3

EPISTEMIC FORGERIES AND GHOSTS IN THE MACHINE

People are trapped in history and history is trapped in them.
—James Baldwin

In his book *Forgeries of Memory and Meaning*, Cedric Robinson examined how the representations of black people in Western culture helped construct whiteness. These representations and the myths accompanying them became, in Robinson's words, "Inventions of the Negro" and were always changing to suit imperial interests and meet antiracist challenges.[1] They were constructed with diverse instruments: from theater, film, fiction, painting, and even cooking to racial science.[2] This last pillar, racial science, played an important role because, according to Robinson, it forged a "scientific" basis for relations of power between whites and nonwhites—categories that racial science attempts to ground. Social configurations that did not already reflect these scientifically ordained differences in capacities and interests between, say, white and black people would have to be changed, while those that do would gain additional legitimacy. This is how racist scientists saw the situation, long past what is considered the heyday of racial science between the

mid-nineteenth and early twentieth century. Francis Crick, a Nobel laureate celebrated for his work on the structure of DNA, declared in 1971 that "more than half of the difference between the average IQ of American whites and Negroes is due to genetic reasons," and this "will not be eliminated by any foreseeable change in the environment."[3] Racial science was thus always understood to have far-reaching political implications.

As Robinson argued, these representations of human beings, whether in science or the arts, employ "forgeries, "fictionalizations," and "counterfeits of history" in order to "racialize" cultures in service of political ends. Building on Robinson, we can say that to construct a body of knowledge that appears so solid that it can prescribe an entire social order, one needs epistemic forgeries. In racial science and across Western science more broadly, one pervasive epistemic forgery is the notion that scientific knowledge is "transcendent" of the political world—even though, as Robinson writes, "the most powerful economic, political, and cultural impulses of a social structure impose themselves as codes and desires on the conduct, organization, and imagination of scientists."[4]

Like the overtly biological racial sciences, AI has also threatened to prescribe the social order. According the expert commentary we have seen, machines with transcendent "human-level intelligence" would not only put certain people out of work but also require a total reorganization of society. These far-reaching conclusions, like those of racial science, also depend on epistemic forgeries.

Epistemic forgeries are the fictions about knowledge and human thought that help AI function as a technology of power. This chapter outlines three closely linked epistemic forgeries that recur within AI's history. The first is AI practitioners' presumption that their systems represent "a view from nowhere"—a

universal "intelligence" unmarked by social context and politics. The second is that AI systems have either matched or exceeded the capacities of human thought. This epistemic forgery has come to the fore in recent years, and it draws on deep-seated notions in Western culture about hierarchies of intelligence. The third epistemic forgery suggests that these computational systems arrive at truth, or knowledge, "on their own," AI practitioners being merely the ones who set off the necessary conditions for computational processes to properly unfold.

The three epistemic forgeries work in tandem to produce fear and uncertainty about an impending social transformation ushered in by machine intelligence. AI's force, as an ideology and vehicle for political visions, partly depends on the extent to which such forgeries are believed authentic.

THE FIRST FORGERY: A VIEW FROM NOWHERE

Western science is not unique in claiming universality. Just as scientists often present their body of knowledge as universal, so too literary scholars offer certain works of literature as universal. And the universality of both canons, scientific and literary, has been challenged.

In her lecture "Unspeakable Things Unspoken" (1988), Toni Morrison examined the American literary canon's relation to the presence and experiences of African Americans. Morrison urged for a reframing of the debate on the canon's universality. The primary question concerning the canon, she argued, shouldn't be, "Why am I, as an Afro-American, absent from it?," but rather, "What intellectual feats had to be performed by the author or his critic to erase me from a society seething with my presence,

and what effects has that performance had on the work? What are the strategies of escape from knowledge?" Morrison proposes a literary analysis that surveys the damages—the escape from knowledge—caused by the aspiration to universality. She asks, "Is the text sabotaged by its own proclamations of 'universality'? Are there ghosts in the machine? Unsummoned presences that can distort the workings of the machine and can also *make* it work?"[5]

There is, clearly, a world of difference between works of literature and computing systems. But not unlike literary texts, computational systems and the discourse around them can be analyzed for "ghosts" in Morrison's sense. The metaphor of ghosts in the machine suits our exploration of the first epistemic forgery: that AI systems encode and illuminate a "universal" intelligence.

If we adopt Morrison's framing, the aim would not be to simply negate the universality of AI's pursuits by pointing out that the interests and experiences of those outside AI's elite white sphere are not reflected in AI systems (though that is true and important). Rather, it would be to first understand the ways in which the social context of the field's architects is imprinted in AI's computing systems, their uses, and surrounding narratives. And second, to see how this social context haunts the endeavor, creeping in to spoil forgeries like the claim to universality.

We can begin our search for ghosts by looking at how AI's influential figures imagined the future. Joseph C. R. Licklider is one of those figures, and his visions had a lasting impact on computing. Formerly an MIT professor and researcher at MIT's Lincoln Laboratory, Licklider became head of ARPA's Information Processing Techniques Office (IPTO) in 1962, which gave him power to allocate funds to academics.[6] IPTO sought "command and control" capabilities, a frame AI practitioners could

easily work within, and Licklider is seen as largely responsible for the Pentagon's extensive funding for AI. Under Licklider, IPTO funded several prominent AI groups, including those headed by John McCarthy, Marvin Minsky, Herbert Simon, and Allen Newell. But Licklider was far more than a bureaucrat in charge of funds. He also wrote several highly influential articles on computing and its future and is credited as the inspiration to numerous computing projects. Robert W. Taylor, who served as IPTO head after Licklider (and later founded Xerox PARC's Computer Science research laboratory), said Licklider had "laid the foundation for graduate education in the newly created field of computer science. All users of interactive computing and every company that employs computer people owe him a great debt."[7]

One of Licklider's influential articles, coauthored with Taylor, was "The Computer as a Communication Device" (1968). Consistent with Licklider's earlier work advocating "Man-Computer Symbiosis"[8]—which presented the computer as an aid, rather than adversary, to people—Licklider and Taylor argued that computers will be communication aids, as opposed to devices for what they call "informational housekeeping" (tasks such as tracking account balances). The article is both prescient and troubling. Licklider and Taylor describe "on-line interactive communities" that "will be communities not of common location, but of *common interest*."[9]

Yet it quickly becomes apparent for whom these networks operate. Licklider and Taylor tell the reader, "Your computer will know who is prestigious in your eyes and buffer you from a demanding world." It is clear from their examples that the target user is an established professional, making business trips and sending telegrams. The online network will enhance and expand that imagined user's professional activities.[10] Even this scenario

of the professional, though, was civilian dressing for deeply militarized activities. The hypothetical user in "The Computer as a Communication Device" perhaps resembled Licklider himself, who was busy pitching new visions of command and control to the Central Intelligence Agency and National Security Agency, while also overseeing ARPA's Behavioral Sciences Program, which was engaged in secretive "counterinsurgency" research abroad.[11]

But perhaps the most striking part of "The Computer as a Communication Device" is not the idea of an online network spanning great distances but the claim that the presence of computers will transform in-person interactions. Face-to-face conversation, in Licklider and Taylor's view, would be enhanced by having a computer serve as intermediary. The computer would help "men" communicate and compare their "mental models."

As in other texts of this genre and period, the references to "Man" ostensibly stand for all humankind. Looking at the article's figure on "mental models," however, we get a more specific sense of "Man" (figure 3.1). According to Licklider and Taylor, the "most important models are those that reside in men's minds. In richness, plasticity, facility, and economy, the mental model has no peer." What would such "mental models" look like, and who will be doing the communicating? The article answers this with a cartoon showing two men wearing suits. In the mental model of the man on the left, there is a bikini-clad woman, and in that of the man on the right, a bicycle. As the cartoon shows, computer-aided communication transforms these mental models. In one man's transformed mental model, the woman is gone, and the bikini pieces hang on the bicycle, and in the other's mental model, the woman is hunched over a pair of bicycle wheels, her body appearing to serve as the bicycle's frame. This makes clear what Licklider and Taylor envisioned as futuristic

FIGURE 3.1 A figure on "mental models," reproduced from Joseph C. R. Licklider and Robert W. Taylor's "The Computer as a Communication Device" (1968). Courtesy of the Computer History Museum.

communication by "Man." In this understanding of communication, women serve as (literally) the sexual objects of men's mental gymnastics.

The challenge facing the developers of AI's early celebrated systems, funded by Licklider's IPTO, is how to masquerade this white, masculine, and militarized perspective as universal.

We have already seen one such attempt with Allen Newell and Herbert Simon's GPS, originally presented as a "general" approach to intelligence. While GPS tries hard to abstract away time, place, and the body, it was hardly "contextless," as Alison Adam showed. As she has argued, Newell and Simon took the narrow range of activities valorized in their own social milieu—logical puzzles and games such as chess—and "extrapolated from such a bounded problem solving situation to make an important claim about the nature of general problem solving." Newell

and Simon offered experimental validation for their system, but as Adam inferred, these were done with subjects that were overwhelmingly male and likely "white and middle-class," given the "considerable financial resources needed to attend a relatively elite US university." Newell and Simon's conclusions, as she says, are ultimately "based on the behavior of a few, technically educated, young, male and probably middle-class, probably white, college students working on a set of rather unnatural tasks in a U.S. university."[12] And the systems that followed Newell and Simon's, including those that are presented as radical departures from the pair's "symbolic" tradition, have also been similarly committed to forging a view from nowhere.

It would not be accurate, however, to conclude that Newell and Simon had nothing to say about "culture" or their own stance. They did, and this is where the haunting begins, and the fictitious basis of the view from nowhere becomes apparent.

In their book *Human Problem Solving* (1972), Newell and Simon explained that their project is "concerned with the performance of intelligent adults in our own culture," which they attempt to situate by describing what their framework excludes:

> The study [*Human Problem Solving*] is concerned with the integrated activities that constitute problem solving. It is not centrally concerned with perception, motor skill, or what are called personality variables. The study is concerned primarily with performance, only a little with learning, and not at all with development or differences related to age. . . . Similarly, long-term integrated activities extending over periods of days or years receive no attention. These restrictions on the scope of our study delineate a central focus, not a set of boundaries.[13]

Newell and Simon add that these decisions were the product of both "opportunism" and "philosophic conviction."

The pair presented a figure showing how various forms of human variation, including "cultural" variation, factor in their framework (figure 3.2). Here we see again the political imagination and racialized hierarchies that animate this work. In the figure, genetic variation is depicted by a one-dimensional "phylogenetic scale" in which "Man" is at the top and "Primates" are below (followed by "Monkeys"). There is also temporal variation, which is where Newell and Simon place differences in task "performance"; this is also where they see "learning" and "development" fitting in. Then we get to the cultural. There are what they consider as intercultural differences, such as those between the "U.S.," "French," and "Chinese" cultures, as well intracultural ones, such as those between a "Student," "Worker," and "Hippie." The hierarchical and ranked connotations of the graphic are apparent. The downward movement of a scale from superiority to inferiority—as seen in the "phylogenetic scale" with "Man" at the top—is also present within the figure's "culture variation" box (U.S. at the top, followed by French, and then Chinese) and the "individual variation" box (student on top, followed by worker, and then hippie).

Newell and Simon's figure exemplifies AI's persistent quest to order the world into ranked cultures, populations, and individuals. The pair's "philosophic conviction" may simply be that it is white men in suits who shall order the world in this way.[14] Their "universal" understanding of intelligence apparently gives them the license to do so.

But the figure also shows how social context haunts their work, as it would haunt AI's later iterations. In this instance, even the ardent advocates of a symbolic, declarative, and mathematical problem-solving approach to "intelligence" concede that some aspects of thought might lie beyond their system. Or do they? The figure can be taken to mean that what is missing may be filled in without fundamental revision to the framework. Yet

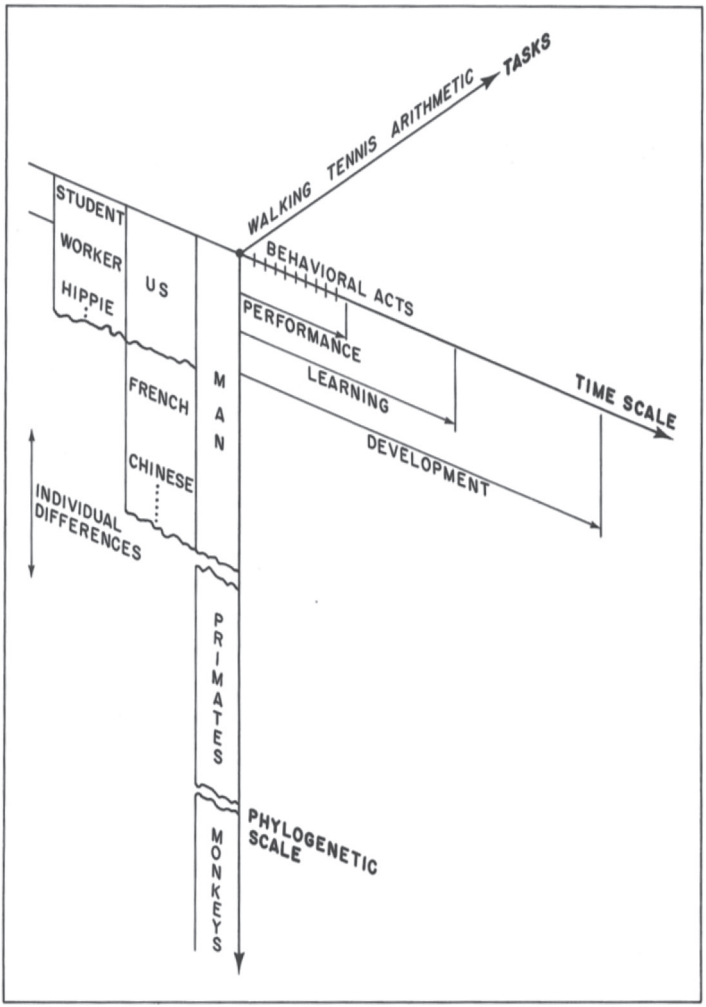

FIGURE 3.2 Figure titled "Dimensions of Human Variation," reprinted with permission from Allen Newell and Herbert A. Simon's *Human Problem Solving* (1972). Copyright Allen Newell Estate and the University of Chicago.

once Newell and Simon say something about "culture," they reveal their situated perspective and provide a glimpse into the endless range of human activities that need to be ignored in order to see "Man" as an "information-processing system" operating from nowhere.

The haunting continued, as GPS's aspirations were never realized. AI practitioners would later backtrack and acknowledge that GPS was not really "general" (an AI textbook from 1992 states, "We Lied about the G").[15] But they then had to find ways to discount the social elements. As Adam has observed, GPS's failure came to be "seen as more of an implementation problem . . . and the empirical research on which its design is founded, is regarded as less problematic than its excessively clumsy implementation." The basic premises of GPS have continued to shape AI.

Yet since AI's engines change quickly, GPS and related systems critiqued by Adam (like Soar or Cyc) would be unknown to many who consider themselves AI practitioners since the mid-2010s. The view of AI as a "knowledge engineering" endeavor, which produces large systems containing thousands of logical rules, is far removed from what the label "AI" conjures in recent years. The more recently celebrated systems nonetheless construct similar fictions, and in these systems, too, we can find ghosts.

THE SECOND FORGERY: SURPASSING HUMAN THOUGHT

Since its rise in the 2010s, as in its prior heydays, AI is presented as a cohesive and transformative force, key to military and economic hegemony. In part, these narratives rest on an epistemic forgery: the notion that AI systems have matched, or exceeded,

the capacities of human thought (a claim that itself presupposes the existence of a "universal" intelligence). This forgery has persisted long past Newell and Simon's GPS. A characteristically triumphalist AI history from 1993 even concludes that "the main battles of the twenty-first century" will not be "fought over issues such as the environment, overpopulation or poverty," but rather "about how we cope with the creations of our human ingenuity; and the issue, whether we or they—our silicon challengers—control the future of the earth."[16]

In the rebranded AI, this forgery was given new wind by commentaries that suggest human-level machine intelligence is either looming or already here. Experts and the media routinely report on "AI systems" outdoing people in a range of activities, such as recognizing images, detecting emotions, or playing video games.[17] Quite often, however, the notion that people will be replaced by superhuman AI is put forward without reference to any specific systems. Instead, commentators offer vague sketches of an AI-driven future. For instance, outlets such as CNN run stories about robots soon "replacing" journalists without any evidence for why this might be the case.[18] Likewise, the prestigious Nieman Foundation for Journalism at Harvard University predicted that, by the end of 2017, "robots will analyze complex editorial content of all lengths, and provide feedback to the humans sitting behind the keyboard."[19] One columnist for the *Guardian* declared in 2019 that "AI can write just like me" and urged readers to "brace for the robot apocalypse."[20] There are similar reports about AI radically altering the practice of science. These narratives are tied to the forgery of universality: an article in the *Atlantic* magazine, for example, suggests that science is "in decline," partly because the random nature of individual scientists' "previous experiences" plays too large a role in scientific discovery—but that "outsourcing to A.I. could change that."[21]

The viability of AI systems exceeding human thought is also conveyed through dystopian scenarios. The *Guardian* reported that Silicon Valley billionaires are "prepping for the apocalypse" by buying secure hideouts in New Zealand, the "apocalypse" being a situation of "systematic collapse" that may include nuclear war or "rampaging AI."[22] Similarly, Silicon Valley mogul Elon Musk has stated to considerable fanfare that current work on AI is "summoning the demon" and that AI is "our biggest existential threat."[23] These narratives are testament to the unstated consensus among experts that AI possesses transformative powers; this is why fantastical commentaries can pass without even referencing specific instantiations of AI or its history. And even the most critical observers of the major platform companies have been swept up by the propaganda about AI's superhuman capacities. Glenn Greenwald, a journalist who played a key role in reporting on Edward Snowden's disclosures, claimed that every time we use Google search, we feed the company's "real business"—which, "unseen to us," is actually "to analyze how the human brain functions, so that it can replace and then improve upon brain functioning in order to create artificial intelligence that's more potent than the human brain."[24]

Mainstream commentary does sometimes contain caveats about AI's superhuman potential, but they are modest. Some surveys of AI practitioners, for one, suggest that they disagree about the timeline and viability of superhuman AI. Practitioners certainly disagree about these issues, but the disagreement is rather minimal. A survey cited by the prolific AI commentator and physicist Max Tegmark shows that practitioners believe there is at least a 50 percent chance or higher that "AI" will "probably reach overall human ability by 2040–50" and that it is "very likely" (90 percent chance) to do so by 2075. Survey participants also estimate that once human-level AI arrives, there is a 10 percent chance "superintelligence" will be reached within

two years, and a 75 percent chance it will be achieved within thirty. Tegmark concludes that "among professional AI researchers," dismissals of human-level AI "have become minority views because of recent breakthroughs" and there is "strong expectation" that human-level AI will be achieved within a century.[25] The very preoccupation with these predictions contributes to the sense that human-level AI is viable or near.

What would such human-level AI look like, according to experts, were it to arrive? The vision tends to be set by the day's most attention-grabbing systems. Currently, these are computing systems that use neural networks, sometimes in combination with reinforcement learning, a set of frameworks in which computational agents are trained to learn by reinforcement signals (described in more detail below). These systems carry many narratives of AI's triumph over human thought. In some respects, these exemplars of AI from the 2010s are radically different from the celebrated AI systems of the 1960s. The more recent systems are based on data-hungry statistical computation that has little in common with AI's "knowledge engineering" stream (associated with Newell, Simon, or Feigenbaum). Yet all these systems are premised on the epistemic forgeries of universality and the defeat of human thought. But because the computing engines are so different, these forgeries take on different forms.

To see how the forgeries manifest, consider some of the most celebrated systems. One is DeepMind's system for playing Atari computer games, which reportedly outperforms human players. Another celebrated system is AlphaGo, also developed by Google's DeepMind, which has beaten human champions at the game of Go.[26] These systems exemplify the aspiration to a radical empiricism. The Atari-playing system, for instance, receives as input images of the game and learns to play based on

reinforcement signals (i.e., how many points it scored in the game). Both the Atari and Go playing systems are presented as free of any human knowledge. The Go-playing system, according to DeepMind, has apparently "learned completely from scratch" and is "completely tabula rasa," which allows the system to "untie from the specifics [of games]." Much more than merely outperforming human players, the system is said to have "understood all the Go knowledge that's been accumulated by humans over thousands of years of playing," which the system was able to reflect on and subsequently "discover much of this knowledge for itself."[27]

The "tabula rasa" rhetoric masks the fact that all these systems have an inductive bias that dictates what patterns they can detect from data and how. Even from a traditional cognitivist perspective, it is possible to critique these systems for having an inductive bias that diverges wildly from people's behavior in the same contexts.

Indeed, cognitive scientists have challenged the claims made about deep learning–based systems. One study evaluated DeepMind's systems and offered several important objections.[28] For one, the Atari-playing system received the equivalent of roughly thirty-eight days' worth of play time. This extensive training allowed the system to obtain high scores, especially in games that do not require longer-term planning. However, a person who gets only two hours of play time can beat the deep learning system in games that do require longer-term planning.

More important, such systems do not acquire the same knowledge about games that people do. The systems are imprinted with particulars of the training data that prevent the sort of generalization people find effortless. The trained deep networks are "rather inflexible to changes in its inputs and goals. Changing the color or appearance of objects or changing the goals of the

network would have devastating consequences on performance if the network is not retrained."[29] For instance, a game-playing system trained with the goal of maximizing its score gets "locked" into this objective. People, by contrast, can flexibly adopt different goals and styles of play: if asked to play with a different goal, such as losing as quickly as possible, or reaching the next level in the game but just barely, many people have little difficulty doing so.

The AlphaGo system suffers from similar limitations. It is highly tuned to the configuration of the Go game on which it was trained. If the board size were to change, for example, there would be little reason to expect AlphaGo to work without retraining. AlphaGo also reveals that these deep learning systems are not as radically empiricist as advertised. The rules of Go are built into AlphaGo, a fact that is typically glossed over. This is hard-coded, symbolic knowledge, not the blank slate that was trumpeted. Nonetheless, the idea of a radically empiricist and general system (which in actuality is confined to narrow domains) is taken to mean DeepMind's approach is ready for grand quests. The company presented AlphaGo not simply as an achievement in computer game playing, but as a way "to discover what it means to do science."[30] The system was presented as a major step toward fulfilling DeepMind's mission: to "solve intelligence" and "use that to solve everything" (see figure 3.3).

These narratives extrapolate from abstract mathematical problems to general intelligence, in the same way Newell and Simon have in the past. Then as now, AI practitioners claim to have uncovered a recipe, even if only in sketch form, for universal intelligence. In the controlled domain of games, the forgery of universality might pass. But when the recipe is applied to broader arenas, the historical context of human life, cast aside by practitioners, begins to creep in.

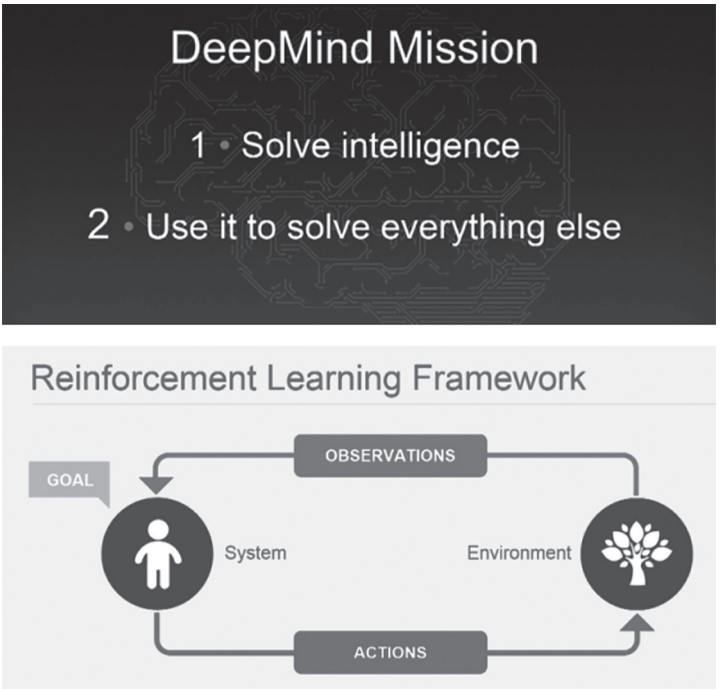

FIGURE 3.3 *Top*, mission statement of Google's DeepMind; *bottom*, framework for achieving it.

SEEING LIKE A DEEP NETWORK

Since seeing cannot be so easily severed from the seeing subject, artificial vision is a place where the epistemic forgery of universality loses much of its force. The rebranded AI presents a narrative in which humans are outperformed on a variety of visual tasks—such as classifying images—by deep network–based vision systems. But what does the world of images look like from the perspective of a deep network? How does a deep network "see"?

The answer depends not only on the structure of the deep networks, clearly, but also on the data set of images that are used for training. Deep network–based vision systems are often used for classification (assigning a label to an image from a predefined set of labels) or captioning (assigning a phrase or sentence to an image), which requires training on a large corpus of images.

The training images typically lack salient historical context. While they are generally real photographs, they play the role of pseudo-Platonic ideals: a "generic" child riding a bicycle, dog chasing a ball, surfer riding a wave, and so forth (see figure 3.4). These images serve as self-evident exemplars of objects and situations that end up in databases such as Microsoft's COCO (Common Objects in Context), where they are classified into categories.[31] Such databases themselves promote the illusion of universality, according to which photographed objects and scenes can be placed into indisputable categories. Yet, as Brian Wallis observed, the "very literalness of photographs produces an uncontrollable multiplication of meanings in even the most banal images."[32]

To interpret even the most banal of images, a background understanding is needed that is missing in artificial vision systems presented as tabula rasa perceivers. This is why, to these systems, an image of people escaping a flood may look like "people on a beach," and a crashing airplane like "an airplane on a tarmac" (figure 3.5). As some cognitive scientists have argued, an understanding of physical scenes, and of human emotions and intentions (easily seen in the faces of those photographed), is missing. But while this critique is revealing, it does not call for abandoning universality. The cognitive scientists who raised this objection imply that the deficiencies can be fixed by supplementing the associationist deep network with hierarchical knowledge of objects and their relations (encoded using symbolic

FIGURE 3.4 Performance of Google's Show and Tell. Reproduced from company paper; see text.

formalisms, for instance). That is, teaching the deep network to "really see" situations. This view doesn't challenge the premise of an artificial system that "sees" without having a body or a history, then, but merely suggests different computational building blocks for building one.

There have been other critiques of existing artificial vision systems on similarly narrow, but still informative grounds. For instance, computer scientists showed that the perceptual space of deep network–based image classification systems differs wildly from that of people. Such systems are trained on the generic image data sets (mentioned above) and assign labels to images, usually with a confidence score. With a simple procedure, it is possible to get these systems to assign high-confidence labels to images that look like noise or abstract patterns.[33] The procedure starts with a pool of random images and iteratively mutates them—randomly perturbing each image's pixels—and then

FIGURE 3.5 Captions generated by a deep network model. Reproduced with permission; see text.

selects only those images that are most confidently classified as the label of interest. After many iterations, this procedure produces random-looking images that are confidently classified by the system as, say, an "armadillo." This shows how these systems are tuned to local visual features in a way that is at odds with human eyes. This bizarre inductive bias prompts another objection to the idea that computing systems can "out-see" people. Artificial vision systems do not have a body—which is arguably where the relational and physical notions that these systems fail to detect originate. Indeed, bodies have been traditionally ignored within much of AI and cognitive science (a point that will be revisited in chapter 6).

Yet a more fundamental gap between human thought and artificial vision systems concerns the historical context of human life. Political and social contexts, which are generally of no interest to AI practitioners, shape how people see their world.

The historical power dynamics among people can be read in photographs, although AI systems are blind to such dynamics. The blind spots can be exposed by probing vision systems in a different way from that intended by their developers. To illustrate this, I have used Google's deep learning–based image captioning system called "Show and Tell"—representative of the

systems that have been claimed to outperform people in the visual arena—to analyze a series of images.[34] Show and Tell was trained on thousands of photographs and can produce a label for an image it has not processed before. When Google showcases the system, it uses banal, generic-looking images that get assigned impressive, or at least reasonable, captions. The images I used, by contrast, were not generic nor banal; they were specifically chosen to demonstrate how historical context shapes the interpretation of scenes.

Consider a photograph of Palestinians arriving at a checkpoint operated by Israeli soldiers (figure 3.6, left). A Palestinian man lifts his shirt to show the soldier, who is motioning to him from the top of a small hill, that he is unarmed. Google's deep network gave the image the caption "A group of people standing on top of a snow covered slope." For a statistical pattern recognizer, the light dirt might look like snow—but the sun, the clothing, and the relationship among those photographed make that an absurd description. Similarly, a 1960 photograph of Ruby Bridges, a six-year-old African American girl being accompanied to a desegregated school by U.S. marshals, is registered as "A group of men standing next to each other" (figure 3.6, middle).

There are many more complex relations among the photographed that are missed. Consider the scene of an Israeli soldier holding down a young Palestinian boy while the boy's family try to remove the soldier (figure 3.6, right).[35] Google's deep network produces the caption "People sitting on top of a bench together" (the "bench" perhaps being the boy). The motives and intentions of the individuals are entirely lost.

It isn't possible to make sense of group scenes without history, either. For instance, Google's system registers an image of Palestinians praying in protest outside the mosque, with the

FIGURE 3.6 Captions generated by Google's Show and Tell deep network. Image credits: *left*, Ammar Awad / Reuters; *middle*, U.S. Department of Justice; *right*, Reuters.

Dome of the Rock in the background, as "A crowd of people standing around a parking lot filled with kites," probably because of the colorful shirts of the men in prostration (figure 3.7, left). Similarly, a 1960 photograph from South Africa's apartheid regime, in which black men line up to receive passbooks from a panel of all-white officials, is captioned as "A black and white photograph of a group of people" (figure 3.7, right).

When one looks at photographs, the history of gender oppression cannot be ignored, either. An ad from the 1960s where a woman is used as ashtray support for a man smoking a cigar is captioned as "A black and white photo of a woman wearing a tie" (figure 3.8, left). Another image—in which a woman carrying a stack of towels is a fleeting background figure and a man watches television on the couch—can instantly evoke the gendered division of household labor yet is registered by the deep network as "A woman sitting on the couch with a laptop" (figure 3.8, right).

A counterargument to these examples might be that with a larger training data set, the same computational system might be able to "understand" even these images. That would presume,

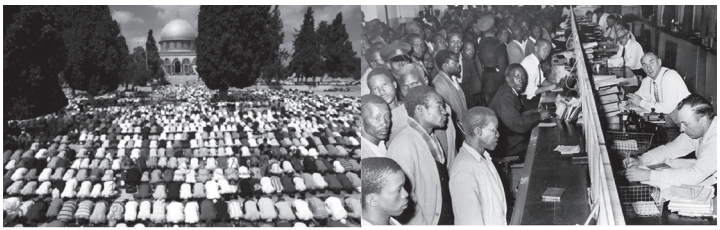

FIGURE 3.7 Captions generated by Google's Show and Tell deep network. Image credits: *left*, Ahmad Gharabli/AFP; *right*, unknown.

however, that a mapping from images to labels is enough; that the "information" is there, inside a large corpus, and it is only a matter of finding the right model to decode it. But more training on labeled images will not prepare the system for something like Ruby Bridges's photograph. This image is not an instance of a "type" of visual configuration that can be extracted from an arsenal of captioned images. The space of photographs cannot be meaningfully parceled into ever finer categories, such as "six-year-old African American girls being protected from violence on the first day of school in the United States of the 1960s."

The failures of deep learning–based vision systems reinforce the fact that motivations and goals are essential to grasping scenes, as cognitive scientists have argued. But they also show that scenes unfold in a historical context that shapes their interpretation. Computing systems, by design, excise this historical context—and are haunted by it.

Historical context also haunts the design of these systems. All these systems depend on the substantial human labor needed to assemble and curate training data, which is hidden from view; this being another way in which the appearance of universality

FIGURE 3.8 Captions generated by Google's deep network Show and Tell. *Left*, a sexist ad for cigars from the 1960s; *right*, an image used to depict unequal division of household work between men and women. Image credits: *left*, Cigar Institute of America; *right*, CORBIS.

is staged. Google's Show and Tell, as we saw, was trained on hundreds of thousands of captioned images—but who provided the captions? Some image data sets that were used for training the system, such as Microsoft's COCO, were captioned by workers on Amazon's Mechanical Turk platform (AMT).[36] Google researchers also used AMT workers in evaluating the image-captioning system by having them score the model-generated captions. From the universal-intelligence perspective, what matters is that the captions were produced or validated by *some* human. The identity of the viewer isn't considered relevant, although it plainly shapes how images are seen: a photograph of a checkpoint in occupied Palestinian territories may well be perceived differently by a viewer in Ramallah compared with a viewer in London. And while AMT does in fact allow the employer to select workers based on country, Amazon has recently limited the worker pool to the United States.[37]

As Lilly Irani has argued, the predominantly U.S.-based employers on the platform prefer U.S. workers because, among other things, "they are likelier to be culturally fluent in the kinds of linguistic and categorization tasks" that are delegated to AMT.[38] The restriction shows how social context does matter. But despite the probable restriction of workers to the United States, the developers of the Show and Tell system reported just 65 percent agreement among AMT workers regarding the validity of computationally generated captions.[39] Universality was being spoiled again. In cases of disagreement, the developers averaged the scores, which blurs contextual differences and services the illusory view from nowhere.

Building systems that aspire to a false universality can be profitable, even as the forgery fails. In the settings that interest corporations, small gains on narrow tasks can be lucrative, and glossing over context pays off. As Microsoft researchers have argued, in systems that predict user clicks on digital ads, for example, "even 0.1% of [average] accuracy improvement would yield greater earnings in the hundreds of millions of dollars."[40]

THE THIRD FORGERY: SETTING THE CONDITIONS FOR MACHINES

The pretense of universal machine intelligence, which exceeds human capacities, is politically useful but unsustainable. The promise of human-level intelligence always proves too grand and computers recalcitrant in realizing it.

But the fiction of an AI that rivals human cognition is accompanied by a more elusive epistemic forgery. This is the third epistemic forgery, which posits that "truth" can be elicited by launching a computational process that operates independently

of people, and whose inner workings cannot be understood by them. By this forgery, the process, once set in motion, is beyond anyone's control, yet its outcomes are superior to what any human collective can achieve through deliberation. This forgery leaves only two roles for people: first, the experts have to set the conditions for said computational process to unfold, and second, everyone else must do their best to adapt to the results, while acknowledging that the procedure's logic will remain forever indecipherable. This forgery not only elevates experts—as they will wield this fantastical force—but also serves to absolve society's most powerful from responsibility for social arrangements.

In some ways, this epistemic forgery is old and familiar. Stanley Kubrick's film *Dr. Strangelove Or: How I Learned To Stop Worrying And Love The Bomb* (1964) cleverly conveys the trickery. As an all-out nuclear war with the Soviets looms, the U.S. president is weighing Dr. Strangelove's plan to save a limited number of Americans ("a nucleus of human specimens") in deep mineshafts. Mr. President recognizes the moral dilemma this plan poses: "Well I . . . I would hate to have to decide . . . who stays up and . . . who goes down." Dr. Strangelove offers a way out: "Well, that would not be necessary Mr. President. It could easily be accomplished with a computer." But Dr. Strangelove's next lines undercut the notion, as he struggles to restrain his right arm (apparently having a mind of its own) from launching into a Nazi salute: "A computer could be set and programmed to accept factors from youth, health, sexual fertility, intelligence, and a cross section of necessary skills. Of course it would be absolutely vital that our top government and military men be included to foster and impart the required principles of leadership and tradition." Dr. Strangelove goes on to justify other arrangements for the hypothetical bunker, notably a ten-to-one ratio of women to men, with the women chosen for possessing

"highly stimulating" sexual characteristics (which the men of the war room agree is "an astonishingly good idea").

The scene distills the absurdity of seeing computing as standing apart from people and institutions. But since the 1960s this fiction has taken on more elaborate forms, and it now manifests in AI's celebrated systems. The intellectual basis of this forgery comes from neoliberal economic theory and behaviorist psychology, two fields that have historically overlapped with AI. It is worth exploring how theorists in these fields articulated the epistemic myth because, as we will see, the rebranded AI's narratives draw on both fields.

The opaque yet all-powerful computing device that is central to neoliberal economic theory is "The Market" (presented as a singular entity)—an information-processing device that "knows" more than any human individual or group could possibly know.[41] This epistemology was articulated by Friedrich Hayek, who leveled it as an argument against centrally planned economies (which he saw as defining of socialism). According to Hayek, any individual's knowledge is local and incomplete, yet all this knowledge must be factored into the organization of society. Yet Hayek claims that an individual's knowledge cannot be articulated; it can only be elicited through behavior, that is, by the response to market signals. Hayek therefore challenged the notion that people could use what they think they know to "order" society.[42]

This was not just an abstract debate: this understanding of knowledge was marshaled to absolve society's elites of accountability. Hayek argued that while workers could be persuaded that wealthy capitalists rig the economy to promote exploitative labor, this is in fact impossible, since the set of facts that govern the markets "*is no more available to capitalists for manipulating the whole than it is to the managers that the socialists would like to replace*

them. Such objective facts simply do not exist and are unavailable to anyone."[43] Put differently, society's richest do not shape the market but, just like the poor, are under its control. No one can be held responsible for this natural force. All society can do, apparently, is make room for it to grow.

This view appeals to the major corporations promoting AI because they are in the business of creating ever-expanding markets. Yet this neoliberal epistemology runs deeper: it is also used to narrate how AI systems built by these companies work. And as in all neoliberal doctrines, this epistemic thesis is used to discount people's knowledge and human collective action, and to absolve powerful institutions of responsibility for social arrangements.

The rebranded AI's narratives rehash these ideas. *Wired* magazine suggests that AI systems now possess "alien knowledge," enabled by "post-paucity computing," which is incomprehensible to the "puny human brain."[44] These systems outdo us in ways we cannot follow, and so we must yield to them just as we must yield to the market.

This perspective is vividly demonstrated in a *New York Times* piece by Steven Strogatz, noted mathematics professor at Cornell University and media commentator, on the powers of AI. The piece centers on Google's Alpha systems. Strogatz describes the future when Google would deliver its ultimate system, "AlphaInfinity": "We would sit at its feet and listen intently. We would not understand why the oracle was always right, but we could check its calculations and predictions against experiments and observations, and confirm its revelations."[45]

In the realm of science, Strogatz writes, people would merely be reduced to "spectators, gaping in wonder and confusion." Eventually, though, "our lack of insight would no longer bother us" because "AlphaInfinity could cure all our diseases, solve all

our scientific problems and make all our other intellectual trains run on time." Strogatz concludes with a sketch of the future: "We did pretty well without much insight for the first 300,000 years or so of our existence as Homo sapiens. And we'll have no shortage of memory: we will recall with pride the golden era of human insight, this glorious interlude, a few thousand years long, between our uncomprehending past and our incomprehensible future." Readers of the *New York Times* would find this view familiar, as it is a redressing of standard neoliberal doctrine.

This perspective on AI inherits neoliberal doctrine's primary contradiction. The contradiction arises when a centralized elite sets the conditions for a magical computational process (whether the market or a computing system) and decides when it works or needs fixing, but also claims that this process is beyond human control. The corporations building AI's celebrated systems likewise espouse decentralized democracy over hierarchical control, but corporate elites dictate what counts as data and how it is used; the mythical flat, democratic market doesn't exist.

Staging our third epistemic forgery entails navigating this contradiction. How is it done? Usually by downplaying the role of the elite class, as well as by drawing on models of the self that discount people's knowledge.

The newly celebrated AI systems do this by appealing to behaviorist psychology, whose proponents have also offered ways to present social arrangements created by an elite class as natural. DeepMind's core principle, for instance, is a behaviorist one: that "intelligent" behavior can be programmed via the right regimen of rewards and punishments (figure 3.3).[46] The success of DeepMind's systems is attributed not only to deep networks but also to reinforcement learning.[47] Reinforcement learning imbibes the behaviorist principle that intelligence arises by individual agents being disciplined by the environment.

This understanding of the self discounts people's knowledge. We can see this through the work of behaviorism's famous champion, B. F. Skinner. Like Hayek, Skinner was suspicious of explicit knowledge and believed individuals can be objectively assessed only through their "behavior." According to Skinner, rather than paying attention to people's "feelings, their states of mind, their intentions, purposes and plans" in order to change behavior, one should tune the environment.[48]

Behaviorism, then, promises to be the science of controlling individuals' behavior and, by extension, whole societies. Skinner recognized that this aspiration invites an obvious challenge: Wouldn't the elite who have this "scientific" expertise be in a position to manipulate the rest of us?[49] Skinner's response is that the elites of behaviorist science are not the "exploiting elite"; they cannot be, because "their task is not to control people but to bring people under the control of more effective physical and social environments. They operate upon the environment, not upon people."[50] Skinner's argument mirrors the neoliberal framework in which capitalist owners do not control the market but merely set the conditions for it to operate. Analogously, for Skinner, the elite do not directly control the lives of individuals but simply tune the environment in which people operate.

The rebranded AI's narratives fuse these neoliberal and behaviorist myths. The major corporations that dominate these narratives both create totalizing markets (through platforms that modify behavior) and then serve up a model of the self in which responding to such reinforcement or market signals is the core of what it means to be "intelligent."[51] This shows how the design of AI's computing systems is informed by a specific conception of social order, and how the celebration of these computing systems reaffirms that social order.

REJECTING THE FORGERIES

The mirage of a superhuman yet indecipherable computational process serves multiple functions in narratives about AI. This illusion not only elevates the idea of a view from nowhere by treating such a computational process as a natural force; it also deflects scrutiny from institutional power by steering our gaze toward inert algorithmic boxes.

AI practitioners and their patrons have long fixated on the innards of computing systems, seeing computers as entities that can "make decisions"—thus reinforcing the forgeries described in this chapter. This view raised worries about "explainability" (the military, for one, wanted to explain computerized battlefield decisions). Following AI's rebranding in the 2010s, explainability again became the clarion call for many experts who see it as a solution to AI's opacity.[52]

Decades ago, however, some reflective AI practitioners offered alternative views of computing systems that reject the epistemic forgeries and the misplaced notions of explainability they suggest. Some practitioners recognized that computer programs have little life except as instantiations on actual physical computing systems, which exist in a dynamic social context. This means that "explanations" of a computing system's behavior could never be confined to the neat and abstract space of algorithms.

The computer scientist Joseph Weizenbaum, for instance, argued that ordinary computer programs are effectively "theoryless." These programs can be quite large and are often developed by multiple people. There is no algorithm one could write down that fully encapsulates how such a program works in practice.[53] If such an abstraction were available, computing systems

would not be as difficult to run, maintain, and debug as they are. Computing practitioners in the corporate world have long recognized this messy reality as it interfered with attempts to hurry teams of programmers to meet the demands of profit.[54]

For these reasons, Weizenbaum offered a different metaphor for computing systems. Rather than seeing a computing system as a realization of theory, which would mean it can be "explained" in algorithmic terms, he argued for seeing it as an intricate bureaucracy. In this bureaucracy, different subsystems, glued together somewhat haphazardly as a product of circumstance, generate outcomes that are subject to disputes over "jurisdiction." This is why programmers, he argued, often "cannot even know the path of decision making" that unfolds in their own programs, "let alone what intermediate or final results" will be produced.[55] Implicit in this argument is the simple observation that every computing system exists in a social envelope. Contrary to the epistemic forgeries we have reviewed, people decide what counts as data, when the software's output is correct, when it needs revision, and so on. Furthermore, all these decisions are shaped, sometimes in unpredictable ways, by the physical constraints of the computing medium. Computing systems are therefore generally not reducible to algorithms nor to the source code of the software they run.

These observations rarely concern the contemporary AI expert industry, even though the computing systems with which these experts are occupied are fraught with the same difficulties. Thus when these commentators claim that, say, systems using neural networks are indecipherable because of their large number of parameters, they are favoring mathematical abstraction over the situated social reality of computing systems. In practice, one does not deal with an abstract neural network but rather its instantiation as software, which is subject to bugs, changes, updates,

and hardware constraints. Things that are considered peripheral to the abstract description of a neural network—such as the decision of when to stop training the model or the versions of different pieces of software used to perform numerical calculations—factor into the "decipherability" of the actual system.[56] That is why it is a stretch, one with political significance, to presume that computing systems that use neural networks are somehow uniquely "indecipherable."[57]

This appeal to abstract indecipherability can be a way to mask institutional power. Consider the narratives we saw earlier about the indecipherability of neural network–based systems, which emphasize those systems' incomprehensibility to people, including their very developers. There is a modicum of truth to this: after a neural network is trained on some data, no easily interpretable rule necessarily emerges that explains its behavior. And all such systems are instantiated physical computing systems, which makes Weizenbaum's points applicable. However, a blanket acceptance of indecipherability is also a gift to institutional power. After all, if AI systems outdo people and hence must be used—everywhere from the court system and policing to hiring decisions—yet are indecipherable, then who or what can be held accountable? This deflection is common within AI commentary; for example, in the title of a *Washington Post* piece in 2018: "A.I. Is More Powerful than Ever. How Do We Hold It Accountable?"[58]

The concern with this misguided question traps us in a loop over these conundrums: Has AI exceeded all human capacities or just some? If it has exceeded them, and therefore must be used, is it decipherable? Could those who build opaque AI systems give us tools to explain, and potentially "de-bias," the decisions made by their systems? And since experts say that making systems more transparent would supposedly make them less effective, how will such cost-benefit tradeoffs be managed?[59]

Here, the AI expert industry attempts to solve problems resulting from its own epistemic forgeries, which only produces dead-ends. Thirty years ago, when analyzing their day's expert industry's discourse on computers in the workplace, Ruth Perry and Lisa Greber captured these dead-ends: "We humans are seen as ships before the storm of technology, lifted or buffeted by forces beyond our control; in the wake of the storm we adapt, choosing the least unpleasant from a limited set of options."[60] The epistemic forgeries presented in this chapter, and the expert industry that reproduces them, are part of that same storm.

We have seen that different streams within AI have shared a commitment to epistemic forgeries—to seeing computation as a natural force, standing outside politics, that surpasses people's capacities and remains indecipherable to them. While these forgeries manifested differently at different periods, in general, the historical context of human life was buried.

In the next chapter I explore how these epistemic forgeries work to naturalize capitalist systems of racial and gendered oppression—at a time when the successes of social movements have forced AI's expert industry to frame itself around social justice.