time in the future.

Within the AI community itself, the perceived failures of grand, monolithic projects such as those referenced above have driven successive waves of architectural innovation, from the "second wave" of artificial neural networks (ANNs) (Rumelhart and McClelland, 1986; Smolensky, 1988), to embodied robotics (Brooks, 1991) and the broader embodied, embedded, enactive, extended, and affective (4EA) cognition movement (Anderson, 2003, Froese and Ziemke, 2009), to deep learning (DL) (LeCun, Bengio and Hinton, 2015). Despite occasional claims to the contrary in the popular press,[2] it is well-known that, so far, all endeavors to achieve human-like AI have failed. Both the size and complexity of the goal of AGI have been seriously and consistently underestimated. Even as DL systems have achieved astonishing practical successes in narrow but important domains, such as those of AlphaFold (Senior et al., 2020; Jumper et al., 2021) and AlphaCode (Li et al., 2022), many now call for re-thinking both the idea that scaling alone will produce an open-domain AGI and the idea of single, monolithic AGIs as sought by the GPS, CYC, or SOAR projects (Dafoe et al., 2020; Marcus, 2020; Brynjolfsson, 2022, Friston et al., 2022).

What, however, is “human-like intelligence”? Is it the same as *human* intelligence (HI) in some individual or collective sense? Is it *general* intelligence (GI)? James (1890) characterized intelligence in terms of adaptability or robustness: the ability to solve some given problem by a variety of means. Laland and Seed (2021) list five prominent aspects of human intelligence – retrospective and prospective memory, tool invention and use, multi-domain problem solving, social cognition, and language – but also point out that each of these appears in some form in many other species. Meloni et al. (2019) emphasize that human intelligence cannot be understood in abstraction from human sensory and motor capabilities and ecological embedding; humans – indeed all organisms – are 4EA systems. Since Damasio (1994) brought it to wide attention, it has become widely accepted that motivation is an integral component of intelligence, but this motivational component is transferred out of the AI system and into the user/trainer even in advanced DL systems. The study of intrinsic motivation and creativity in humans has been more closely coupled to developmental robotics than to the pursuit of AGI (Kaplan and Oudeyer, 2007; Oudeyer, Baranes and Kaplan, 2013; Cangelosi and Schlesinger, 2015); indeed, social neuroscience as a whole has been more closely coupled to social robotics than to the pursuit of AGI. Intrinsic motivation is central to the proposal of Friston et al. (2022) to base AI research on the free energy principle (FEP) (Friston, 2010; 2013): the FEP is a principle of uncertainty minimization, which it characterizes as the primary motivation of all systems, living or not, that interact with an external environment (Ramstead et al., 2022).

Concepts such as intrinsic motivation – or of affect generally – are generally considered to be psychological concepts. How much of human *psychology* is built into the concept of human *intelligence*? How much of human psychology needs to be included in “human-like” intelligence? Is HI, under some suitable definition, the same as GI? How “general” does GI need to be? What kind of psychology, beyond some general motivational

mechanism, is needed for GI? These questions are seldom addressed explicitly, and when they are, the answers tend to be vague and contentious. Human intelligence is often just defined by pointing: it is whatever (most) humans have. General intelligence is often defined in terms of computability. However, "capable of computing any Turing-computable function, up to resource constraints" clearly will not do for AI's purposes, as then a laptop would count as an AGI.[3] The Turing test will not do, as laughable claims to have "passed it" demonstrate; indeed the Turing test was probably never intended to be criterial for intelligence (again see Dietrich et al., 2021)[4]. Specific abilities like chess-playing or solving undergraduate physics problems will not do, because, obviously, they are not general. General claims for attributes like creativity or flexibility or robustness are, in the absence of a characterized embodiment and task environment, only pointers and are scarcely better defined than HI itself. Hence, while it is reasonably clear that no human-like AIs or AGIs yet exist, it is less clear why. On the one hand, we – not just the AI community but the entire mythopoetic tradition of artificial humans (see Brynjolfsson, 2022 for numerous examples) – grossly underestimated how hard the problem of replicating HI is. We do not know how HI works, either at the algorithmic level or at the level of the neural (and more generally, bodily) implementation (Melloni et al., 2019). We do not, for example, know what concepts are, what categorization is, what semantic relevance is, and on and on (Margolis and Laurence, 1999; Dietrich et al., 2021). On the other, we do not know how "human-like" something needs be to have GI. In particular, we do not know how human-like the psychology of a GI needs to be. In this conceptual vacuum, failures can be recognized, but criteria for success are not just ill-defined from an engineering perspective, but rather deeply and still philosophically controversial.

Whitby (2003) is, moreover, not alone in claiming that even if artificial HI (AHI) or AGI is or has been a primary goal of AI — the other primary goal

being technological utility — it is a mistaken goal. Brynjolfsson (2022) has recently argued that the proper goal of AI, for economic and moral as well as scientific and technological reasons, is not to duplicate HI but to exceed it in specific, targeted areas. We support this critique for reasons outlined already in Dietrich et al. (2021, 2022). To be blunt, why re-invent the wheel? Humans are not in short supply, so why try to replicate HI? As for GI, it is not clear that humans have it. There are many problems humans appear unable to solve, despite generations of trying; many of these are in the ethical, social, economic, and political spheres where embodiment and motivation play at least as large a role as "thinking" in the traditional sense. Hence if human psychology is intrinsic to HI, it is not clear that replicating HI is even on the path to AGI. And, again, much AI research fails to distinguish between HI and AGI, conflating the two and complicating the discussion.

In what follows, we will first expand on the above blunt critique, arguing in Sect. 2 that replicating HIs with human-like psychology is deeply immoral, and in Sect. 3 that such systems would not lie on the path to AGI. We will then, in the remainder of the paper, argue for reframing the question. We start with the fact that humans – indeed all organisms, even bacteria (Stal, 2012) – have "extended minds" (Clark and Chalmers, 1998) in the straightforward sense of employing stigmergic memories (Fields, Glazebrook and Levin, 2021), i.e. memories written on the environment, such as pheromone trails, grocery lists, or any messages passed to another agent whose memory can be relied on in the future. Humans and many other organisms also employ parts of the environment as tools to solve novel problems, and humans (and some other organisms) design and build tools when found objects are insufficient (Visalberghi et al., 2017). One can, indeed, regard AI systems as such tools. Human problem solving is, moreover, typically a *collective* endeavor; humans use *each other's* intelligence when their own is insufficient by itself (De Jaegher and Di

Paolo, 2007; De Jaegher and Froese, 2009; Dubova, Galesic and Goldstone, 2022). Humans (almost) always operate, in other words, with a composite (HI, OI), where OI is some "other intelligence" that may be quite minimal (a piece of paper, a way-marker) or quite sophisticated (a smartphone, a laptop, one or more colleagues). Humans, in other words, almost always operate with greatly extended minds. The proper goal of AI is, in this case, not to replicate HI but to maximize (HI, OI), as indeed Dafoe et al. (2020), Brynjolfsson (2022), and Friston et al. (2022) have also argued from their various perspectives.[5] AI is, therefore, properly a composite discipline, one that seeks both to understand HI well enough to characterize its weaknesses, and to develop OIs that compensate for these weaknesses. AI is, in this sense, continuous with human – indeed hominin – engineering practice since the invention of the hand axe. It is *dis*continuous with this tradition, however, in attempting to build systems that are not just tools, but in an important sense *colleagues* (Fields, 1987). As colleagues, OIs need not just intelligence but psychologies. We argue, as an answer to our title's question, that we want AIs to be psychologically like (most of) us in a particular and generally neglected way: AIs need to be, and be motivated to be, *team players*. The fact that humans perform best as team players has been largely neglected until the past two decades; see Graesser (2018) for review. In particular, AIs need to be good *diverse* team players, capable of working with both humans and other artifacts, regardless of capabilities or architectures of the latter. AIs must be smart enough to know when they cannot solve a problem alone, and smart enough to ask for help. They must, moreover, have good enough theory-of-mind (ToM) (Frith and Frith, 2005; Carlson, Koenig and Harms, 2013) capabilities to ask the right kind of

system for help. They need, in particular, theories of *our* minds and include our cognitive strengths and weaknesses. They also need the capability to design and build a system they need to help them, just as humans (sometimes) do. Such AIs will work not *for* us but *with* us, or perhaps we will work with them. We conclude that any feasible AGI will be a composite human (or humanity)-in-the-loop system that, if it is to be of value, will be capable of solving problems that neither humans, nor current (HI, OI) systems, can solve alone.[6]

2 Because It Is There

Before proceeding to offer and critique potential definitions of HI and GI, it is useful to ask: Why strive to build artificial human-level intelligences (AHIs) at all? Why would an AHI ever have been a goal of AI? We can suggest four kinds of reasons. First, there is the mythopoetic reason: "Because we will have created our equals in the universe — we will no longer be alone." There is a more basic, curiosity-driven reason: "Because it is an obvious challenge and goal." One is reminded of George Mallory, who when asked, in 1923, why he wanted to climb Mount Everest, responded with: "Because it is there." There is the scientific reason: "Because to build a machine as smart as we would tell us a lot about how it is that we are smart." There is, finally, an engineering or overtly economic reason: "To do work that is too dangerous, too expensive, or otherwise inefficient or undesirable for humans to do." The influence of the last three of these reasons on AI research is well documented; we suspect that the first has exerted a more subtle and implicit influence from Turing onward.

None of these, however, are sufficiently good reasons. As soon as a human-like psychology is included in the idea of an AHI, they all raise immediate moral questions.

The primary problem with the scientific reason is that it gets the flow of information backwards: it assumes that we can replicate an extraordinarily-complex system without knowing how it works. One can obviously build a fire without knowing any chemistry (but not without knowing that sticks burn and rocks do not), but one cannot just happen to build a human-level intelligence and then reverse engineer it to find out how we work. This becomes obvious as soon as psychology is included in the mix, so it is useful to examine why it did not appear obvious in the early days of AI. AI was conceived as a discipline in the aftermath of World War II when behaviorism enjoyed its maximum influence. For a behaviorist, a functional specification of desired behavior is sufficient; indeed a functional specification is all that is relevant, even in principle. To scientists brought up on the idea that thinking – or at least the best thinking – was logical, the idea that building a machine that could prove theorems (Newell and Simon, 1956) would be building an AGI (and hence automatically an AHI) might seem natural. The problem with the scientific reason is that this psychologically-naive way of thinking has persisted and has exerted enormous influence on the culture and pedagogy of the field. The discipline-wide pivot toward ANNs in the 1980s (see Rumelhart and McClelland, 1986; Smolensky, 1988 and other foundational papers), for example, did not incorporate 1980s cognitive neuroscience, but rather simplified models of neurons based conceptually on those of McCulloch and Pitts (1943). There is still no convincing evidence that biological neuronal networks employ error back-propagation (as widely used in deep learning; see e.g. Wright, 2022), though see Millidge et al (2022) for evidence that predictive coding systems may approximate back-propagation. Moreover, ANNs bear only the most abstract resemblance to “neuromorphic” computing systems that aim to functionally replicate neurons (see Schuman et al., 2017; Tang et al., 2019 for recent reviews). Hence even if they are very successful in solving problems, as e.g. AlphaFold (Senior et al., 2020; Jumper et al., 2021) undoubtedly is, current deep learning (DL) systems

cannot be expected to tell us anything of interest about human cognition. Emulation is not explanation. Indeed both Marcus (2020) and Friston et al. (2022) make this point in their respective critiques of the current state of AI; Melloni et al. (2019) do the same from the perspective of neuroscience.

An obvious potential counterexample to the above is developmental robotics. Here, however, the flow of motivating theory is in the human (or animal) to AI direction: the goal is to build robots that undergo developmental processes, including motivational development and "learning how to learn," that we largely understand from prior work with humans and other animals (see e.g. Cangelosi and Schlesinger, 2015). Experimental platforms like the iCub are just that: experimental platforms. They are not, and are not intended to be, artificial children.[7]

Demonstrable success in building an AHI with an even minimally human-like psychology for scientific reasons would clearly raise ethical issues; indeed Institutional Review Boards (IRBs) could be expected to step in well before success was demonstrable.[8] The ability to register stress is widely recognized as foundational to even the most basal psychologies, being evident even in bacteria (e.g. Fields, Glazebrook and Levin, 2021). In the language of the Free Energy Principle (FEP), stress is uncertainty, and hence the fundamental motivator of cognition (Friston et al, 2022). An AHI with sufficient psychology to have human-like intelligence can, therefore, be expected to register stress, and even the counterfactual stress – stress induced by the imagination of future events – that constitutes suffering. Such a system would be one for which the notion of well-being is relevant.

[7] Though this statement summarizes current practice, see e.g. Moravec, H. (1990) *Mind Children*. Cambridge, MA, Harvard University Press for a more radical, posthumanist projection.

These issues come into even greater focus when we consider the curiosity-driven (or to be less charitable, boredom-avoidance) reason to build AHIs with human-like psychology. This kind of reason works well for climbing mountains and other risky challenges, *but the risk has to be to oneself*. This would obviously not be true for a machine with human-level intelligence. In this case, we would be forcing risk on to something else: the machine. And there lies the problem. A machine with human-like psychology would be able to suffer (from all kinds of things, just like we do), and would probably fear death (just like we do). So, building one because we need a challenge is immoral. Compare: curiosity, boredom-avoidance or needing a challenge is an inappropriate reason – though it often the actual reason, stated or otherwise – when deciding to get pregnant and have a human baby or when deciding to get a dog. Being bored or loving a challenge is not a morally acceptable reason to take on being responsible for another life. Indeed, its immorality is obvious.

What about the mythopoetic reason? It seems high-minded, but it, too, suffers from immorality. This is obvious when it is noted that the mythopoetic reason is about *us*. Building a human-level intelligence puts a feather in *our* caps. But what does it do for the intelligence thus created, *for the other being*? As with the "because it is there" reason, it places such an intelligence at serious risk of being a curiosity, an exhibit, a pet, or a slave of some sort. The goal of "aligning" AI with human values (Markus and Davis, 2019; Stray, 2020; Han et al., 2022), for example, explicitly renders AIs subservient to human goals and desires, including our human desire for control. Indeed, it is here that the mythopoetic and engineering reasons overlap: building an artificial human-like worker is in fact building a slave, as Čapek's *R.U.R.* (1920/2001) makes perfectly clear and Bryson (2010), who rejects our moral-hazard position in favor of an *a priori* assumption that AIs will not be conscious, explicitly advocates. McEwan's *Machines Like Me* (2019) provides a recent counterpoint: the "artificial

humans" are both smarter and more moral than we are, and commit suicide out of despair. It is difficult not to think of this when contemplating the sex robots, war robots, or nurse robots conceived of as industrial products (see Sullins, 2012, 2013a, 2013b, 2014, 2017). These carry on a long tradition: slavery is as old as socially stratified civilization and continues robustly today. According to the United Nations International Labor Organization, 40.3 million people are now enslaved (see, Hodal, 2019; ILO Report, 2017). This number does not include all the dogs, cats, horses, and agricultural animals that live horrible lives due to human wants and interests. The history of slavery and other forms of exploitation suggests that once we start building *intelligent* machines, thinking of them as slaves, as unpaid servants, etc., will come naturally.[9] On any position that acknowledges moral hazard, immorality of a monstrous size would then ensue.

Hence, we return to the notion of reinventing the wheel, noting that in the case of AHIs, the reinvention is not only pointless but cruel. This, of course, is *why* AHIs have mythopoetic status, as Hollywood continually and tiresomely reminds us. Not noticing this can only be considered a massive failure of science-society communication on the part of the AI community.[10]

3. Human Intelligence Is Not General Intelligence

[10] We have focused here on moral consequences for the AI systems themselves. There are obviously also moral consequences for us. One could argue that the principles outlined in the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems have thus far been recognized mainly by their violation. Creating AIs that replicated the worst of human morality, for example, would obviously be grossly immoral; see e.g. Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell (2021) On the dangers of stochastic parrots: Can language models be too big? In: Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pp; https://doi.org/10.1145/3442188.3445922 or Birhane, A. and J. van Dijk (2020) Robot rights?: Let's talk about human welfare instead. AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 2020, pp. 207–213; https://doi.org/10.1145/3375627.3375855.

We now turn, as a second preliminary, to the question of how, precisely, to characterize HI and GI. Abstracting from psychology, it is relatively straightforward to characterize an ideal GI: an ideal GI is a system that can solve any problem that it can recognize as a problem, up to resource and computability constraints (but see Ji et al., 2021 for an argument that Turing computability can be exceeded). In the increasingly-popular language of the FEP (Friston, 2010; 2013), an ideal GI is a system in the limit as time-averaged Bayesian surprisal (i.e. net prediction error) approaches zero – any momentary upticks in Bayesian surprisal could be quickly dealt with by problem-solving. This characterization has a trivial special case, the case in which the "GI" inhabits an environment in which interesting problems – problems that are unanticipated and require multi-step problem solving –never arise. Interesting GIs, and hence interesting AGIs, would be systems somewhere in the vicinity of this limit of low average surprisal that inhabit interesting, problem-rich environments. "GI" in practice is, therefore, a continuum, not a bright line. We will, in what follows, only be interested in GIs that can recognize (and hence solve, up to constraints) at least all the problems that we can recognize.

Human intelligence is often held up as an exemplar of General Intelligence, though with the proviso that HI may also be less than some feasible GI (e.g. Goertzel, 2014). Setting aside gods, the Western philosophical tradition since Descartes tends to regard HI as the *only* extant exemplar of GI. Human intelligence does, indeed, exhibit significant generality. Humans can abstract, categorize, deduce, draw conclusions, dream up counterexamples, explain, infer, intuit, reason, and combine all of these very quickly into a thought. Human languages are syntactically complete. Humans are good enough at computation to have invented the theory of computation, including its metatheory. Humans are claimed by some to be more powerful than Turing machines, though see Dietrich et al.

(2022, §2.1) for a thorough criticism of Lucas' (1961; 1996) classic version of this argument.

Is, however, HI general in the sense intended by GI? Can humans solve all the problems, up to resource and computability constraints, that they can recognize as problems? There are clearly human-niche problems that humans have not yet solved, many of them quite serious and long-standing. Are these problems – e.g. the problems of peaceful co-existence, population control, and environmental degradation – solvable by humans, even in principle? What does "in principle" mean here?

To ask what "in principle" means is to raise the problem of how the social and affective components of human cognition – in short, the components that render us 4EA systems – both enable and constrain problem-solving capability. Hence it requires including human psychology, particularly motivational psychology, in our notion of HI. Asking about actual human problem-solving capability also raises the question of variation, not just of some testable measure of problem-solving ability along one or more dimensions but of core capabilities such as imagination, intrinsic motivation, memory, attention management, or event-oriented "mental time travel" in one or both directions. Proponents of AHI are, we can assume, always intending to replicate "the best" examples of HI, not just in game-playing challenges but across all applications of intelligence. Is, however, the idea of "the best" HI even coherent? It is implausible that one human could be "the best" in every problem-solving domain in which humans are capable, including not just theoretical and practical but also *moral* problem solving. Hence "the best" must be an abstraction, an idealized combination of "bests" in different domains. Is this, however, a coherent idealization? Does it make sense, even in principle, to assume such an idealization for *human* intelligence?

This is not a question for speculation, but a question for experimental psychology and neuroscience. It is a question of means along axes of variation, and whether they overlap, either in fact or in some plausible ideal. Consider the Big 5 personality dimensions (Digman, 1990): openness, conscientiousness, extraversion, agreeableness, and neuroticism. A "typical personality" would be someone within some fixed distance (e.g. 1 sigma) of the mean on each of these dimensions. If the means of the distributions are sufficiently separated in the population, however, no such "typical personality" would exist. It is not clear, moreover, that a mean value on any of these dimensions is optimal, or even whether an "optimal" personality can be defined in any context-independent way.

No dimensional analysis of human core cognitive capabilities with the level of acceptance of the Big 5 exists. However, analyses of variation in everyday, uninstructed experience (Heavey and Hurlburt, 2008), autobiographical memory (Fan et al., 2022), mental imagery (Milton et al., 2021), and general cognitive functions (Kanai and Rees, 2011) all suggest the existence of broad distributions across the human population. The idea of broad variation is reinforced by studies of variation along spectra associated at their extremes with autism and psychosis (e.g. Crespi and Badcock, 2008) and the correlation of such variants with default behavior and career choices (see Fields, 2011, for review), and by studies of variation along spectra associated with empathy and sociopathy, and hence with moral capability (Sapolsky, 2017). Such studies reinforce the everyday observation that humans who are very good in one domain (science, art, persuasion, etc) may be very poor in others (social relations, decision making, empathic caring, etc). They suggest that even the idea of a "neurotypical" human may be of little use outside of the narrow, clinical context in which it originated.[11] If this is the case, however, "human intelligence" may not be coherently definable for individual humans, groups of humans, or even idealizations of (groups of) humans. It may be at best an

informal notion, a vague summary of a list of general capabilities – e.g. the five listed by Laland and Seed (2021) – that characterize most humans to some extent or other. Even setting issues of general psychology aside, such a weak notion of HI can have only minimal relevance to any useful idea of AGI. Indeed, in the space of all possible intelligences – something we may not even be equipped to conceive of in any detail – the component spanned by all possible varieties of human intelligence, including moral intelligence, may be small.

One can also ask what happens if HI, under even a vague definition, is nudged *at the population level* in the direction of some hypothesized GI. This introduces a social psychology question: what is the range of social behavior that can reasonably be considered *human* social behavior? This question is particularly pressing in the moral sphere, where humans, particularly groups of humans, exhibit undesirable characteristics with very long evolutionary histories (e.g. Wrangham and Peterson, 1997; Sapolsky, 2017). Han et al. (2022), for example, speak of the "moral progress" of humans as a species or population as an essential component of AI alignment. How soon would such a "nudged" intelligence at the population level, in the moral or any other sphere, cease to count as a *human* intelligence? Would a "human" society that no longer elevated immoral individuals to positions of leadership, and refused to follow – refused to *enthusiastically* follow – the commands of such individuals still be recognizably human? A society in which such a change was implemented would be a historical novelty. Humans have a great tolerance for the prodigal, but stretching human psychology too far towards an ideal can generate an "uncanny valley" (e.g. Saygin et al., 2012) on the suprahuman side.[12] It is not clear that we would be capable of regarding a suprahuman GI – even one that just maximized known human capabilities simultaneously – as fully human at the individual level; here again McEwan (2019) is a

useful study of this question. Recognizing suprahuman capabilities at the broad social level as “still human” may prove even more difficult.

4. A New Goal for AI: Composite Intelligence

The extraordinary diversity of cognitive and affective capabilities across the human population is increasingly seen as selectively advantageous at the group level and hence as maintained over deep evolutionary time (Nettle, 2006; Holmes and Patrick, 2018). This has an obvious correlate: optimal problem solving will typically be achieved by groups, not individuals (Graesser et al., 2018; Dubova, Galesic and Goldstone, 2022). The parallel between this social-scale phenomenon and the requirements for cooperation between phenotypically-diverse individuals in the construction of a multicellular organism (Strassmann and Queller, 2010) are similarly obvious, leading to the proposal that *all intelligence is fundamentally composite or collective* (Levin, 2021; 2022, Fields and Levin, 2022). We therefore suggest that “human intelligence” is properly thought of as a composite (HI, OI), where OI is some “other intelligence” that may be human but may be as simple as a physical system (a notebook, a laptop) supporting stigmergic memory. Conceiving of human intelligence as composite in this way, we will argue, reframes the goal of AI away from replacement (of HI) and towards augmentation (of the composite (HI, OI)), as suggested by both Brynjolfsson (2022) and Friston et al. (2022). This new goal is already being pursued in the context of human-robot collaboration (Vysocky and Novak, 2016; Franklin et al., 2020), but has yet to be taken up broadly within mainstream AI.

If the proper goal of AI is not to replicate HI (whatever that would mean), but rather, as we suggest, to maximally complement communities of diverse HIs, then AI is free to pursue one of its most distinctive

characteristics: its *difference* from HI. We elaborate on this in the two sections below. We then turn to a critical way in which successful AIs need to be like humans: they too need to be team players. In particular, they need to be capable of "playing" on diverse teams, e.g. teams including both humans and other, very different AIs. Being a capable diverse-team player requires capabilities that AIs currently do not have, or have only rudimentary versions of (compare, e.g., Kraus, 1997 and Dafoe et al., 2020 on cooperative problem-solving capabilities). It requires, in particular, both robust models of the self and others and a capacity to care about the goals of both oneself and others. It requires, in other words, both *theory of mind* and *empathy* (Doctor et al., 2022).[13] Team-capable AIs need, in particular, the abilities to recognize when a problem they are trying to solve is too hard, to determine what "OI" they need to approach for assistance, and to locate, teach, design and build, or otherwise find that OI. They need to be "like us" in having an ability to creatively supplement their own intelligence. When employed in procedural, technical, or abstract domains such as law, engineering, or science, they also, clearly, need to be like us in the ability to explain what they are doing and why, and hence to explain why they need help from some OI, artificial or human. Diverse-team AI, in other words, requires explainable AI (XAI) (Arrieta et al., 2020; Samek et al., 2021). Here again, ToM skills are critical (Taylor and Taylor, 2020).

4.1. Minding the gaps: understanding where HI fails

Computers were first developed as fast, accurate calculators. This responded to a specific need: although humans (mainly women) were employed as "computers" until well into the 1960s, humans are not, aside from a few spectacular exceptions, fast, accurate calculators. Robots were first developed as tireless, reliable, accurate performers of repetitive mechanical tasks. This, too, responded to a specific need: humans have

been employed (or forced as slaves) to perform repetitive mechanical tasks since the invention of agriculture, but humans are not tireless, reliable, accurate performers of such tasks. Successful applied AI, in general, does not replace humans in unnecessary tasks, or in tasks that humans are good at. Successful applied AI replaces humans in necessary tasks that humans are relatively bad at.[14] For example, autonomous-vehicle control systems will eventually replace human drivers because humans are, by and large, bad drivers – humans are often distracted, discourteous, and notoriously disrespectful of rules. While autonomous systems do not yet have, for example, sufficient pattern-recognition ability to detect hazards humans can detect (e.g. Nyholm, 2020), these abilities can rationally be expected to improve. Humans, on the other hand, cannot rationally be expected to become less distracted, more courteous, and more respectful of the rules of the road than they now are. The future replacement of human by AI drivers is controversial, however, not just because bad driving is still lucrative, but because bad driving is still enjoyable.

Reframing AI as maximizing the capability of (HI, OI) systems transfers the need to understand where and how HI fails – or where and how HI is nonoptimal – from outside the purview of AI to centrally within it. Systems that actively monitor the attention of human users in critical settings such as the cockpit provide an example (Lutnyk, Rudi and Raubal, 2020). Fortunately, since Tversky and Kahneman (1974) and Simon (1982), the systematic study of human problem-solving failure has become a mainstream component of both cognitive psychology (for reviews, see Masgood, Finegan and Walker, 2004; Benjamin, 2019) and operations research (Endsley, 2012). Compensating for cognitive biases and coping with ubiquitous motivated reasoning constitute major, largely-unrecognized, opportunities for AI. AIs can be successful in these areas precisely to the extent that they are *not* like us.

A particular challenge in this regard, one that bears on the discussion of ToM below, is the deeply-ingrained human resistance to evidence and imperviousness of beliefs to argumentation (Henriques, 2003; Mercier and Sperber, 2011, 2017; Lewandowsky and Oberauer, 2016). AI "assistants" capable of counteracting these tendencies would be playing, in fact, the role of advisors or mentors. Such capabilities are far beyond current AI, which indeed sometimes reinforces existing biases instead of countering them, but are needed if (HI, OI) systems are to approach GI as a goal.

4.2. Maximizing (HI, OI)

As noted earlier, humans have never worked alone. Intelligent problem solving is a social affair. Even great scientists and mathematicians who seemed to have worked alone have "stood on the shoulders of giants." All of human culture, in this regard, serves as a shared stigmergic memory.

It is, therefore, not surprising that AI systems have had their greatest success not by replicating human capabilities, but by offering supra-human capabilities to human teams. AI systems are not alone in this: memory systems such as books offer supra-human capabilities, as do essentially all technological devices. Where AI systems excel is in offering supra-human attention, learning, inference, and problem-recognition capabilities. Aircraft autopilots, for example, are valuable because they have superior attention and faster problem recognition, and can take faster inference-driven corrective action, than (most) human pilots. Autopilots can fail spectacularly, but do so less often than humans do. These are the characteristics looked for in all autonomous-vehicle applications, with on-the-fly learning a bonus.

Learning comes to the fore in deep learning systems, particularly in scientific systems such as AlphaFold (Senior et al., 2020; Jumper et al., 2021). Such systems have now been deployed in many settings and domains. Their performance clearly exceeds that of teams of humans, even teams of humans equipped with expensive apparatus.

While an autopilot functions in some sense as a real-time colleague, a system such as AlphaFold does not. A human operator sets a goal and (effectively) leaves; AlphaFold works to find a solution, and then informs the operator. This is "collaboration" only in a diachronic sense: one collaborator sets the goals (or gives the orders) while the other collaborator does the work. Restricting the human role to goal-setting is reminiscent of standard scientific computing in the pre-interactive, batch-job era. It is reflected in the goal of *fully* autonomous vehicles, perhaps with a batch controller at some distant location. It only works in settings in which the goals can be fully specified in advance.

It is not clear whether this diachronic model, in which humans and AI systems each solve their problem components alone, with minimal communication, is capable of optimizing (HI, OI) capabilities, just as it is not clear that such a diachronic interaction can optimize the performance of human teams. While the broad, overall objectives of a project may be specified in advance, a synchronic model in which humans and AIs work jointly and interactively on each (major) aspect of a problem may be required, especially in cases where creative solutions are needed. Such real-time collaboration may include collaborative identification of intermediate goals and negotiation of intermediate problem-solving strategies. Humans and AIs may sometimes work on separate parts of a problem independently, just as human collaborators do. They may sometimes need to brainstorm, just as human teams do. While diachronic models still require significant advances for optimal performance even in

appropriate domains – AI systems still need better failure or inadequate-generalization detectors and XAI capabilities – synchronic models can be expected to require substantially better ToM (both self- and other-directed) and communication capabilities as discussed below.

A further issue for (HI, OI) problem solving, one that touches on the ethical concerns raised above, is that of power. Humans at present exercise complete control over resources, and can simply turn off the power if they do not like or agree with what an AI colleague is doing. While in cases of conflict this may remain a valuable last resort (here HAL (Clarke, 1968) comes to mind), such lopsided control remains ethically troublesome (again assuming moral hazard as above) in all other situations. Human control of resources has a flip side: the potential for AI control of – and ability to destroy – critical knowledge. Current DL systems already approach this level of control, particularly systems that learn autonomously in an open environment. Hence safe-guards are needed that prevent both humans (by accident or by intentionally "pulling the plug") and AI systems (out of spite, perhaps) from destroying hard-to-acquire or mission-critical data obtained by DL or other automated means. Procedures for resolving conflicts and preventing stalemates will, one can expect, be just as necessary for human —AI problem-solving teams as they are for purely human teams.

4.3. AIs need *umwelten* to be diverse-team players

What is it about the team, the group, or the community that enhances problem solving? One answer is that in addition to knowledge and skills, each participant brings a certain *point of view* to the problem-solving process. Each participant brings to the problem-solving event individual perceptions and interpretations of the problem, the problem's context, and

its consequences. Working out how to accommodate each of these points of view is a key component of solution-finding.

Biology has a technical term available to what we just called a point of view: *umwelt* (von Uexküll, 1957). 'Umwelt' often translated as "life-world" refers to the world experienced by a particular organism, as it is experienced by that organism. Similar organisms living in similar niches will have similar experiences, but the umwelt of each individual is unique to that individual. Radically differing individuals will have radically differing umwelten. All umwelten are unique because the individuals are unique, not just structurally and functionally, but historically and experientially. The idea of an umwelt thus both extends and personalizes the traditional idea of meaning. While it is clear that AI systems need meaning (Froese and Taguchi, 2019), each also needs its own umwelt.

The difficulty of understanding another organism's umwelt underlies Nagel's (1974) famous reflection on the experiences of bats. Understanding the umwelten of other organisms is, however, part of any biologist's job description, just as understanding the umwelten of diverse other people is a crucial requirement for living in human society. It involves not just understanding what another organism can perceive and do, but critically, what another organism is capable of remembering or caring about (Levin, 2021; 2022). As Nagel's work emphasizes, understanding another being's umwelt in this 3rd-person sense is not the same as experiencing it oneself. While (most) humans have the empathetic and imaginative skills to at least approximate another human's experiences, this may not translate even to other mammals, let alone other organisms in general. Hence in practice, "understanding" the umwelt of another is a matter of understanding capabilities.

Artifacts, including current AI systems, are not generally considered to have umwelten, at least in part because they are not generally considered conscious (for extensive discussion, see Dietrich et al., 2021). However, if "umwelt" is read as task environment – a reading quite consistent with its usage in biology – AI systems and even ordinary non-AI computing systems have umwelten. Understanding how AIs can function as members of diverse teams, however, requires understanding their umwelten, including what they detect about their environments, what actions they can take on their environments, and what they can care about. This includes, in particular, what they can detect about, and how they reason about, their coworkers on the team, whether these are humans or other AI systems.

It is often, moreover, assumed about both other organisms and machines that "the environment" is our environment, that they share our umwelt as well as being participants in it. This is, implicitly or sometimes explicitly, an assumption that our human umwelt is "objective" or observer independent. This is, of course, a contradiction in terms: an umwelt is organism- and even individual-specific by definition. Considering the umwelten of other animals, or of plants or even microbes, makes it clear how differently they perceive even the physical world; when the extensive human virtual world is included, the differences are even more stark. The same lessons apply to AIs. Even if an AI system can "see" the same "objects" that we do, we cannot assume that it identifies those objects in the way that we do, that it assigns the same properties to them that we do, or that they have the same meaning or significance to the AI system that they have to us.

Deep learning systems provide a timely example of the need for thinking clearly about the umwelten of artifacts. Informally, we think of the "world" of AlphaFold as comprising protein sequences and structures.

These are, after all our inputs and outputs of interest. AlphaFold, however, knows nothing about proteins; its world is a world of correlations between bit strings, bit strings that it divides only into inputs and outputs. This naiveté about the semantics – effectively, the background knowledge – that we assign to these bits strings is in part an advantage: AlphaFold can "see" patterns that we cannot. It is, of course, also the deep source of the XAI problem. AlphaFold encodes protein sequences in a much higher-dimensional representation than we use, detects relationships in that high-dimensional representation that we do not and perhaps cannot encode in our lower-dimensional representations, and does not have the semantic knowledge needed to describe its representation in our language.

Word-association learners such as GPT-3 (and more recently, ChatGPT) provide a similar example. The world of GPT-3 is not language, and certainly not conversation, though it is often interpreted as such. The world of GPT-3 is a world of correlations between words and phrases, as its easily-revealed lack of semantic knowledge illustrates (e.g. Floridi and Chiriatti, 2020). To claim that GPT-3's evident knowledge of its world gives it insight into our world (as suggested, e.g. by Chalmers, 2020) is simply a mistake (Bender et al., 2021). It confuses GPT-3's umwelt with ours.[15]

The XAI problem for DL systems stems from the fact that we are not DL systems, and so we cannot make sense of DL system training sets – of indeed, machine-learning (ML) training sets in general – in the way that DL systems can. It is exacerbated by the fact that training is (not necessarily phenomenal) experience; identical systems with different training sets cannot be expected to compute the same function. The umwelten of ML systems, especially ML systems that learn autonomously, are unique, just as they are for organisms. Absent a principled theory capable of assigning semantics to arbitrary functions (see Marcianò et al., 2022 for an example

of what such a theory could look like), XAI for DL systems is effectively experimental cognitive psychology, as Taylor and Taylor (2021) suggest.

If AI systems are to become diverse-team players, one of the first requirements that must be addressed is expanding their umwelten to include us, and any other team members with which they are to cooperate. Other-system identification is a common feature of distributed AI systems. In the multi-agent system described by Steels (2001), for example, a language is evolved by a collection of distributed, identified agents; unlike in the case of GPT-3, this language has semantics *for the agents themselves* (at least in some sense). Security-system issues, e.g. trust, are clearly relevant in any such setting, as are representations of other agent's goals and abilities (Dafoe et al., 2020). Here again, the analogy between AI and biology is obvious (Levin, 2021; 2022; Fields and Levin, 2022).

5. Concluding thoughts - whither AGI?

We have argued here that a considerably broader vision than "replicating human-level intelligence'' is needed to approach AGI. As discussed in §2 and §3 respectively, the very idea of AHI is fraught with ethical difficulties, and "human intelligence" may not even be a well-defined target. It is not, in summary, clear that HI is even on the path toward AGI.

One important aspect of HI, however, clearly is on the path to AGI: the ability to participate in diverse-team problem solving. While some of the capabilities for team participation have been developed in the context of distributed AI systems, much work remains to be done. Understanding the experienced worlds – the umwelten – of AI systems and other artifacts will be key to developing the ToM and empathic or caring capabilities needed for effective teamwork, particularly in synchronic settings. The

increasingly-deep analogies, and in cases of hybrid bio-AI systems, explicit overlaps, between biological and AI versions of, and approaches to, these problems can be expected to be increasingly consequential.

From a practical perspective, we are most interested in human-defined problems, human-set high-level goals, and the capabilities of human-in-the-loop teams. As AI systems become increasingly capable, the extent to which humans remain "in charge" of all aspects of problem solving may start to change. We may, for example, develop systems that can recognize problems that we cannot. The XAI problem, in this case, becomes the problem of whether they can explain to us not just what they are doing and why, but even what problem they are working on. Should this ever occur, AI will indeed have taught us something deep about human intelligence.

Conflict of interest: The authors declare no conflicts of interest relevant to the current work.

Citations

Anderson, J. R. 1993. *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Anderson, M. L. 2003. “Embodied cognition: A field Guide.” *Artificial Intelligence* 149: 91-130.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser et al. 2020. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.” *Information Fusion* 58: 82-115.

Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell 2021. ”On the dangers of stochastic parrots: Can language models be too big?” In *Conference on Fairness, Accountability, and Transparency* (FAccT ’21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pp. https://doi.org/10.1145/3442188.3445922

Benjamin, J. D. 2019. “Errors in probabilistic reasoning and judgment biases.” In *Handbook of Behavioral Economics*, Vol. 2, ed. B. D. Bernheim, S. DellaVigna and D. Laibson, pp. 69-186. Amsterdam: North Holland.

Birhane, A. and van Dijk, J. 2020. “Robot rights?: Let's talk about human welfare instead.” *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, February 2020, pp. 207–213. https://doi.org/10.1145/3375627.3375855

Brooks, R. A. 1991. “Intelligence without representation.” *Artificial Intelligence* 47: 139–159.

Brynjolfsson, E. 2022. “The Turing trap: The promise & peril of human-like artificial intelligence.” *Daedalus* 151(2): 272–287.

Bryson, J. J. (2010) “Robots should be slaves.” In: Wilks, Y. (Ed) *Close Encounters with Artificial Companions*, ed Y. Wilks, pp. 63–74. Amsterdam: John Benjamins.

Cangelosi, A. and M. Schlesinger 2015. *Developmental robotics: from babies to robots*. Cambridge: MIT Press.

Čapek, Karel 1920/2001. *R.U.R.* (Translated by Paul Selver and Nigel Playfair) Dover Publications.

Carlson, S. M., M. A. Koenig and M. B. Harms 2013. “Theory of Mind.” *WIRES Cognitive Science* 4: 391-402.

Chalmers, D. 2020. “GPT-3 and General Intelligence.” *Daily Nous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers (assessed 15 June, 2022).

Clark, A. and D. Chalmers 1998. “The extended mind.” *Analysis* 58(1): 7–19.

Clarke, A. C. (1968) *2001: A Space Odyssey.* New York: New American Library.

Crespi, B. and C. Badcock 2008. “Psychosis and autism as diametrical disorders of the social brain.” *Behavioral and Brain Sciences*, 31: 241–320.

Dafoe, A., E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson and T. Graepel 2020. “Open problems in cooperative AI.” *NeurIPS 2020 Cooperative AI Workshop*. Available as arxiv:2012.08630 [cs.AI].

De Jaegher, H. and E. Di Paolo 2007. “Participatory Sense-Making: An enactive approach to social cognition.” *Phenomenology and the Cognitive Sciences* 6(4): 485-507.

De Jaegher, H. and T. Froese 2009. “On the role of social interaction in individual agency.” *Adaptive Behavior* 17(5): 444-460.

Dietrich, E. 2011. “Homo sapiens 2.0: Building the better robots of our nature.” In *Machine Ethics*, ed. M. Anderson and S. Anderson, Cambridge: Cambridge University Press.

Dietrich, E., C. Fields, J. Sullins, B. Van Heuveln and R. Zebrowski 2021. *The Great Philosophical Objections to Artificial Intelligence: The history and legacy of the AI Wars*. London: Bloomsbury Academic Press.

Dietrich, E., C. Fields, J. Sullins, B. Van Heuveln and R. Zebrowski 2022. “The AI wars, 1950-2000, and their consequences.” *Journal of Artificial Intelligence and Consciousness*, 9(1): 127-151.

Digman, J. M. 1990. “Personality structure: Emergence of the five-factor model.” *Annual Review of Psychology*. 41: 417–440.

Doctor, T. O. Witkowski, E. Solomonova, B. Duane and M. Levin 2022. “Biology, Buddhism, and AI: Care as the driver of intelligence.” *Entropy* 24: 710.

Dubova, M., M. Galesic and R. L. Goldstone 2022. “Cognitive science of augmented intelligence.” *Cognitive Science* 46: e13229.

Endsley, M. R. 2012. “Situational awareness.” In: *Handbook of Human Factors and Ergonomics*, 4th Ed., ed. G. Salvendy, pp. 553-568. Hoboken, NJ: John Wiley.

Fan C. L., S. Simpson, H. M. Sokolowski and B. Levine 2022. “Autobiographical memory.” *Oxford Handbook of Human Memory* (in press).

Feigenbaum, E. and J. Feldman 1963. *Computers and Thought.* New York: McGraw-Hill.

Fields, C. 1987. “The computer as tool: A critique of a common view of the role of intelligent artifacts in society.” *Social Epistemology* 1(1): 5–25.

Fields, C. 2011. “From ‘‘Oh, OK’’ to ‘‘Ah, yes’’ to ‘‘Aha!’’: Hyper-systemizing and the rewards of insight.” *Personality and Individual Differences* 50: 1159–1167.

Fields, C., J. F. Glazebrook and M. Levin 2021. “Minimal physicalism as a scale-free substrate for cognition and consciousness.” *Neuroscience of Consciousness* 7(2): niab013.

Fields, C. and M. Levin 2022. “Competency in navigating arbitrary spaces as an invariant for analyzing cognition in diverse embodiments.” *Entropy* 24: 819.

Floridi, L. and M. Chiriatti 2020. "GPT-3: Its nature, scope, limits, and consequences." *Minds and Machines* 30: 681–694.

Franklin, C. S., E. G. Dominguez, J. D. Fryman and M. L. Lewandowski 2020. "Collaborative robotics: New era of human–robot cooperation in the workplace." *Journal of Safety Research* 74: 153-160.

Friston, K. J. 2010. "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11: 127–138.

Friston, K. J. 2013. "Life as we know it." *Journal of the Royal Society Interface* 10:20130475.

Friston, K. J., K. J. D. Ramstead, A. Kiefer et al. 2022 "Designing ecosystems of intelligence from first principles." Preprint arxiv:2212.01354.

Frith, C. and U. Frith 2005. "Theory of mind." *Current Biology* 15(17): R644–R645.

Froese, T. and S. Taguchi 2019. "The problem of meaning in AI and robotics: Still with us after all these years." *Philosophies* 4: 14.

Froese, T. and T. Ziemke 2009. "Enactive artificial intelligence: Investigating the systemic organization of life and mind." *Artificial Intelligence* 173: 466–500.

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

Graesser, A. C. et al. 2018. “Advancing the science of collaborative problem solving.” *Psychological Science in the Public Interest* 19: 59–92.

Han, S., E. Kelly, S. Nikou and E.-O. Svee 2022. “Aligning artificial intelligence with human values: Reflections from a phenomenological perspective.” *AI & Society* 37: 1383–1395.

Heavey, C. L. and R. T. Hurlburt 2008 “The phenomena of inner experience.” *Consciousness and Cognition* 17: 798–810.

Henriques, G. 2003 “The Tree of Knowledge system and the theoretical unification of psychology.” *Review of General Psychology* 7(2): 150–182.

Hodal, Kate 2019. “One in 200 people is a slave. Why?” *The Guardian*. https://www.theguardian.com/news/2019/feb/25/modern-slavery-trafficking-persons-one-in-200

Holmes, A. J. and L. M. Patrick 2018. “The myth of optimality in clinical neuroscience.” *Trends in Cognitive Sciences* 22(3): 241–257.

International Labor Organization Report 2017. *Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. September 19th. Print: 978-92-2-130131-8[ISBN], Web PDF: 978-92-2-130132-5 [ISBN]. https://www.ilo.org/global/publications/books/WCMS_575479/lang--en/index.htm

James, W. 1890. *The Principles of Psychology.* New York: H. Holt and Company.

Ji, Z., A. Natarajan, T. Vidick, J. Wright and H. Yuen 2021. “MIP*=RE.” *Communications of the ACM* 64(11): 131–138.

Jumper, J., R. Evans, A. Pritzel et al. 2021. “Highly accurate protein structure prediction with AlphaFold.” *Nature* 596: 583–589.

Kageki, Norri 2012. "An uncanny mind: Masahiro Mori on the uncanny valley and beyond." *IEEE Spectrum*. New York City: Institute of Electrical and Electronics Engineers.

Kanai, R. and G. Rees 2011. “The structural basis of inter-individual differences in human behaviour and cognition.” *Nature Reviews Neuroscience* 12: 231–242.

Kaplan, F. and P.-Y. Oudeyer 2007. “In search of the neural circuits of intrinsic motivation.” *Frontiers in Neuroscience* 1: 225–236.

Kraus, S. 1997. “Negotiation and cooperation in multi-agent environments.” *Artificial Intelligence* 94: 79–97.

Laird, J. E. 2012. *The Soar Cognitive Architecture*. Cambridge: MIT Press.

Laland, K. and A. Seed 2021. “Understanding human cognitive uniqueness.” *Annual Review of Psychology* 72: 689-716.

Lenat, D., M. Prakash and M. Shepherd 1986. “CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.” *AI Magazine* 6(4): 65–85.

LeCun, Y., Y. Bengio and G. Hinton, G. 2015. “Deep learning.” *Nature* 521: 436-444.

Levin, M. 2021. “Life, death, and self: Fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms.” *Biochemical and Biophysical Research Communications* 564: 114–133.

Levin, M. 2022. “Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds.” *Frontiers in Systems Neuroscience* 16: 768201.

Lewandowsky, S. and K. Oberauer 2016. “Motivated rejection of science.” *Current Directions in Psychological Science* 25(4): 217--222.

Li, Y., D. Choi, J. Chung et al. 2022. “Competition-level code generation with AlphaCode.” *Science* 378: 1092–1097.

Lieberman, Debra and Patrick, Carlton (2018) *Objection: Disgust, Morality, and the Law.* Oxford: Oxford University Press.

Lucas, J. R. 1961. “Minds, machines, and Gödel.” *Philosophy* XXXVI: 112-127.

Lucas, J. R. 1996. “Minds, machines, and Gödel. a retrospect.” In *Machines and Thought: The Legacy of Alan Turing*, Vol. 1, ed. P. J. R. Millican and A. Clark, pp. 103 124. Oxford: Oxford University Press.

Lutnyk, L., D. Rudi and M. Raubal 2020. “Towards pilot-aware cockpits.” *Proceedings of the 1st International Workshop on Eye-Tracking in Aviation,* ETH Zürich, pp. 121–127.

Marche, S. 2022. “Artificial consciousness is boring.” *The Atlantic*, 19 June 2022. https://archive.ph/2roRf (Accessed 25 June 2022).

Marcianò, A. et al. 2022. “Quantum neural networks and topological quantum field theories.” *Neural Networks* 153, 164–178..

Marcus, G. 2020. “The next decade in AI: Four steps toward robust artificial intelligence.” Preprint arxiv:2002.06177.

Marcus, G. and E. Davis 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust.* New York: Pantheon.

Margolis, E. and S. Laurence 1999. *Concepts: Core Readings*. Cambridge: MIT Press.

Masgood, T., A. Finegan and D. Walker 2004. “Biases and heuristics in judgment and decision making: The dark side of tacit knowledge.” *Issues in Informing Science and Information Technology* 1: 0295–0301.

McEwan, I. 2019. *Machines Like Me*. London: Jonathan Cape.

Melloni, L., E. A. Buffalo, S. Dehaene et al. 2019. “Computation and its neural implementation in human cognition.” In *The Neocortex*, ed. W. Singer, T. J. Sejnowski and P. Rakic, pp. 323-346. Cambridge, MA: MIT Press,

Mercier, H. and D. Sperber 2011. “Why do humans reason? Arguments for an argumentative theory.” *Behavioral and Brain Sciences* 34: 57--111.

Mercier, H. and Sperber, D. 2017. *The Engima of Reason*. Cambridge, MA: Harvard University Press.

Millidge, Beren, Alexander Tschantz, and Christopher L. Buckley 2022. "Predictive coding approximates backprop along arbitrary computation graphs." *Neural Computation* 34(6): 1329-1368.

Milton, F., J. Fulford, C. Dance, J. Gaddum, B. Heuerman-Williamson, K. Jones, K. F. Knight, M. MacKisack, C. Winlove and A. Zeman 2021. "Behavioral and neural signatures of visual imagery vividness extremes: Aphantasia and hyperphantasia." *Cerebral Cortex Communications* 2: 1–15.

Moravec, H. 1990. *Mind Children*. Cambridge, MA: Harvard University Press.

Nagel, T. 1974. "What is it like to be a bat?" *Philosophical Reviw* 83(4): 435–450.

Nettle, D. 2006 "The evolution of personality variation in humans and other animals." *American Psychologist* 61(6): 622–631.

Newell, A. and H. A. Simon 1956. "The logic theory machine." *IRE Transactions on Information Theory* 2(3): 61–79.

Newell, A. and H. A. Simon 1961. "GPS, A program that simulates human thought." In *Lernend Automaten*, ed H. Billing, pp. 109-124. Munich: R. Oldenbourg.

Newell, A. and H. A. Simon 1976. "Computer science as empirical inquiry: Symbols and search." *Communications of the ACM* 19(2): 113-126.

Nilsson, Nils J. 2006. “Human-level artificial intelligence? Be serious!” *AI Magazine* 26(4): 68-75.

Nyholm, S. 2020. *Humans and Robots: Agency, Ethics, and Anthropomorphism.* London: Bowman & Littlefield.

Oudeyer, P.-Y., A. Baranes and F. Kaplan 2013. “Intrinsically motivated learning of real world sensorimotor skills with developmental constraints.” In *Intrinsically Motivated Learning in Natural and Artificial Systems*, ed. G. Baldassarre and M. Mirolli, pp. 303–365. Berlin: Springer.

Penrose, R. 1989. *Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford: Oxford University Press.

Pinker, S. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: Technology Press and John Wiley.

Ramstead, M. J., D. A. R. Sakthivadivel, C. Heins et al. 2022. “On Bayesian mechanics: A physics of and by beliefs.” Preprint arxiv:2205.11543.

Rumelhart, D. E., J. L. McClelland and the PDP Research Group 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Samek, W., G. Montavon, S. Lapuschkin, C. J. Anders and K.-R. Müller 2021. "Explaining deep neural networks and beyond: A review of methods and applications." *Proceedings of the IEEE* 109: 247-278.

Sapolsky, R. 2017. *Behave: The Biology of Humans at our Best and Worst.* New York: Penguin.

Saygin, A. P., T. Chaminade, H. Ishiguro, J. Driver and C. Frith 2012. "The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions." *Social, Cognitive and Affective Neuroscience* 7(4): 413–422.

Schuman, C. D., T. E. Potok, R. M. Patton, D. Birdwell, M. E. Dean, G. S. Rose and J. S. Plank 2017. "A survey of neuromorphic computing and neural networks in hardware." Preprint arxiv:1705.06963v1 [cs.NE].

Senior, A. W. et al. 2020. "Improved protein structure prediction using potentials from deep learning." *Nature* 577, 706–710.

Shapiro, S. C. 2000. "SNePS: A logic for natural language understanding and commonsense reasoning." In *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, ed. L. M. Iwanska and S. C. Shapiro, pp. 175-195. Menlo Park, CA/Cambridge, MA: AAAI Press/MIT Press.

Simon, H. A. 1982. *Models of Bounded Rationality*. 2 vols, Cambridge, Mass.: MIT Press.

Smolensky, P. 1988. "On the proper treatment of connectionism." *Behavioral and Brain Sciences* 11: 1-23.

Stal L. J. 2012. “Cyanobacterial mats and stromatolites.” In *Ecology of Cyanobacteria II: Their Diversity in Space and Time*. ed. B. Whitton, pp. 65–125. Berlin: Springer.

Steels, L. 2001. *The Talking Heads Experiment*. Berlin: Language Science Press.

Stray, J. 2020. “Aligning AI optimization to community well-being.” *International Journal of Community Well-Being* 3: 443–463.

Sullins, J. P. 2012. “Robots, love and sex: The ethics of building a love machine.” *Affective Computing*, *IEEE Transactions on Affective Computing* , vol.3 issue 4, pp. 398-409.

Sullins, J. P. 2013a. “An ethical analysis of the case for robotic weapons arms control.” In *Proceedings for the 5th International Conference on Cyber Conflict*, ed. K. Podins, J. Stinissen, M. Maybaum. Tallinn, Estonia: NATO CCD COE Publications.

Sullins, J. P. 2013b. “Roboethics and telerobotic weapons systems.” In *Philosophy and Engineering: Reflections on Practice, Principles and Process*, ed. Dian P. Michelfelder, Natasha McCarthy, David E. Goldberg.

Sullins, J. P. 2014. “Deception and virtue in robotic and cyber warfare.” In *The Ethics of Information Warfare*, ed. Luciano Floridi and Mariarosaria Taddeo. Berlin: Springer.

Sullins, J. P. 2017. “Robots, sex, and love.” In *Philosophy*: *Technology,* ed Beavers, Anthony, pp. 217-243. Farmington Hills, MI: Macmillan Reference USA.

Tang, J., F. Yuan, X. Shen, Z. Wang, M. Rao, Y. He, Y. Sun, X. Li, W. Zhang, Y. Li, B. Gao, H. Qian, G. Bi, S. Song, J. Yang and H. Wu 2019. “Bridging biological and artificial neural networks with emerging neuromorphic devices: Fundamentals, progress, and challenges.” *Advanced Materials* 31: 1902761.

---

[1] We use the term “artificial intelligence’’ (AI) to denote artifacts that exhibit intelligence, as has been commonplace since the 1950s. We do not mean to imply that there is anything “unnatural” about AI by this usage. Under AI’s main goals, Wikipedia lists “Artificial General Intelligence” first. Searching for “human like artificial intelligence” on Google Scholar yields nearly 3 million hits (9 June, 2022).

[2] For example, from the Daily Mail, 18 May 2022: “DeepMind, a British company . . . may be on the verge of achieving human-level artificial intelligence (AI). Nando de Freitas, a research scientist at DeepMind and [a] machine learning professor at Oxford University, has said 'the game is over' in regards to solving the hardest challenges in the race to achieve artificial general intelligence (AGI).” (See, https://www.dailymail.co.uk/sciencetech/article-10828641/Googles-DeepMind-says-close-achieving-human-level-artificial-intelligence.html.)

[3] Another version of identifying GI with universal Turing computability is to claim that the laptop would be a GI if it were running the right algorithm — a GI algorithm. It is natural and plausible to view AI history as the search for the right algorithm (Dietrich, E., C. Fields, J. Sullins, B. Van Heuveln, B. and R. Zebrowski (2021) *Great Philosophical Objections to Artificial Intelligence: The history and legacy of the AI Wars*. London: Bloomsbury Academic Press). It is also a plausible interpretation of AI history that the existence of this right algorithm is an article of faith rather than a search for something scientifically predicted. Compare this to the case of the Higgs Boson, where the criteria of success were well-defined in advance.

[4] The recent controversy concerning Google’s LaMDA system show that some people still treat the Turing Test as criterial for sentience. Marche, S. (2022) Artificial consciousness is boring. *The Atlantic*, 19 June 2022 (https://archive.ph/2roRf; Accessed 25 June 2022) provides a thoughtful analysis.

[5] Ironically, claims of AI capabilities are often claims about the combined capabilities of an AI system and its human “users” or coworkers; claims of AI language understanding are a case in point. Even the much-hyped and much-deplored ChatGPT is a human-in-the-loop system: without a human to organize a training set and a human to ask questions, it does nothing. One could, of course, imagine future large language models that (presumably incrementally) train themselves and engage in conversation in the absence of human intervention.

[6] This characterization is obviously human-centric and hence optimistic. Given what we know about the evolution and psychology of ethics and morality (e.g., Lieberman, Debra and Carlton Patrick (2018) *Objection: Disgust, Morality, and the Law.* Oxford University Press), it seems possible that future AI systems could be not only our intellectual superiors, but also our moral superiors. We might even consider such machines sufficiently superior

Taylor, J. E. T. and G. W. Taylor 2021. “Artificial cognition: How experimental psychology can help generate explainable artificial intelligence.” *Psychonomic Bulletin & Review* 28: 454–475.

Turing, A. 1950. “Computing machinery and intelligence.” *Mind* 59: 433-460.

---

to be suitable replacements for ourselves (Dietrich, E. (2011) Homo sapiens 2.0: Building the better robots of our nature. In M. Anderson and S. Anderson, (eds.), *Machine Ethics*, Cambridge University Press). We consider this “posthumanist” point further below.

[8] The absence of effective ethical oversight outside academia has become an increasingly political issue, one to which initiative such as the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (https://standards.ieee.org/industry-connections/ec/autonomous-systems/) are belated responses.

[9] For a famous example of this in popular culture, see episode 9, season 2 of Star Trek: The Next Generation titled “The Measure of a Man.” Snodgrass, M. (Writer) & Scheerer, R. (Director) (1989, February 13).

[11] “Neurotypical” has become something of a stand-in for the discredited term “normal” and the even less popular notion of “human nature” (Pinker, S. (2002) *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking). It is revealing that much of what evolutionary psychology – the discipline that believes in it – has to say about human nature is nothing to be proud of.

[12] Uncanny valleys are demonstrated by the many humanoid robots who, in their failure to look and act convincingly human, leave us with a feeling of the monstrous. See, for example, Vyommitra, built by the Indian Space Research Organization to accompany Indian astronauts on space missions. https://en.wikipedia.org/wiki/Vyommitra. Also, see Kageki, Norri (2012). "An Uncanny Mind: Masahiro Mori on the Uncanny Valley and Beyond". *IEEE Spectrum*. New York City: IEEE.

[13] As pointed out by a referee of this paper, these characteristics are sufficient for team problem solving in some domains, e.g. in the case of dog-human search and rescue operations.

[14] Here, clearly, we mean intellectually successful, not just commercially successful. While their contribution to accessibility can be lauded, the language skills of chatbot telephone receptionists are, it is almost universally agreed, thus far a terrible failure of AI despite their commercial success.

[15] There is quite of bit of complicated epistemology and metaphysics lying beneath the surface here. The culprit complicating things is the notion of *shared umwelten* or the notion of a *shared umwelt.* How do we develop such a thing? How does the idea even occur to us? Is it possible for humans and computer programs to share an umwelt? What evidence would support the conclusion that two or more umwelten are being shared? Quine’s famous and difficult *indeterminacy of translation* thesis (Quine, W. V. O. (1960).

Tversky, A. and D. Kahneman 1974. “Judgment under uncertainty: Heuristics and biases.” *Science* 185: 1124–1131.

Visalberghi, E., G. Sabbatini, A. H. Taylor and G. R. Hunt 2017. “Cognitive insights from tool use in nonhuman animals.” In *APA Handbook of Comparative Psychology*, Vol. 2, ed. J. Cell, pp. 673–701. Washington, DC: Am. Psychol. Assoc.

von Uexküll, J. 1957. “A Stroll through the worlds of animals and men.” In *Instinctive Behavior*, ed. C. Schiller, pp. 5–80. New York, NY: International Universities Press.

Vysocky, A. and P. Novak 2016. “Human-robot collaboration in industry.” *MM Science Journal* 2016 (June): 903-906.

Whitby, B. 2003. “The myth of AI failure.” CSRP 568, University of Sussex.

Wrangham, R. and D. Peterson 1997. *Demonic Males.* New York: Houghton, Mifflin and Company.

Wright, L.G., T. Onodera, M. M. Stein et al. 2022. “Deep physical neural networks trained with backpropagation.” *Nature* 601: 549–555.

Word and Object. Technology Press and John Wiley. Cambridge, MA) is relevant here. Quine’s thesis raises immediately a serious question: In order to even state the indeterminacy of translation thesis between say person A and person B, one has to assume a shared umwelt between A and B. But such a shared umwelt is precisely what Quine’s thesis is undermining. So, in a very strong sense, to doubt communication we have to first communicate. There is much work here to be done, but it is beyond the present scope.