Introduction

Neither deep learning, nor other forms of second-wave AI, nor any proposals yet advanced for third-wave, will lead to genuine intelligence. Systems currently being imagined will achieve formidable reckoning prowess, but human-level intelligence and judgment, honed over millennia, is of a different order. It requires "getting up out" of internal representations and being committed to the world *as world*, in all its unutterable richness. Only with existential commitment, genuine stakes, and passionate resolve to hold things accountable to being in the world can a system (human or machine) genuinely refer to an object, assess ontological schemes, distinguish truth from falsity, respond appropriately to context, and shoulder responsibility.

Does that doom AI? No. Automated reckoning systems will transform human existence. But to understand their capacities, liabilities, impacts, and ethics, and to understand what assemblages of people and machines should be assigned what kinds of task, we need to understand what intelligence is, what AI has accomplished, and what kinds of work require what kinds of capacity. Only with a carefully delineated map can we wisely choreograph the world we are developing—the world we will jointly inhabit.

This book is intended as a contribution to that cartographic project. It is written out of a belief that the rise of computing and AI is of epochal significance, likely to be as consequential as the Scientific Revolution—an upheaval

that will profoundly alter our understanding of the world, ourselves, and our (and our AIs') place in that world. So encompassing is the reconfiguration being catalyzed by computing and AI that we need fundamentally new ontological and epistemic frameworks to come to terms with it—new resources with which to ask such ultimate questions as *who we are*, *who to be*, *what to stand for*, *how to live*.

With an eye toward such questions, this book develops some intellectual tools with which to assess current developments in AI—especially recent advances in deep learning and second-wave AI that have led to such excitement, anxiety, and debate. The book will not assess specific projects, or recommend incremental advances. Instead, I will adopt a general strategy, in pursuit of two aims:

1. To unpack the notion of intelligence itself, in order to understand what sort(s) we humans have, what sort(s) AI aims at, what AI has accomplished so far, what can be expected in the foreseeable future, and what sorts of tasks can responsibly be assigned to systems currently constructed or imagined.
2. To develop a better understanding of the underlying nature of the world, to which I believe all forms of intelligence are ultimately accountable.

The second concern gives the book a strongly ontological flavor. In the end, I argue (i) that the deepest reason for the failure of first-wave AI was the untenable ontological world view on which it was founded, (ii) that the most important insight of second-wave AI is the window it gives us onto an alternative ontological perspective, and (iii) that the nature of reality implies that constructing anything warranting the label "artificial general intelligence" (AGI) will require developments far beyond and quite different

from those imagined in either first- or second-wave AI,[1] involving various forms of committed, participatory engagement with the world.

I use **judgment** for the normative ideal to which I argue we should hold full-blooded human intelligence—a form of dispassionate[2] deliberative thought, grounded in ethical commitment and responsible action, appropriate to the situation in which it is deployed. Not every human cognitive act meets this ideal—not even every conscious act of every person we call intelligent. I claim only that judgment is the standard to which human thinking must ultimately aspire.[3]

Judgment of this sort, I believe, is a capacity we strive to instill in our children, and a principle to which we hold adults accountable. It is an achievement that far transcends individuals—a resource that has been forged, over

1. Or any that have been proposed for third-wave AI.

2. "Dispassionate" (and "disinterested" when I use the term) in its original sense of being fair, unbiased, open-minded, and free of prejudice. I neither mean nor intend to suggest that judgment should (or can) lack in care or commitment. On the contrary, as argued in chapter 11 and again in chapter 13, judgment must be simultaneously passionate, dispassionate, and compassionate.

3. This is not to suggest that judgment, as a regulative ideal, is remote from either consciousness or experience. One of the most important accomplishments of the historical development of culture, I believe, in diverse forms around the world, is to have established standards of judgment as a background condition on what it is to be a responsible adult. The fact that contemporary rending of the fabric of public discourse (perhaps abetted by digital technologies) is so widely decried stands witness to the fact that such norms have not been forgotten, even if they appear to be under threat.

Coming to understand what it is or would be to hold the human condition accountable to such an ideal is an added benefit of documenting what AI systems are, and what they are not.

thousands of years and in diverse cultures,[4] as a foundation for rationality, thought, and deliberative action, into which individuals must be recruited. It need not be articulate, "rationalistic," or independent of creativity, compassion, and generosity—failings of which (especially formal) logic is often accused. Rather, by judgment I mean something like what we get at (or should be getting at) when we say that someone "has good judgment": a form of thinking that is reliable, just,[5] and committed—to truth, to the world as it is.

With judgment in view as the ultimate goal of general intelligence, I examine the history of artificial intelligence, from its first-wave origins in what Haugeland dubbed "Good Old Fashioned AI" (GOFAI) to such contemporary second-wave approaches as deep learning. My aim is neither to promote nor to criticize—but to understand. What assumptions underlie the various technologies we have constructed? What conceptions of intelligence have been targeted at each stage? What kinds of success have been achieved so far, and what can be expected in the future? What aspects of judgment will contemporary AI systems reach, and what aspects have they not yet begun to

4. The development of judgment, we can safely presume, was accomplished without requiring alteration in DNA or neural architecture.
5. Philosophical readers may balk at the inclusion of justice and ethics in a norm on truth. As will become increasingly clear, I take *accountability to the world* to be something of an "ur-norm" that underlies not only truth but ethics, care, and compassion as well. To see how and why that is true, however, requires understanding what the "world" is, how ontology and truth arise, how existential commitment is a precondition for ontology, and so on. An argument for this metaphysical position is beyond the compass of this book; I will take up that project elsewhere, but for an initial glimpse see my *On the Origin of Objects* (Cambridge, MA: MIT Press, 1996).

approach? And to up the ante, and in order to bring into view one of the most important issues facing us today: can articulating a conception of judgment provide us with any inspiration on how we might use the advent of AI to raise the standards on what it is to be human?

I use the term **reckoning** for the types of calculative prowess at which computer and AI systems already excel—skills of extraordinary utility and importance, on which there is every reason to suppose computers will continue to advance (ultimately far surpassing us in many cases, where they do not do so already), but skills embodied in devices that lack the ethical commitment, deep contextual awareness, and ontological sensitivity of judgment. The difference between reckoning and judgment, which I will argue to be profound, highlights the need for a textured map of intelligence's kinds[6]—a map in terms of which to explain why reckoning systems are so astonishingly powerful in some respects, yet fall so spectacularly short in others.

Four caveats. First, the book is not a comparison of humans and machines. I see no reason to doubt that it may someday be possible to construct synthetic computational systems capable of genuine judgment. Or perhaps equivalently: if, as may happen, we construct synthetic creatures capable of evolving their own civilizations, or of incrementally participating in ours, nothing in my argument

6. Psychology has developed detailed conceptual maps of the human psyche, distinguishing such capacities as cognition, sensation, memory, and so on. As noted in chapter 2, the conception of intelligence on which AI was founded was very general, not making any such theoretical distinctions. I believe the kind of map we need in order to assess AI, moreover, is of a different order from that provided by psychology—in part because computers are opening up vast regions unoccupied by humans or nonhuman animals.

militates against the possibility that, in due course, such creatures might themselves evolve judgment—much as we have, and perhaps with no more explicit understanding of their capacities than we have of our own. Nor am I arguing that cyborgs or other human-machine assemblages, *for that reason alone*, will be challenged in regard to their capacity for authentic judgment. My claims are just two: (i) the systems we are currently designing and building are nowhere near that point; and (ii) no historical or current approaches to AI, nor any I see on the horizon, have even begun to wrestle with the question of what constructing or developing judgment would involve.

Yet attempting to reach this conclusion by drawing a distinction between DNA- and silicon-based creatures would be a grave mistake, in my view—chauvinist, sentimental, and fatally shallow. Rigor demands that we articulate a space of possible kinds of intelligence in terms of which AIs, humans, and nonhuman animals can all be evenly and nonprejudicially assessed.

Second, as will progressively emerge, judgment in the sense I am defending it is an overarching, systemic capacity or commitment, involving the whole commitment of a whole system to a whole world. I do not see it as an isolable property of individual capacities; nor do I believe it is likely to arise from any particular architectural feature—including any architectural feature missing in current designs. Readers should not expect to find specific architectural suggestions here, or recommendations for technical repair. The issues are deeper, and the stakes higher, than can be reached by any such approach.

Third, I am fully aware that the conception of judgment I will defend does not fit into any standard division between "rational thought," on the one hand, and "emotion"

or "affect," on the other. On the contrary, one of my aims is to unsettle reigning understandings of rationality, in part to break it up into different kinds, but also to suggest that reason in its fullest sense—reason of any sort to which we should aspire—necessarily incorporates some of the commitments and compulsions to action that are often associated with affectual states. These moves arise out of a larger commitment: if we are to give the prospect of AI the importance it deserves, we must not assume that time-honored conceptions of rationality will survive unscathed.

Fourth, although I take seriously many of the critiques of first-wave AI articulated in the 1970s, this book is by no means intended to be an updated treatise along the lines of Dreyfus's *What Computers Can't Do*.[7] On the contrary, one of my goals is to develop conceptual resources in terms of which to understand what computers *can* do. In fact, the entire discussion is intended to be positive. I do not plead for halting the development of AI, or argue that we should bar AI systems from functioning in situations of moral gravity. (When landing in San Francisco, I am glad the airplane is guided by sophisticated computer systems, rather than by pilots peering out through the fog looking for the airport.) I am also not worried, at least here, about whether AI systems will grow more powerful than we humans, or that they will develop their own consciousness. And I take seriously the fact that we will soon need to learn how to live in productive communion with synthetic intelligent creatures of our own (and ultimately their) design.

Two things do terrify me, though: (i) that we will rely on reckoning systems in situations that require genuine judgment; and (ii) that, by being unduly impressed by

7. Hubert Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper & Row, 1972).

reckoning prowess, we will shift our expectations on human mental activity in a reckoning direction. Current events give me pause in both respects. The calls to which I believe we should respond, and to which I hope this book will draw our attention, are: (i) that we learn how to use AI systems to shoulder the reckoning tasks at which they excel, and not for other tasks beyond their capacity; and (ii) that we strengthen, rather than weaken, our commitment to judgment, dispassion, ethics, and the world.