Phenomenology and the Cognitive Sciences (2023) 25:289–310
https://doi.org/10.1007/s11097-023-09942-x

"Where lies the grail? AI, common sense, and human practical intelligence"

William Hasselberger[1] · **Micah Lott**[2]

Accepted: 5 October 2023 / Published online: 31 October 2023

Abstract

The creation of machines with intelligence comparable to human beings—so-called "human-level" and "general" intelligence—is often regarded as the Holy Grail of Artificial Intelligence (AI) research. However, many prominent discussions of AI lean heavily on the notion of human-level intelligence to frame AI research, but then rely on conceptions of human cognitive capacities, including "common sense," that are sketchy, one-sided, philosophically loaded, and highly contestable. Our goal in this essay is to bring into view some underappreciated features of the practical intelligence involved in ordinary human agency. These features of practical intelligence are implicit in the structure of our first-person experience of embodied and situated agency, deliberation, and human interaction. We argue that spelling out these features and their implications reveals a fundamental distinction between two forms of intelligence in action, or what we call "efficient task-completion" versus "intelligent engagement in activity." This distinction helps us to see what is missing from some widely accepted ways of thinking about human-level intelligence in AI, and how human common sense is actually tied, conceptually, to the ideal of practical wisdom, or good (normative) judgment about how to act and live well. Finally, our analysis, if sound, also has implications for the important ethical question of what it means to have AI systems that are aligned with human values, or the so-called "value alignment" problem for artificial intelligence.

Keywords Artificial intelligence · Practical intelligence · Common sense · Wisdom · Value · Alignment

✉ William Hasselberger
hasselberger@ucp.pt

Micah Lott
micah.lott@bc.edu

[1] Institute for Political Studies, Catholic University of Portugal, Lisbon, Portugal

[2] Department of Philosophy, Boston College, Chestnut Hill, MA, USA

1 Introduction: the holy grail for artificial intelligence

The creation of machines with intelligence comparable to human beings—so-called "human-level" intelligence—is often regarded as the Holy Grail of Artificial Intelligence (AI) research.[1] Indeed, in their extensive "State-of-the-Art Review," Philip Jansen and his co-authors suggest that this goal is constitutive of the field: "AI can be defined as the science and engineering of machines with capabilities that are considered intelligent by the standard of human intelligence."[2] The language of "human-level AI" is powerful and evocative. It conjures up images of science fiction characters like C3PO in *Star Wars*, Data in *Star Trek*, or the tragically-fated "replicants" in *Blade Runner* – characters who are just as intelligent as you and I (if not more so), and who interact as fellow members (albeit eccentric ones) of the human social world. And while it's very difficult for laypersons to grasp how advanced AI systems work, phrases like "human-level intelligence" suggests an end result that anyone can understand. We all have an intuitive sense of what human intelligence is, rooted in our first-person experience of living, acting, and interacting with other humans. So talk of human-level AI can be a way of saying: "Take your basic understanding of human intelligence, and imagine an AI system that is essentially *like that*."

Some prominent figures in the world of AI research believe that the Holy Grail is not too far away. Shane Legg, for example, a co-founder of Google's AI company Deep Mind, predicts that "Human-level AI will be passed in the mid-2020s."[3] Others are more skeptical.[4] But recent advances in large language models (LLMs), such as OpenAI's GPT-4 and Google's LaMDA, have generated a new wave of speculation that the day might not be far off when we create human-level AI. Microsoft researchers Sebastein Bubeck, Varun Chandrasekaran, et al., suggest that GPT-4 is "is a significant step" towards an "Artificial General Intelligence" that is "at or above the human-level."[5] OpenAI states that the company's ultimate goal is to create "Artificial General Intelligence," or "AI systems that are generally smarter than humans," while ensuring this technology "benefits all of humanity."[6] Others even herald—or forewarn—of progress towards superhuman, "God-like" AI.[7]

[1] See Margaret Boden: "General intelligence is still a major challenge, still highly elusive. AGI is the field's Holy Grail." *Artificial Intelligence* (Oxford, UK: Oxford University Press, 2016), 19.

[2] Philip Jansen, Stearns Broadhead, Rowena Rodrigues, David Wright, Philip Brey, Alice Fox, & Ning Wang. (2019). SIENNA D4.1: State-of-the-art Review: Artificial Intelligence and robotics (V.04). Zenodo. https://doi.org/10.5281/zenodo.4066571, 5.

[3] Cited in Melanie Mitchell, *Artificial Intelligence* (New York: Farrar, Straus and Giroux, 2019), 40.

[4] Brian Cantwell-Smith, *The Promise of Artificial Intelligence* (Cambridge: MIT Press, 2019); Gary Marcus, *Rebooting AI* (New York: Pantheon, 2019); Melanie Mitchell, "Why AI is Harder Than We Think," *arXiv*: 2104.12871 (2023); Erik Larson, *The Myth of Artificial Intelligence* (Cambridge: Harvard University Press, 2021); and AUTHOR CITATION REMOVED.

[5] Bubeck, et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv* (2023).

[6] https://openai.com/blog/planning-for-agi-and-beyond

[7] Ian Hogarth "We must slow down the race to God-like AI" *Financial Times* Magazine, April 12, 2023. See also Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2016).

Our focus in this essay is not on when human-level AI might be achieved, or how it will be achieved, or even whether it is possible. Rather we are interested in a question that is more basic: *what*, exactly, is "human-level intelligence"? Intelligence itself is a notoriously elusive concept, and unless we have a reasonably well-defined conception of human intelligence, and one that is faithful to the complexity and richness of the phenomenon itself, this idea can't play the role it's meant to be playing in our conversations about AI – namely, providing a sense of the ultimate goal of AI research, and a regulative ideal for assessing machine cognition. Even worse, if we are operating with a distorted, partial, or superficial conception of human-level intelligence, we will confuse ourselves about where AI research is heading, and how close it might be to getting there.

We believe that much recent writing on AI is deficient in just these ways. Discussions of AI often lean heavily on the notion of human-level intelligence to frame AI research, but then rely on conceptions of human capacities that are sketchy, one-sided, and philosophically loaded and highly contestable.[8] As Melanie Mitchell puts it, we are confused about the long-term goal of AI because the field's "central notion—intelligence—remains so ill defined."[9]

Our goal in this essay is to bring into view some underappreciated features of the *practical* intelligence involved in ordinary human agency, what some would call our human "common sense." These features of practical intelligence are implicit in the structure of our first-person experience of embodied and engaged agency, deliberation, and human interaction, and they have received attention in the philosophy of action. We argue that spelling out these features and their implications reveals a fundamental distinction between two forms of intelligence in action, or what we call "efficient task-completion" versus "intelligent engagement in activity." This distinction helps us to see what is missing from some widely accepted ways of thinking about human-level intelligence in AI, and what would truly be required for an AI system to be intelligent in a way comparable to a human being. Finally, our analysis, if sound, also has implications for the important ethical question of what it means to have AI systems that are "aligned" with "human values"—the so-called "value alignment" problem in AI.

Before moving to our main discussion, there are two points to note. First, writers sometimes distinguish between human-*level* and human-*like* intelligence.[10] The basic idea is that a machine might perform the same (range of) intelligent tasks as a

[8] We discuss examples of this tendency among prominent authors below. See also Ray Kurzweil, *How to Create a Mind* (New York: Penguin 2013); and Stuart Russell, *Human Compatible* (New York: Penguin, 2019). Two important exceptions to this are Cantwell Smith, *The Promise of Artificial Intelligence*, and Mitchell, "Why AI is Harder Than We Think," both of whom offer extensive and insightful examinations of the nature of human intelligence in relation to the possibility of AGI. However, both Cantwell Smith and Mitchell largely focus on theoretical rationality – i.e., human intelligence with regard to representing or depicting the world. They have much less to say about specifically *practical* rationality – i.e., the intelligence humans display in action, situational perception, and deliberation that leads to action. Our focus here will be on this sort of practical intelligence.

[9] Mitchell, *Artificial Intelligence* (2019), 6.

[10] See, e.g., Murray Shanahan *The Technological Singularity* (Cambridge, MA: The MIT Press, 2015) 80–84.

human being but in a way that relies on different underlying processes from the ones that humans employ when performing those tasks. For example, an artificial neural network might classify faces using pattern-matching algorithms and predictive models derived from large datasets in a way that does not mirror the specific underlying physiological and psychological mechanisms by which humans recognize faces. Nothing we say here will assume that human-level intelligence must be human-like in this sense. Our focus will be on the nature and variety of intelligent activities that human beings perform in everyday contexts – i.e., the character of practical intelligence as manifested in familiar, even mundane, human actions. What matters, in our view, is not sameness of underlying physical or computational processes. What matters is whether a machine is able to perform the same sort of intelligent activities, or manifest the same sort of practical intelligence, as a human being, regardless of underlying causal mechanisms, computational architecture, or material constitution. Our question is: what would it *mean* for a machine to perform those activities and manifest this kind of "human-level" practical intelligence?

Second, many writers use "human-level AI" as synonymous with "artificial general intelligence" (AGI), where this refers to machine intelligence that can be applied to an open-ended range of tasks coextensive with the range to which human intelligence can be applied. Others use "human-level AI" as synonymous with "strong AI," where this refers to machines with genuine minds, truly capable of thought, understanding, and other mental states, as opposed to mere simulations of thought, understanding, and mind (or "weak AI").[11] We do not object to either terminological practice. However, we take there to be a precedence to the notion of *human* (-level) intelligence. Human intelligence is our paradigm for general intelligence, and human thinking is our paradigm for real (rather than simulated) thinking—at least in the context of AI research. As Bubeck, et al., put it in their study of GPT-4: "We use AGI to refer to systems that demonstrate broad capabilities of intelligence, including reasoning, planning, and the ability to learn from experience, and *with these capabilities at or above human-level*."[12]

This is significant. We do not start from some determinate notion of "general intelligence" and then discover, lo and behold, that humans happen to have this general intelligence. Likewise, we don't start with some general idea of "real minds" and then find, lo and behold, that humans happen to have minds in this strong sense. Rather when we attend to the human case, we find a form of intelligence that it is natural to describe as *general* rather than *task-specific*, and *authentic* rather than merely *simulated*. Both "general" and "strong" are locutions meant to indicate what

[11] The term AGI was popularized by Shane Legg and Ben Goetzel. See Susan Schneider *Artificial You* (Princeton: Princeton University Press, 2019) 9, and Mark Coeckelbergh *AI Ethics* (Cambridge: MIT University Press, 2020). 64–66. For the strong/weak distinction, see Mitchell (2019), 40–41, 44–45. This distinction was first made in John Searle's seminal article, "Minds, Brains, and Programs," in *Behavioral and Brain Sciences* 3 (1980), 417–24.

[12] "Sparks of General Intelligence: Early Experiments with GPT-4" *arXiv:2303.12712v5* (2023), emphasis added. Human intelligence here serves as a paradigm, or benchmark, from which the notion of AGI acquires meaning and content.

is peculiar and distinctive about the thought and intelligence of human beings, and they derive their conceptual contours from the human case.[13]

Let's now consider in more detail some common ways that literature on AI has portrayed the Holy Grail of the field: the creation of machines with human-level intelligence.

2 Beyond the Swiss army knife model of intelligence

Are contemporary computer systems becoming as intelligent as humans? In one sense, the answer is obviously and unambiguously *yes*. It's been over twenty-five years since IBM's Deep Blue computer defeated the world-champion Gary Kasparov in chess, and over six years since Deep Mind's AlphaGo system defeated the world-champion Lee Sodel in Go—later yielding the still-more powerful, self-trained AlphaZero system, capable of besting humans at multiple games, including chess and Go, and the MuZero system, which uses "reinforcement machine learning" and self-play to master various games without even being pre-programmed with their rules.

Of course, machines have been out-performing humans at difficult tasks for centuries, from power looms, to steam engines and Gatling guns. What is remarkable about programs like Deep Blue and AlphaZero is the intellectual difficulty of activities like chess or Go. After all, elephants can lift much more than humans, and horses can run faster, but no other animal has the sort of intelligence required for chess or Go. The influential AI pioneers Allan Newell, J.C. Shaw, and Herbert Simon went so far as to claim that, "If one could devise a successful chess machine, one would seem to have penetrated to the core of the human intellectual endeavor."[14] On the Newell-Shaw-Simon view, we approximate human intelligence in a machine when we get the machine to match or exceed human performance at a particularly "smart" task—i.e., a task that demands a high degree of intelligence in humans, like master-level chess or Go.

Even so, no one today is tempted to say that AlphaZero is just as intelligent or "smart" as a human being, or that MuZero possess human-level intelligence. An obvious reason for this is that these programs can only perform well at a single category of task—game-playing—and, moreover, the tasks within this category (games) take place only within a bounded and unambiguous "toy" world. In the toy world of a game like chess or Go, the objective of all action (the so-called "success condition") is precisely specified, and the set of possible actions, along with the set of potentially relevant factors are strictly limited. Humans, on the other hand, can perform indefinitely many different kinds of activity, and we act

[13] To see this point vividly, just consider the classic Turing Test. The philosopher John Haugeland argued that the "anthropomorphic prejudice, 'human chauvinism,' is built into our very concept of intelligence. This concept, of course, could still apply to all manner of creatures; the point is merely that it's the only concept we have—if we escaped our 'prejudice,' we wouldn't know what we were talking about." Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge: MIT Press, 1985), 5.

[14] Allan Newell, J. C. Shaw, and Herbert Simon, "Chess-Playing Programs and the Problem of Complexity," *IBM Journal of Research and Development* 2 (1958), 320–35.

within the real world—an unbounded physical and social world brimming with rich cultural meanings, uncertainty, ambiguity, and unpredictable and open-ended change. The range of possible human objectives and activities is not pre-fixed, nor is the range of features of the world that may be relevant to those objectives and activates. Newell, Shaw, and Simon apparently over-estimated the applicability of narrow chess skill beyond the "toy" world of the game and, at the same time, underestimated the complexity and sophistication of human intelligence as manifested in the concrete world.

He may have lost the chess match, but Kasparov could do many things with intelligence that Deep Blue could not. He could make a sandwich, write a book, form a family, speak out against corruption and authoritarianism in Russia, and much besides.[15] This might tempt us to think that we will achieve human-level AI when we create AI systems that can perform sufficiently many "smart" tasks as well as, or better than, a human being. Call this the **Swiss Army Knife Model** of human-level intelligence: Just as a Swiss Army knife contains numerous tools for various tasks, so what is special about humans is the large number and broad range of intelligent activities that we can perform. On this picture, the capacity for performing some suitable number and range of different, independently specifiable intelligent tasks—tasks *A, B, C*…(etc.)—constitutes the "general" intelligence that is characteristic of human beings. AlphaZero, on this picture, made a small step towards human-level intelligence when it mastered *three* games, rather than just one.

In the literature on AI, it sometimes sounds like the Swiss Army Knife Model is the picture of human-level intelligence at work. Nick Bostrom, for instance, defines "Human-level AI" as "An AI that can match the intellectual performance of a typical human being in all practically important domains."[16] And a key reason that Large Language Models like GPT-4 strike us as impressive is that they can learn so many different tasks, from writing computer code to writing rap lyrics about ancient philosophers and even passing the Bar exam. As Bubeck and colleagues report on their experiments with GPT-4, "The combination of the generality of GPT-4's capabilities, with numerous abilities spanning a broad swath of domains, and its performance on a wide spectrum of tasks at or beyond human-level, makes us comfortable with saying that GPT-4 is a significant step towards AGI."[17]

The Swiss Army Knife Model surely captures something important about human intelligence. But a mere Swiss Army Knife Model is implausible as an account of human intelligence, and many reject it outright. For instance, Ben Lorica and Mike Loukides write:

[15] As Shanahan says, "a human being is a generalist, a jack of all trades. A human chess champion can do a whole lot more than just play chess." *The Technological Singularity*, 3.

[16] Bostrom, *Superintelligence,* 408.

[17] "Sparks of General Intelligence," 4.

> You can add narrow AIs *ad infinitum* (a car could have a bot that talks about where to go; that makes restaurant recommendations; that plays chess with you so you don't get bored), but a pile of narrow intelligences will never add up to a general intelligence. General intelligence isn't about the *number* of abilities, but about *integration* between those abilities.[18]

Likewise, Murray Shanahan insists that general intelligence requires more than "a certain critical mass of skills."[19] It's not enough to be a "multi-specialist."[20] Why not? "The multi-specialist," argues Shanahan, "is going to get stuck as soon as it has to face a problem that is outside of any of its areas of expertise, an inevitable occurrence in an ever-changing world."[21] Human-level general intelligence requires capacities to integrate various skills (perceptual, sensorimotor, cognitive, social, emotional capacities), learn new skills, and cope with fluid, ambiguous, and unpredictable real-world environments.[22]

According to Shanahan, endowing a machine with properly human-level general intelligence will require two features that go beyond the multi-specialist Swiss Army Knife model: *common sense* and *creativity*.

Common sense is a well-known and long-standing challenge for AI systems.[23] Common sense functions as an umbrella term for all the unspoken intuitive knowledge, reasoning capacities, and various skills (physical, perceptual, psychological, cultural, and social capacities) that normal human adults bring to bear, often without even realizing it, in dealing intelligently with people and the world.[24] Shanahan characterizes "common sense" as "an understanding of the principles of operation of the everyday world, in particular the physical and social environment."[25] Common sense, for Shanahan, is our tacit, intuitive grasp of "everyday physics" and "everyday psychology."

And creativity, says Shanahan, is "the ability to innovate, to generate novel behavior, to invent new things or devise new ways to use other things."[26] Taken together, common sense and creativity define the generality of human intelligence:

[18] Ben Lorica and Mike Loukides "What is Artificial Intelligence?" O'Reilly: https://www.oreilly.com/radar/what-is-artificial-intelligence/ (2016) (emphasis added).

[19] Shanahan, *The Technological Singularity*, 5.

[20] Ibid.

[21] Ibid.

[22] Cf. Lorica and Loukides (2016): "A general intelligence would have the ability to follow multiple pursuits and to adapt to unexpected situations.".

[23] Recognition of this issue goes back at least to John McCarthy's early paper "Programs with Common Sense," *Symposium on the Mechanization of Thought Processes,* National Physical Laboratory (Teddington: 1958) and Hubert Dreyfus, *What Computers Still Can't Do* (Cambridge: MIT Press, Revised Edition, 1992).

[24] As Yejin Choi puts it: "One of the fundamental limitations of AI can be characterized as its lack of commonsense intelligence: the ability to reason intuitively about everyday situations and events, which requires rich background knowledge about how the physical and social world works." Yejin Choi, "The Curious Case of Commonsense Intelligence" *Daedelus* (2022), 139–155 (quote at 139).

[25] Shanahan, *The Technological Singularity*, 6.

[26] Ibid., 7.

> Creativity and common sense complement each other. Creativity enables the individual to come up with new actions, but a commonsense understanding of the everyday world is needed to anticipate the consequences of those actions. On the one hand, creativity without common sense…is nothing more than thrashing in the dark. On the other hand, common sense without creativity is inflexible. But an intelligence that can wield both is a powerful thing.[27]

As an example of intelligence combining creativity and common sense, Shanahan describes an instance in which a New Caledonian crow devised, without training, a novel way to bend a wire in order to acquire food. And humans go much further in this regard than crows: such a combination of creativity and common sense is manifested across the whole range of human activities, from the mundane to the magnificent, and "human-level artificial intelligence must display a similar blend of common sense and creativity if it is to perform comparable feats."[28]

A similar picture is suggested by Susan Schneider when she writes that "AGI is intelligence that, like human intelligence, can combine insights from different topic areas and display flexibility and common sense."[29] Hector Levesque, too, argues that human-level general AI (or, as he puts it, "Real AI") requires "common sense," a capacity to deal intelligently and creatively with "new and unfamiliar situations… [C]ommon sense enters the picture when we are forced to act in situations that are sufficiently unlike the patterns we have seen before."[30]

In our view, these thinkers are right to reject a mere Swiss Army Knife Model of human-level general intelligence. We agree with Lourica and Loukides that the integration of various skills, and not simply the quantity or sophistication of skills, is central to human intelligence. We agree with Shanahan, Schneider, Levesque, and others, that the complex capacities we refer to as *common sense* and *creativity* are crucial elements of distinctively human intelligence. However, the terms "common sense," "creativity," and "integration," are nearly as abstract and ill-defined as the generic notion of "general" or "human-level" intelligence, which they were meant to specify. We have an unarticulated sense of what these terms mean in daily life and action. But we need a better understanding of the nature and relationship of common sense, creativity, and skill-integration in ordinary human practice.

To make progress on this, we propose to look more carefully at some fundamental features of human practical intelligence: the kind of everyday competence and know-how we exercise in perceiving, understanding, and responding to particular situations as embodied and self-conscious agents.[31]

---

[27] Ibid., 8.

[28] Ibid., 9.

[29] Schneider, *Artificial You*, 9. Cf. Boden, *Artificial Intelligence*, 109–110.

[30] Hector J. Levesque, *Common Sense, The Turing Test, and the Quest for Real AI* (Cambridge: MIT Press, 2017), 5.

[31] All features we highlight have been studied extensively in the philosophy of action. However, their relevance to questions of "human-level" intelligence in AI research has yet to be properly appreciated.

3 What humans do: the shape of ordinary intelligent performance

In ordinary practice, we draw the contrast between "intelligence" and "non-intelligence" in different ways. One intuitive (all-or-nothing) contrast is between those entities that count as intelligent agents at all and those that do not: humans (and other animals) are intelligent agents, and rocks are not. But once we are talking about the category of intelligent agents, there is a further contrast between things that such agents "do" that count as intelligent and other things that do not. On any given day, you do many things with intelligence – e.g., make an omelet, write an email, play the piano, converse with a friend and seek her advice. All such actions are "intelligent" in a broad sense that distinguishes them from other things that you might be said to "do"—such as digesting your lunch, exerting gravitational force, or sneezing—that don't manifest your rational agency at all.

Within the broad category of intelligent human actions, some actions are also intelligent in a still more specific sense that involves a kind of practical success, excellence, or achievement. For example, to describe a military commander as showing "great intelligence" on the field is not just to characterize the commander as an intelligent agent (unlike a rock), and not just to say that what he did counted as intelligent action (unlike a sneeze). It's to say that the commander's actions non-accidently achieved a kind of practical success or excellence in the situations he faced. In this more specific sense, acting intelligently contrasts with acting in ways that are stupid, foolish, mechanical, thoughtless, etc. To describe an action as stupid, foolish, thoughtless, etc., is to characterize it as deficient, as manifesting some failure of intelligent agency.

Our focus here is intelligence in this more specific sense of success or excellence in action in a specific circumstance. But we are not only interested in the exceptional success of outstandingly skilled persons, experts, or "geniuses." Rather our focus is on the everyday, commonsensical intelligence of ordinary human actions in real-world settings. Let's consider, then, certain intuitively recognizable features of intentional human action.

First, when humans do something intelligently, we know what we are doing under some description. If you are intelligently making an omelet or writing an email, you know that you are making an omelet or writing an email.[32] Second, when we act intelligently, we take appropriate means to our ends. That is, an agent acting intelligently needs to know not only what she's doing (under some description) but also how to do it. Suppose you are making an omelet, and you decide to use nine-volt batteries instead of eggs, or you heat the frying pan only after you've removed the eggs from the pan. Such actions lack intelligence because these are not effective ways to bring about the goal of having an omelet. Likewise, in ordinary (not desperate) circumstances it is foolish to sell your wedding ring in order to buy eggs for an omelet. Even if selling your wedding ring is an effective means to your goal, its

[32] Closely related to this point, when we act intelligently, we are able to *say* what we are doing, under some description. These points go back to Elizabeth Anscombe's seminal text in action theory *Intention* (Cambridge: Harvard University Press, 1957), 5ff.

clearly not an intelligent—i.e., contextually appropriate—way to go about achieving that goal.

In commonsense intelligent action, humans are sensitive to means that are both effective and minimally appropriate to the situation. This basic point is pretty obvious—so much so that humans, when they deliberate practically, typically ignore a vast range of causally effective but contextually inappropriate means to their ends. A closely related point about practical deliberation about means and ends is this. Humans intelligently pursue not only a broad range of narrowly-defined tasks, whose conditions of success are precisely known to the agent at the outset. Humans also intelligently pursue many tasks that are themselves broadly-defined, in the sense that what counts as "success" in the task is not precisely or concretely known at the beginning, but must be clarified over time. An example of a narrowly-defined task might be: "locking the door." Examples of broadly-defined intelligent activities might include: "hosting a good party," "having a meaningful conversation", "writing an excellent paper", "being a good father," or, perhaps most broadly, "living a worthwhile human life." Showing practical intelligence in these latter cases involves more than taking effective means to some precisely-defined objective (or success condition) that is known to the agent beforehand. In the case of broadly-defined activities, like "hosting a good party" or "living a good human life," we exhibit intelligence in the way that we specify, or make concrete, what counts as appropriately realizing the goal. We might still say that doing this well amounts to taking the appropriate "means" to our "end," but the nature of the means-ends relationship is different. The means partly constitute the end, rather than operating as steps on the way to a distinct and separately known end. Hence, practical deliberation in these cases involves involves specifying, clarifying, what would fittingly realize or constitute the end itself.[33]

Third, when humans act intelligently, we know (and can say) something about how things are going – i.e., how well or poorly our performance is proceeding, how close we are to reaching our goal, accomplishing our task, or completing the performance (or some part of it). Again, there are obvious limits to this ability. In challenging and temporally-extended activities, we may be unsure of whether things are going well or poorly. But having some at least tacit awareness of how our activity is unfolding is an essential feature of ordinary intelligent human performance. This is because intelligent human performance is, at least to some extent, within the agent's control. And if you have absolutely no awareness of how things are going with your own action (i.e., if you cannot even "monitor" your performance as it is unfolding), then you cannot really be in control of your own activity—cannot really be an intelligent agent.

Finally, and crucially, in intelligent human action, the agent has some grasp, whether tacit or explicit, of why they are doing what they are doing. That is, as

[33] For an account of practical thinking and the "specification" of ends in deliberation, see David Wiggins, "Deliberation and Practical Reason," in *Needs, Values, Truth* (Oxford, UK: Oxford University Press, 1998).

Elizabeth Anscombe points out, in acting intentionally, the agent has some understanding of the point or purpose or value of their unfolding action and can offer some answer to the question "why are you doing that?"[34] Why am I cleaning the bathroom? *Because I'm hosting a party tonight and want the house to look nice.* Why am I skiing? *For my health, and for fun.*

Our teleological grasp of the point, purpose, or value of our own intelligent performances applies, first, to the *type* of action we are doing. We know something about the point, in general, of (say) making omelets: they are nutritious, easy to make, taste good, and so forth. But our sense of the point, purpose, or value of our intelligent performances is not exhausted by such generalities: it also applies, second, to the *specific instance* of what we are doing, in their *specific circumstances*. That is, we have a grasp of why doing such-and-such particular action is a good idea, here and now. For example, in making an omelet, I don't just know why people generally make omelets, I know why it makes sense (or is reasonable, or worthwhile) to make an omelet in my specific situation: e.g., it's breakfast time, I'm hungry, I've got eggs in the fridge, and would like an omelet.

Both the general and situationally-specific grasp of the purpose or reason why we are doing something are important for intelligent action. To see this, suppose someone habitually makes omelets without any sense of the point or purpose of omelet making, in either the general or in the specific sense. He just goes through the motions, setting aside the results, over and over, with no idea *why* he is doing this. Even if the obsessed-omelet-maker's omelets are 'successfully' made (e.g., indistinguishable from the omelets at the diner down the street), this case would not be a full-fledged instance of commonsense practical intelligence, but deficient—something like a bizarre obsession.

Now imagine someone understood the point or value of omelet-making in general (taste, nutritional value, etc.) but, on a given occasion, continued with the activity of omelet-making even as the fire-alarm blared and the house burned down around him, refusing to leave until he had finished his omelet. Again, even if his omelet making was in some sense skillful, his overall action would not be intelligent in the intuitive, practical sense—but extremely foolish, negligent, and extremely irrational. Our ordinary grasp of human intelligence in the performance of any given type of activity always has this embedded or contextual dimension: intelligent action is aptly sensitive and responsive to the concrete circumstances or setting of action.[35]

[34] As Anscombe famously puts it, intentional actions are processes to which a "certain sense of the question Why? Is given application." The "why?" question, according to Anscombe, seeks the agent's *reasons* for action, her sense of the point or purpose of so acting. Anscombe, *Intention*, 9ff.

[35] Many of these features of ordinary human intelligence have been emphasized and explored by researchers working on *situated cognition*. See Robbins, P., & Aydede, M. (eds.) *The Cambridge Handbook of Situated Cognition* (Cambridge, UK: Cambridge University Press, 2008) and Ataizi, M. "Situated Cognition" in Seel, N.M. (ed) *Encyclopedia of the Sciences of Learning*. (Boston, MA: Springer, 2012) 3082–3084.

4 Mere task-completion *versus* intelligent engagement in activity

We are now in a position to make a fundamental distinction between two things that we might have in mind when we speak about *acting intelligently*. Both of these concern "intelligence" as a kind of practical achievement or excellence (rather than "intelligence" in the broader senses of intelligent versus non-intelligent entities, or of actions versus non-actions). The nature of the achievement in each case, however, is very different.

A) **Efficient Task-Completion:** An agent acts intelligently, in this sense, when it performs well with regard to a specific, narrowly defined activity. Performing well amounts to acting in such a way as to bring about the distinctive result, product, or end-state that defines success in that activity. For example, you play chess intelligently, in this sense, if you can reach a checkmate against a skillful opponent via a sequence of legal moves, perhaps with some agility or originality. Practical "success" is understood in isolation from any other activities, or any contextual features of the setting of the task in question.
B) **Intelligent Engagement in Activity:** An agent engages intelligently in an activity when, in a given context, it decides appropriately *when, where, how*, and *why* to engage in a particular activity, relative to the situation, to the nature of the activity, and to other possible activities. You play chess intelligently, in this wider sense, if you grasp when, where, how, and why to play chess, in relation to all the other things that you might do, or need to do, and in relation to what else is happening in your situation.

Depending on the task and your level of skill therein, (A) might be fairly easy to achieve, or very difficult. And depending on the task, achieving competence might be easy for a human being but hard for machine, or vice-versa. Whatever the task, however, (B) requires something that (A) does not. Unlike mere task-completion, intelligent engagement in any activity requires an active understanding of the whole situation in which the activity is unfolding, or what we will call holistic practical awareness. This term refers to the human capacity to discern what is practically relevant in a concrete situation at a given time and place—that is, what is salient for a good decision, *all things considered*. This might sound like a tall order, but (B) is not rare among human beings. Taken together, (A) and (B) specify a familiar and plausible ideal of "common sense" or "good judgment": (A) knowing *how* to perform specific, separately defined tasks (e.g., reading, playing chess, making omelets, etc.) and (B) making sound choices about *what* to do in situated contexts.

Earlier we imagined someone making nutritious and tasty omelets while his house burns down. This is a case of efficient task-completion (A) without intelligent engagement (B). While he might know that "breakfast is the most important meal of the day," and that "omelets should be cooked at such-and-such heat," our omelet-maker lacks adequate holistic practical awareness of the situation. He fails to grasp what really matters here and now – namely, that omelet-making is not the thing to be doing, since even a *very* good omelet is not worth losing your home or your life.

It would be a mistake to suppose that the holistic practical awareness involved in (B) is something we need only during moments of conscious deliberation or difficult decisions among alternative options. Even when we act fluently, without deliberation, there is always a tacit and intuitive background—a practical framing or horizon—to whatever we are doing in the given situation. This situational understanding is implicit in what we immediately notice and pay attention to, and what we ignore, and it orients us as we modulate between different activities by foregrounding in our experience certain features of the situation as salient, as "calling for" our focus and response.

Importantly, a human being's situation is not simply defined by the physical and social layout of her immediate environment. A person's situation also includes her ongoing activities and relationships. To see this, consider two different ways in which you can be "doing" something: (1) as a discrete action or (2) as an ongoing activity or project. Suppose you are at a dinner party, and you overhear someone say, "I'm teaching kindergarten." You don't suppose that they are, at that very moment, engaged in some specific stretch of teaching (as if there were kindergartners somewhere in the room that you hadn't noticed). Instead, we naturally take this person to be saying that teaching kindergarten is one of their ongoing, temporally extended activities, or projects. In all likelihood, they are answering a question like, "What do you do for work?" or "What are you up to these days?".[36]

At any given moment, we humans are always engaged in many such ongoing activities or projects. That is, at any time of the day, there are numerous things that a human being could truthfully be described to be "doing" in sense of their ongoing life-activities – e.g., raising a child, reading a novel, learning the piano, saving for retirement, eating plenty of fruits and vegetables, earning a living, grappling with the loss of a loved one, learning to be less anxious about small stuff, and so on. This structural feature of human agency means that whenever we decide to perform a specific action, it is always against the backdrop of many other specific actions we could be doing, each of which would contribute to some ongoing activity that matters to us.

Hence, whenever we undertake a specific action, we need some tacit or explicit grasp of why *this* is what we should be doing, here and now, instead of something else, and how engaging in this activity relates to everything else we are also "doing" in the ongoing sense – e.g., how eating this sandwich, or reading this book, or changing this diaper, fits into the fabric of all our activities, as part of *the single human life we are leading*. A human agent's background understanding of her ongoing life activities need not be conscious and explicit at the moment of action: it can simply be shown in the immediate, pre-reflective sense an agent has of some action as relevant (or irrelevant) in the context, and how she modulates, intuitively, between a wide range of potential activities, given the changing situation.

Hence, the idea of a human agent's "situation" includes both a narrower and a broader sense. On the one hand, your situation is your immediate context: what is happening here and now in your environment, including the means for action that are currently at your disposal. On the other hand, your situation extends out from the immediate context to

[36] Of course, it's possible they are answering the question, "What are you doing tomorrow morning at 10am?" in which case the answer *does* point to a discrete action.

include much else in your unfolding life: your ongoing activities, projects, relationships, and commitments. Intelligent engagement in an activity requires an adequate grasp of the significance of "your situation" in both the narrower and broader senses.

This may, yet again, sound exalted and highly intellectual. But our capacity to properly perceive and understand situations (in both the narrow and broad senses) is a common feature of human agency and is often phenomenologically immediate—i.e., it doesn't necessarily require effortful discursive reasoning. For example, you recognize that here and now is a good time to go for a walk with your grandmother, in this park, for this length of time, wearing this jacket and shoes, and so on. That recognition depends on your sense of the immediate context, as well as your sense of the walk as part of your friendship with your grandmother, and also seeing that friendship in relation to the background context of the other ongoing activities and projects that, together, make up your unfolding life. That is, you recognize the situational appropriateness of going for a walk with your grandmother against your tacit, background sense of the value and significance of your relationship with her in *your life as a whole*, with your many distinct objectives and projects.

Holistic practical awareness is our term for the human capacity to understand the significance of a situation, both in its immediate contextual features and its relation to the wider background of one's ongoing activities in life as a whole. Human beings are able to manifest this practical, situational understanding, on-the-fly, as they respond to changing circumstances. This capacity enables human agents to act "intelligently" in a sense that involves more than merely successfully completing abstractly defined tasks. We can understand the nature of a situation—including the meaning and relative worth of different potential activities therein—and can engage intelligently in activities on the basis of that wider understanding.

5 AI, AGI, and human-level intelligence

We began with various intuitively recognizable features of the practical intelligence at work in ordinary human agency. These features are typically taken for granted, but they become apparent in their absence (as in the case of the obsessed-omelet-maker). By attending more closely to these features and their implications, we get a better sense of what "human-level intelligence" really involves. This includes what we have called *intelligent engagement in activity* and *holistic practical awareness* of a situation. Drawing on this account, let's now return to the long-sought Holy Grail of AI: creating machines with human-level general intelligence. Three initial points stand out.

5.1 The distinction between (A) *Task-Completion* and (B) *Intelligent Engagement* provides a helpful way for thinking about the success and limits of current "narrow" AI.

Powerful special-purpose AI systems like AlphaZero are masterful at task-completion (A), but blind to the wider significance and contextual setting of the activity, including the activity's relationship to other activities, in the world. AlphaZero, for

instance, could play super-human-level Go (moving inexorably towards the game's success-condition) while the venue where the game is being played, or the data-center where the system's servers are housed, burned to the ground from an easily extinguishable fire. AlphaZero's performance would be extremely intelligent in the sense of (A), efficient completion of the "task" of Go, but utterly mindless in the sense of (B), engaging in the activity, in its wider situational setting and relationship to other activities. Any *human* who played Go while the venue burned down around him would show extreme foolishness or irrationality. That is because human intelligence is more than mere task-completion, whatever the task. Human practical intelligence involves judging the significance of the activity and the situation, in wider frame of an ongoing life composed of a wide range of activities, projects, and ends.

It may be useful to compare our claim to a more familiar one in the field of AI. AI researchers refer to "long-tail" problems for AI systems in real world environments: there is an enormous range of low-probability, unexpected, and disruptive situations which an AI system could encounter in actual situations in the real world.[37] A self-driving car, for example, encounters a flash rainstorm and mud washing out the road, or a police cordon and people directing traffic by hand, or graffiti drawn on the road, or children chasing a ball across the road. None of these things are very likely: they are in the "long-tail" of low-probability distributions. But creating an autonomous system that can be trusted to operate safely and intelligently in the real world—much less creating human-level general AI—requires that the system be capable of recognizing and coping with a massive (perhaps indefinite) range of unanticipated and unusual long-tail statistical phenomena.

This point is congenial with our argument, but our claim is more fundamental. Intelligent engagement (B), the commonsense competence humans show in everyday agency, isn't just an ability to deal well with some long-tail phenomena in successfully completing a task: it involves understanding the nature and significance of the task itself, including the task's importance relative to a wide (perhaps indefinite) range of other tasks, within concrete situations, in both the narrow and broad senses of the "situation."

Why is this a challenge for AI, as it currently stands? Advanced game-playing AI systems, even those capable of superhuman performance like AlphaZero, do not understand *why* humans engage in the "task" (inappropriately so-called) of game-playing at all, i.e., what significance this form of activity has in our way of life. This lack of understanding is sometimes revealed, starkly, by the un-humanlike errors made by AI systems that are otherwise competent at a task. For example, a reinforcement-learning based AI, meant to achieve super-human mastery of the video game Tetris, "simply learned to pause the game to evade losing."[38] Whatever this algorithm "learned," it was not how to *play a game* in a manner expressing the point or purpose of playing games. There is a penumbra of meaning that surrounds and distinguishes game-playing as a unique form of human activity.[39] AlphaZero, MuZero and other game-playing AI systems have no grasp of how the concept of

[37] See, e.g., Mitchell, *Artificial Intelligence*, 115–116.

[38] R. Geirhos, J.H. Jacobson, et al., "Shortcut Learning in Deep Neural Networks," *arXiv*:2004.07780 (2020).

[39] For a brilliant historical account of the complex social significance of games in human culture, see Johan Huizinga, *Homo Ludens: A Study of the Play-Element in Culture* (New York: Beacon Press, 1971).

playing a game differs from the task of making predictions from prior data and optimizing for an outcome; nor of how the idea of a game relates to concepts like leisure, play, contest, and pleasure; nor of how different games afford distinctive kinds of enjoyment (physical, intellectual, imaginative, etc.); nor of how games fit into particular cultural traditions; nor of why children play so much and in so many different ways (from peekaboo and pretend to Pokémon).

A special-purpose game-playing AI system doesn't know when and why it's situationally appropriate to *play to win* and when and why it's appropriate to play gently or graciously (e.g., when teaching a child how to play, or playing against a murderous dictator). A game-playing AI system does not know how game-playing relates to other kinds of activity that are central in human life (e.g., work, rest, parenting), or when it would be appropriate to simply give up game-playing and do something entirely different (e.g., when the venue is burning down). A game-playing human being, on the other hand, understands what a game is—i.e., the significance and meaning of a game and how games relate to and differ from other activities—and shows intelligence in the way they engage in game-playing activities in context, or is culpable for failing to exercise that intelligence.

Hence, the distinction between intelligence as mere task-completion and as intelligent engagement in activity reveals something important about the difference between human practical intelligence and current narrow (task-based) forms of AI. Narrow AI systems, even those capable of super-human performance on multiple tasks, like AlphaZero, perform activities "intelligently" only in a highly abstract sense, denuded of their conceptual meanings and their relationships to other activities, and detached from the situational settings in which these activities are embedded in the real world and lives of agents. Put differently, narrow AI systems perform *tasks* but do not engage in *activities* at all, if we mean by "activities" the active components of an unfolding life (a life that is shaped by one's awareness of the significance and relative worth of an indefinite array of pursuits). This may be obvious on reflection, but it is underappreciated in the prominent literature on AI.

5.2 The distinction between A) *Task-Completion* and (B) *Intelligent Engagement* helps us to see more clearly what would be missing in a Swiss Army Knife AI system

Such a system can call upon its large range of narrow (task-specific) forms of machine intelligence. However, doing so intelligently, as opposed to arbitrarily or stupidly, requires that the AI system be able to correctly judge *that* some narrow skill is needed in a specific situation, *which* particular narrow skill is necessary (e.g., object recognition, document-search, or chess-playing?), in combination with which *other* skills, and in what *order* in time and logical precedence. Even the commonplace ability to stop and answer a request for directions from a passerby on a busy street corner requires more than a mere aggregate of separately identifiable skills: it requires the wider practical capacity to "integrate" skills, as Lorica and Loukides put it. But this is not a feature of any one particular skill of the same order as chess-playing, text generation, or object recognition. It's not a Swiss Army Knife-type

capacity to perform a single task that can be defined separately from other tasks and from specific contexts of engagement. Instead, human-level general intelligence is shown in the everyday capacity to understand the significance of a concrete situation in relation to an indefinite range of possible tasks and ongoing life activities.

Hence**,** the truly interesting phenomenon of "integration" in human intelligence is not just the integration of technical skills for the performance of any given, narrowly defined task (like the integration of braking, steering, shifting-gears, responding to the flux of traffic, etc., while driving). More importantly, human intelligence involves modulation between, and integration of, various ends/goals, values, projects, norms, and practical considerations overall. That sort of active integration among ends, values, projects, and overall practical considerations, is required by (B), intelligent engagement in activity.

5.3 The discussion so far sheds light on the capacities of common sense and creativity that are necessary for human-level intelligence.

As shown by the example of the Caledonian crow, Shanahan takes creativity to be the capacity to generate novel behaviors—innovative *means*—to achieve pre-given given objectives or ends. This is true and necessary, but insufficient. Human creativity is shown not just with regard to *means* – as might be the case if we only had (A) in view. Instead, humans are creative in the ways they conceive, order, select among, and specify their *ends*. Humans think creatively about what their goals and objectives mean: how their goals should be understood, visualized, properly pursued, and (re) imagined, in context. Again, that is part of (B), intelligent engagement in activity.

Many working in AI construe common sense as a kind of descriptive-theoretical knowledge, loosely modeled on the natural sciences. Common sense, says Shanahan, is "an understanding of the *principles of operation* of the everyday world," including "everyday physics" and "everyday psychology," that enables us to predict the likely outcomes of different courses of action.[40] This framing of the issue of common sense is familiar in the field.[41] But this conception of common sense – as body of theoretical-descriptive knowledge akin to a "folk" science – is too narrow.

Intelligent engagement in activity (B) captures an element of common sense that is not just about how to depict or represent the world, given some pre-fixed objective. The failure of common sense displayed by the obsessed-omelet-maker, whose house burns down around him, is not primarily a failure to predict or understand how the world works. It is a failure to recognize what *matters*. Human beings show common sense in recognizing what, in a given situation, is most important, and why—and this includes a sense of the relevance and relative value of one activity or objective over others. Common sense is not just theoretical (or propositional) in form, but also practical, and irreducibly so. By "practical thinking," we mean thinking (cognition, discernment, understanding) that directs itself to the all-things-considered question

[40] Shanahan, Ibid., 8 (emphasis added) and 57.
[41] See, e.g., Choi, "The Curious Case of Common Sense Intelligence.".

of *what to do and how to live*. This includes what is sometimes called "ethical reasoning"; but practical thought is broader than "ethics" in the narrow (modern) sense of what we owe others as fellow humans, or as rational beings, or as subjects of pleasure and pain.[42] Practical thought, thus, aims at action in the broadest sense.

6 Machines, human values, and wisdom

The Holy Grail project of building machines with "human-level" general intelligence raises questions and concerns that not just theoretical but are also ethical in nature. Our analysis of human practical intelligence in terms of intelligent engagement and holistic practical awareness has two ethical implications worth emphasizing.

First, consider the challenge of ensuring that AI systems are properly aligned with "human values"—what Stuart Russell and others call the "Alignment Problem." Iason Gabriel offers a helpful initial characterization of the challenge:

> The goal of AI value alignment is to ensure that powerful AI is properly aligned with human values (Russell 2019, 137)…The challenge of alignment has two parts. The first part is technical and focuses on how to formally encode values or principles in artificial agents so that they reliably do what they ought to do… The second part of the value alignment question is normative. It asks what values or principles, if any, we ought to encode in artificial agents.[43]

In thinking about AI value alignment, we can begin by distinguishing between (a) operating *in accordance* with a value, and (b) *valuing* or having a value *as your own*. For instance, many cars beep if the driver or passengers are not wearing their seat belts, as a reminder and incentive to buckle up. This feature of the car's design reflects a concern for human safety. We might say that the car has, in this respect, been designed to operate in accordance with the value of human safety, or that a concern for this value has been "put into" the car's way of functioning by its designers. But it would be misleading to say that the car values human safety, or has that value as its own, in anything like the way a person "has" values (e.g., values the safety of her children in the car).

Turn now to current special-purpose AI systems that pursue a specific goal, such as winning a game. Insofar as these systems have a goal, and they guide themselves effectively toward that goal, it becomes more tempting to say that these systems have values of their own, at least in some sense. For whatever the system has been designed to achieve – e.g., winning at chess – serves as a kind of goal,

[42] On the nature of ethics, morality, and the wider space of practical thinking, see Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge: Harvard University Press, 1985) and Philippa Foot, *Natural Goodness* (Cambridge: Oxford University Press, 2001).

[43] Iason Gabriel, "Artificial Intelligence, Values, and Alignment" *Minds and Machines* (2020) 30:411–437, quotes at 412–413. As Gabriel points out, there are a variety of things that we might try to align AI systems *with*: instructions, expressed intentions, revealed preferences, informed preferences, well-being, values. Gabriel does a nice job of showing that these are not equivalent as targets of AI alignment, and he makes a strong case for focusing on values.

or purpose, or "good" that the system is after. Of course, the goal is still "external" in the sense that it is put into the system by something external: its creators. But insofar as the system guides itself, in flexible and possibly unexpected ways, toward its goal, it might seem natural to say that the system values winning, in a way that is not true of the example of the car seatbelt detection system.

Whatever might be the merits of speaking this way, we should keep in mind that special-purpose AI systems do not value things, or "have values," in the manner that human agents do. This is not just because these systems don't feel anything, or because they lack phenomenal consciousness and self-awareness of what they are up to. Rather, it is because the way humans value any specific activity – our distinctive way of having values – involves grasping the value of various things in relation to one another, as aspects of our unfolding life as a whole.

For instance, it's plausible to suppose that the activity of omelet-making can be valuable (worthwhile) because of how it promotes the value of human health, as well as values like friendship (shared meals) and skillful activity (the craft of cooking). But any coherent and sensible grasp of the value of omelet-making must involve the recognition of its value in relation to *other* human values, needs, goals, and interests. Consider our obsessive omelet-makers once again. We might jokingly say, "Boy, they really value omelets!" But in a more important sense someone who compulsively makes omelet after omelet for no further reason, or who makes omelets as the house burns down, does not really count as valuing omelets or omelet making – at least not in the distinctly human way of valuing thing, where valuing and having values is a meaningful and characteristic aspect of living a human life integrating plurality of goods.

We suggest that it is a mistake to think that special-purpose AI programs "have values" in the relevant sense and, hence, it is misleading to describe of the Alignment Problem as a matter of somehow imbuing AI systems with the "right values" rather than the "wrong values." Rather, we should remain keenly aware of how different the concepts are of "valuing" and "having values" in the cases of AI systems, on the one hand, and human beings, on the other. In the paradigm sense of valuing things – the sense that captures the familiar and distinctive way that humans have values – AI systems in anything like their current form do not have *any* values at all, neither the right nor the wrong ones. AI systems optimize for certain (restricted) outcomes, not, like humans, leading an unfolding life in light of a conception of a plurality of interrelated values or goods.

What about value alignment in an AI system with "human-level" intelligence? Our earlier discussion does not show that such a thing is impossible. But it does suggest that it would be a mistake to suppose that we might first create a human-level AI without any particular values and then in a second, separable step imbue that AI system with the "appropriate" values. The problem with this two-step picture is that our core, ordinary notion of practical intelligence already presupposes a certain range of intelligible values—it is not "value neutral." Going after some things rather than others, or seeing certain things as beneficial or harmful, or taking certain things to be good and others bad, is necessary for acting in meaningful and intelligible ways. And acting in ways that are meaningful and intelligible

is a (minimal) condition for acting in ways that are intelligent, where that refers to a kind of success.

6.1 An adequate account of human-level intelligence suggests that the true Holy Grail of AI research might be the creation of *wise* machines.

If the kitchen is on fire, it is a matter of common sense that one stop making the omelet and get help; if a dear friend walks through the door crying unconsolably, stop making the omelet and attend to the friend; if the police knock…etc. The human capacity to recognize these features and grasp their practical meanings in situ is what we've called holistic practical awareness. This is more than simply using background knowledge to make classifications and predictions and successfully complete a narrow activity. Ordinary human intelligence involves "reading" a situation correctly in the normative sense of picking up on the most salient and meaningful features, ignoring an infinite number of irrelevant details, and adjudicating between different ongoing activities and different possible ends/goals.

Once we have this picture of human intelligence in view, we can see the truth in Samuel Taylor Coleridge's claim that, "Common sense in an uncommon degree is what the world calls wisdom."[44] On an influential view of wisdom, the wise person has sound judgment about the conduct of human life as a whole, not just in a limited area (such as a specific task, skill, or profession). This is wisdom in the sense of Aristotle's *phronesis*, or practical wisdom. A wise person has a sense of what goods, and what evils, matter more than others, both in general for human beings, and for her in the particular situation she faces. She doesn't just find effective means to her ends (no matter what they are). She also has sensible or good ends: she perceives and judges correctly the worth and value of different goals. In this way, practical wisdom is the capacity for acting well overall—or all things considered—not simply in one domain or with regard to one type of consideration.[45]

If something along these lines is the right way to think about wisdom, it seems that common sense is not different in kind from wisdom. For common sense, too, is matter of good judgment about the conduct of human life. Like wisdom, common sense requires a grasp of normative appropriateness: knowing which things matter more than other things, both in general and in the specific case (and not just descriptive-theoretical principles of "everyday physics" and "everyday psychology").

[44] Coleridge, *The Literary Remains of Samuel Taylor Coleridge*, Volume III (London: William Pickering, 1836), 151.

[45] The literature on practical wisdom, from antiquity to the present day, is vast. For a few more recent discussions, see Daniel Russell, *Practical Intelligence and the Virtues* (Oxford: Oxford University Press, 2019); Stephen Grimm, "Wisdom," *Australasian Journal of Philosophy* 93 (2016), 1–16; Gavin Lawrence, "The Deep and the Shallow," in *Philippa Foot on Natural Goodness*, ed. Hacker-Wright (Cham: Palgrave, 2018), 187–256. See also: Philippa Foot, *Natural Goodness*. For another appeal to *phronesis* in the context of AI, see Cantwell-Smith *Promise*, 110–111. Although Cantwell-Smith's focus is somewhat different from ours (and less concerned with specifically *practical* intelligence), we take his account of judgment to be compatible with, and complementary to, the view we have spelled out in this paper.

Common sense—and its perfected form, wisdom—embodies a view of how to live well as a human being: what things really are important (and what is trivial) in a human life, what makes sense (and what is senseless), why someone should (or shouldn't) do something in a particular circumstance. Hence, common sense and wisdom both involve perceiving, judging, and acting sensibly, understanding the various meanings implicated in a situation, and assessing and balancing different (and sometimes competing) activities, ends, and objectives in a whole life.

What the wise person does well is the same thing that the person who exercises "mere" common sense is already in the business of doing. So if the Holy Grail of AI research is human-level intelligence, and if human-level intelligence must be spelled out with reference to common sense, it would seem that the real aspiration of AI research should be machines with excellent common sense – i.e., wise machines.[46]

Again, we don't take a stand on here on whether or not it is empirically possible to create human-level AI, and so by extension we don't take a stand on whether it is possible to create wise machines. But if our argument is on the right track, the notion of machine *wisdom* deserves much more investigation than it has received by AI researchers.

7 Conclusion

To connect the various threads of our argument: we take intelligent engagement in activity (intelligence B above) to be primary in the explanation of human-level intelligence. The conception of an abstracted notion of task-success is actually derivative from the more fundamental idea of intelligent engagement in activity in context, which is the paradigm of ordinary human performance, and an element of common sense. Common sense, in turn, involves more than descriptive-theoretical knowledge: it involves good (normative) judgment about what is intelligible, what is good, what is most important, what can be ignored, in a concrete circumstance and, more broadly, in the conduct of human life. This is continuous with the ideal of human wisdom.

At this point, we are left with a number of unanswered questions. Can common sense or wisdom, in the sense of good normative judgment about conduct in situations and life as a whole, be programmed on computers? Could we construct, whether through human-coding, or reinforcement learning, or other machine learning techniques, some comprehensive model of *all* human objectives, activities, and their relations of relevance and priority in *all* possible contexts—so as to automate the components of normative judgment in human common sense? Is affective and embodied cognition necessary for the possession of common sense/wisdom? Are there local or domain-specific forms of general common sense/wisdom? For example, could a care-giver robot exhibit something like "domestic-care wisdom," without an ability to practical cope with very different circumstances? Are the experiential and felt dimensions of emotions—hence, phenomenal consciousness—somehow necessary for common sense/wisdom?

[46] For doubts about the possibility of creating wise machines, see Joshua P. Davis. "Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere)" *Oklahoma Law Review* 51 (2019) 51–89.

A final, and particularly thorny question is this: *if* "human-level" intelligence has the structure we describe above, i.e., one tied to a holistic evaluative awareness of an unfolding life and continuous with the ideal of wisdom, *should* we aim to build machines with *this* kind of intelligence? We may ask, in other words, whether this is an ethically appropriate "Holy Grail" for the field.

Funding Open access funding provided by FCT|FCCN (b-on). No special funding was received for this manuscript.

Data availability Not applicable.

Declarations

This manuscript is not currently under review at any other journal.

Ethical approval Not applicable.

Informed consent Not applicable.

Statement regarding research involving human participants and/or animals Not applicable.

Competing interests The authors have no relevant financial or non-financial interests to disclose.