Optimal Brain Damage: Theorizing our Nervous Present

Johannes Bruder and Orit Halpern

Introduction

The COVID 19 pandemic has seemingly naturalized the relationship between computation and human survival. Digital systems, at least in the Global North, sustain our supply chains, labor, vaccine development, public health, and virtually every manner of social life. Nowhere has this link become more powerful then at the intersection of statistics, artificial intelligence and disease modelling.

Early in the pandemic hopes were high that tracing apps, mobility data and statistical models might give the predictions and models for human behavior and social interactions to stop the progress of COVID-19. All over the world, scientists gathered to pool their knowledge in the modelling of complex processes and provide guidelines for measures to contain the spread of the virus. For instance, a group of like-minded scientists formed the Independent Scientific Advisory Group for Emergenics (IndieSAGE), an organization providing independent scientific advice to the UK government and public on how to minimize deaths and support Britain's recovery from the COVID-19 crisis.[1] Somewhat surprisingly, the group did not simply rely on epidemiological models but used one of their member's hard-won knowledge about interactions of neurons in the human brain to infer on how social interaction could influence the spread of the virus in the UK (Friston *et al.* 2020a, Friston *et al.* 2020b). 'Here, we use a ubiquitous form of model', the authors write, 'namely a mean field approximation to loosely coupled ensembles or populations. In the neurosciences, this kind of model is applied to populations of neurons that respond to experimental stimulation… Here we use the same mathematical approach to model a population of individuals and their response to an epidemic' (Friston *et al.* 2020a). These models made neural nets proxies for human behavior in order to model disease spread during the pandemic.

Fig. 1. Tweet by Anthony Costello, Director of Global Health at University College London and member of IndieSAGE. https://twitter.com/pritjohn/status/1248662106758578177.

Those who took notice of the resulting models were divided about the question of whether neuroscientists would be well-suited to modeling the onslaught of the virus. When Anthony Costello, Director of Global Health at University College London and like Friston member of IndieSAGE tweeted about the model, he received many skeptic comments. Those who defended the efforts of Friston and colleagues, however, provided clues as to why a neuroscientist who would typically model neuronal, cognitive processes in the brain is an excellent fit to help halt the pandemic. 'Because of [Friston's] understanding of the dynamics of complex systems mathematically', one commenter replied. 'The brain, the body, and indeed the spread of viruses are all complex causal networks' (see Figure 1).

What these three "networks" have in common from a modling standpoint is that attempts at modeling and predicting the interactions of their elements is plagued by a characteristic uncertainty, and by the economic problems this uncertainty creates. Models of complex systems quickly become unfathomable, simply because unfolding dynamics are hard to predict. For brains, bodides, and societies, it is therefore important to reduce uncertainty by focusing on specific parameters or metrics.

In fact, Friston and colleagues' SEIR Covid-19 model is thus an attempt to model of uncertainties, and it is designed to allow for swift interventions when networked interactions get out of control and cause damage to brains, bodies, and societies. One of the model's central, theoretical elements are so-called Markov blankets, which are statistical definitions of a system's boundaries and go back to AI pioneer Judea Pearl (see e.g., Pearl 1988). Markov blankets are include all the information that the network needs to survive. What is not contained within is technically expendable and can be forgotten—from a modelling standpoint. Friston considers Markov blankets as a central rationality of human cognition: our brains decide which information is relevant for us to respond to events in our environment. Information that appears to be irrelevant for survival is ignored in favor of an efficient response to the situation, especially in times of crisis. . If the brain—as Friston argues—is considered a good, general model of complex network dynamics, these principles apply also to societies confronted with financial crises or the current pandemic. 'The assumption that connectivity is always a good thing is for me so

naive', Friston consequently said in an interview. 'From the point of view of that delicate self-organization that enables these Markov blankets that constitute ourselves, or a society, or an ecosystem to survive, connectivity is the killer… It's basically putting energy into a system and literally boiling it—and destroying all that delicate structure' (Wing Kosner 2019).

While Friston's efforts are but one model, the popularity of his ideas in inspiring epidemiologists, politicians, and neural network developers demonstrates the possibility that ideas of neuroscience and mental health can also then become architecture for computing and social policy. Such ideas of keeping systems healthy by reducing connectivity could lead to legitimize the closing of borders and the expulsion of all things unnecessary, if deployed in certain contexts.

There is much at stake in this account. Using neuro-science knowledge to model disease epidemics might appear unintuitive, or even foolish. But the greater scientific interest in networks, and the way that brains are considered possible models of non-linear network dynamics, hints to an underlying epistemology of contemporary neoliberal governmentality that is reflected also in the design of neural networks.

The links between neuroscience and machine learning or artificial intelligence are currently often historicized, with reference to the purely statistical nature of contemporary algorithmic systems (see e.g., Arnold & Tilton 2021). But we argue that the idea and ideology of cognition do in fact bridge the wide gaps between brains and neural networks, and between populations of neurons and populations of human individuals when it comes to questions of governance and systemic health.[2] Indeed, Friston's theories are considered as cornerstones of current and future artificial intelligence.[3] He himself emphasizes his fruitful exchanges with, and the general compatibility of his approach to dynamic causal modeling with Canadian AI forerunner Geoffrey Hinton's Helmholtz machines.

The idea that all complex systems are, or can be, modeled as networked populations of decision making 'neurons' hence sits at the heart of this paper. This entanglement—and we are not using this term lightly—of neuroscience and machine learning has a long history.

We offer a genealogy of this abstract understanding of population dynamics, which we call '*the neural imaginary*'. The neural imaginary is the idea that *populations* of neurons can be aligned with the behavior of *populations* of humans, and that models which abstract from the nature of a population's elements can explain such seemingly incomparable phenomena as learning, financial crises, and the spread of a virus during the current pandemic. As we hope to demonstrate, the seemingly dated and insignificant neural imaginary thus has enormous impact on the future management of planetary populations and life through technological means, for it suggests transferring knowledge about how the brain protects itself against the effects of information overload to neural networks and societies.

To make this argument we will trace the neural imaginary through the links between the work of the Canadian neuroscientist Donald O. Hebb, and its later influence in neo-liberal economic thought and on the development of neural networks for machine learning. We are tracing an epistemic shift in the comprehension of cognition and decision making. Within fields as varied as economics to machine learning, there emerged a new model of that believed that networked systems could accomplish acts of evolution, change, and learning impossible for individual neurons or subjects—minds, machines, and economies should therefore be governed to change and deal with the unexpected. Their understandings, we argue, were symptomatic of a broader cultural change in how minds and machines were understood at the time (Hayles 1999, Kay 2001).

Most critically, these new models of systems imagined volatility and change as necessary for healthy systems and sources of productivity and growth—whether in brains, markets, or societies. By the late 1980s, shock—whether through sensory deprivation, fake data, wrong information, viruses, noise, or sensory overload –was conceived as both unavoidable and potentially productive. Hebb's experiments with sensory deprivation, however, had proven that the dynamic instability of networks can also have catastrophic effects. Striking a happy medium between plasticity and stability, between exposing to and protecting from shock has since become a centerpiece of research on the brain and on networks.

In the second part of this paper, we turn to the designs of contemporary neural nets, and techniques such as 'optimal brain

damage' (LeCun, Denker, and Solla 1990), which respond to Hebbian concerns about plasticity and stability by turning homeopathic doses of shock and trauma into experimental techniques devised to *protect* neural networks in the long run. In this case, cognitive bias—the cognitive science correlate of Markov bankets—appears as a mechanism that allows the network to become more efficient and survive. While discussions revolving around ethical AI are currently often focused on eliminating bias in, and creating inclusive data sets, we argue that cognitive bias is inherent to the very idea of machine learning and AI. We thus ask a more general question that pertains to the epistemology and rationality behind the SEIR Covid 19 model and contemporary neural networks: what is the effect of trying to govern populations of individuals mainly by managing the 'health' of the network?

Stochastic Brains

In 1949, the Canadian neuroscientist Donald O. Hebb announced a new conception of the mind. 'It is impossible', he wrote, 'that the consequence of a sensory event should often be uninfluenced by the pre-existent activity [of the neurons]... the problem for psychology is no longer to account for the existence of set but to find how it acts and above all to learn how it has the property of consistent, selective action…' (Hebb 1949: 6). Neurons, Hebb argued, are not static relays of data, merely completing stimulus-response reactions. Rather, he forwarded a stochastic understanding of the brain and intelligence. When synapses fired in concert this increased the probability of cognition. In neuro-science, the finding was summarized as: 'Cells [i.e. neurons] that fire together, wire together.'

Such concepts of plastic networked minds were not solely the inventions of lone psychologists. Hebb was among many to ponder the dynamic mechanics of the brain. In 1943, the McCulloch-Pitts model of the neural net was introduced, and Hebb apparently was influenced by this research and cybernetics. The neural net was perhaps the first logical demonstration of how neurons could theoretically (at least) physically compute logical problems, proving that psychic processes could emerge from physiology. The model was an enormous reduction from real brains, but it inspired a new concept of minds as both machinic and programmable (Halpern 2014).

Like the original neural nets that inspired them, Hebbian networks were also theories of memory and storage. He elaborated that these networks, now labelled 'Hebbian synapses' were synchopated in time and could be trained:

> Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability... When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. (Hebb 1949: 62)

The model posited that neurons that fire in temporal relationships to one another (syncopated although not synchronous) 'strengthen' their relationship –the more they repeat the action, the stronger the net. Neuronal nets are thus 'weighted' statistically. The more often they fire together, the more likely they will do so in the future; they are *learning*. (Hebb 1949: 62-65)

Hebb's idea of synapses implied that what was stored in a brain, the content of perceptions, memories, and cognitive commands, was not the result of an infinite data base of stored information (this was not a theory of the infinite archive of the Freudian unconscious) but rather comprised of patterns, or 'nets' of relations between neurons. The archive was one of patterns not of stimuli. Certain stimuli would trigger networked pathways that collaboratively created an action of thought or behavior. One did not store every image of a cat for example, but rather stored a pattern that would trigger upon the stimulus of a cat. Brains store a process or an architecture, not specific pieces of data.

Such abstract notions of minds that anticipate contemporary deep-learning (as we will demonstrate), did not come from nowhere. Hebb was working with individuals who had suffered injuries to the brain; a problem that was of increasing concern during and after the Second World War. In his research he documented how different cognitive functions might return over time even though parts of their brains were injured. Victims of stroke and accidents all appeared capable, over time, of regaining functions initially lost with the injury. Hebb even found that often cognitive skills and new modes of action could be re-learned by the injured subject, and he assumed this was the

result of the neurons finding new connections circumventing the injury. Correlating these observations with studies of neurons, EEGs, and other rather theoretical and imperfect (by our standards) efforts to visualize neuronal action, Hebb came to the conclusion that networks of neurons are capable of learning by reorganization. Cognition, he concluded, was networked and neurons assembled in certain arrangements might be capable of functioning in ways that were un-anticipatable from their discrete biology or location.

Hebb intended his work as an attack on psychological testing, particularly the racist Binet IQ tests, genetic determinism, and behaviorism (Hebb 1949, Hebb 1937, Hebb 1942, Hebb and Penfield 1940, Hebb 1938). In particular, he opposed the concept that individuals were tied to their biology and upbringing. People, Hebb was convinced, would not store specific pieces of discrete unrelated data but develop cognitive mechanisms in response to their environment—and these mechanisms can be changed through the physical rewiring of networks of neurons.

The derivative corollary of this theory of neuro-plasticity is that our environments re-time our neurons and change memory, cognition, and perception in the same go. Brains could be trained, their nets re-syncopated. Thus, even with physiological changes to the brain (such as an accident) the model posited that new synchronizations and probabilities would develop, allowing a new net of co-activating neurons to emerge. The brain could learn and change at a physiological, neuronal, scale (Hebb 1949: 62).

Hebb's research firmly reconfigured notions of cognition as ecological (the result of interactions between individuals and their environments), made the environment itself a medium for design or technical crafting, and fundamentally transformed understandings of memory and minds as networked and storing populations of patterns or nets rather than discrete data points. In isolation, none of these points appear that important, but collectively and in conversation with similar innovations in economics and computer science we can trace the rise of an emerging 'neural imaginary'.

Machinic Cognition

Hebb's idea of a self-organizing, and perhaps even, evolving intelligence reflected and advanced ideas also found in economics and in the emerging fields of artificial intelligence. For example, in the preface of his often overlooked, early book "The Sensory Order," the famous neo-liberal economist and founder of the Mont Pélerin society, Friedrich Hayek, opened arguing that,

> Professor Donald Hebb's *Organization of Behavior*… contains a theory of sensation which in many respects is similar to the one expounded here; and in view of the much greater technical competence …as I am concerned more with the general significance of a theory of that kind than with its detail, the two books, I hope, are complementary rather than covering the same ground. (Hayek, 1952, location 71, Kindle Edition)

Hayek claimed this relation on the grounds that he felt that there might be a different utility of Hebb's theory, not for reprogramming individual psyches, but for modelling emerging self-organizing phenomena.

The Sensory Order captured the idea that intelligence is networked—whether composed of neurons or human individuals—and that it consists in the capability of populations to adapt to their environment by reorganization. Hebbian notions of psychology, for Hayek, as well as for behavioral economists, and organizational managers (such as Herbert Simon) were the inspiration and model for understanding the human mind and decision-making practices as networked rather than sovereign and individual. These ideas of mind mirrored the idea of a self-organizing market, network, or system. Such concepts also reflected and advanced a broader understanding of intelligence emerging at the time in many locations from organizational management to finance to Buckminster Fuller-esque conceptions of synergetics and systems (Fuller 1975, Halpern 2015, Simon 1955, Mirowski 2002).

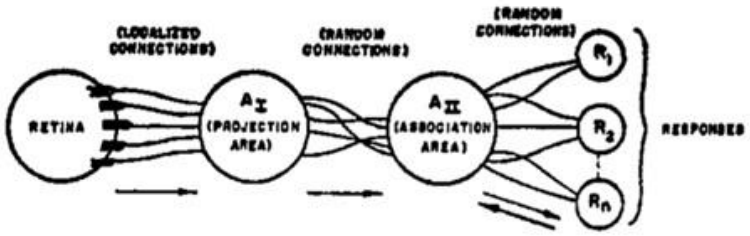

Fig. 2. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," Psychological Review 65, no. November (1958): 388.

In machine learning similar trends were visible. This neural-based, and increasingly neo-liberal, understanding of cognition and intelligence served, for example, as direct inspiration for the first artificial neural network, Frank Rosenblatt's perceptron. Rosenblatt proposed that learning, whether in non-human animals, humans, or computers, could be modeled on artificial, cognitive devices that implement the basic architecture of the human brain. (Rosenblatt 1962)

In his central paper detailing the idea for the perceptron, Rosenblatt wrote that this concept, 'has been heavily influenced by Hebb and Hayek…' (Rosenblatt 1962: 5). Crucially, Rosenblatt's model depended upon a net of neuron-like entities among which associations would be established whenever a sensory organ was triggered by external stimuli (Rosenblatt 1958: 288-89).

The key to learning for the neural net approach was exposure to a 'large sample of stimuli', so that those stimuli which 'are most "similar" . . . will tend to form pathways to the same sets of responding cells' (386-408). As Rosenblatt stressed, this meant approaching the nature of learning 'in terms of probability theory rather than symbolic logic' (388). For Rosenblatt, only stochastic models might teach us about mental properties. Cognition was fundamentally probabilistic. The central tenet of his approach, therefore, is that neurons are mere switches or nodes in a network that classifies cognitive input—intelligence emerges only on the level of the population and through the patterns of interaction between neurons.

We emphsaize the idea that Rosenblatt aspired to a stochastic form of learning, because it emphasizes how learning became a matter of statistical pattern seeking out of data sets, or perhaps

to return to neo-liberal ideas, "markets" of information. It was precisely because perceptrons require training data that decision making became a matter not of consciousness or individual experience, but of networked systems and populations (of neurons).[6] What we are stressing in making these correlations is how these new ideas about decision making through populations reformulated economic, psychological, and computational practices and experimental methods. In doing so, the ideal of networked intelligence, modelled through Hebbian synapses and perceptrons, became the dominant ideology that made human cognition, economic decision making, and machine learning commensurate and part of the same system. This epistemic model of the neuron as the central element of cognition and memory thus signaled a seismic change in the history of intelligence and the conception of agency and decision making. Prediction, or control, became a pressing problem for researchers, economists, and engineers deploying these models. If networks change in response to their environment, how can they be immunized against traumatic events?

Opportune Shocks

Early in his work, Hebb remarked that the 'stability' of learning was sometimes maladjusted to 'perception'. Once a net is trained, how does it maintain its training and not constantly change in accordance with new data? Systems that are too sensitive to new inputs become unstable and lose stability of 'meaning'. Rosenblatt (1962) discovered that errors in weighting might propagate and exacerbate errors, while positive feedback might lead to oscillation and instability. Much of the perceptron model is dedicated to correction of errors including through back-propagation.

This was later labelled the 'sensitivity-stability' problem (see e.g., French 1997). Neural network researchers only refracted a broader discourse repeated by cyberneticians, political scientists, social scientists and economists: what if networked feedback loops fed the wrong positive feedback (for example in nuclear confrontations) leading to network instability and even terminal failure (Halpern 2015, Edwards 1997)?

What we might find surprising, however, is how this seeming terminal problem, became a new-found capacity in nets. In his now infamous sensory deprivation study, Hebb unearthed this

volatile nature of neural nets. The study was funded by the Canadian Defense Research Board. While this research has gained infamy as the progenitor of soft torture in the CIA, its initial goal was far more banal (Klein 2007, McCoy 2006). It was to examine the 'monotony' of the contemporary work environment, and its impacts on attention. Radar operators and other people working in the newly electronic workspace were known to suffer extreme 'boredom', inattention, and depression. To test the monotony of the modern work environment, twenty-two male student volunteers were recruited to lie in a chamber designed to induce 'perceptual isolation' (Heron 1957: 52-53; Crist 2015). The experimental theory correlated the increase of electronic data with sensory deprivation. We might extrapolate that implicitly boredom and information overload were assumed to be related; which is to say too much data given in certain environments might be the same as no data at all.

To ensure maximum 'boredom', the students wore a translucent plastic visor that emitted diffuse light to prevent 'pattern vision', as well as cotton gloves and cardboard cuffs that covered their arms from elbow to fingertips to eliminate or at least reduce tactile stimulation. A U-shaped foam rubber pillow helped dampen auditory stimuli, but according to reports and to the histories of the experiment an air conditioner in the ceiling remained on 24 hours a day that masked small sounds. Intermittently the participants were given verbal and written test for cognitive acuity and memory, and were also made to listen to a battery of recordings with counter-scientific, supernatural, and superstitious propaganda. Afterwards, individuals were examined for their attitudes towards supernatural phenomena and compared with their response to the same questions before the experiment. Individuals had hallucinations and suffered impaired cognitive functioning. By the end many participants seemed to believe in ghosts, and the supernatural, and no one lasted more than four days. The study appeared to demonstrate a way to impact people's thinking without ever touching their bodies (Croft 1954; Heron 1957: 52-53). When adjoined to theories of networked cognition and neuro-plasticity it appeared that brains could be remotely programmed, from afar, through suggestion and environmental manipulation of data. Hebb himself labelled it 'torture', an observation that found concrete realization in the CIA's Cold War interrogations (Crist 2015, McCoy 2006).

After the study, debates have raged over whether participants suffered from too little data, or too much (the sounds and stimulus of the containment are also forms of stimulus potentially). For psychologists, and an army of trainers after Hebb, information overload increasingly became a norm and an expectation; witness the rise of training regimens for the shocks of contemporary life aimed at teaching the subject to concentrate, manage and filter excess data (now labelled 'stress') such as yoga, immersion tanks, self-care, and apps for sleep, concentration, and 'mindfulness', all which supposedly arose from Hebb's research (Lilly 1956, Lilly and Gold 1996, Crist 2015).

Psychologists were not alone in suddenly finding the new programmable nature of the brain and the network an opportunity. Since the 1970s, flash crashes, noise trading, and exponentially leveraged positions have been core concerns but also opportunities in markets now understood as arbitrators of information (Black 1986, Mackenzie 2014, Summers 1990).

Shock As Technology: Optimal Brain Damage

Shock—whether through sensory deprivation, fake data, wrong information, viruses, noise, or sensory overload –was reconceived as unavoidable, with great implications for our present. Healthy brains and neural networks would therefore have to be equipped with a range of mechanisms that govern the effects of volatility on the respective system to allow for continuous change and adaptation while avoiding catastrophic breakdowns. Drawing on Hebb and Hayek, biologist and Noble Prize winner Gerald Edelman offered the idea in 1993 that there are 'major fluctuations in the physiologically detected boundaries of the neural territories and maps to which these connections contribute' (1993: 116). These fluctuations would point towards a sort of 'neural Darwinism' or neuronal group selection that occurs through experience. According to Edelman, neurons would huddle in collectives of variable size and structure to yield adaptive behavior in the organism. As a result, individual nervous systems differ to an extent that 'far exceeds that which could be tolerated for reliable performance in any machine constructed according to current engineering principles' (115).

Indeed, the proponents of artificial neural networks struggled to build perceptive machines that could at least simulate some capabilities of the human brain. Just a few years earlier, psychologists Michael McCloskey and Neil Cohen had identified a 'sequential learning problem' in neural networks (McCloskey & Cohen 1989). It occurs when many of the weights[7] that contain a system's knowledge are modified during the process of learning new tasks or adapting to sudden changes in the environment. For instance, a neural net might lose the ability to play Pac-Man after being trained on Space Invaders, for it purposefully 'forgets' information that it seemingly does no longer need. Older knowledge is over-written, and the network appears to be out of sync with the world.

What McCloskey and Cohen discovered was that the generally productive volatility could also have catastrophic effects and result in a traumatic loss of memory—which is why the sequential learning problem is now also referred to as 'catastrophic forgetting' (Kirkpatrick et al. 2017). To respond to this dilemma, a group of researchers at the AT&T Bell Laboratories experimented with homeopathic doses of shock and trauma, to yield adaptive behavior in artificial neural networks. In 1989, Yann LeCun, now Facebook's chief AI scientist, John Denker and Sara Solla published an article that was fittingly titled 'Optimal Brain Damage' and proposed to improve the speed, efficiency, and reliability of learning in artificial neural networks by selectively (and experimentally) deleting weights that hold unnecessary knowledge and slow the network down. 'The basic idea of OBD', the authors write, 'is that it is possible to take a perfectly reasonable network, delete half (or more) of the weights and wind up with a network that works just as well, or better' (LeCun, Denker & Solla 1989: 598).

Optimal brain damage riffed off a process that comes naturally to human brains, and by which synapses are removed if they are no longer needed, allowing for new connections to be made. In human brains, this so-called synaptic pruning occurs especially between early childhood and adulthood, and during phases of rest and sleep. It is considered as some form of mental hygiene by which the brain consolidates its map of the world. 'One of the essential functions of sleep is to take out the garbage', neuroscientist Gina Poe writes in a pertinent paper, 'erasing and "forgetting" information built up throughout the day that would clutter the synaptic network that defines us' (Poe 2017: 464).

Optimal brain damage was conceived to automate this process and shrink the network to a) make running it more efficient and b) to optimize its learning curve. In other words, catastrophic forgetting is prevented by purposeful forgetting. 'The main idea is that a "simple" network whose description needs a small number of bits is more likely to generalize correctly than a more complex network, because it presumably has extracted the essence of the data and removed the redundancy of it', Le Cun and colleagues explain (1989: 604). The bigger the model, the harder it is for this model to learn new things, and the easier it breaks. Optimal brain damage hence proposes a heuristic that selects the *smallest* among a number of more or less successful models (Gorodkin et al. 1993).

This points towards the idea that some parts of the network are more significant than others—while those considered essential for the functioning of the system as a whole must be protected, others would seem to be expendable or even detrimental to the network's operations. Crucial in this regard is the definition of redundancy or conversely, 'saliency.' LeCun, Denker and Solla targeted those parameters 'whose deletion will have the least effect on the training error' (LeCun, Denker & Solla 1989, 599). This principle still dominates evaluation in the domain of machine learning—and it ultimately suggests that the 'health' of the tool—not its faithfulness to the worldconstitutes the ultimate ratio of epistemology.

What sits at the heart of optimal brain damage are the principles of neuro-economy that Rosenblatt had insisted on in his paper on the perceptron. Drawing on Hebb and Hayek, he offered that one of the major flaws of the deterministic systems proposed by Marvin Minsky and others is that they had no concept of internal change and reorganization—which is why these machines must forever fail in reproducing any sort of biologically plausible processes of cognition (Rosenblatt 1958). LeCun, Denker and Solla took inspiration from this cornerstone of the neural imaginary to engage with the stability-plasticity dilemma—yet, to what effect? Optimal brain damage is an experimental and continuous process that arguably does not aim at the most comprehensive and diverse representation of the world by the network. Rather, it takes up Hayek's idea of the market as an epistemological principle and institutes a Darwinist process of neural selection, which turns shock and trauma into a technology that ultimately protects the network at the expense of connectivity and diversity.

While the technology of optimal brain damage might come across as somewhat obscure, it has influenced an abundance of cutting edge 'pruning' techniques and other mechanisms that relinquish the neuro-darwinist language of optimal brain damage. Such systems activate only specific parts of the network during learning to protect others (see e.g., Masse, Grant, and Freedman 2018). As a result, every node of the network can be involved in dozens of operations, but with a unique set of peers for each individual task. The latter is important to make sure that networks do not sprawl. 'Intuitively, you might think the more tasks you want a network to know, the bigger the network might have to be,' says David Freedman, professor of neurobiology at University of Chicago and co-author of a paper on mechanisms that prevent catastrophic forgetting. 'But the brain suggests there's probably some efficient way of packing in lots of knowledge into a fairly small network. When you look at parts of the brain involved in higher cognitive functions, you tend to find that the same areas, even the same cells, participate in many different functions' (Mitchum 2018).

Such ideas might sound promising from the standpoint of a neural network engineer, yet they hint at the problems that arise due to the neural imaginary: that it legitimates shock and bias with recourse to economic criteria and the cognitive principles that purportedly come natural to the human brain. To elaborate these concerns, we would like to return to the present and ongoing discussions that revolve around the unfathomable size of current neural networks.

Conclusion

In a paper accepted for the 2021 FAccT Conference, Google's former AI ethicist Timnit Gebru and her colleagues Emily Bender, Angela McMillan-Major, and Margaret Mitchell argue that language models such as the hyped GPT-3 can get 'too big' to be sustainable (Bender et al. 2021). The model was developed by OpenAI (a group founded by figures like Elon Musk) and has an incredible 175 billion parameters.. As Bender and colleagues' as well as the original authors of the model (Brown 2020) offer, the sheer size of the model presents environmental issues, for its training and operation devours unimaginable amounts of energy. But of greater concern to our argument is the stochastic and potentially uncertain nature of such systems.

The authors of the GPT-3 model already announced upon releasing the model that its size made it un-amenable to modelling, and that discrimination and bias was always a possibility. Bender and colleagues now label these models as 'statistical parrots' that throw racist slurs and hate speech back at us. Initial tests done with the model on sample data sets utilizing basic categories such as sex, race, and religion found regular biases. For instance, words like 'Muslim' had higher probabilities of being paired with terms like 'terrorist' and similar negative correlations appearing with 'Jew', 'black' and so forth. The authors also noted concerns over the inability to know what future biases might occur (since there are billions of parameters, the data sets always grow, and language evolves), or what part of the model creates the bias; all problems for legal accountability or verification of discrimination--at least under current US law (Tom B. Brown 2020).

Furthermore, concern has already been openly voiced about how such a system whose impacts are unpredictable due to size will interact with older infrastructures, such as those of advertising, and 'filter' bubbles that channel users to higher paying sites. How would we manage the problem of society and social biases that are historical if we cannot even represent and visualize the data sets upon which these systems are being trained (Tollefson 2021)? GPT-3 illustrates a new dilemma; between the need to have ever larger models to account for a complex world and the importance of representation, which legitimates abstraction and even 'pruning', to produce coherence and responsibility for humans.[8]

Optimal brain damage and its underlying, market-oriented epistemology would seem to provide a way out of the current dilemma that model size proliferates. This is to say that optimal brain damage and other pruning or compression techniques exhibit how shock and trauma can serve a stability function in our present, in fact, even a mechanism to preserve meaning and value in markets as well as in language processing. Reading about the progress made in machine learning, one might even be tempted to believe that neural networks could become a tool of more just and democratic representation and learning. Our concern, however, is that pruning, optimal brain damage, and other techniques originating in the neural imaginary ultimately protect the tool instead of the population. They represent inward-facing techniques, conceived primarily to improve the (mental) health of the network. In optimizing damage for

efficiency there is the question of what is being so carefully excluded or forgotten?

A team of researchers from Google and the Montréal-based MILA Québec AI Institute performed tests on heavily pruned and compressed networks which show top-line performance by commonly used benchmarks. In a preprint, they remark that the surprising accuracy of these networks comes at a heavy cost. 'Compression disproportionately impacts model performance on the underrepresented long-tail of the data distribution,' the authors write (Hooker et al. 2020). That is, it may amplify existing algorithmic bias for sensitive tasks such as face recognition (Buolamwini & Gebru 2018) or health care diagnostics (Esteva et al. 2017), hence pruning is likely at odds with fairness objectives.

Where artificial neural networks become arbiters of social realities, the techniques and mechanisms employed to make network and model more efficient and resilient to shock add another possible source of distortion that may catastrophically affect the social relations they are supposed to represent. Pruning amounts to at once over-emphasizing certain knowledges and purposeful forgetting that may ultimately come to the detriment of the diversity of knowledges in our world as represented through the network. After all, the neural imaginary draws on a system that learns extraordinarily well but is also known to aggravate cognitive bias to help us survive.

Our concerns with the neural imaginary hence go beyond its role in neural networks. The author Naomi Klein has labelled our current situation a 'pandemic shock doctrine', referring to the pandemic as an accelerator of cloud-driven labor at home –a 'screen new deal' to replace a 'green new deal', she contends (Klein 2020). We offer that this is a new moment in the history of shock. This digital shock doctrine consists in the naturalization of shock and trauma as *protective mechanisms* that the systems purportedly rely on, and that optimal brain damage as a discourse exemplifies.

For us, these experiments in the design of neural networks are not isolated sidenotes in the history of science. Rather, we argue these are the symptoms of a neuro-imaginary discourse that structures much of the contemporary relationship between economy, epistemology, and artificial intelligence. Most pressingly, the oft-lamented opacity of algorithms is a

discourseemployed to fuel a fantasy of cartesian perspective long absent from any model of the brain within machine learning or finance. Like the politics of 'Making America Great', our own ethics discussions are haunted by a fantasy to return to a past when our networks could be controlled from central command centers, and borders could be selectively closed at will. This is an imaginary that ignores or seeks to dispel the stochastic nature of our machine learning networks.

On one hand we seek localized optimization, the culling or elimination of 'connections' to minimize risks and make networks predictable and representable, and on the other hand we indulge in the fantasy of an ever evolving and seemingly inevitable expansion of computing into life. The noisy unpredictability of vast networks that threaten value and meaning (think Gamestop) vacillate with the fantasy of managed destruction and localization that breeds reactionary politics and isolationism (think QAnon).

Markets, and now reactionary politics, seek volatility without diversity. This dialectic feeds ourpolitics and decisions. Shock has been normalized to be managed through our electronic networks. Politicians, reactionary movements, and financial markets forced to contend with volatility and uncertainty in the current pandemic, make the move to create only islands of permitted volatility, which foster homogenous violence or dispossession. Surrounded by our nervous nets, we face inward, unable to recognize that there is a world outside. The imperative is to maintain the network's internal health at the cost of the world. So the stock market, and similar systems are maintained as healthy, while people of color and the poor die from COVID. If history is the only remedy for situating data, then the automated amnesia of our present can only result in damages that we should never label "optimal" or acceptable.

Acknowledgements

The authors thank Robert Mitchell for his contribution to many of the ideas involving Hayek and Rosenblatt. We also thank the anonymous reviewers who so generously contributed constructive criticism to this piece. Our research was supported by the Swiss National Science Foundation, through the Sinergia Grant *Governing Through Design* (CRSII5-189933).

Notes

1. See http://www.independentsage.org.

2. Robotics pioneer Rodney Brooks contemptuously referred to this as "the new cerebral blind alley" (Brooks et al. 2012).

3. See e.g., https://www.wired.com/story/karl-friston-free-energy-principle-artificial-intelligence/.

4. For histories of reason and rationality, as well as the economic decision maker see: Mirowski (2002), Paul Erickson (2013), Foley (2002).

5. The discussion of Hayek and Rosenblatt is indebted to Robert Mitchell with which Orit Halpern is writing a forthcoming book "The Smartness Mandate".

6. When studying perceptrons, the "object of analysis is an experimental system which includes the perceptron, a defined environment, and a training procedure or agency" (Rosenblatt 1958).

7. Weights are parameters that represent the strength of a connection between two units of a network.

8. GPT-3 and its predecessor GPT-2 (and similar models used by Google and Microsoft) have thus been blamed with the exacerbation, if not creation of, both reactionary politics and extreme discrimination against groups (Tom B. Brown 2020).

References

Arnold, T. and Tilton, L. (2021) "Depth in Deep Learning. Knowledgeable, Layered, and Impenetrable", in *Deep Mediations. Thinking Space in Cinema and Digital Cultures*, (eds.) K. Redrobe and J. Scheible. Preprint: https://statsmaths.github.io/pdf/2020-deep-mediations.pdf.

Bender, E. et al. (2021. "On the Dangers of Statistical Parrots: Can Language Models Be Too Big?", *Conference in Fairness, Accountability, and Transparency (FAccT '21)*. (March 3rd-10th): Doi: 10.1145/3442188.3445922.

Black, F. (1986) “Noise”, *The Journal of Finance* 41. No. 3: 529-543.

Brooks, R. et al. (2012) “Turing Centenary: Is the brain a good model for machine intelligence”, *Nature* 482: 462-3.

Brown, T. et al. (2020) "Language Models are Few-Shot Learners", *Computer Science* 1 (May 28): https://arxiv.org/abs/2005.14165.

Bruder, J. (2018) “Where the sun never shines. Emerging paradigms of post-enlightened cognition”, *Digital Culture & Society* 4. No. 1: 133-153.

Buolamwini, J. and Gebru, T. (2018) ”Gender shades: Intersectional accuracy disparities in commercial gender classification”, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. (Eds.) Friedler, S. and Wilson, C.: 77–91.

Crist, M. (2015) ”Postcards from the Edge of Consciousness”, *Nautilus* 27 (August 6).

Croft, F. (1954) "Look What Utter Boredom Can Do", *Macleans* (May 15):18-19.

Edelman, G. (1993) “Neural Darwinism: selection and reentrant signaling in higher brain function”, *Neuron* 10. No. 2:115-25. doi: 10.1016/0896-6273(93)90304-a.

Esteva, A. et al. (2017) “Dermatologist-level classification of skin cancer with deep neural networks”, *Nature* 542. doi: 10.1038/nature21056.

Edwards, P. (1997) *The closed world computers and the politics of discourse in cold war america*. Cambridge, Massachussetts: MIT Press.

Foley, D. (2002) "The Strange History of the Economic Agent", *Unpublished presentation to the General Seminar at New School for Social Research* (December 6th).

French, R. (1997) “Pseudo-recurrent connectionist networks: An approach to the ’sensitivity-stability’ dilemma”, *Connection Science* 9. No. 4: 353-379.

Friston K. (2012) "The history of the future of the Bayesian brain", *NeuroImage* 62. No. 2: 1230–1233. https://doi.org/10.1016/j.neuroimage.2011.10.004.

Friston K. et al. (2020a) "Dynamic causal modelling of COVID-19", *Wellcome Open Research* 5: 89. https://doi.org/10.12688/wellcomeopenres.15881.2.

Friston K. et al. (2020b) "Effective immunity and second waves: a dynamic causal modelling study", *Wellcome Open Research* 5: 204. https://doi.org/10.12688/wellcomeopenres.16253.2

Fuller, B. (1975) "Synergetics: Explorations in the Geometry of Thinking", *Macmillian Co. Inc.* (Accessed January 13th): http://www.rwgrayprojects.com/synergetics/synergetics.html.

Galison, P. (1994) "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision", *Critical Inquiry* 21: 228-266.

Gorodkin, J. et al. (1993) "A Quantitative Study of Pruning by Optimal Brain Damage", *International Journal of Neural Systems* 4. No. 2: 159-169. https://doi.org/10.1142/S0129065793000146.

Halpern, O. (2005) "Dreams for Our Perceptual Present: Temporality, Storage, and Interactivity in Cybernetics", *Configurations* 13. No. 2: 36.

Halpern, O. (2014) "Cybernetic Rationality", *Distinktion: Journal of Social Theory* 15: 223-238.

Halpern, O. (2015) *Beautiful Data: A History of Vision and Reason Since 1945.* Durham: Duke University Press.

Hayek, F. (1945) "The Use of Knowledge in Society", *The American Economic Review* XXXV (September): 519-530.

Hayek, F. (1952) *The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology.* Chicago: University of Chicago Press.

Hayles, N. (1999) *How we became posthuman virtual bodies in cybernetics, literature, and informatics.* Chicago, Ill.: University of Chicago Press.

Hebb, D. (1937) "The innate organization of visual activity I. Perception of figures by rats reared in total darkness", *Journal Genetic Psychology* 51: 101-126.

Hebb, D. (1938) "Studies of the organization of behavior: Changes in the field orientation of the rat after cortical destruction", *Journal of Comparative Psychology* 26: 427-444.

Hebb, D. (1942) "The effect of early and late brain injury on upon test scores, and the nature of normal adult intelligence", *Proceedings of the American Philosophical Society* 85: 275-292.

Hebb, D. (1949) *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley.

Hebb, D. and Penfield, W. (1940) "Human behavior after extensive bilateral removal from the frontal lobes", *Archives of Neurology and Psychiatry* 44: 421-438.

Heron, W. (1957) "The Pathology of Boredom", *Scientific American* 196. No. 1: 52-57.

Hooker, S. et al. (2020) "What Do Compressed Deep Neural Networks Forget?" *arXiv* (13 July): https://arxiv.org/abs/1911.05248.

Kay, L. (2001) "From Logical Neurons to Poetic Embodiments of Mind: Warren McCulloch's Project in Neuroscience", *Science in Context* 14. No. 4: 591-614.

Klein, N. (2007) *The Shock Doctrine: The Rise of Disaster Capitalism.* New York: Picador.

Klein, N. (2020) "Screen New Deal", *The Intercept* (May 8th): https://theintercept.com/2020/05/08/andrew-cuomo-eric-schmidt-coronavirus-tech-shock-doctrine/.

LeCun, Y. et al. (1989) "Optimal Brain Damage", *Advances in neural information processing systems 2*: 598-605.

Lilly, J. (1956) "Mental effects of reduction of ordinary levels of physical simuli on intact healthy persons", *Psychiatric Research Reports of the American Psychiatric Association* 5: 10-28.

Lilly, J. and Gold, E. (1996) *Tanks for the Memories: Flotation Tank Talks*: Gateway Books and Tapes.

Mackenzie, D. (2014*) A Sociology of Algorithms: High-Frequency Trading and the Shaping of Markets.* Edinburgh: University of Edinburgh Press.

Masse, N. et al. (2018) "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization", *Proceedings of the National Academy of Sciences* 115. No. 44: E10467-E10475. doi: 10.1073/pnas.1803839115.

McCloskey, M. and Cohen, N. (1989) "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem", *Psychology of Learning and Motivation* 24: 109–165. doi:10.1016/S0079-7421(08)60536-8.

McCoy, A. (2006) *A question of torture: CIA interrogation, from the Cold War to the War on Terror*. New York: Metropolitan/Owl Book/Henry Holt and Co.

Mirowski, P. (2002) *Machine dreams: economics becomes a cyborg science.* New York: Cambridge University Press.

Mirowski, P. (2006) "Twelve Theses concerning the History of Postwar Neoclassical Price Theory", *History of Political Economy* 38: 344-379.

Mitchum, R. (2018) "Brain-inspired algorithm helps AI systems multitask and remember", *UChicago News* (October 16th): https://news.uchicago.edu/story/brain-inspired-algorithm-helps-ai-systems-multitask-and-remember.

Erickson, P. et al. (2013) *How Reason Almost Lost Its Mind:The Strange Career of Cold War Rationality*. University of Chicago Press.

Pearl, J. (1989) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann Publishers.

Poe, G. (2017) “Sleep is for Forgetting”, *Journal for Neuroscience* 37. No. 3: 464-473.

Rosenblatt, F. (1958) “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”, *Psychological Review* 65 (November): 386-408.

Rosenblatt, F. (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington D.C.: Spartan Books.

Simon, H. (1955) “A Behavioral Model of Rational Choice”, *The Quarterly Journal of Economics* 69. No. 1: 99-118.

Simon, H. (1992) “What is an “explanation” of Behavior?" *Pscychological Science* 3. No. 3: 150-161.

Summers, A. and Lawrence, H. (1990) “The Noise Trader Approach to Finance”, *The Journal of Economic Perspectives* 4. No. 2: 19-33.

Tollefson, J. (2021) “Tracking QAnon: how Trump turned conspiracy-theory research upside down”, *Nature* (February 4th).

Wiener, N. (1961) *Cybernetics; or Control and communication in the animal and the machine.* New York: MIT Press.

Wing, A. (2020) “Karl Friston takes on the pandemic with the brain’s arsenal”, *Dropbox Blog: Work in Progress* (May 10th): https://blog.dropbox.com/topics/work-culture/karl-friston-takes-on-the-pandemic-with-the-brain-s-arsenal.