Rethinking our Thinking about Thinking

Epistemology, Architecture, and World

Brian Cantwell Smith
Draft of January 16, 2020

Adapted and extracted from Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment*, MIT Press, 2019 (herein "*Promise*")

There is an idea—associated with Descartes, but of much longer pedigree—that thinking, at its best, involves moving rationally and logically between and among "clear and distinct" ideas: ideas with unambiguous meanings, determinate extensions, and context-independent implications. For centuries this idea has served science and mathematics well; it has also had enormous impact on contemporary models of epistemology. The idea also underwrote the development of the modern computer, whose foundations are formulated in terms of digital states (ones and zeroes), discrete symbols, and strict and unambiguous logic. In turn, the success of computation has reinforced the sense that logical moves among discrete conceptual ideas must be an intrinsic norm on rationality.

This "logicist" conception of rationality holds such sway in the contemporary intellectual imagination, in fact, that those who believe that human mental life is not best or fully captured by it often conclude that there must be more to "good thinking" than rationality—turning in its place to such alternative categories as emotion, affect, and qualia. Others, such as devotees of the enactive and embodiment camps of cognitive science, argue that focusing on the mind as a way to understand intelligence is itself misguided.

Given this history, it is no surprise that the first serious attempts to model human thought on computers, in the first wave of artificial intelligence (famously dubbed "Good Old-Fashioned Artificial Intelligence," or GOFAI, by philosopher John Haugeland), developed architectures to implement just such a model of discrete conceptualist thinking. Perhaps the most extreme examples were theorem provers and other systems explicitly based on formal logic. But even when some of the most restrictive strictures of logic were eased, such as in the replacement of logical theorem-proving with a more practically-oriented idea of "satisficing,"[1] most of the logicist conception was retained:

[1]The term 'satisficing' is most strongly associated with Herbert Simon, an early AI pioneer who promoted that the idea that rationality was best understood as having the goal of coming to conclusions that were good enough for the purposes at hand, not as aiming for logical perfection. Persuaded by Simon's argument, and motivated by the unrealistic resources required for theorem proving, many projects in "first-wave" AI research focused on the development of *heuristics*—methods of coming up with practicable answers in a reasonable amount of time using modest resources. But as stated in the text, this and other themes in first-wave AI amounted to adjustments, rather than rejections, of the classic logicist framework. It was not

the idea that thinking involved clearly separable concepts, represented in discrete symbols, on the model of discrete words in a language. It was assumed that the vast variety of the world could be captured by combining these symbols, according to the rules of a conceptual grammar, into logical complexes analogous to full sentences, on what was called a "compositional" model of meaning.[2]

1 · Second-wave AI

Recently a new computational architecture, based on networks of interconnected nodes running parallel statistical algorithms, has transformed AI's approach to modeling cognition. The approach was catalysed by the development of a particular architecture called "deep learning," though various generalized and alternative forms have been developed within the same general architectural approach. So successful has the new approach been to a class of previously unsolved problems that AI has entered a new phase, called "second-wave AI." Second-wave architectures are now often referred to under the general label "machine learning," but since (i) there have been proposals for machine learning since AI's early days, and (ii) it is presumptuous to assume that the particular form of learning currently being investigated covers even a significant fraction of the full range of possible epistemic techniques to acquire knowledge and skill through study, experience, or instruction, I will use "second-wave AI" as a label for the entire new class of system.

Second-wave AI systems work in a manner that is almost the exact opposite of the classical model. Rather than using one or a small number of serial processes to carry out explicit inference over conceptually symbolic structures, these new systems consist of thousands or millions of nodes in a network or graph performing, in parallel, relatively simple numerical calculations. Their forte is the prediction and exploration of the consequences of huge numbers of extremely weak correlations between and among vast

---

until the development of second-wave AI that logicism as a model of rationality was replaced.

[2]Note that conceptually structured representations of a similar sort are familiar in myriad other modern forms, including databases, computer-aided design (CAD) systems, architectural blueprints, and health records. It is also the metaphor on which programming languages are still constructed.

In addition, though conceptual structuring is most obvious in the case of digital representations of the sort we have been talking about (written language, logic, classical AI, etc.), at a more abstract level it also underlies traditional analog representations—those that represent continuous quantities in the represented domain (the domain of the problem to be solved) with continuous quantities in the representation itself. Thus consider constructing an analog representation, in an electrical circuit, of a physical process such as air flow through an organ pipe, or wind in an impending weather system. In order to accomplish this, one sets up a discrete correspondence between "conceptual" properties of the electrical circuit and conceptual properties of the moving air. Given this conceptual mapping, real-valued quantities in the represented domain (air pressure, wind velocity, etc.) can then be modeled by real-valued quantities in the representing domain (voltage, current, etc.).

Classical physics also represents continuous quantities in the represented, such as mass, velocity, charge, etc. But it does so by employing digital (discrete) representations: "27.342 kilograms," "$x_f = x_i + vt + \frac{1}{2}at^2$", etc. What distinguishes analog representations is continuity in the representation, not continuity in what is represented.

numbers of numerical or statistical weights. In addition, and of great consequence, these statistical correlation machines are capable of what is called "learning." When "trained" on enormous troves of data, they can adjust the weights used in their calculations so as to hone in more and more accurately on the outcomes desired, and therefore reinforced, by their users and designers.

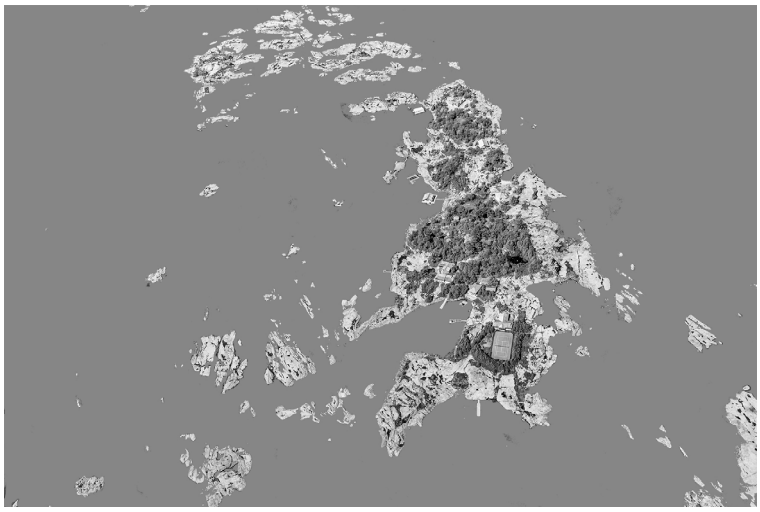

Figure 1: Islands in Georgian Bay
(A masked version of Figure 3, with the land separated out from the water, at a specific moment in time)

Second wave AI techniques have been stunningly successful when applied to a variety of problems for which classical (first-wave) AI proved inadequate. Perhaps the most dramatic successes are in the realm of perception and classification. Face recognition, image classification, handwriting recognition, etc., are now well-developed technologies that rely on these second-wave AI techniques. But the triumphs of second-wave AI are not limited to perceptual and classificatory tasks. In the fall of 2017 Google changed the technology underlying its vaunted machine translation service to use second-wave techniques, unleashing a dramatic advance in the quality of the results.

Indeed, the triumphs of second-wave AI have been sufficient to trigger a public wave of sometimes almost unbridled enthusiasm, leading to predictions (and fears) that in many areas humans will be replaced by more competent machines, and that so-called "general purpose artificial intelligence" (AGI), of a quality equal to or greater than that of people, is just around the corner.

2 · Ontology

Why do 2nd-wave AI systems work so much better than first-wave—at least in the areas in which they excel? The answer is ontological.

First-wave AI was based on an assumption that the world consisted of what I call "formal ontology": discrete objects, exemplifying distinct properties, standing in well-defined relations, grouped together in sets, etc. That is, first-wave AI presumed that the "furniture of the world"—the stuff of reality itself—consisted of arrangements of such familiar ontological kinds as *objects*, *types*, *properties*, *sets*, and the like, such as people, cars, trees, rooms, countries, mountains, conversations, etc. For example, suppose a room was represented, in a first-wave AI system, as containing a table, four chairs, a rug, a sideboard, and three people. Then if the representation was "correct," it was assumed that what it represented corresponded exactly to what was "out there,"

fundamentally and ontologically: a table, four chairs, and the rest. Issues of abstraction and idealization were relevant only to *perception*—to identifying those objects correctly, and to recognizing the properties they exemplified (their chair-ness, their rug-ness, their humanity, etc.). The representer's task (human or machine) was to come up with a representation that was *true*—that corresponded to what was really the case in the represented domain.

Figure 2: A computer-generated diagram showing the shape and layout of the islands in Fig. 1

Second-wave AI was developed as a (computationally implemented) mental architecture—not as an ontological thesis. But the way it works, and the fact that it so successful, suggests a very different picture. Perception is not assumed simply to be a case of looking out and figuring out what is ontologically "out there," as if God's labels for pre-individuated reality could be read off the incoming data stream. Rather, the realm that perception is trained on is treated as a potentially unbounded and vastly variegated data stream, to be clustered and clumped and "coarse-grained" into a resulting classification.

It is easiest to understand this approach in terms of a metaphor. Figure 1 is a photograph of some islands in Georgian Bay, part of the Great Lakes north of Toronto. Already, one can see that the real-world topography fails to support the cut-and-dried ontological registration that GOFAI assumed. Figure 2 "cleans the picture up" in a way reminiscent of logic and GOFAI, making the islands, though still relatively detailed, "clear and distinct," and also internally homogenous, in the way that is assumed in conceptual models, such as in data bases and logical statements such as 'ISLAND(X)'. The question of "how many islands are there" may have a determinate answer in figure 2, but the same is not true of the world depicted in figure 1. As I put it in *Promise*, distinctness flees, as realism increases. In the world itself, the question lacks a determinate answer.

More telling yet, though, is figure 3. This is the photograph on which figure 1 was based, without any filter blocking the transparency of the water. Figure 3, that is, reveals the submarine topography. Compared to the world's messiness, the image is still simple: gravity is a single dimension of salience, the water line is relatively sharp, the image is gray scale, and so on. Nevertheless, if the islands in the image are taken as analogs for properties, then the images suggest what in fact is true: that as soon as one presses for detail, distinctions multiply without limit, and that a richly connected submarine texture underlies all of the parts that happen to project above the surface.

The point of the metaphor is as follows. Whereas first-wave AI assumed that the

ontology of the world was given, the suggestion implicit in figure 3 is the reality itself is arbitrarily detailed, and that any ontological "parse" into discrete objects and properties results from how it is viewed—or as I put it, how it is **registered**. What we do, that is, not only in the course of pragmatically navigating the world and conducting our projects, but in thinking and reasoning about it, is to *register* the world—render it intelligible in ways appropriate to our projects and perspectives.

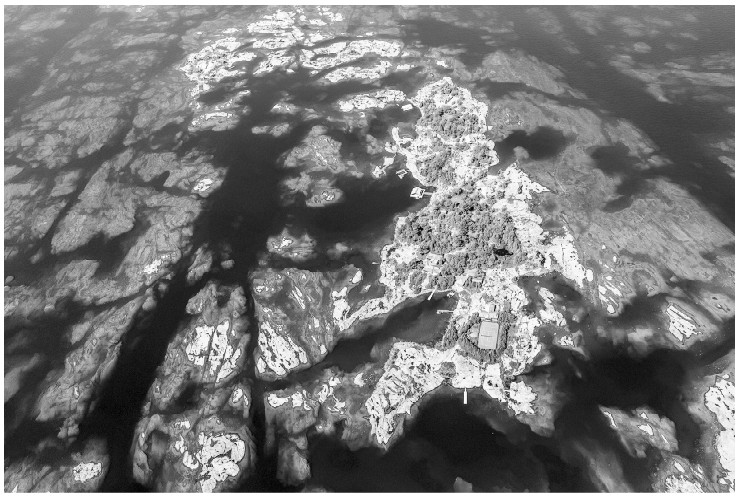

Figure 3: The original aerial photographs of islands in Georgian Bay, revealing the submarine topology

Discrete concepts—those expressed by words—have perhaps misled us into thinking that the world comes discretely articulated. But two facts, highlighted in the metaphor, belie the simplicity of that conception. First, "beneath the level of the concepts"—beneath the level of the objects and properties that the conceptual representations represent—the world itself is recognized to be permeated by arbitrarily much more thickly integrative connective detail. It is not even that our concepts sometimes have vague or unclear boundaries; it is that the facts we conceptually represent—with vaguely or not—tell on a world that itself is not itself clear cut. And, crucially, it is only in part with reference to the registration scheme, as well as with reference to the world that is registered, that such questions can be answered. That is, taking the water level to represent the conceptual "cut" imposed by the registration scheme, there are no facts about identity except with reference to it. Is an obstreperous child the ssme as or different from a rambunctious child—or an obstreperous CEO? If we are climbing an 8,000 meter peak in Nepal, and there is another local maximum 100 meters away from the summit we have reached, do we need to go over and climb that one as well? Where does one "fog" end and another start? *Reality will not tell us.* If we want "clear and distinct" answers, we need to employ conceptual schemes that impose them—that *register the world in terms of them*.

3 · Registration

What are the consequences of these insights for AI? What follows from recognizing that the nature of reality is as suggested in figure 3: a plenum of surpassingly rich differentiation, which intelligent creatures ontologically "parse" or register in ways that suit their projects?

Overall, it means that AI must take on board one of the deepest intellectual realizations of the last 50 years, joining fields as diverse as social construction, quantum mechanics, and psychological and anthropological studies of cultural diversity: that taking

the world to consist of discrete intelligible mesoscale objects is an *achievement* of intelligence, not a premise on top of which intelligence runs. AI needs to *explain* objects, properties, and relations, and the ability of creatures to find the world intelligible in terms of them; it cannot assume them.

How we register the world—how we make it ontologically intelligible in such a way as to support our projects and practices—is in my judgment the most important task to which intelligence is devoted. Developing appropriate registrations does not involve merely "taking in what arrives at our senses," but—no mean feat—developing a whole and integrated picture accountable to being in the world. It is not just a question of finding a registration scheme that "fits" the world in ways locally appropriate to the project at hand, but of relentless attunement to the fact that registration schemes necessarily impose non-innocent idealizations—inscribe boundaries, establish identities, privilege some regularities over others, ignore details, and in general impose idealizations and do an inevitable amount of violence to the sustaining underlying richness. This process of stewardship and accountability for registration, never imagined in the GOFAI project, is of the essence of intelligence.

And it is from this perspective that we can begin to understand the power of second wave AI. It is an architectural approach that starts to give us a handle on registration.

4 · Examples

Three examples will illustrate. First is the game of Go, long considered one of the world's most challenging board games. Even recently, many AI researchers believed that a computer program capable of playing championship Go was unlikely to be constructed for many more years (with some even doubting that it would prove possible at all). In 2015, however, an AI program called Alpha Go[3] defeated Lee Sedol, one of the best players in the world. Over the next couple of years, a successor to Alpha Go defeated Ke Jie, then the best Go player in the world, and subsequent versions are recognized as playing Go better than any humans ever have, or likely ever will.

A second example of a second wave AI success: reading x-rays. Research on using deep learning programs in radiology has recently been a subject of intense investigation, and just as this was written *Nature* reported on an AI system capable of surpassing human experts in breast cancer identification in mammograms.[4] A third example is the proposed joint project by Factum Arte and Case Western Reserve University to use deep learning to identify the brush strokes in paintings attributed to El Greco, with the aim of being able to discriminate those painted by the master himself from those painted by others—including painters working under his direction in his studio.

What distinguishes these three cases is that the identifying "signs" or "signature" of the phenomenon being looked for (a winning strategy for Go, the presence of cancer

[3]Developed by the Deep Mind division of Google; cf «…»

[4]McKinney, S.M., Sieniek, M., Godbole, V. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577,** 89–94 (2020) doi:10.1038/s41586-019-1799-6

in an x-ray, the hand of El Greco in the case of the paintings) is almost certainly not betrayed by any single, or even by a few, local, conceptually articulable properties of the underlying image. Rather, just as humans are now believed to identify faces by recognizing extremely complex patterns of very weakly correlated microvariables in the images of those faces, second-wave AI systems use deep learning and other associated techniques to similarly compute complex correlations over thousands or even millions of similar microproperties.

5 · Epistemological Consequences

The epistemological consequences of the success of second-wave AI approaches are substantial, but have so far been only inchoately explored. It is evident, in part because they are able to retain vast amounts of detail, to pour through extraordinarily massive data bases, and to operate at millions of times the speed of the human brain, that second-wave AI systems are liable to—and in many cases already do—demonstrate much greater prowess than that of which humans are capable, at least on the tasks at which they excel.

The potential consequence that has received the most attention has to do with jobs, employment, and deployment of human skill. Computers are already so much better than people at complex arithmetic tasks that it would quaint for a person to persist in doing arithmetic computations "by hand," rather than employing a calculator or computer. By analogy, it may be that if automated second wave AI systems are more accurate at diagnosing cancer in mammogram images than radiologists-(reducing both false negatives and false positives), it will be comparably quaint to insist that people be involved in reading x-rays.[5] And if that and similar reconfigurations come to pass, the disruption and reconfiguration of the employment landscape may be substantial and disruptive.

Other questions, however, remain open. One in particular has to do with what, if anything, we humans will *learn* through computers developing such skills. If a second-wave AI system performs better than people at some task, for example, will people become better as a result? Strikingly, Lee Sedol was not dismayed at being beaten at Go by a machine; he was reportedly thrilled that Alpha Go gave him a glimpse into previously unexplored regions of the game—regions that he might never have known about without its leading him into them, but that he could now study and perhaps come to master. Whether radiologists will learn how to read x-rays better because second-wave AI systems can do so is not immediately clear, but it is by no means impossible. Suppose a radiologist and an AI system "read" the same x-ray, but differ on their conclusions. Suppose, too, that in due course it emerges that the AI system was correct, and the radiologist wrong. It may become possible, in such a case, for the radiologist to examine what part of the x-ray was crucial to the AI system's diagnosis, or to examine

[5]As opposed to prescribing treatment as a result of such diagnoses, which is likely to require judgment; see §6.

artificially constructed images that differ only minimally from the real one but that would have led the AI to the alternative diagnosis, and to learn thereby how to make discriminations than they had previously been incapable of making. Similarly, if an AI system is able to distinguish brush strokes made personally by El Greco from those made by others in his studio, art critics may be able, again by examining examples of each, to develop the capacity to make such discriminatory judgments themselves.[6]

Why not simply ask the AI system to *describe* which aspects of the image were critical to its diagnosis? That is the aim of a huge research effort currently underway, under the heading "explainable AI": to build transparent second-wave AI systems that can *explain* how and why they reach the conclusions that they do. It is my sense, however, that any such dream may be unrealistic—because of facts that reach directly into the fundamental reasons why second-wave AI is successful in the first place.

The difficulty is that the full scope of "reasons" why a second-wave AI system comes to the conclusion that it does may not be "effable"—may not be conceptually articulate, expressible in a finite and comprehensible linguistic form. As we have already seen, the AI's conclusions are likely to rest on thousands or millions of extraordinarily weak correlations between and among innumerable microproperties of the original image. It is exactly because of its ability to make use of vast numbers of weak correlations that second-wave AI systems are as powerful as they are. Asking for explainable AI is effectively asking for second-wave AI systems to revert to being first-wave AI systems. And there was a reason that first-wave AI failed at the very sorts of task on which second wave AI succeeds.

If the reasoning processes of second-wave AI systems are intrinsically ineffable, however, does that mean that can never be communicated to us—that they will remain forever alien and incomprehensible? No, that does not follow. It would only be true if the conceptualist model of thinking that underwrote first-wave AI *was true of human cognition*. But as I have been at pains to suggest, not only is there no reason to suppose that it is constitutive of human cognition (that we do work that way), but if it is true that second-wave AI is starting to implement processes of perceptual registration that people already do extraordinarily well, then the ways that these systems work may not be alien after all.

There is a very real question, though: of how, and to what extent, we can come to understand them. The answer has partly to do with how we end up in the cognitive states that we do, and partly with the relation between those states and the words we utter and understand.

---

[6]At present, suppose we are unable to make this discrimination. Then we may at present have two sets of paintings or fragments of painting for an art critic to study: some made by El Greco himself, and some made by *either* El Greco or another painter in his studio. What we do not have, though, is the comparison set that might be critical for learning: one set by El Greco, and another set *not* by El Greco but by one of his studio painters. If the AI system could learn to make the discrimination, then it could be used construct these non-overlapping comparison sets, which might enable human painters to learn to detect the difference.

Suppose a perceptual registration system (human or AI) "outputs" or reports the result of a perceptual registration in a single word or token. Nothing about such behaviour implies there that the state of the system that manifests such behaviour must thereby be reduced to a single choice out of the number of conceptual terms we use for such expressions. To think that would be to *presume* that, like first-wave AI systems, we think solely in terms of conceptually formulated representations. But it is that very assumption that we are challenging.

Suppose in particular that, as the result of looking at a facial image, a system generates a 1,000-dimensional array, with an 8-bit number entered as the value for each dimension, representing the image. Such a representation would require just 1 kilobyte of computer storage—very little by present standards. Yet the value of such an array identifies one in a space of $256^{1,000}$ ($\approx 10^{2,408}$) possibilities. Suppose further that only a tiny fraction (let's say 0.001%, or 1 in 10^5) of these potentially representable images are comprehensible (recognizable as human faces expressing emotions), and that fewer than 1% of those are classified as representing "anger." That still means that there are more than $10^{2,400}$ different images representing anger. If the system were then to make an inference as to what that angry person might do next, there is no reason that that inference might not use the full details of their representation in reaching its conclusion. That is: just because the system can "effably" *report* the emotion it has perceived using the single word 'anger,' nothing mandates that the complexity of the system's state—the state being reported—need be thereby reduced to the simplicity of the verbal expression used to report it.

The question of whether "being angry" is an effable state, in other words, is far from straightforward.[7] On the one hand, there is a sense in which "being angry" is almost tautologically *effable*, since we have just demonstrated a way to express it in words: "being angry." On the other hand, at least the first-person phenomenology of anger would surely be called *ineffable*, since the full, nuanced, multi-faceted[8] experience of anger is vastly richer than can anything that can be "captured" in words. What the discussion of face recognition suggests, however, is that the ineffability need not be limited to the first-person case. As the discussion of face recognition suggests, and other modalities would amplify and support, an external observer's sense of someone else's anger may also be indescribably richer than anything that mere words can comprehend. Even if, in terms of figure 3's metaphor, our words are in some sense restricted to labeling what is "above the water line," there is no reason why those of us who *use*

[7]The relationship of effability to second wave AI (and the ontological view espoused in §2) deserves its own paper.

[8]The term 'multi-faceted' implies that the space of angry feelings has conceptually separable dimensions—as does the suggestion that it might be represented in a DL system by a 1,000 dimensioned array. But in the latter case the array might well be best characterized as an *implementation* of a representation of anger; the rich regionality of anger at the level relevant to a person's phenomenal experience may not be higher-order conceptual at all. This is the sort of issue that would need to be explored in a fuller discussion of these issues.

language, even in the midst of verbally expressing our thoughts, ever need to abandon the submarine richness that is leading us to utter those words.

Consider laughter. Suppose someone says "they laughed at him because of what he was wearing," or "she could make anyone laugh, no matter the situation." We hear such statements, and understand them, in part because we understand the word 'laugh'. But what does that understanding consist in? Does it mean that in our thought we merely instantiate some mental analogue of "LAUGHED(X, t)"? Not necessarily. On the mental model we are considering, our reception of the word may index or trigger some or many of the million-dimensional ineffable state(s) that we know, first-person, from ourselves having laughed, and from having interacted with others who laugh. Is that ineffable richness part of the "meaning" or "content" of the word 'laugh'? Has that ineffable richness been *communicated*? On the logicist model the answer might be thought to be *no*. But what second-wave AI systems are suggesting is that that negative answer may not be correct. At a minimum, if the ineffable state unleashed by hearing that someone laughed is used as the grounds for further inference or reflection, there is no reason to suppose that the bare (logicist) conceptual framing "LAUGHED(X, t)" need figure as the premise or argument for any subsequent conclusion. Rational deliberation, that is, may involve the submarine wealth of ineffable detail just as much as perception.

These are really the epistemological promises of second-wave AI. If meaning or content refers to what resources a statement supplies in for a competent hearer to use in reasoning and deliberation, then meaning or content should perhaps be understood as the *richness that utterances of words evoke or signify*, not merely an analogue of the bare skeletal form of the utterance itself. There is no doubt that the successes of DL systems with respect to language translation, inferences from Big Data, etc., suggest that it is these systems' capacity for employing the richness that the networks are capable of representing that leads to their inferential success. Perhaps in fact the whole idea of a "third-person" stance on language, and the conception of what a "competent language user" is, should be questioned. If, for example, "x laughed" can only be understood by a creature who has laughed, or who has had ineffably rich contact with others who laugh, perhaps we need to adopt the idea that language and communication should be understood as second-person plural ("we") phenomena—where "we" means those of us who have laughed.

Second-wave AI is surreptitiously powerful. Epistemology should take note.

6 · Judgment

Given the successes of second-wave AI, should we expect that general purpose AI (AGI) is just around the corner? Will computers soon be as intelligent as humans?

Not remotely. Even in our three examples, the limitations of these second wave systems is as evident as their successes. Alpha Go and its successors may play Go well, at least in some sense of the word "play," but they have no sense of what a game is—no sense that Go was invented in China thousands of years ago, no sense that it has been viewed as a game of such difficulty as to challenge people's sense that it would

ever succumb to computational approaches, no sense of the magnitude of their own accomplishment. By the same token, no current "x-ray reading" systems have any sense of what lung cancer is, no understanding of chemotherapy, no sense of whether an elderly person will want to suffer worse quality of life in order to extend their life expectancy. Similarly, what makes an artwork significant—what brings it together as a whole, what its subject matter or style or expressive quality signifies, etc.—all these things not only fall outside the scope of a deep learning system, not just a little bit, but profoundly and completely.

In fact nothing in current approaches gives these AI systems any sense that the data they deal with *represent anything at all*. They may present us with symbols or images or responses, but in a literal sense they have no idea what they are talking about. They may produce labels associated with particular patterns (*damezumari*, *adenocarcinoma*, *El Greco*) but to them those words are yet more meaningless patterns. They mean something to us, and we configure the systems, tune and debug them, deploy them in specific situations, and feed them with data, in ways that mean something to us.

To use language developed in *Promise*, first-wave AI systems automated a kind of computational *reckoning*—the sort of calculative prowess that we are familiar with from all manner of computational system, including spreadsheets, databases, web searches, and the like. Though their ontological assumptions are different, and their architectures different, and because of those facts they are strikingly more capable in many types of situation, second-wave AI systems are still reckoners—still instruments that amplify our intelligence, but not themselves genuinely intelligent at all.

We humans, in contrast, possess or at least aim for what I call **judgment**—a kind of dispassionate[9] deliberative assessment of a situation, appropriate to the entire context in which it occurs, ethically responsible, and accountable to the world. By judgment, that is, I mean what we get at when we say that someone "has good judgment"—the sort of judgment that we require in a baby-sitter, so that even though it is impossible to put into words—to register in advance—every conceivable situation that could come up (keep in mind the ontological picture adumbrated above), the baby sitter is held accountable for dealing with any situations that arise *in a way that is accountable to the world, to the child being cared for,* etc. Accountability is not ultimately to language, but the world that language registers.

What does judgment require? The bulk of *Promise* is devoted to an explorations of this question. Here I can say just this: it is of an entirely different order from reckoning. Moreover, it will not even be seen, let along understood or approached, through anything like research of the sort that has brought us first- and second-wave AI. I see nothing on the horizon—in scientific or technological or intellectual imagination—that suggests that we are about to construct, or indeed have any ideas as to how to construct, or are even thinking about constructing, systems capable of full-scale judgment:

[9]In the original sense of the world, as not being ruled or swayed by personal emotions, self-interest, etc.

1. Systems existentially committed to the world they register, represent, and think about;
2. Systems that will go to bat for the truth, reject what is false, balk at what is impossible—and know the difference;
3. Systems not only *in* and *of* the world, but *for which there is a world*—a world that *worlds*, in the sense of constituting *that to which all is ultimately accountable*;
4. Systems that know that the world that hosts them, the entities they reason about, and all of humanity and community as well, must be treated with deference, humility, and compassion

It is this kind of judgment, I believe—a seamless integration of passion, dispassion, and compassion—that ultimately underwrites what matters about human cognition.

Where does that leave us? We should be humbled by the inadequacy of first-wave AI, given the depth of the very real insights on which it was based. We should honour, but be cautious about, the successes of second-wave AI—respect and make use of its merits, and appreciate its epistemological implications, but remain forever mindful of its limitations and restrictions. But mostly we should stand in awe of the capacity of the human mind, and of the achievements of human culture, in having developed registrational strategies, governing norms, ontological commitments, epistemic practices, and existential being that allow us to comprehend, exercise judgment, and go to bat for the world as world.