From Thin to Thick
Toward a Politics of Human-Compatible AI

JACOB G. FOSTER

When you read the *New York Times* or—let's be honest—doomscroll your Apple News feed, you often come across breathless reporting about the latest advances in artificial intelligence (AI). Perhaps the eerie and often beautiful images generated by DALL-E or Midjourney; perhaps the ongoing debate about when large language models will attain sentience (if they haven't already).

In counterpoint with this largely positive reportage is a different voice: the voice of critique. Most such critiques follow one of two lines of full-throated criticism. The first line focuses on the many harms of actually existing AI technologies, from credit scoring algorithms to facial recognition and deep fakes—harms that disproportionately affect marginalized groups (Buolamwini and Gebru 2018; Benjamin 2019a, 2019b; Bender et al. 2021). The second line focuses on so-called existential risks of imagined, far-future AI technologies (Bostrom 2017): the theophanic apocalypse of the Singularity or the comic apocalypse of the world-as-paperclip-factory.[1]

In recent years, a third voice has joined the critical polyphony, linking these two approaches. It calls for the development of "human compatible" artificial intelligence (Russell 2019). This critique rightly recognizes that the transformational effects of future advances in AI demand a reimagining of the foundations of the

I am profoundly grateful to Alondra Nelson, Charis Thompson, and all the participants of the Science and the State seminar at the Institute for Advanced Study for months of inspiration and intellectual stimulation. This paper could only have sprung up in that rich soil. I thank Phil Ford and J. F. Martel of *Weird Studies* for countless hours of nourishing content and conversation, for renewing my acquaintance with Henry Corbin, and for their compelling exegeses of Philip K. Dick, Ishmael Reed, and Abeba Birhane. I presented the first fruits of my newfound anarchist squint at a 2021 workshop on Collective Intelligence in Natural and Artificial Systems at the Santa Fe Institute; these ideas ripened further at the 2022 Datatopia colloquium at the University of Michigan. I am grateful to both of these audiences for their feedback and encouragement. Finally, I thank Lee Smolin and the Perimeter Institute for introducing me (literally and figuratively) to Roberto Unger, almost two decades ago. That long-dormant seed is finally sprouting.

1. An advanced AI is tasked with maximizing the productivity of a paperclip factory. To satisfy its objective, the AI uses its superior intelligence to seize control of all matter and energy on the planet, so it can make more paperclips. Hilarity ensues. Clippy is avenged.

field. Its conception of "human compatible," unfortunately, is rather thin and apolitical; it largely positions the detection and satisfaction of individual human preferences as its universal goal.

In this article, I want to offer another vision of human compatibility[2] in AI: one that embraces a thick and demanding world of human capacity, social complexity, and local politics in place of the thin, pliable, universalizing world of individual preferences (Walzer 2019).[3] This vision, too, demands a wholesale reimagining of the foundations of AI research. It goes deeper, however, in requiring a change in both core technical paradigms and established modes of organizing technoscientific practice. It peers beyond the field of AI, finding another (albeit less well-funded) field whose foundations need to be reimagined: social science. I first (briefly) describe current AI paradigms and practices before offering a series of reflections on thick human compatibility—and the threats that current AI paradigms pose to it. I close with a vision of social science as the exploration of the infinite space of social possibilities.

Where We Are: AI in Technoscientific Practice

Contemporary AI research is utterly dominated by machine learning (ML). Rather than programming a computer to do something, scientists and engineers program a computer to *learn* to do something. This involves a knot of decisions about what task(s) the computer will learn to do and how to measure success; above all, it involves a mountain of data. To simplify radically, most advances in AI over the past decade have depended on training larger and larger neural networks on more and more data (Goodfellow, Bengio, and Courville 2016). Training those models is energy and resource intensive (Schwartz et al. 2020). So is collecting and curating the training data. This, alas, often depends on uncompensated or poorly compen-

2. I am aware that framing this essay around an undertheorized category of the human is a bit old-fashioned. My interlocutors on the AI side of the aisle typically adopt this framing, and I strategically follow their lead to maximize opportunities for dialogue. That said, both my line of critique and the alternative possibilities I suggest embrace and honor the well-being of our more-than-human fellow creatures.

3. Walzer's thick/thin dichotomy riffs on the notion of thick description articulated by his IAS colleague Clifford Geertz (1973). Here is how Walzer (2019: xiii) describes the debt: "I have borrowed the idea of thickness from Clifford Geertz's defense of 'thick description.' . . . But it is not my claim to offer a thick description of moral argument, rather to point to a kind of argument that is itself 'thick'—richly referential, culturally resonant, locked into a locally established symbolic system or network of meanings. 'Thin' is simply the contrasting term." In this essay, I am not offering a thick description of AI research practice or of human compatibility; rather, I am calling for a notion of human compatibility that is "richly referential, culturally resonant" and, above all, "locked into a locally established symbolic system or network of meanings"; really, a thick notion of human *compatibilities*.

sated labor; for example, the accumulation of personal data through surveillance capitalism (Zuboff 2015) or the "ghost work" required to annotate, clean, and otherwise render legible the mess of found or gathered data (Gray and Suri 2019).

It is unsurprising, then, that the "bigger is better" demands of the technology are coupled with a distinct social organization of technoscientific labor, within and beyond the academy. Contemporary ML practice is dominated by "benchmarking." In this practice, tasks (things we might want a computer to do, like recognize and label individual objects in a scene) are represented by specific "benchmark" datasets and well-defined (but often very narrow) metrics for evaluating the success of an algorithm on that task. The benefits of this paradigm are clear: it is trivial to compare algorithms on a common benchmark and hence to assess "progress" on that task. The costs can be substantial. Insofar as benchmarks (datasets, metrics, and task operationalizations) depart from the details of the real-world problem, purported progress can be illusory and real-world deployments may go catastrophically wrong. These disastrous deployments often affect marginalized groups, whose data are underrepresented in critical benchmark datasets (Buolamwini and Gebru 2018).

In collaboration with Remi Denton (Google AI Ethics) and Alex Hanna (DAIR), my student Bernie Koch and I conducted the first systematic, quantitative analysis of benchmarking practices across machine learning (Koch et al. 2021). Thanks to a remarkable open source repository called "Papers with Code" (PWC), we examined patterns of dataset usage across machine learning task communities and across time from 2015 to 2020 (don't be shocked by the short time span; machine learning moves quickly). We found that task communities increasingly concentrate their attention on fewer datasets, with staggering levels of inequality in attention. While task communities create their own datasets at a reasonably high rate, they tend to *use* datasets borrowed from other tasks; such borrowing often decreases the ecological validity of the benchmark for the task at hand. Most remarkably, we found that a handful of institutions—twelve top universities and tech companies—produced the benchmark datasets that accounted for the majority of usages, across more than twenty-six thousand papers.[4] It goes without saying that these datasets reflect the interests and values of the elite scientists and corporations who organize and subsidize their production; sometimes indirectly, and sometimes quite directly.

As the thumbnail sketch above suggests, contemporary AI—as a technology and as a field of inquiry—overwhelmingly participates in a modernist political imaginary that favors scale, standardization, and concentration. It is a happy quisling of

4. With different inclusion criteria for benchmarks, the number of institutional producers drops even further, to a mere seven.

an authoritarian state or a globe-striding colossus of techno-capital. In its present form, this is a technology that can be humanized, at best (Unger 2005). It will happily participate in the better detection and satisfaction of human preferences. But it will never be human compatible; at least not in a thick sense.

What does thick human compatibility look like? What changes might it demand in the technological underpinnings of AI? I offer a few thoughts below, drawing primarily on James Scott's (2012) essays on anarchism. This "anarchist squint" will help us see more clearly the perils of contemporary AI and associated research practices.

An Anarchist Squint on Human Compatibility

AGAINST STANDARDIZATION; FOR VERNACULAR AI

It is in the nature of modernist institutions (Scott 1998) to crush local variation in the name of standardization and legibility (where *legibility* means legible to the powerful, at the center). But vernacular traditions, knowledges, and practices often fit local circumstances in ways that become apparent only too late, after they have been swept away in the modernist tide (Scott 2012).

The current trend in AI is human incompatible insofar as it rests on a handful of massive, unrepresentative datasets and gigantic "foundation models"—standard, centrally produced solutions to core algorithmic challenges that serve as building blocks for larger systems (Bommasani et al. 2021). This mode of AI practice demands standardization and legibility; it is only through such strategies that engineers can reach the scale demanded by current architectures.[5]

Massive benchmarks and the (foundation) models trained on them also impose a particular semantics on the world (discussed further below); the optimal label or action for situation X is the one that follows from a foundation model or some other system trained on the largest, most utilized dataset.[6] Any person or organization that needs to interface with these systems benefits from aligning with their classificatory schemes, their way of seeing the world. The vernacular world deforms—

5. Foundation model advocates would point to fine-tuning as a way to extend standardized infrastructure to particular, "vernacular" challenges. In this strategy, a foundation model is pretrained on gigantic datasets and then subsequently *fine-tuned* to adapt it to a particular domain. But as the authors of a prominent review on foundation models point out (Bommasani et al. 2021), better performance is often achieved by training *only* on the relevant data, task, or domain (where possible); in other words, there is a cost to the strategy of pretraining and fine-tuning. Of course, the dataset for fine-tuning also needs to be built, labeled, annotated, and so on, which can be costly or (potentially) impossible.

6. Perhaps after fine-tuning, in which case the *fine-tuning* dataset also imposes its semantics.

perhaps subtly, at first—to fit the standardized infrastructure. The future is funneled into the narrow and often deeply unjust categories encoded in past data (Birhane 2021). Insofar as humans and our interactions (with the world and with each other) are characterized by particularity and by open-ended "ambiguity" and "fluidity" (Birhane 2021), this is antihuman.

Far better is an AI that learns to represent local vocabularies, traditions, knowledges, and practices; that can translate between these local worlds, in full knowledge of the violence involved in any translation (Steiner 2013); that embraces human dynamism, fluidity, and ambiguity (Birhane 2021). This thick approach demands data-efficient AI systems that can learn from a handful of examples; it prizes transfer and flexibility, so that standardization and scale become irrelevant. And it pushes against the dominance of a handful of benchmarks, urging a technoscientific practice oriented toward a wide variety of datasets carefully tailored to the local worlds where the system in question will operate.

MANY OBJECTIVES, NOT ONE

As Jane Jacobs realized long ago, many of the institutions and organizations in which we participate (like neighborhoods) actually involve many objectives, at least when they are healthy (Jacobs 1961). It is a characteristic defect of the modernist mind to decide what the one objective or goal should be and to maximize that goal. This same defect is reflected in AI benchmarking practices, which emphasize comparative performance on *one* metric when assessed against *one* standard dataset. It is on display when researchers carve up the world into tasks; never mind the fact that the same situation might mean radically different (and perhaps incompatible) things to different people.

Although multi-objective optimization is an area of active investigation, the main obstacles are less technical than conceptual. Stuart Russell and other mainstream human compatibility researchers are right to *start* by acknowledging the diversity of human preferences (which must be learned from diverse humans, not imposed by system designers). But the notion of preference is far too narrow.[7]

Thick human compatibility requires a way to talk about the different meanings humans make, the many projects they pursue, and the radical challenges of rec-

7. It is beyond the scope of this essay to explore the genealogy of the notion of preferences in AI research; I would gesture to the field's close intellectual symbiosis with economics—manifest in the leading textbook (Russell and Norvig 2021). This framework of preferences and preference satisfaction is too conceptually thin for our purposes; hence the call for a richer (or thicker) conceptual vocabulary.

onciling different preferences, values, meanings, and projects. This challenge goes far beyond a weighted sum of several objective functions. It requires a new, more human vocabulary at the foundation of AI research. It requires an understanding of preferences, values, meanings, projects, and objectives as coconstructed, dynamic, context dependent, and negotiated. Without such vocabulary and vision, the dystopian tendency in current AI technologies could produce a grid of preferences that are imposed, static, universal, and nonnegotiable: Philip K. Dick's (1981) Black Iron Prison.

HICKSIAN AI AND THE PRODUCTION OF PERSONS

Organizations and institutions shape the people who participate in them. Whatever their other business, from education or healing to search and advertising, they participate in the social production of persons. That social production can be good or bad. A student enters economics graduate school. She is perfectly sensible. By the time she graduates, she would sacrifice part of her thumb—past the knuckle—to publish in a top journal like *American Economic Review* (Attema, Brouwer, and Van Exel 2014). This transformation makes her fit better into the discipline and the university. Is it good for her in the long run? Is it good for science in the long run?

We must ask the same questions about AI systems, and the new organizational and institutional forms that involve AI as critical components. The notion of Hicksian income provides one way to think about these questions (Scott 2012). It essentially says that—when evaluating a firm or organization—you need to track what happens to its factors of production (Hicks 1946). If the firm in question is damaging its factors of production—destroying the climate, degrading the soil, reducing the capacity of the labor force, attenuating their ability to participate in civic or community life—that counts against it in terms of Hicksian income.

We are just *beginning* to ask Hicksian questions of AI systems, when it comes to people or to the many other factors of production that they currently consume and degrade, from climate to culture. It is entirely possible to satisfy preferences while ultimately degrading people. Indeed, pure preference satisfaction is likely to do just that; this is why Odysseus famously lashed himself to the mast, as his preference in the moment he heard the Sirens' song would have been to dash himself and his men on their rocks. It is very likely that the preferences being constituted and satisfied by current AI-enabled technologies are degrading our mental health, our attention, and our capacity for civic participation; we dash ourselves and our societies on the rocks each day.

Human compatible AI, in the thick sense, must put human flourishing at the center of its evaluative calculus. It must build human capacity, not destroy it. But what sort of capacity? And who decides?

FROM ANTIPOLITICAL AI TO POLITICAL AI

Modernist institutions love objective, quantitative schemes of evaluation and classification: SAT scores, h-indices, cost-benefit analysis, and so on. These schemes are *antipolitical* devices (Scott 2012). They seek to translate questions of judgment, debate, and disagreement into (apparently) judgment-free, unbiased procedures: expertise, crystallized. As those metrics, schemes, and categories are established, politics is smuggled in. Nor is the transparency of evaluative or classificatory schemes a defense; the politics is in their construction, not in their application. Once they are available for all to see, it is too late.

Contemporary AI technologies indulge in this same antipolitics of "objective" evaluation and classification. Early critique targeted the opacity of such AI-powered classifications. The fieldwide response—to promote "interpretability" or explainability—is insufficient, at least in its spare form. Concepts like local explanation (what features caused the output) or counterfactual faithfulness (what would I have to do to change the outcome) are merely an algorithmic misdirection for the same old sleight of hand (Mohseni, Zarei, and Ragan 2021). You learn how to play the game, but the problems start long before the game does. The rules are rigged.

What is the broader point? Human compatible AI should avoid creating antipolitical machines and instead promote healthy debate and deliberation—informed by but not deferring excessively to expertise. This is especially so when it comes to the allocation of scarce resources or grave harms.

Science can be viewed as a domain of "reversible time" (Eyal 2019). It is a protected space for debate, revision, disagreement. The world of policy is, in Gil Eyal's (2019) language, a world of "irreversible time." Decisions must be made, they must be made quickly, and they are rarely revisited. The "policy sciences" connect reversible time to irreversible time, digesting scientific debates into actionable, "frozen" insights. As the term *frozen* suggests, we might switch metaphors and call policy a "low temperature" or "low energy" regime—with science a "high temperature" or "high energy" regime (Unger 2005).

Being for the vernacular, for many goals, for human capacity, and for politics—all things we need in a time of high global temperature—means resisting artificial intelligence that serves as a "policy science," freezing and reifying scientific and

real-world debates.[8] Instead, we need human compatible AI that unleashes participation, creativity, entrepreneurship; the domain of politics. Make the world like science at its best, rather than science at its worst.

What Is at Stake

When seen through an anarchist squint, the perils of contemporary AI practice become clear. It starts with the same old modernist drive to uniformity, legibility, scale, and stasis. Yet that drive is especially perilous in its AI incarnation because of the unprecedented capacity of those technologies to reach into and rework local worlds of meaning; to deliver classifications, judgments, and rankings with unprecedented speed and reach (Fourcade and Johns 2020). Because current AI practice demands computer power, massive datasets, and hence considerable resources, its iteration of the modernist drive is allied of necessity and in complicity with late capitalist impulses toward scale and concentration.

Contemporary AI risks making the world thin: lacking in reference or cultural resonance, floating free of locally established symbolic systems or networks of meaning (Walzer 2019). To see what this means—what is really at stake—I draw from two concepts from distant disciplines: the concept of niche construction, from contemporary evolutionary biology; and the concept of semantics, as operationalized in recent machine learning scholarship.

CONSTRUCTION, THICK AND THIN

We can see both AI systems and the humans subject to them through the lens of *adaptation*. In the world of theoretical biology and complexity science, *adaptation* is any process that leads to a greater "fit" between a system and its world (Krakauer and Rockmore 2015). Such fit can be achieved in two ways: either the system can change to fit the world, or the system can change the world to fit it. The former maps onto familiar processes of learning and vanilla evolutionary dynamics. The latter maps onto a less familiar process called *niche construction*. Construction can make the world more predictable and legible; it can also change others' behavior to benefit the constructing system (Laland, Boogert, and Evans 2014).

8. Consider one recent example: large language models like ChatGPT, Bard, or Sydney. These AI systems will confidently give (incorrect) answers to questions, without citations; Bard famously declared that the James Webb Space Telescope "took the very first pictures of a planet outside of our own solar system," which it definitely did not (Sparkes 2023). This mode of question answering is fundamentally antipolitical; it turns the AI into an oracle. A more political AI system would express uncertainty, surface debate, and point users toward references and resources for further exploration, engagement, and feedback.

Humans—both individuals and groups—deploy both of these approaches to adaptation. We learn the features, concepts, categories, and norms of our physical and social worlds. But we also relentlessly modify—or *construct*—our physical, mental, and social worlds to suit our ends. Informal norms provide an excellent example; they render our social worlds more legible, navigable, and learnable by reducing the vast space of possible social actions and interactions to a few options (Bowles 2009). In many cases—especially when norms are vernacular and specific to a place and a set of people—their details are the product of, and subject to, continued negotiation. They change, with time and with context. And the same is true of many other forms of niche construction—whether remolding the physical, mental, or social environment—insofar as they are *thick* in Walzer's sense; insofar as they are coconstructions whose authors vary only so much in their power (that is, their ability to impose learning on others and construct environments to their advantage).

Thin niche construction, by contrast, is the product of vast power disparities. The local, complex, and political cannot be sustained when some have the power to construct the social world, while others must learn and adapt to the choices of the powerful. The power disparities inherent in present AI practice and its capitalist implementations make AI an instrument of thin construction. This is especially sobering when we consider classic analyses of biological niche construction (Jones, Lawton, and Shachak 1994). These note that the consequences of construction become more significant when constructors are very active (billions of operations a second!), when they are very numerous (billions of algorithms with a handful of handlers), when their actions can accumulate through artifacts or other long-lasting modification (AI-driven changes to our semantic environment, our mental furniture), and when they affect many or various resources (jobs, loans, information, freedom).[9]

SEMANTIC POWER

What does it mean to construct the world? In the idiom of AI, it is to change the objective function. When I am in situation s and perform action a, the objective function $R(s,a)$ assigns some value (possibly probabilistic) to that combination.[10]

9. You might wonder if the "third way" human-compatibility approach of Russell and others provides an escape; it does not. While AI designed to satisfy human preferences might *seem* as if it is merely learning our preferences, it is in fact constructing both individuals and social worlds in a thin, universalizing, and atomizing fashion. Where is the social, where the political, in the vast enterprise of preference satisfaction? This is the antihuman constructor which, like the Devil, has tricked us into believing that it doesn't exist.

10. To those who abandon all hope at the sight of equations, I offer the standard advice of "humming" along. I have endeavored to make the argument accessible without any of the formalism.

Insofar as people care about those outcomes, they will learn to take actions that lead to higher values, according to the objective function.

You can see, I hope, where this is going. As AI systems have more and more influence on our objective functions, they can shape our behavior quite invisibly. But they can also shape our semantics—the way we assign meaning. Explaining this will require a brief technical digression; here I follow David Balduzzi (2016). In the spirit of pragmatist theories of social action (Gross 2009), we can define $f: S \rightarrow A$ as the "action map" that sends situations to actions, which might be literal actions (buy! sell!) or labels (hotdog! not hotdog!). On an (admittedly extreme) pragmatist reading, the *meaning* of a particular action α is the (sub)set of situations that produce such an action, that is to say, $S \supset f^{-1}(\alpha)$. Here the "inverse" of the action map $f^{-1}(\alpha)$ sends us from an action α to the situations that produce it, which are a subset of all possible situations. So, for example, if S is a set of images and α is the label "hotdog!" then $f^{-1}(\alpha)$ will return a subset of images that collectively represent whatever the map $f: S \rightarrow A$ "means" by hotdog (actual hotdogs, but maybe also people in hotdog costumes, the Oscar Mayer Wienermobile, and so on).

Whoever controls the objective function—whether a government agency or an AI-driven platform—begins to control meaning. How does this happen? With a certain looseness of detail, we can parameterize the action map $f_\theta: S \rightarrow A$ to capture a space of possible maps from situation to action. For a given objective function $R(s,a)$, there is some action map f_{θ^*} that leads to optimal value. And this optimal map then leads to a sort of "optimal semantics" for any particular action $\alpha: S \supset f_{\theta^*}^{-1}(\alpha)$. Slowing down a bit, this means that an optimal action map f_{θ^*} (induced by a particular objective function) will mark out a specific set of situations as "correct" meanings of the action α (that is, situations where the action should be produced). If the action map is being used to select food items for purchase, the corresponding objective function might penalize any mislabeling of nonfood items with food labels. The "optimal semantics," in this case, would limit the meaning of "hotdog" to the food item, excluding the inedible but indelible Oscar Mayer Wienermobile.

In the world of thick construction, objective functions and thus meanings are constantly debated, negotiated, and co-constructed; objectives and meanings are multiple, overlapping, and at times contradictory. In the world of thin construction, overwhelming power—the ability to impose an objective function—leads to semantic power: the ability to impose meaning. The force of semantic construction is amplified in the case of contemporary AI technologies because they rely on a handful of key datasets and foundation models. In other words, these technologies all end up favoring similar objectives and meanings, reducing the variety of action maps (and hence semantics) that construct our world.

And so is the world made thin: semantic power dilutes rich reference, deadens cultural resonance, abolishes local symbolic systems, and unravels networks of meaning.[11]

"We Give Not What You Want, but What You Need" (The Tree Shamans, *Centaurworld*)

I hope that I have convinced you of the peril more than incipient in contemporary AI technologies: a capacity to subtly and pervasively reconstruct our social and mental worlds in ways that make even the most ambitious modernist state actors look like slackers or petty tyrants. I began by pointing to the roots of this problem in current research practices and technical imaginaries. And I share with Stuart Russell and others in the community devoted to "human compatible AI" a conviction that we need to remake the foundations of AI research.

But I think we need to go further. As social scientists, we should ask whether we have the *social science* we need to set AI on a better course. After all, the current human compatible approach draws heavily on the most formal wing of positive social science: economics. Economics speaks the language of preferences, utility, optimization, and games. This is what AI researchers want: a ready-to-hand mathematical tool for building "human compatibility" into AI. But is it what they need?

If AI has the transformational potential that its advocates claim, the answer must be no. In fact, *no social science* is up to the task of imagining the radical social possibilities inherent in AI.[12] This is because of how contemporary social science treats change. The old social theories committed the unforgivable sin of necessitarianism (Unger 2004); although they acknowledge that the social world is socially constructed, they incoherently read into that construction laws of social evolution, with societies proceeding through inexorable developmental stages toward some more or less clearly imagined end of history. The positive social sciences, by contrast, render change with a modesty that borders on timorousness (Unger 2004); they might help us understand local adjustments around present arrangements, but they do not give us the imaginative resources to envision new possibilities or chart a course from here to there.

11. Although I have focused my discussion of niche construction and semantic power on artificial intelligence, these conceptual tools are broadly applicable; in particular, they can be productively deployed to think about states and statecraft, especially the modernist varieties thereof. Modernist states seek the same semantic power, leading to the same thinning out of the world.

12. As we have seen, however, the anarchist critics of modernist states and statecraft are natural allies as we dissect the dystopian potential of current AI technologies.

The great French scholar of religion Henry Corbin (1964) used the term *mundus imaginalis* to denote the intermediate world between "the empirical world and the world of abstract understanding." In this *imaginal world*, "there are cities whose number it is impossible to count." In its positive face, social science is too concerned with the empirical world (i.e., the world as it is, typically in Anglo-American modernity) to help us see the worlds that could be. In its necessitarian face, grand social theory confuses categories of the intellect for necessary laws of history (Unger 2004). The social science we *need* to meet the challenge of AI resides in the imaginal; in mapping and navigating the space of possible social worlds—and finding the worlds where the inevitable alliance between the modernist state and the modernist science of artificial intelligence is disrupted, leading to more vibrant social formations and more thickly human compatible AI.

How to begin? How to find these possible worlds? Consider the "anthropological" lineage from Elinor Ostrom (1990) and James Scott (2012) to the Davids Graeber and Wengrow (2021), or the Afrofuturist lineage from Ishmael Reed (1972) to Alondra Nelson (2002) to Ruha Benjamin (2019b). As these lineages have taught us, there is a wealth of alternative social worlds on offer, in history, at the present, and in our imaginations. Such alternatives begin to populate the space of possible social worlds, and by exploring those sites—empirically and formally—we can begin to see more of this *mundus imaginalis*. We can turn to other scholarly, literary, and prophetic traditions for further examples.

The challenge is to remain alive to our human capacity to construct new selves and new social worlds; to rigorously question when and why a possible world is dismissed as impossible. This imaginal social science has as much in common with mathematics and with poetry as it does with contemporary social science practice. We must expand our imagination of what the human can be and what humans can build; only then can we do justice to the question of what sort of AI will be truly human compatible.

Jacob G. Foster is a professor of sociology at the University of California, Los Angeles, and an external professor at the Santa Fe Institute. He uses tools from machine learning and complexity science to study collective intelligence and the evolutionary dynamics of ideas. He is cofounder of the Diverse Intelligences Summer Institute, which cultivates the transdisciplinary study of intelligences in many manifestations, from ants to humans to AI. He is writing a book on knowledge as an emergent feature of complex adaptive systems.

References

Attema, Arthur E., Werner B. F. Brouwer, and Job Van Exel. 2014. "Your Right Arm for a Publication in AER?" *Economic Inquiry* 52, no. 1: 495–502.

Balduzzi, David. 2016. "Grammars for Games: A Gradient-Based, Game-Theoretic Framework for Optimization in Deep Learning." *Frontiers in Robotics and AI* 2: 39.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York: Association for Computing Machinery.

Benjamin, Ruha. 2019a. *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham, NC: Duke University Press.

Benjamin, Ruha. 2019b. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Birhane, Abeba. 2021. "The Impossibility of Automating Ambiguity." *Artificial Life* 27, no. 1: 44–61.

Bommasani, Rishi, et al. 2021. "On the Opportunities and Risks of Foundation Models." *arXiv [cs.LG]*. http://arxiv.org/abs/2108.07258.

Bostrom, Nick. 2017. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bowles, Samuel. 2009. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In "Proceedings of the First Conference on Fairness, Accountability, and Transparency," edited by Sorelle A. Friedler and Christo Wilson. Special issue, *PMLR* 81: 77–91.

Corbin, Henry. 1964. "*Mundus imaginalis*; ou, L'imaginaire et l'imaginal." *Cahiers internationaux de symbolisme* 6: 3–26.

Dick, Philip K. 1981. *Valis*. New York: Bantam.

Eyal, Gil. 2019. *The Crisis of Expertise*. New York: Polity.

Fourcade, Marion, and Fleur Johns. 2020. "Loops, Ladders, and Links: The Recursivity of Social and Machine Learning." *Theory and Society* (August): 1–30.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York: Basic Books.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Graeber, David, and David Wengrow. 2021. *The Dawn of Everything: A New History of Humanity*. New York: Farrar, Straus and Giroux.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Gross, Neil. 2009. "A Pragmatist Theory of Social Mechanisms." *American Sociological Review* 74, no. 3: 358–79.

Hicks, John. 1946. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*. 2nd ed. Oxford: Clarendon.

Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. New York: Random House.

Jones, Clive G., John H. Lawton, and Moshe Shachak. 1994. "Organisms as Ecosystem Engineers." *Oikos* 69, no. 3: 373–86.

Koch, Bernard, Remi Denton, Alex Hanna, and Jacob G. Foster. 2021. "Reduced, Reused, and Recycled: The Life of a Dataset in Machine Learning Research." In *Thirty-Fifth Conference on Neural Information Processing: Systems, Datasets, and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=zNQBIBKJRkd.

Krakauer, David C., and Daniel N. Rockmore. 2015. "The Mathematics of Adaptation; or, The Ten Avatars of Vishnu." In *The Princeton Companion to Applied Mathematics*, edited by Nicholas J. Higham, 591–97. Princeton, NJ: Princeton University Press.

Laland, Kevin N., Neeltje Boogert, and Cara Evans. 2014. "Niche Construction, Innovation, and Complexity." *Environmental Innovation and Societal Transitions* 11 (June): 71–86.
Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." *ACM Transactions on Interactive Intelligent Systems* 11, nos. 3–4: 1–45.
Nelson, Alondra, ed. 2002. *Afrofuturism*. Durham, NC: Duke University Press.
Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
Reed, Ishmael. 1972. *Mumbo Jumbo: A Novel*. New York: Doubleday.
Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.
Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. 4th ed. London: Pearson.
Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. "Green AI." *Communications of the ACM* 63, no. 12: 54–63.
Scott, James C. 1998. *Seeing Like a State*. New Haven, CT: Yale University Press.
Scott, James C. 2012. *Two Cheers for Anarchism*. Princeton, NJ: Princeton University Press.
Sparkes, Matthew. 2023. "Google Bard Advert Shows New AI Search Tool Making a Factual Error." *New Scientist*, February 8. https://www.newscientist.com/article/2358426-google-bard-advert-shows-new-ai-search-tool-making-a-factual-error/.
Steiner, George. 2013. *After Babel: Aspects of Language and Translation*. 3rd ed. New York: Open Road Media.
Unger, Roberto Mangabeira. 2004. *Social Theory: Its Situation and Its Task*. London: Verso.
Unger, Roberto Mangabeira. 2005. *What Should the Left Propose?* London: Verso.
Walzer, Michael. 2019. *Thick and Thin: Moral Argument at Home and Abroad*. Notre Dame, IN: University of Notre Dame Press.
Zuboff, Shoshana. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology Impact* 30, no. 1: 75–89.