[3]

The neural network at its limits

Lucy Suchman

The connections between neural networks read as technologized physiology or as physiologized technology, depending upon the direction of analogic travel, signal the longer history of traffic across the disciplinary boundaries of life and systems sciences (see Dhaliwal, this volume). Recognition of those connections is not the endpoint of inquiry, however, but the starting place. This chapter begins from that place to revisit the question of how sameness/difference is made between the brain and computer, not to resolve the question metaphysically or philosophically but rather to see how analogic thinking informs the projects of neuroscience and of computational neural networks, and where it breaks. I follow two threads in the rhetorical fabric of the neural network; the entanglements of technological analogy and physiological modeling in the work of neuroscience, and the retreat from model to inspiration when the alignment of brain and computer breaks in the technical work of computer science. I consider what the brain/computer analogy enables in the fields of neuroscience and computational neural networks, and what we might learn by foregrounding the moves that the protagonists of this story make when they reach the analogy's limits. More specifically, I examine how the brain/computer analogy breaks differently for feminist neuroscientist Gillian Einstein (who follows her research problem from the brain to sexed/gendered bodies and their worlds) than it does for neural networks researcher Geoffrey Hinton (who is committed to sustaining the analogy, even as he acknowledges its limits).

My argument in brief is that the enduring commitment that informs the project of computational neural networks – and its embrace by computational neuroscience – is to cognitivism. The sense of cognitivism in this context is a theory of intelligence based in a correspondence between mental representations formed in the brain/mind, and a world taken to stand outside of it. Located inside the bounds of the skull (Star 1992) the brain/mind is receptive to its surrounding world in the form of inputs (rendered for purposes of processing either symbolically or statistically) and responsive through its outputs. The literature from recent anthropology, science and technology studies, feminist theory and kindred fields critiquing cognitivism and articulating its alternatives is by now extensive.[1] Relevant arguments center on the inseparability of cognition from the lived experience of embodied persons-in-relation, with one another and with culturally and historically constituted social and material worlds. In these theorizations and associated empirical studies the intelligibility and significance of social/material worlds is variously reproduced and transformed in practice, rather than given to individual brains/minds. The politics of practice, in the sense of how social ordering is enacted and with what differential consequences, are fundamental.

The cognitivist frame systematically relegates all that doesn't fit to the status of epiphenomena beyond the bounds of the science.[2] The commitment is not simply programmatic, in other words, but what enables the perpetuation of the analogy of reasoning to computation. Sailing within the winds of cognitivism has led to a tacking back and forth between a logic or rule-based symbolic approach, regarded as "abstract reasoning," and a statistically based approach, figured as "deep learning." Posited as a remedy to the limits of the symbolic approach, the statistical approach now begins to show its limits. As recognition of those limits grows, the only course available within the closed world of cognitivism is a partial return to the symbolic, in the form of an imagined new symbolic/statistical hybrid.[3] This vaguely articulated synthesis promises a solution that enables practitioners to stay within the boundaries of the cognitivist frame. In agonistic dialogue with these developments are a set of critical engagements that insist on attention to the different material and cultural histories of organic and artificial intelligences, where intelligence is understood not as brain-centered cognitive functions but as practices enlivened in and through ongoing relations of collective world-making, for better and worse.

Articulating the Analogy

The figure of the neuron has been extensively traced within science and technology studies (STS), following its attachment as a prefix to a range of fields from neuroeconomics (Schüll and Zaloom 2011) to neurocultures (Ortega and Vidal 2011).[4] Brosnan and Michael (2014, 681) observe that the designation by the United States Congress of the 1990s as the "decade of the brain," as part of a wider cerebral turn among science and technology funders internationally, resulted in the channeling of resources into neuroscience research. Their study of what they characterize as the enactment of the 'neuro' in a neuroscience laboratory in the UK builds upon literature in the sociology of expectations (Brown and Michael 2003), attuned to the performative effects of promissory rhetoric in the institutionalization of technoscientific projects. More specifically, they analyse the promise of translation despite systemic boundaries between lab work and its clinical application, and the multiplicity of the neuro's enactment within as well as across those boundaries. The neuron in this context, they propose, doesn't so much cohere spatially as an entity as it is made to adhere temporally to the imagined future that will translate its materializations in the lab into efficacious interventions in the clinic.[5]

Another line of analysis within STS examines the biological reductionism of neuroscience (Dumit 2004; Martin 2004; Ortega and Vidal 2011). Vidal (2009) traces the genealogy of the "cerebral subject" in its positing of the brain as the essential organ of personhood, with the body as its vessel. This brain-centrism, he argues, is central to the figuration of modern humanity since the 17th century, based in individualism and a self-consciousness separable from body and world. "The idea that 'we are our brains,'" Vidal observes, "is not a corollary of neuroscientific advances, but a prerequisite of neuroscientific investigation" (2009, 7).[6] Debates about mind as reducible to brain, or as emergent, are longstanding however novel their manifestations in contemporary (computational) neuroscience. Moreover, while the object of laboratory work is the organic matter of biology, laboratory work is equally involved in practices of abstraction, as "the squishy stuff of the brain becomes a subject of graphic comparison, sequential analysis, numerical measure, and statistical summary" (Lynch, 1988: 273). Quantification and its associated denaturing smooth the path for the neuron's travel from the biological laboratory to the research 'laboratory' of computational neuroscience, with its own promise of translation from research to application.[7]

The limits of the neural analogy are less important in the case of artificial intelligence than the analogy's power as what Dennett (2013) first named an intuition pump for technical projects. Gary Marcus, a sympathetic critic of the project of computational neural networks, insists that the aim should not be for machines to literally replicate the human brain (whatever a literal replication of organic matter in machinery could mean). After all, he observes, the human brain is "deeply error prone, and far from perfect" (2018, 21), problems that by implication might be avoided in the design of computational machines. Yet there are, Marcus acknowledges, many areas in which the human retains an advantage. In a passage exemplifying a biotechnical imaginary that at once naturalizes the artificial and posits the organic as always already technological, Marcus suggests that:

> A good starting point might be to first try to understand the innate machinery in human minds, as a source of hypotheses into mechanisms that might be valuable in developing artificial intelligences (2018, 21).

In this time of computational ascendance, it is hardly surprising that the "innate machinery" of the brain would be seen as:

> a broad array of reusable computational primitives – elementary units of processing akin to sets of basic instructions in a microprocessor – perhaps wired together in parallel, as in the reconfigurable integrated circuit type known as the field-programmable gate array (Marcus et al 2014).

While seeming to reject the binary of nature versus culture, these are rhetorical moves that naturalize the technological, rather than articulating either differences or relations between the organic and the machinic. The imaginary of the wiring diagram as a network exploits that term's etymology as naming an "open textile fabric tied or woven with a mesh for catching fish, birds, or wild animals alive," along with its expansion in the 19th century to reference "any complex, interlocking system." [8] Once rematerialized as canals, railways and other infrastructures for transport, the figure of the network is well positioned for its translation into mid-20th century technologies of communications and information, along with the installation of calculation as a universal process, and the attendant erasure of the specificity of constitutive entities and relations.[9] The neuron can then be figured as a logic gate that determines the operation of synaptic connections (see Halpern 2022, 335).

With this backstory in mind, I consider how the analogy of technology and physiology mediates the project of computational neural networks and laboratory neuroscience, exemplified in the narratives that two contemporary practitioners offer regarding the logics and trajectories of their respective techno/scientific practices. As indicated my focus is on how each conceptualizes their field of experimentation, and what direction their projects take as they encounter that field's methodological limits.

The neural Network in the Work of Computer Scientist Geoffrey Hinton

Geoffrey Hinton is widely recognized as a founder and leading researcher in the subfield of artificial intelligence devoted to computational neural networks, and more specifically to so-called convolutional neural networks or deep learning (see

Lepage-Richer, this volume). An emeritus Distinguished Professor at the University of Toronto, recipient of prestigious awards, and former Vice President and research fellow at Google, Hinton's aim "is to discover a learning procedure that is efficient at finding complex structure in large, high-dimensional datasets and to show that this is how the brain learns to see."[10] With degrees in experimental psychology and artificial intelligence, Hinton sits at the intersection of two laboratory-based approaches to human cognition.

In a conversation with then *Wired* Editor in Chief Nick Thompson (2019), Hinton offers this explanation of a neural network:

> **GH:** You have relatively simple processing elements that are very loosely models of neurons. They have connections coming in, each connection has a weight on it, and that weight can be changed through learning. And what a neuron does is take the activities on the connections times the weights, adds them all up, and then decides whether to send an output. If it gets a big enough sum, it sends an output. If the sum is negative, it doesn't send anything. That's about it. And all you have to do is just wire up a gazillion of those with a gazillion squared weights, and just figure out how to change the weights, and it'll do anything. It's just a question of how you change the weights.

This response to Thompson's request to explain what neural networks are exemplifies the slippery rhetorics of biotechnical translation. Implicitly taken as a question about computational neural networks, the answer begins with reference to "processing elements that are very loosely models of neurons." The gesture here toward computational entities as models of organic ones is qualified to suggest that the relation is more analogy than strong claim. The connections between elements have "weights," a familiar term in computational vernacular that refers to a set of mathematized values, which Hinton explains change through "learning." Here a process associated with organic life is used as a technical term referring to the computational adjustment of values based on better and worse results according to a pre-specified outcome. The "neuron" reappears in the next sentence as an agent that "decides," based on calculations (over processes now rendered as "activities") between binary alternatives. The power, Hinton explains, comes from a combination of the number of such processing elements and the techniques for manipulating the numerical values. The sleight of what is necessarily a human hand appears in the figure of the "you" who works out how to adjust those weights, a momentary shift in figure and ground from system to programmer that at once effects a difference between them and reveals their interdependence.[11]

Thompson follows with the question: “When did you come to understand that this was an approximate representation of how the brain works?” to which Hinton responds:

> **GH:** Oh, it was always designed as that. It was designed to be like how the brain works.
>
> **NT:** So at some point in your career, you start to understand how the brain works. Maybe it was when you were 12; maybe it was when you were 25. When do you make the decision that you will try to model computers after the brain?
>
> **GH:** Sort of right away. That was the whole point of it. The whole idea was to have a learning device that learns like the brain, like people think the brain learns, by changing connection strings.

While Thompson’s question suggests that the computational neural network offers at least an approximate model for the workings of the brain, Hinton’s response flips that relationship to position the brain as a model for the computational neural network. It is this inversion that enables the shift from model to inspiration, as when pressed by Thompson later in the interview on what is clearly a computational technique that deviates from any processes evident in the brain. Hinton demurs:

> **GH**: I’m not doing computational neuroscience. I’m not trying to make a model of how the brain works. I’m looking at the brain and saying, “This thing works, and if we want to make something else that works, we should sort of look to it for inspiration.” So this is neuro-inspired, not a neural model. The whole model, the neurons we use, they’re inspired by the fact that neurons have a lot of connections, and they change the strengths.

This reassertion of the difference between brains and computers belies Hinton’s earlier statement in the same conversation that “we are neural nets. Anything we can do they can do.” It is the polysemy of the term “neuron,” referring alternately to the processing elements of the computational neural network and to the physiological entities that are the objects of laboratory neurosciences, that enables Hinton’s statements about brain/computer relations to shift seamlessly between analogy and identity, inspiration, and model.[12]

As the conversation moves onto the topics of consciousness, learning, and finally Hinton’s ‘four theories’ of dreaming, things become increasingly ungrounded. With respect to the educational implications of neural networks Hinton opines:

> And we know now … you can just put in random parameters and learn everything. If we really understand what's going on, we should be able to

> make things like education work better. And I think we will. It will be very odd if you could finally understand what's going on in your brain and how it learns, and not be able to adapt the environment so you can learn better … And once [computational] assistants can really understand conversations, assistants can have conversations with kids and educate them.

The figure of learning is ubiquitous in both symbolic AI and neural networks discourse, traceable to the original Dartmouth Summer Research Project proposal (McCarthy et al 1955: 1) "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." Conceptualized as a process located inside the brain, learning here is serviced by an environment that needs to be represented or modelled before it can be registered. It follows that it should be possible to re-engineer the environment to optimize its alignment with the brain's processing requirements. Defined technically, learning in the context of neural networks references optimization of the computational processes required to produce a 'correct' output, where the latter is a function of determinations made by humans considered to have relevant knowledge in a target domain. But in its capacities as a more floating signifier, learning stands as the holy grail for so-called artificial general intelligence, or what Marcus (2018) characterizes as "a human-like flexibility in solving unfamiliar problems."

In his critical review of progress in so-called deep learning, "a statistical technique for classifying patterns, based on sample data, using neural networks with multiple layers," Marcus identifies what he names '10 challenges' to the field (Marcus 2018). First among these is reliance on large amounts of data, necessitated by what he characterizes as the inability of neural networks to grasp abstract relationships. As an example, Marcus points to the ease with which his human readers, presented with the concept of a 'schmister' defined as a sister over the age of 10 but under the age of 21, can identify whether they themselves have a schmister. "In learning what a schmister is," he writes, "in this case through explicit definition, you rely not on hundreds or thousands or millions of training examples, but on a capacity to represent abstract relationships between algebra-like variables" (2018: 6). Granted that our ability to make sense of the concept of 'schmister' is not based on thousands or millions of training examples, just how, we might ask, is it a capacity to represent abstract relationships? This reader, in any case, thinks immediately of my own very lively and embodied sisters on hearing that word. And in what sense is this a relationship between algebra-like variables, beyond the numbers 10 and 21, unless the words sister and schmister are treated as associated text strings and not associated persons? The idea of abstract in this context already presupposes that knowledge equals the correct classification of instances into categories, ignoring other modes and references, not least to lived relations of kinship.

Marcus observes that the ‘deep’ in deep learning refers not to profundity, but rather to the number of middle level or hidden layers of computational processes (introduced for greater efficiency). The actual shallowness of neural networks is a factor in their widely cited brittleness, which leads to some acknowledgment that they do not in fact engage in learning in anything like the human sense. Marcus (2018: 8) - points out that the neural network trained to play the video game Breakout, which famously “realized that digging a tunnel through the wall is the most effective technique to beat the game,” did not just fail to observe the rules. Rather, he notes, the system has no perception of digging, tunnels, walls, games, or rules, but only of pixels mathematized in such a way that, given a predetermined objective, it deploys statistical techniques to optimize the likelihood of that output on any given round. Moreover, neural networks have no way of dealing with what Marcus characterizes as “prior knowledge” (often referred to as “common sense”), a problem that cognitivists trace back to the lack of ‘abstract concepts.’[13] Read more broadly, neural networks have no qualitative understanding of entities and their relations, an unsolved problem for AI that now leads Marcus to suggest the need for ‘hybrid models’ and at least a partial return to representational or symbolic processes (Marcus 2018: 20). The premise that abstractions can be represented by symbols only reframes the problem as one of reference, however, while continuing to beg the question of what Lave (2011: 115) calls “knowledgeability in practice,” including the inseparability of learning from the concerns of everyday life.

Most relevant for the concerns of this essay, Marcus points to what he describes as a “culture in machine learning that emphasizes competition on problems that are inherently self-contained” (2018: 12). It is this reliance on closed worlds that sidesteps the recurring and unsolved problems arising from the limits of the computational sensorium. Neural networks require for their operation a dataset of machine-readable inputs amenable to analysis for statistical correlations that fit a set of pre-determined outputs. Quantitative advances in the scale of datasets and the speed of processing do nothing to address the infamous difference between correlation and causation, or the place of the latter in our reasoning about indeterminate relations and effects. Framed in terms of the brittleness or narrowness of computational neural networks, prevailing discourses implicitly reference a transfer model of learning that has been thoroughly critiqued by scholars like Lave (1988, 2011). Marcus characterizes this as “the problem of generalizing outside the training space,” but Lave (2011, 155) rather draws on Stuart Hall’s figure of “rising to the concrete” (2003, 131) to name the ability to bring generalized theory into relation with the specificities of ongoing, cultural/historical worlds. On this understanding, the reliance of computational neural networks on closed worlds is not just a symptom of their inability to deal with ‘novelty,’ but of a much more fundamental difference between computation and knowing in/as practice.

Trans-Boundary Explorations in the Neuroscience Research of Gillian Einstein

The limits of computational neural networks are managed through the effective closure, for practical purposes, of the worlds in which they are designed to operate. The efficacy of systems is measured through their performance on a small set of industry-standard “benchmark” tasks and registered as relative scores on competitive “leader boards” (Bender et al 2021, 618, Inioluwa et al 2021). The promise of translation from laboratory to application that underwrites investments in computational neural networks is sustained through a combination of massive datasets and associated compute power, along with the over-representation of the scope of resulting capabilities. Most salient to this discussion, as we have seen in the case of Geoffrey Hinton, the strategic ambiguity of claims for the brain as a model for computational neuroscience enables retreat to the brain as an inspiration in the face of untranslatable differences between organisms and machines.

The figure of the neural network in the biological neurosciences is similarly inflected by twentieth century technological imaginaries, and within the confines of the laboratory the neurosciences operate according to regimes of experimentation that offer another form of self-referential closure.[14] On the margins of mainstream neuroscience, however, an alternative project of translation is underway that builds on feminist critiques of heteronormative figurations of brains, bodies, and worlds, to articulate a relational ontology of biological difference (Haraway 1989, Martin 1991, Fausto-Sterling 2012). To explore the implications of that project for the case of neural networks, I turn to the work of feminist neuroscientist Gillian Einstein, head of the Einstein Laboratory for Cognitive Neuroscience, Gender and Health at the University of Toronto.[15] In a paper titled ‘Situated Neuroscience: Exploring Biologies of Diversity,’ Einstein describes her project as one of “research into the nervous system that would give voice to areas of research previously silenced, uncover pockets of ignorance – not just ‘knowledge gaps’ – [and] turn expectations about the essentialism of biology on its head” (2012, 150).

To concretize what that could mean, Einstein reflects on her investigation of the neurobiological effects of the traditional North East and West African practice of

Female Genital Cutting (FGC). These effects, she hypothesized at the outset of her inquiry, might be more broadly constitutive of associated 'normal/desirable' women's bodies:

> the result of the involvement of the central nervous system (CNS) would be to embody the tradition affecting the way women with FGC walked, carried themselves, and generally, experienced the world through their bodies thus, in effect, embodying their culture. I wondered specifically if the purpose of the tradition was to instantiate a corporeal difference in the CNS between male and female that wasn't present without the procedure (151).

To pursue this hypothesis, Einstein realized, would require a series of extensions to the prevailing methods of experimental neuroscience. To begin with, there were no existing investigations of FGC that traced its effects neurobiologically. This "pocket of ignorance" in her view is sustained by the 17th century model of the body as a machine with independent parts or systems (158). Within this model, she observes:

> the brain still sits privileged atop our polarized body with other body systems arrayed like arms, legs, and trunk on a marionette's strings – to be pulled and moved by the brain. Information comes in. The brain processes it. An action is generated and then carried out by the peripheral nervous system. The rest of the body responds. On this view, the brain is the CEO of the body. Perhaps because of this the body itself has not been considered knowledgeable and thus, has not been thought to have its own narrative (159).

In contrast, Einstein insists that:

> the brain isn't the only nervous system the body has. Other nervous systems are hard at work interacting and being affected by the rest of the body. The spinal circuits and the peripheral nervous system – nerves, receptors, and far-flung neurons – as in the retina, dorsal column nuclei and enteric nervous system – all contribute to what the cerebral cortex 'knows' … This underscores the point that body, brain, and society are in a reciprocal relationship mutually affecting each other … the world writes on the whole body (160).[16]

Beyond a commitment to tracing whole body connections, then, Einstein's hypothesis required a more radical expansion of her methods, beyond bodies to worlds. This took her outside the laboratory into a collaboration with 14 Somali Canadian immigrants to Toronto (she emphasizes that they were positioned in the research as colleagues rather than subjects), in whose native country the practice of FGC still affected 98% of women at the time of her study. Einstein's collaborators, she is careful to point out, are not meant to be representative of all Somali women, or women who have experienced FGC: "Most were abroad visiting, studying, or working when the war broke out [in the late 1980s and early 1990s], and they simply never went home. They are healthy, engaged, energetic women with a particular sense of their place in the world" (163). And just as there is no singular Somali culture or category of woman, she emphasizes, there is a multiplicity of FGC.

The brain played an important role in the study, not only insofar as Einstein was interested in brain/body/world interconnections, but also methodologically in the openings that a shift to neurobiology afforded in her conversations with her interlocutors. As Einstein explains:

> I was able to start out the conversation with each woman by saying that I was not interested in her genitals; I was interested in her brain. Redirecting the questions from the genitals – a site of silence in cultures practicing FGC – to the brain – a site not previously considered but privileged in the popular imaginary, allowed participants to talk about their circumcision as well as placing the topic in what for them was a respectful space (161).

The results of the study confirmed FGC's whole-body effects for these women, including those expected in the initial hypothesis (Perović et al 2021). At the same time, the study participants did not see themselves as disabled or unwell, and some expressed pride in what they had endured. While Einstein is mindful of the small size and specificity of her study population, she underscores its significance as further evidence for the inseparability of the brain from the body's multiple and interconnected nervous systems, and the inseparability of 'pain' in both its measurable and its experiential forms from neurobiology's cultural embodiment.

More generally Einstein insists that the body has no independent parts, and rather than the privileged brain "the nervous system is an integrator of and integrated with the entire body and the world … Thus, a practice that affects one part of the body will be owned by the entire body or, embodied through the interconnections of all body systems and the environment" (2012, 158). Einstein expresses her "love" for the endocrine system, observing that through the medium of the blood it works as "a huge unifying effector; people call it a modulator, but if you were really committed you could say that it was the starter, not the modulator. You could take any part of the body that's been studied as a system unto itself and realize that it's a modulator for the entire body and modulated by it" (Interview, May 2022). Breakdown, moreover, is an opening to new knowledge: while the "small lie" that systems are distinct may have some predictive power, Einstein observes, “where it doesn’t work is a good sign of where things are really interconnected” (ibid.).

Einstein conceptualizes the brain in terms of neural ‘circuitry’ but for her the connections are not only profoundly and complexly embodied, but also change with experiences not reducible to weighted inputs (2012, 162). Asked to talk about the limits of the neural network analogy in computational neuroscience, Einstein points to the premise that the top layer of the computational neural network is analogized to input receptors, while the middle layers determine, through back propagation and in unaccountable ways, the system’s output. Neuroscientists, in contrast, care about what is in between:

> In the brain, a network consists of many neurons in many brain areas interconnecting. Whether you think of it as hierarchical or parallel processing, there are identified, individual neurons in a network. In the primary visual cortex, or any cortex, we like to think anyway that we know what kind of neurons are in layer one, layer two, where the neurons in layer three project to, where layer four gets their input, that is, the wiring diagram. As I understand current brain analogies in computer science models, one doesn’t need to know about the specificities those layers, because back propagation doesn’t require that we understand the details of what’s between input and output. But in the brain, I think it does require an understanding of these intricacies to model how an actual brain processes input. As well, for the internal connections of a given brain region, there are external inputs from other brain regions that might be getting completely different information about the world (hormonal, sensory, etc.) (Interview, May 2022).

We might recall that Hinton’s narrative differentiates neurons exclusively on the basis on relative weights; all neurons are in a sense commensurable. Einstein’s account, in contrast, indicates different classes of neurons (based on location, which dictates function) and her comments above regarding the endocrine system suggest that modulators of intelligence are diffused throughout the nervous system and difficult to ‘address’ in a fixed way.[17] While the intersections of neuroscientific and electrical engineering imaginaries are evident in the figure of the ‘wiring diagram’, for Einstein it is the methodological implications of the computational claims that are most troubling. She admits to her own investments in the laboratory methods of neuroscience, with its goals of “taking things apart to see what they do,” and confesses that in her research she is less disturbed if there is a brain change that doesn’t show up behaviorally, than she is if there is a behavioral change for which there is no observable brain change.[18] When asked how she accounts for the strength of this commitment she explains:

> I believe behaviour is organized by brain circuits. What might ultimately result in behavioural changes might manifest early in the brain. While there may be a brain change that isn't measurable or for which the measure hasn't been discerned – and therefore, we don't 'see' it, if there is a measurable brain change for which a behaviour hasn't yet been observed, I'm not surprised. It takes a lot of neurons to organize a behaviour. So, some neurons may have changed but not enough yet to yield a behavioural change (Interview follow up, January 2023).

For Einstein experimentation is not a mechanistic or reductionist project because she sees the neuron as inseparable from its relations, both within and beyond the body. Methodologically, this poses the ongoing challenge of holding together what she characterizes as "a big world and a minute world" (interview, May 2022). The former is the person/body/world involved in studies of sex, gender, and women's health, while the latter is the microscopic world of the laboratory sciences. The 'immeasurable results'[19] of qualitative research and the measurement systems of experimental science are often difficult to connect: in Einstein's work on cases of female genital cutting self-assessments of chronic pain using standard indices, physiological indicators, and reported experience from more extended qualitative interviews fail to align in any simple way (Perović et al 2120), and premenstrual syndrome shows measures of hormone levels that do not correlate with mood (Romans et al 2012). When the connections do appear, she observes, research interlocutors' familiarity with received narratives of brain/body relations, and their attunement to popular framings of the problems, further complicate the project of determining cause and effect (Interview, May 2022).

At the close of our conversation Einstein comments that she often thinks about the standard scientistic drive to discover something generalizable and wonders how this can be done without taking modulating systems like the endocrine (or immune, etc.) system and context into account:

> I don't think that the general scientific goal of producing something 'fundamental' is going to work very well for real-world biological systems. In real biology, difference and variation are what is 'fundamental'. To find the 'true' we need to restrict our claims to the exact conditions/organisms/modulated state under which the study was carried out. If someone homogenizes difference and glosses over variation, they will lose what makes these systems tick. I think we need to be modest in the face of the brain. We can learn something about CA1 pyramidal neurons in aging female Sprague Dawley rats that have had their ovaries removed but this doesn't really tell us about even the same neuronal type in another species of rodent and certainly not in humans. I tell my students that what

> we are learning and reporting on is a particular phenomenon, in this animal (human or non-human), at this age, in a given context. Perhaps to really know something true is to restrict our claims, and then begin to compare those claims with others.

The sense of fundamental that Einstein resists here is one that works to delimit the insides and outsides of the research object – specifically in the case of computational neuroscience the brain as neural network – by rendering other systems as epiphenomenal and so extraneous.[20] For Einstein, in contrast, the biological brain is taken as inseparable from the body-in-the-world, and a neuroscience capable of fundamental insight requires methods that expand to incorporate their object's constitutive relations. Rather than containing her research object, in other words, she is committed to reopening and reconfiguring its boundaries as her understanding of it deepens.

What We Might Learn When the Neural Network Imaginary Encounters its Limits

Lepage-Richer (2021, 200) adopts the trope of 'adversarial epistemology' to argue that knowledge claims for neural networks are "historically contingent on a larger techno-epistemic imaginary which naturalizes an understanding of knowledge as the product of sustained efforts to resist, counter, and overcome an assumed adversary." In the case of computational neural networks, the adversary is not only a brain that hides its secrets, but also competing propositions for how those secrets might be disclosed. It might seem ironic that Marcus and others (see Olazaran 1996) attribute the strength of commitment on the part of neural network proponents less to either data or logic than to historic lines of struggle and animus within the computing community, beginning in the 1960s when Marvin Minsky and Seymour Papert drew up their critique of Frank Rosenblatt's 'perceptron' (Marcus 2022). The re-emergence of neural networks in the 1980s was framed explicitly as an irreconcilable alternative to the symbolic approach.[21] While the symbolic approach was characterized as having reached a dead end, resulting in the so-called AI Winter of the 1980s, Marcus now proposes a new 'hybrid' way forward, identified as a 'neurosymbolic' approach (Marcus 2022).[22]

In other words, as the capacities of so-called deep learning approaches reveal their limits there is a return in computational neuroscience to the idea that intelligence requires the manipulation of symbols.[23] Marcus (2022) characterizes symbol manipulation as involving two essential ingredients: "sets of symbols (essentially just patterns that stand for things) to represent information, and processing (manipulating) those symbols in a specific way, using something like algebra (or logic, or computer programs) to operate over those symbols.” Once the brain is understood to be involved in symbol manipulation, and symbols translated as code (strings of binary digits or bits), neural processes read as symbol manipulation are ready made for computational operations. As Marcus explains: “Symbols offer a principled mechanism for extrapolation: lawful, algebraic procedures that can be applied universally, independently of any similarity to known examples. They are (at least for now) still the best way to handcraft knowledge, and to deal robustly with abstractions in novel situations” (2022).[24]

Deep learning approaches hoped for the ‘emergence’ of intelligence given sufficient data along with sophisticated techniques for the detection of potentially meaningful correlations. Yet in their fundamental assumptions and commitments, deep learning and symbolic approaches share more than they offer up in the way of alternatives. While one relies on statistical analysis and the other on the encoding of algorithms that determine computational operations (so-called ‘rules’), both have already translated cognition into a problem of computation before the research begins. Whether a product of stochastic processes or ‘abstract reasoning,’ comprehension is to be achieved through operations enabled by the brain’s capacity to ‘recognize’ and translate input from an externalized world into manipulable numbers, translated in turn into appropriate output.

The brain/mind as an abstract reasoning machine, on this logic, then needs to be put into interaction with a ‘real world’ understood as outside and separate from it. For biology, the connections of brain/body/world are made through relational processes intrinsic to organisms/environments, while for computational neuroscience these are interfaces to be designed. The now well-developed critique of this form of nature/culture dualism is too extensive to rehearse here.[25] But crucial for the purposes of this essay is the premise that there is an inseparable relation between cultural/historical articulations of the real and the real worlds that we as humans inhabit. The delineation of brains and computational systems, and the articulation of sameness and/or difference between them, is not an innocent matter of objective observation but rather a project of worlding[26] in which all of us engaged in the discussion are implicated. And the stakes are not confined to the laboratory, but have political, economic, and material consequences for how we draw the boundaries of the human and of other/more than human relations.

The moral of these stories that I hope to draw out concerns the limits of a commitment to containment and closure in both theory and method. I have suggested that the commitment to cognitivism, *inter alia*, sustains the closed world logics of laboratory computational sciences. That commitment seems to leave practitioners with little recourse, when encountering the theoretical and methodological limits of their practice, other than a return that promises a new and salutary synthesis. The alternative, I have suggested, is something closer to the critical technical practice envisioned by Agre (1997: 23) when he wrote:

> Instead of seeking foundations it would embrace the impossibility of foundations, guiding itself by a continually unfolding awareness of its own workings as a historically specific practice. It would make further inquiry into the practice … an integral part of the practice itself. It would accept that this reflexive inquiry places all of its concepts and methods at risk. And it would regard this risk positively, not as a threat to rationality but as the promise of better ways of doing things.

As the work of Gillian Einstein makes evident, the fulfilment of Agre's call requires a tolerance for some mess and incompleteness in one's knowledge making practices, and humility regarding one's knowledge claims. In the hands of a critical practitioner, encounters with the contingency and partiality of knowing are taken not as a sign of a failure that needs to be hidden, but of the irremediable openness of worldly relations. Those relations involve modes of learning that are deep not in the sense of the multiplication and ingenious manipulation of homogeneous arrays of numbers, but through their implication in practices of ongoing and heterogeneous world-making.

I am deeply grateful to Gillian Einstein not only for her feminist technoscientific practice but also for her generous engagement in the development of this chapter. Thanks also to Théo Lepage-Richer, Ranjodh Singh Dhaliwal, Andrew Clement and Ben Gansky for critical and sympathetic readings of an earlier draft.

Notes

[1] For indicative works see Dreyfus 1992, Goodwin 1994, 2017, Hutchins 1995, Lakoff and Johnson 1999, Lave 1988, Lynch 1993, Myers 2015, Suchman 2007b.

[2] This could be seen as an extension of the strategy adopted by early 20th century neuroscientists as described by Star (1992, 213), where "difficulties that could not easily be addressed by some physical or medical model were relegated to 'mind'-related lines of work, such as psychiatry and psychology." The cognitivist project of computational neural networks is to render mind as a technology, relegating the specificities of the brain to the realm of neuroscience while maintaining the position of bodies and worlds as outside its bounds.

[3] On the historical connections between closed worlds and cognitivism in the context of geopolitics and computing see Edwards 1996.

[4] As well as a critical characterisation of the widespread embrace of the neuro across the behavioral, social and biological sciences, Vidal and Ortega (2011) identify neuroculture as a "practical field that bridges advances in the neurosciences or brain sciences … with

health, military, criminological and other regulatory policies that seek new forms of individualised risk management and social control …neuroculture is not simply a question of the power or persuasive appeal of the neurosciences within the laboratory or clinic, but of their wider social, cultural, political and economic salience and significance about the future of humanity and potential for its optimisation."

[5] Brosnan and Michael note the call for a "critical neuroscience" that, with little reference to practice-oriented scholarship in other fields, recovers bodies and worlds through reference to cognate moves within the cognitive sciences, beginning with Varela et al (1991). Promoting this call, Slaby and Gallagher (2015) posit the existence of what they name "cognitive institutions," of which they propose science in general, and neuroscience in particular, as exemplars. Such institutions, "through various practices and rules, shape our cognitive activity so as to constitute a certain type of knowledge, packaged with relevant skills and techniques" (2015, 35).

[6] Or at least of the mainstream of neuroscientific investigation: this is not to say that it couldn't be otherwise, as will be suggested in the second half of this chapter.

[7] In biochemistry to denature something is to 'destroy the characteristic properties of (a protein or other biological macromolecule) by heat, acidity, or other effect which disrupts its molecular conformation' (*Oxford Dictionary of English)*. As in biochemistry, denaturing as a process of decontextualization is unstable, subject to disruption by interactions that can't be contained.

[8] https://www.etymonline.com/word/network

[9] For an eloquent exposition of the differences that matter between calculation, as a form of reckoning, and judgment see Smith 2019.

[10] https://www.cs.toronto.edu/~hinton/. I return to the figure of learning below.

[11] More specifically, it is the programmer who defines the system's 'objective functions,' which in turn determine the relative 'correctness' of its outputs, where both are constrained by what a computational system can do. Hinton explains to Thompson that growing dissatisfaction with the labelled data required for back propagation has led to a greater commitment to novel approaches to so-called 'unsupervised learning'. Rather than explicit classification of input, unsupervised learning is the detection of correlations in computationally legible inputs (for example pixels in the case of computer vision) translated mathematically as 'feature detectors,' which in turn become the input data for subsequent layers of the network until an effective 'data model' is generated. While effectiveness in this context still means that system output is aligned with results assessed as meaningful by associated humans, the generation of that output is further automated and less reliant on human labor.

[12] There is of course no inherent contradiction between inspiration and model, as these could easily work in a complementary fashion. The issue is one of accountability, specifically the way in which reversion to the status of inspiration operates here as a hedge on the stronger claim to be engaged in modelling.

[13] Ben Gansky (personal communication) points out that Marcus' "prior knowledge" deficit, supposedly solvable through further encoding of abstract concepts, posits a free-floating corpus of knowledge retrievable on demand. Extensively critiqued within feminist theory, this conceptualization ignores realities of learning and intelligence as always specifically situated in lived experience and positionality within a sociotechnical landscape. For a critical examination of the premise that knowledge can be represented as a corpus of common-sense knowledge see Adam 1998.

[14] Traveling outside of the laboratory, neuroscientific technologies of imaging and techniques of diagnosis have inspired expansive therapeutic projects, not least in partnership with pharmaceutical industries (see for example Dumit 2004).

[15] For a fuller biography see https://einsteinlab.ca/about-us/our-lab/gillian-einstein/

[16] This argument is developed further in Brown et al 2022.

[17] I'm grateful to Ben Gansky (personal communication) for this point. For an argument regarding addressability as a core requirement for all approaches to computing see Dhaliwal 2022.

[18] She adds with a laugh "I only tell that to my best friends" (Interview, May 2022)

[19] 'Immeasurable results' is the title of a painting by Lynn Randolph, which provides the frontispiece for Haraway 1997.

[20] An analogous case might be theories of language that position context as complicating rather than constitutive of communication. See discussion in Suchman 2007a, chapter 7.

[21] As a notable exception Marcus (2022) cites Hinton's 1990 *Connectionist Symbol Processing*, as "explicitly aimed to bridge the two worlds of deep learning and symbol manipulation"; a project that, Marcus observes, Hinton subsequently abandoned. "When deep learning reemerged [after another brief winter in the early 2000s] in 2012, it was with a kind of take-no-prisoners attitude that has characterized most of the last decade."

[22] Yann LeCun, similarly, has recently advocated a 'bold new vision' for AI involving a return to a "cognitive architecture" that includes symbolic reasoning, planning, and "common sense" (Heikkilä and Heaven 2022).

[23] Marcus (2022) cites as a turning point in over-optimism regarding the power of neural networks the 2021 NetHack 'challenge,' hosted at NeurIPS 2021 as a partnership between Facebook (now Meta), AI Research, AI Crowd, Oxford, UCL, and NYU, and supported by sponsors Meta AI, and DeepMind. The most recent within the tradition of closed, game-world competitions, the challenge was won by Team AutoAscend with a non-neural net, symbol-manipulation based approach. Accessed May 26, 2023. https://nethackchallenge.com/report.html.

[24] Arguably neural networks are already also symbol-manipulating systems, insofar as they rely on curated databases as their training materials.

[25] Primary references in feminist theory include Barad 2007; Butler 1993; Haraway 1989, 1991, 1997; see also Latour 1993.

[26] 'Worlding' is a term introduced within contemporary anthropology to emphasize the ongoing discursive and material practices through which the (always relational) entities that comprise specific cultural and historical realities are enacted, as well as the limits to translation among them. See de la Cadena and Blaser 2018.

References

Adam, Alison. 1998. *Artificial knowing: gender and the thinking machine*. New York: Routledge.

Agre, Philip. 1997. *Computation and human experience*. New York: Cambridge University Press.

Barad, Karen. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, North Carolina: Duke University Press.

Bender, Emily, Gebru, Timnit., McMillan-Major, Angelina, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT' 21*. Accessed May 26, 2023. https://dl.acm.org/doi/10.1145/3442188.3445922.

Brosnan, Caragh and Mike Michael. 2014. Enacting the 'neuro' in practice: Translational research, adhesion and the promise of porosity. *Social Studies of Science, 44*(5), 680-700.

Brown, Nick. and Mike Michael. 2003. A sociology of expectations: Retrospecting prospects and prospecting retrospects. *Technology Analysis & Strategic Management* 15(1): 3–18.

Brown, A., Karkaby, L., Perovic, M., Shafi, R., Einstein, G. 2022. Sex and Gender Science: The World Writes on the Body. In: *Current Topics in Behavioral Neurosciences*. Springer, Berlin, Heidelberg.

Butler, Judith. 1993. *Bodies that matter: on the discursive limits of "sex"*. New York: Routledge.

de la Cadena, Marisol and Mario Blaser (Eds.). (2018). *A World of Many Worlds*. Durham and London: Duke University Press.

Dennett, Daniel. 2013. *Intuition Pumps And Other Tools for Thinking*. New York: W. W. Norton.

Dhaliwal, Ranjodh Singh. 2022. On Addressability, or what even is computing? *Critical Inquiry, 49*(1), 1-27.

Dreyfus, Hubert. 1992. *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Dumit, Joe. 2004. *Picturing Personhood: Brain scans and biomedical identity*. Princeton: Princeton University Press.

Edwards, Paul. 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, MA: MIT.

Einstein, Gillian. 2012. Situated Neuroscience: Exploring Biologies of Diversity. In R. Bluhm, A. J. Jacobson, & H. L. Maibom (Eds.), *Neurofeminism* (Vol. New Directions in Philosophy and Cognitive Science., pp. 145-174). London: Palgrave Macmillan.

Fausto-Sterling, Ann. 2012. *Sex/Gender: Biology in a Social World*. New York and London: Routledge.

Goodwin, Charles. 1994. Professional Vision. *American Anthropologist, 96*(3), 606-633.

Goodwin, Charles. 2017. *Co-Operative Action*. Cambridge: Cambridge University Press.

Hall, Stuart. 2003. Marx's Notes on Method: A 'reading' of the 1857 introduction. *Cultural Studies, 17*(2), 113-149.

Halpern, Orit. 2022. The Future Will Not Be Calculated: Neural Nets, Neoliberalism, and Reactionary Politics. *Critical Inquiry, 48*(2), 334-359.

Haraway, Donna. 1989. *Primate visions: gender, race, and nature in the world of modern science*. New York: Routledge.

Haraway, Donna. 1991. *Simians, cyborgs, and women: the reinvention of nature*. New York: Routledge.

Haraway, Donna. 1997. *Modest _Witness @Second_Millenium.FemaleMan_Meets_OncoMouse™: Feminism and Technoscience.* New York: Routledge.

Heikkilä, Melissa and Will Douglas Heaven. 2022. Yann LeCun has a bold new vision for the future of AI. MIT Technology Review, June 24. Accessed May 26, 2023. https://www.technologyreview.com/2022/06/24/1054817/yann-lecun-bold-new-vision-future-ai-deep-learning-meta/.

Hutchins, Edwin. 1995. *Cognition in the wild*. Cambridge, Mass.: MIT Press.

Inioluwa, Deborah Raji, Bender, Emily, Paullada, Amandalynne, Denton, Emily, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. https://arxiv.org/pdf/2111.15366.pdf.

Interview by Lucy Suchman of Gillian Einstein. May 26, 2022, follow up January 29, 2023. Unpublished.

Lakoff, George and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Latour, Bruno. 1993. *We have never been modern*. Cambridge, Mass.: Harvard University Press.

Lave, Jean. 1988. *Cognition in practice: mind, mathematics, and culture in everyday life*. Cambridge; New York: Cambridge University Press.

Lave, Jean. 2011. *Apprenticeship in Critical Ethnographic Practice*. Chicago and London: University of Chicago.

Lepage-Richer, Théo. 2021. Adversariality in Machine Learning Systems: On Neural Networks and the Limits of Knowledge. Chapter 7 in *The Cultural Life of Machine Learning: An incursion into Critical AI Studies* Roberge, Jonathan and Castelle, Michael (eds). Palgrave.

Lynch, Michael. 1988. Sacrifice and the transformation of the animal body into a scientific object: Laboratory culture and ritual practice in the neurosciences. *Social Studies of Science* 18(2): 265–289.

Lynch, Michael. 1993. *Scientific practice and ordinary action: ethnomethodology and social studies of science*. New York: Cambridge University Press.

Marcus, Gary. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*

Marcus, Gary. 2022. Deep Learning is Hitting a Wall. *Nautilus*. March 10. https://nautil.us/deep-learning-is-hitting-a-wall-238440/, accessed October 2022.

Marcus, Gary, Marblestone, Adam, and Thomas Dean. 2014. The atoms of neural computation. *Science*, 346 (6209), 551-552.

Martin, Emily. 1991. The Egg and the Sperm: How science has constructed a romance based on stereotypically male-female roles. *Signs: Journal of Women in Culture and Society, 16*(3), 485-501.

Martin, Emily. 2004. Talking Back to Neuro-reductionism. In H. Thomas & J. Ahmed (Eds.), *Cultural Bodies: Ethnography and Theory* (pp. 190-211). New York: Whiley & Sons.

McCarthy, John, Minsky, Marvin, Rochester, Nathaniel, and Claude Shannon. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Accessed May 26, 2023. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

Myers, Natasha. 2015. *Rendering Life Molecular: Models, Modelers, and Excitable Matter* Durham: Duke University Press.

Olazaran, Mikel. 1996. A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science, 26*(3), 611-659.

Perović, Mateja, Jacobson, Danielle, Glazer, Emily, Pukall, Caroline, and Gillian Einstein. 2021. Are you in pain if you say you are not? Accounts of pain in Somali–Canadian women with female genital cutting. *PAIN: The Journal for the International Association for the Study of Pain*. 162 (4): p 1144-1152, April 2021.

Romans, Sarah, Clarkson, Rose, Einstein, Gillian, Perović, Mateja, and Stewart Donna. 2012. Mood and the Menstrual Cycle: A Review of Prospective Data Studies. *Gender Medicine* 9(5): 361-384.

Schull, Natasha and Caitlin Zaloom. 2011. The shortsighted brain: Neuroeconomics and the governance of choice in time. *Social Studies of Science, 41*(4), 515-538.

Slaby, J. and Gallagher, S. 2015. Critical Neuroscience and Socially Extended Minds. *Theory, Culture & Society, 32*(1), 33-59.

Smith, Brian Cantwell 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press.

Suchman, Lucy. 2007a. *Human-Machine Reconfigurations: Plans and Situated Actions, revised edition.* New York: Cambridge.

Suchman, Lucy. 2007b. Feminist STS and the Sciences of the Artificial. In E. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies, Third Edition* (pp. 139-163). Cambridge, MA: MIT Press.

Star, Susan Leigh. 1992. The Skin, the Skull, and the Self: Toward a Sociology of the Brain. In A. Harrington (Ed.), *So Human a Brain*. Boston, MA: Birkhäuser.

Thompson, Nicholas. 2019. An AI Pioneer Explains the Evolution of Neural Networks. *Wired*, May 13. Accessed May 26, 2023. https://www.wired.com/story/ai-pioneer-explains-evolution-neural-networks/.

Varela, Francisco, Thompson, Evan, and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT.

Vidal, Fernando. 2009. Brainhood, anthropological figure of modernity. *History of the Human Sciences, 22*(1).

Vidal Fernando and Franciso Ortega. 2011. Approaching the neurocultural spectrum: An introduction. In: Ortega F, Vidal F (eds) *Neurocultures: Glimpses into an Expanding Universe*. Frankfurt am Main: Peter Lang, pp. 7–27.