6 — Assessment

How well do machine learning and second-wave AI deal with the four critiques of GOFAI?

6a · Neurological

The neurological critique may be sufficiently addressed. Inspired by what we know about low-level neural organization, contemporary machine learning architectures do to some extent mimic the brain. As suggested in chapter 3, though, the importance of that architectural similarity is not entirely clear. For one thing, it would be premature to assume that what matters for our brains epistemic power is our *general* neural configuration—shared by all higher mammals. That is all that current architectures mimic. Second, as programmers know, it is easy enough to implement one kind of architecture on top of another, albeit sometimes at a significant performance cost. While it is unlikely that evolution would have engaged in architectural mapping at the level of low-level neural circuits, the verdict is still out for the higher level forms of reasoning that are distinctive of our (so far) uniquely human capacities, which seem unlikely to be a simple consequence of general configuration. Third, there is the fact that no one knows whether working in the way our brains do is the only or even best route to general intelligence.[1]

1. If, as may turn out to be the case, the ontological structure of the world is such as effectively to require massively parallel networks for

Nevertheless, the parallelism of ML architectures, and perhaps their statistical abilities to deal with networks of probabilities, will likely be of lasting significance, perhaps particularly at the perceptual level (though it is ironic that ML systems currently take so much computational power to train, given that anything neurally realistic must be slow[2]).

6b · Perceptual

Regarding the perceptual critique, ML systems again seem closer to the mark. Their impressive performance on face ⌜recognition⌝ tasks, for example, is telling. As usual, moreover, the lesson is as much ontological as architectural. According to our best current understanding, it turns out that what makes faces distinctive are high numbers of complex weakly correlated variances across their expanse—the very sorts of feature that ML architectures are suited to exploit—rather than the presence of a few gross characteristics.[3] In general, visual ⌜recognition⌝ of scenes, and allied tasks such as ⌜reading⌝ X-rays, voice identification, and so on, are the types of task at which ML most evidently

---

interpretation, then that would be a reason for AIs to have such architectures—but the logical structure of such an outcome would be that AIs and brains would be similar for the same reason, not per se that AIs need to mimic brains.

2. See the discussion of Feldman's "100-step rule" in chapter 3, note 3 (pp. 23–24).

3. One might wonder whether there may not be a small number of characteristics in terms of which people can be recognized and identified, characteristics that we theorists have not yet discovered—or even that recognizing those characteristics is exactly what successful ML systems do. But if it takes a ML-type architecture to extract those characteristics from records of the physical light profiles reflected from faces, the point may be moot.

excels.[4] Successes in these realms are a substantial part of what has fueled contemporary excitement about the power of second-wave AI.

Still, it would again be premature to extract sweeping conclusions from these results on their own. It is sobering, to take just one example, that many of today's ML image recognition algorithms can be defeated by what to humans seem trivial changes in the images with which they are presented.[5] In the next chapter I will suggest one reason why this might be so.

6c · Ontological

Regarding ontology,[6] the issue is trickier (I will consider

4. *Pace* the standard caveat that calling such accomplishments "recognition" is a possibly culpable shorthand for saying that they are able to learn and repeat mappings between images of individual entities and some other computational structure associated with them.

5. For examples of such "adversarial" examples, see, for example, Athalye et al., "Synthesizing Robust Adversarial Examples," *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden, *PMLR* 80, 2018).

As argued throughout the text, the fact that these adversarial examples work is evidence that current-generation systems may not in fact doing be *perception* at all, in the sense of dealing with an image of a distal situation—but rather something closer to mere image pattern matching, which we interpret as perception. That is, at best it is ⌜perception⌝.

6. As has been clear since the beginning, I use the term "ontology" in its classical sense of being the branch of metaphysics concerned with the nature of reality and being—that is, as a rough synonyms for "what there is in the world." (I defer discussion of the relation between ontics and onto*logy*, technically its *study*, to another occasion.) As in so many other cases, the term "ontology" has unfortunately come to be redefined, in contemporary computational contexts, to refer to structures that *represent* reality: classes, data structural types, concepts, etc.—licensing such otherwise inscrutable constructions as

the epistemological critique presently). Machine learning is not committed to any particular ontological story, so only indirect conclusions can be drawn. As already mentioned, moreover, discussions of second-wave AI typically focus on "uninterpreted" internal configurations (patterns of weights, activation strengths, transformations, etc.), obscuring whatever assumptions are being made about the nature of the world those configurations represent—about which different researchers, moreover, undoubtedly have different views. In addition, the relentless pace of ongoing research makes ML an unstable target of analysis. Still, it is not too early to say that the successes of second-wave AI provide evidential support both for the ontological critique itself and for the metaphysical view introduced in chapter 3.

It is in the realms of perception and action that machine learning most obviously overcomes the limits of formal ontology, and attends to the "subconceptual" terrain suggested in chapter 3's figure 6 (p. 34). When fed with data obtained directly from low-level sensors—visual pixels, haptic signals, and so on—ML systems have vastly improved on the levels achieved in GOFAI, even achieving levels competitive with human performance. This is enabled by a number of factors, including continuous patterns of weights, which allow for incremental adjustment and training, and sufficiently high dimensionality to "encode" all kinds of subtlety and nuance.[7] The resulting systems are particularly im-

"creating an ontology," and "ontological engineering." It is the world itself I am interested in here; I defer questions about its representation to considerations of epistemology.

7. It can be argued that the values of discrete pixels impose a "formal" grid on the array of impinging radiation, and therefore that the data are not genuinely continuous, but even if that were relevant a stream of values does not per se implicate an object, and the value of any such

pressive in not being defeated by intermediate cases, in accommodating noisy data, and in being robust in the face of ambiguity (as well, of course, as being able to be trained)—all of which capacities rely on the fact that they do not have to categorize and discretize their inputs at the outset.

In part, the accomplishments of contemporary systems stem not just from their being oriented toward a wealth of subconceptual detail, but also from their ability to store and work with it, rather than merely attending to it when initially presented—especially to integrate large amounts of information extracted from it into adjusted weights in the activation networks. This ability to ingest vast amounts of detail gives them a leg up on some perceptual tasks, and is critical in allowing them to move beyond what is humanly possible.

Our visual systems, too, seem capable of processing staggering amounts of low-level visual data when immediately presented with it, but it is less easy to imagine that we can store anything close to all of it, once effective coupling with the input is removed. Yet saying anything definite about human information retention is difficult, given our current ignorance about exactly how the brain works. Artists and visually-oriented people display stunning recognition ability for faces and individual scenes not recently encountered, for example—a facility that suggests not only nonconceptual but also informationally dense memories or predictive fabrics of expectation.

Still, in a point of considerable significance not only to AI but also to cognitive science and philosophy, it is commonly assumed that the role of perception, in the human case, is to take in the vast complexity of the perceptual

readout is highly context sensitive, affected by incident illumination, the camera's position and orientation, and numerous other factors.

input and to output a "conceptual parse" of it—a conceptual parse of what is "out there," that is—arrayed in terms of familiar (and effable) ontological categories, no longer "burdened" by the wealth of detail that led to it. Once the perceptual input is categorized, that is, it is assumed, on the GOFAI model in particular and in many (especially analytic) philosophical models of mind, that an intelligent system can *discard the detail that led to that abstractive categorization*, and that reasoning or rationality from that point forward can operate purely in terms of the categories (i.e., purely in terms of sentences, propositions, or data structures categorically framed). This assumption fits into a general story that human categorization is at least in part a technique for avoiding information overload—abstraction in order not to swamp the capacity of the brain.[8] It is also, as we will see, the idea that underlies the Cartesian desire for "clear and distinct" ideas.

The success of second-wave AI suggests that reasoning need not work that way—and that it may not even work that way in humans.

One way to depart from the classical "discard the details" approach to classification is to avoid categories altogether. While we humans may classify other drivers as cautious, reckless, good, and impatient, for example, driverless cars may eschew discrete categories and chunking entirely, in favor of tracking the observed behavior of every single car

8. As technology advances, one might imagine that computational memory will be less easily swamped than our own. Time will tell, though it is sobering that even at today's state of the art, storing high-resolution video streams of all video cameras in operation remains challenging. Still, prospects of incredibly dense computational storage (e.g., DNA-based) will allow us to store many orders of magnitude more information that we do at present. How that will affect our environment and the fate of AI systems no one yet knows.

ever encountered, with that data then uploaded and shared online—participating in the collective development of a profile of every car and driver far in excess of anything humanly or conceptually graspable. Or to consider a different case, BlueDot, a Toronto startup,[9] collects worldwide travel itineraries, including billions of airline itineraries a year, to aid tracking and predicting the global spread of infectious diseases. Whereas traditional epidemiology rests on discrete categories or characteristics (middle-aged man, cancer sufferer, abuse survivor, etc.), no technical reason prevents a ML system from tracking all individual medical records, and dealing solely with vastly dimensioned vectors of real numbers, without any evident need to compartmentalize the data. The promise of "personalized" medicine, medical records with individuals' DNA sequences, and so on, may similarly "get in underneath the categories," to impressive effect.[10]

Even if a network does ⌜classify⌝ something, moreover—as a person, intersection, political dispute, war zone, whatever—it need not do so classically. Nothing in these architectures requires that, in "selecting" some conceptual category, the system must discard the trove of detail that

9. https://bluedot.global

10. There will be analytic challenges in how we understand such systems. Whereas a traditional diagnosis might be phrased as "you have a 52% chance of having melanoma" on a frequentist interpretation (i.e., because 52% of the people in this or that group that you are now identified with have developed melanoma), that may not be an appropriate way to cast a ML system's conclusion. It is not that probabilities will not pertain, just because a system is not dealing with groups; in fact most ML architectures are defined in terms of probabilities. The probabilistic diagnoses they come up with, however (or that we derive from their calculations) may require interpretation in something more like an epistemic measure of certainty: "I am 52% confident that *you in particular* have melanoma, based on what I know."

led to that result—detail that may provide information about the warrant for the classification, inflect it with ineffable shading and modulation, relate it to other concepts (neighboring islands), and so forth. In fact the very claim that the system *has* classified something may merely be a statement on our part, as external observers, that the patterns of weights and activations are "within the region" associated with the discrete labels "person", "war zone," and so on. Unless the system is required to make a hard-edged choice[11] among discrete alternatives, that is—such as to output a discrete token or word, corresponding to our human categories—even the distinction between whether or not a system has classified something need not be sharp.

Moreover, the success of ML systems in cases of simple reasoning shows that retaining and working with the statistical details and correlations derived from a "submarine" ontological perspective can convey substantial inferential power (making the reasoning for that reason at least partially nonconceptual). It is exactly such capacities that empower the widely touted era of Big Data. What is transformative about the present age is not just that we have access to mountains of conceptually represented facts, but that we have developed computer systems with predictive and analytic power enabled by their ability to track correlations and identify patterns in massive statistical detail, without having to force-fit those patterns of relation into a small number of conceptual forms.

11. Technically this should be ⌜choice⌝, but it would be pedantic to mark every possible instance of the distinction. Plus, we all employ what Dennett would call an "intentional stance" in our characterizations of computers (Dennett, *The Intentional Stance*, Cambridge, MA: MIT Press, 1987). I will mark just those cases where it is most important that we resist the tendency to attribute more capacity to the system than is warranted.

> No one of these facts about the successes of second-wave AI is ontologically determinative; none provides invincible evidence of how the world is. But the more successful these systems grow, the more compelling the argument that the "coarse-graining" involved in interpreting the world through articulated concepts and discrete objects (i.e., interpreting it through the lens of formal ontology) is an information reduction strategy for purposes of calculation, reasoning, or verbal communication, rather than corresponding to any definite prior in-the-world discretization. Yes, we may talk as if the world were ontologically discrete; and yes, too, we may believe that we think that way. It seems increasingly likely, however, that such intuitions[12] reflect the discrete, combinatorial nature of language and articulation more than any underlying ontological facts, and also more than the patterns of tacit and intuitive thinking on which our articulations depends.[13]

12. A natural suggestion, from ML architectures, is that intuitions are (distributed) patterns of weights or activations formed in the high-dimensional networked representations of the richly interconnected submarine topologies of the world that underlie our concepts. The common difficulty we have "expressing" them may reflect the fact that words and discrete concepts are excessively bulky, insensitive tools with which to capture their ineffable modulation and subtlety.

13. These lessons are increasingly recognized within AI itself. Rich Sutton, a founder of computational reinforcement learning and leading ML scientist, recently put it this way: "We have to learn the bitter lesson that building in *how we think we think* does not work in the long run. … [T]he actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries. All these are part of the *arbitrary, intrinsically-complex, outside world*. … [T]heir complexity is endless." (Rich Sutton, "The Bitter Lesson," http://www.incompleteideas.net/IncIdeas/BitterLesson.html, emphases added.)

In the face of second-wave AI, in sum, Descartes's idea that understanding must be grounded on "clear and distinct ideas" seems exactly backwards. The successes of ML architectures suggest that a vastly rich and likely ineffable web of statistical relatedness weaves the world together into an integrated "subconceptual" whole.[14] That is the world with which intelligence must come to grips.

✦ ✦ ✦

As always, care must be exercised when drawing such ontological lessons from the current state of the art.

First, we typically feed ML algorithms with data that are already processed, and to that extent "postconceptual": sex or gender selected from a short list of discrete possibilities, experience measured as various forms of unidimensional scalar, videos of traffic at what we humans classify as "intersections," even light intensity coming from some preclassified direction, and so on. Even if it looks on the surface as if the ML system is dealing with the world at a pre- or nonconceptual level, that is, there are many ways in which human conceptualization can sneak it—in ways unaccompanied by clarificatory subconceptual detail.

Questions need to be asked about the origin, appropriateness, bias, and so forth, of all such groupings and factorings—indeed, about the full range of data sets on which such systems are trained. If a ML system is given pixel-level detail about images of people or plants, it may not need to make a binary ⌜decision⌝ about whether some plant is a bush or a tree, or whether a person is brown or white. But if it is being trained on databases of images that have been categorically tagged by human observers, any subcategorical subtlety and traces of prejudicial nuance will in all

14. See my "The Nonconceptual World," unpublished manuscript.

likelihood have been lost, and the system is liable to fall, without ⌜knowing⌝ it, into derivative patterns of bias and prejudice. If fed data from Twitter, Facebook, and similar sources, for example, ML systems famously inherit and reproduce patterns of racism, public shaming, false news, and the like—all without, as it were, batting an eyelash.

We humans are of course affected by the discourses in which we participate, too, but one can at least hope that humans will bring a critical or skeptical attitude to such sources in a way that ML systems are as yet unable to do. Such reflective critical skills are exactly of the sort that I argue cannot arise from ever-more-sophisticated second-wave techniques. Instead, they will require what I will call full judgment.[15]

A second reason for caution in interpreting the success of second-wave AI stems from the fact ML systems are increasingly dedicated to sorting inputs into categories of manifest human origin and utility. To the extent that they are designed to mesh with our categories, even if they retain subconceptual detail, they will nevertheless thereby adopt, and be affected by, those categories' interests, utility, and bias. And if the outputs are discrete categorical

15. As mentioned in the introduction, it is possible that if second-wave AI were used as a basis for a class of synthetic creatures (perhaps along the lines of Sony's Aibos) that were themselves able to evolve, then over a long period of time such creatures might (as we did) eventually develop full-blooded rationality and judgment. See the discussion of "creatures" in chapter 10. The point is just that such a capacity would depend on their forming cultures and community, making and living by commitments, being governed by norms, going to bat for the truth, etc. They would not be capable of judgment, if indeed they ever reached that stage, merely in virtue of supervening on second-wave AI techniques. An *explanation* of their thereby developed normative capacities, therefore, would necessarily advert to more than their being just machine learning or second-wave systems.

classifications, as suggested above, and if we design the systems based on our classical myths about the nature of classification, the abundance of detail on which they rest, and thus any subtleties about the origins and appropriateness of such categorizations, are likely to be lost.

A third caution concerns a topic of some contemporary urgency: increasing calls for ML systems to "explain" their actions may prove to be curiously perverse. The abilities that the systems are being pressed to ⌜explain⌝ may be powerful exactly because they do *not* arise from the use of the very concepts in terms of which their users now want their actions accounted. The pressure to develop "self-explaining" or "interpretable" neural networks, that is, may inadvertently decrease their performance, and drive them toward unwarranted reliance on binary or discrete categories, toward implicit or even explicit reliance on formal ontology—may drive them, that is, back toward the epistemological and ontological inadequacies of GOFAI.

What should we make of all this? While the state of current research is messy, and clear-cut conclusions difficult to draw, I believe three ontological morals can be drawn.

1. The classical assumption of a discrete, object-based "formal" ontology is not a prerequisite of machine learning and other second-wave AI techniques. On the contrary, the success of ML systems, particularly on perceptual tasks, suggests a different picture: that the world is a plenum of unbelievable richness, and that the familiar ontological world of objects, properties, and relations, represented in articulated conceptual representations, is very likely "the world taken at a relatively high level of abstraction," rather than the way that the world is.

2. Much of ML's power stems from its ability to track correlations and make predictions "underneath" (i.e., in terms of vastly more detail than is captured in) the classificatory level in terms of which such high-level ontology and conceptual registration is framed.
3. The fact that ML systems are increasingly being targeted toward domains that have been ontologically prepared by—and targeted for—humans, typically in conceptually structured ways, inevitably leads these systems to inherit both the powers and limitations of human approaches, without any critical faculties in terms of which to question them. It is these factors that are giving rise to the widely discussed (but inappropriately described) phenomenon of "algorithmic bias."[16]

Years ago,[17] as noted in chapter 3, I outlined a picture of the world in which objects, properties, and other ontological furniture of the world were recognized as the results of registrational practices, rather than being the pregiven structure of the world. The picture is useful in terms of which to understand both the failures of GOFAI and the successes of machine learning. It depicts a world of stupefying detail and complexity, which epistemic agents *register*—find intelligible, conceptualize and categorize—in

16. It is the data—not only its form and its content, but also attendant factors about its selection, use, etc.—that is the primary locus of bias in machine learning results. The algorithms that run over the data are undoubtedly not innocent—requiring data sets to be formed in particular ways, etc. But most examples of bias cited in the press and literature are due more to skewed data than to culpable algorithm. We need critical assessments that properly tease apart the respective contributions of these two dimensions of ML architectures.
17. *On the Origin of Objects*, 1996.

order to be able to speak and think about it, act and conduct their projects, and so on. Most importantly, the view was developed from the ground up to take seriously the fact that most of the world—indeed, the world *as* world—outstrips the reach of effective access, necessitating disconnected, semantic representations that, for reasons of complexity, necessarily abstract away from most of its suffusing detail. It leads to a picture of conceptual prowess as most relevant to the intelligibility of relatively more distal situations, and nonconceptual skills as particularly appropriate for the suffusing detail of the immediately nearby. As I put it in another context:[18]

> "I sometimes think of objects, properties, and relations (i.e., conceptual, material ontology) as the long-distance trucks and interstate highway systems of intentional, normative life. They are undeniably essential to the overall integration of life's practices—critical, given finite resources, for us to integrate the vast and open-ended terrain of experience into a single, cohesive, objective world. But the cost of packaging up objects for portability and long-distance travel is that they are thereby insulated from the extraordinarily fine-grained richness of particular, indigenous life—insulated from the ineffable richness of the very lives they sustain."

Both the successes and limitations of first- and second-wave AI make eminent sense in terms of this picture—predictable characteristics of architectures beginning to extract the rich but radically simplifying registrations fundamental to perception and cognition.

18. "The Nonconceptual World," unpublished manuscript.

6d · Epistemological

What then about epistemology—subject matter of the remaining GOFAI critique?

Here the rubber finally meets the road. Two major issues, to be addressed in the next chapter, stand in the way of AI's reaching anything that can truly be called *thinking*. Both have to do with what is involved in holding thinking and intelligence accountable to the fabulously rich and messy world we inhabit. One, relatively straightforward, involves reconciling the first- and second-wave approaches—taking advantage of their respective strengths, and moving beyond at least some of their limitations. This integrative goal is starting to be recognized, and to be suggested as a necessary ingredient for third-wave AI.

The other challenge is more profound. I do not believe that any current techniques, including any yet envisaged as a subject matter of AI research, even recognize the importance of this second issue, let alone have any idea of what would be involved in addressing it. Explaining it will take us into realms of existential commitment and strategies for dealing with the world as world.