The Simulation of Intelligence and Creativity: On the Foundations of Machine Learning

Christian Georg Martin

Introduction

Ever since its inception in the 1950s the research program of AI has been marked by a profound ambiguity which is still with us today. The proposal for the 1956 "Dartmouth Summer Research Project on Artificial Intelligence," to which that program owes its name, was based on the "conjecture that every aspect of learning or any other feature of intelligence can be in principle so precisely described that a machine can be made to *simulate* it."[1] In 1957, Frank Rosenblatt, the forefather of the deep learning approach to AI which has regained prominence in recent decades and is currently the field's leading AI paradigm,[2] characterized the perceptron, the first artificial pattern recognition system imitating the human brain, as "*a model* of a system which is primarily concerned with the recognition of the forms, sounds, and other stimuli which make up the ordinary physical world, as we know it through our senses."[3] While on the one hand conceiving of AI as a *simulation* or *model* of human intelligence, the forefathers of AI, on the other hand, viewed the creation of "*fully intelligent* machines" as imminent.[4] The same ambiguity can

1 John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955 [1955]," *AI Magazine* 27, no. 4 (2006): 12, https://doi.org/10.1609/aimag.v27i4.1904.

2 See Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* (Random House, 2020), 7–9.

3 Frank Rosenblatt, *Two Theorems of Statistical Separability in the Perceptron* (Cornell Aeronautical Laboratory, 1956), 2.

4 Rosenblatt, *Two Theorems*, 5.

be observed at present, for instance, insofar as large language *models* such as ChatGPT are often credited with thought, meaningful speech, and creativity.

It only makes sense to speak of a "model" or a "simulation" if there is a difference in kind between the model or simulation and what it models or simulates: a model or simulation is not "the real thing." Accordingly, if a machine could indeed be granted thought, understanding, or creativity, it would not just be a model or simulation thereof. In many cases, it is easy to tell a model or simulation and its object apart. There is no temptation, for instance, to confuse a climate model run by a computer with climate itself, i.e., the actual weather conditions on earth over a period of time. In other cases, however, a simulation might resemble its object in ways that give rise to confusion. Such confusion can also be deliberately created so as to illicitly substitute an established practice with one that merely simulates it. The simulation of democratic procedures in a nascent authoritarian state, for instance, is designed to conceal the fact that the state in question isn't a democracy any more. If a human practice is being replaced by a simulation of it, what ultimately results is deskilling: an impoverishment of the capacity to engage in the original activity.

As I shall argue, the deep neural networks underlying contemporary AI can only provide us with more or less impressive simulations of intelligent activity. That something is a mere simulation does not mean it couldn't be useful. However, if the output of AI amounts to a simulation of intelligent and creative activity, this raises the question of what kinds of *subservient* use *we* can or should put it to within our own intelligent and creative activities. The question of how to distinguish between potential use and abuse of AI within human practice is vastly complex and deeply variegated, depending on the particular activity in question and its place within our forms of life. To dispel some of the fog that currently surrounds the deep learning approach to AI, it seems helpful to compare an elementary and pervasive example of intelligent activity of ours with its machine learning counterpart so as to precisely explain why the latter amounts to a mere simulation of the activity in question. Accordingly, rather than aiming to compare human and artificial intelligence, broadly speaking, the present contribution contents itself with confronting them with regard to an elementary example, the use of ordinary concepts such as *red* or *inside*.

When reflecting on what humans or machines *can do* we compare *capacities* rather than particular *occurrences*. Many of *our* capacities are self-constituting, i.e., we learn them by doing and deepen them by way of ongoing exercise. Deep neural networks, in turn, acquire their capacities in a process of training. A capacity is a potential to engage in a certain kind of activity that can be exer-

cised on an indefinite number of occasions. The capacity is defined by what it is the capacity for, i.e., by examples of its successful exercise. All kinds of things might in fact go wrong when a capacity is exercised, and its exercise will then be flawed. Nonetheless, what the capacity is for can only be grasped by recourse to what is achieved if things go well. It is therefore misguided to compare human and artificial intelligence, as is usual in the machine learning community, by comparing *average* scores on a certain kind of task. Instead, one needs to ask with an eye on the particularly felicitous exercise of a human capacity whether a machine could in principle do *that*.

The following comparison between intelligent activity on our part and the output of deep neural networks is conceptual rather than observational. We dispose of a certain understanding of our own intelligent activities by virtue of engaging in them rather than based on observing ourselves doing certain things. It cannot happen, for instance, that I'm baking a cake or getting married without me knowing that I am. Such knowledge is not based on observation of an independent object, but is internal to the activity known. It is philosophy which clarifies and deepens the self-knowledge that is inherent to our intelligent activities such as thinking, speaking, or artistic creation. Such clarification is required since the inchoate self-knowledge inherent to our intelligent activities tends to be confused.

It might require observation to find out whether a machine can in fact do what we designed it to do. It does not require observation to find out whether a machine that has only apparently been designed to engage in full-blown intelligent activity might in fact exhibit such activity. For intelligent activity does not just happen to occur. If a machine had been designed to randomly print letters on sheets of paper, we could know *by way of reflection* that this machine does not produce meaningful texts. We would not have to compare its actual output with meaningful texts. Analogously, we might recognize by way of reflection that deep neural networks do not use concepts, but rather simulate their use. Thinking otherwise would then amount to a confusion, which this essay seeks to highlight. It is structured into three parts. Part one sheds light on human thought by clarifying what concepts are and what using them amounts to. In part two it is argued that deep neural networks can only simulate conceptual activity. The third and final part exhibits the challenge that AI poses to us, namely to distinguish between use and abuse of machine generated simulations of intelligence within our human form(s) of life.

Human Thought: The Use of Concepts as Involving Reason and Creativity

Traditionally, the idea that we are intelligent beings has been spelled out by conceiving of ourselves as rational animals or finite thinkers.[5] An animal is a creature whose cognitive access to its environment depends on that environment appearing to it by way of the senses. As animals, we are finite insofar as we do not know everything all at once. That we are *rational* animals means that we aren't lost in ever-changing environments but have the ability to expand our acquired understanding to unforeseen situations that we thereby integrate into a unified horizon in which *anything* that might occur to us can be placed. As such, we are creatures who live in a *world*. To integrate unforeseen circumstances into our world-view we can neither treat them as entirely novel and incomparable, nor can we simply assimilate them to situations that are already familiar. We steer through the unknown by way of using concepts. Concepts are representations which allow us to recognize a unity between otherwise different situations. The rose and the sunset, for all their difference, can both instantiate the concept *red*. Using concepts is thinking. Concepts allow us to anticipate future situations such as the next red sunset. However, they do not make us infinite thinkers: grasp of a concept does not allow us to partition "all possible" situations into those that fall under the concept at hand and those that don't. Our grasp of a concept is always partial in that it doesn't rule out unforeseeable situations that defy the concept as we understand it: situations in which we no longer know how to apply the concept as it was hitherto understood and in which we thus run the risk of losing it, not knowing what to say. It is part and parcel of the capacity to use concepts to come to terms with such situations by *expanding* a concept in a way that allows it nonetheless to be applied to a situation which defies its usage as hitherto understood. The successful *modification* of a concept in light of an unforeseen situation that at first defies its application is an act that is both creative and rational. It is *creative* in that it involves doing something novel and original that does not simply result from the application of conceptual resources already available. For it is precisely these resources which have proven insufficient when faced with the given situation. The novelty is *rational* insofar as the successful modification

5 Matthew Boyle, "Essentially Rational Animals," in *Rethinking Epistemology: Volume 2*, ed. Günter Abel and James Conant (de Gruyter, 2012), 395–428.

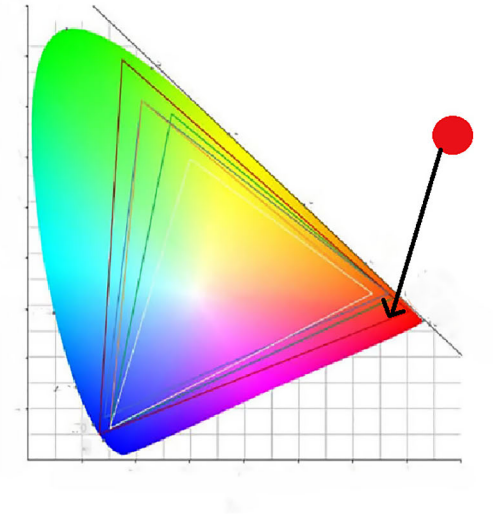
of a concept in light of that situation isn't arbitrary but can be justified *in retrospect*, insofar as it allows us to overcome the conceptual predicament and to thus find a way out of the dead end which our previous understanding of the concept had led us into.

Concepts thus have a more complex texture than one might imagine: they involve an inner articulation insofar as they record critical junctures of their application which motivated their expansion. Not all of those who grasp a certain concept have an equally refined understanding of the junctures it incorporates. However, even those whose understanding of a concept is relatively limited have the general ability to move from one stage to another, i.e., to expand their understanding in a way that is both creative and rational. What the capacity for conceptual expansion amounts to can best be seen by looking at examples. As our example we can take the run-of-the-mill concept *red*. If we looked at *water* or *number* instead, we would arrive at essentially the same results. Reflecting on how we apply the concept *red* and how we have learned to expand it in the face of situations that at first seemed to defy its application will reveal that we tend to imagine concepts and conceptual activity in ways too simple to do justice to the intricacies of actual usage.

At first sight, it might seem that there must be some sort of shared ingredient that corresponds to the concept *red* on the part of the things that fall under it. However, things can be red in different ways, to which different shades of red correspond. No shared ingredient, then! Accordingly, it might seem more appropriate to view the concept *red* as delineating a certain *region* within a "space" whose dimensions are given by three axes of possible variation: hue, saturation, and brightness. Any point falling within a certain somewhat blurry region of this color space would count as red. Applying the concept *red* to a thing encountered in real life would accordingly amount to placing that thing, or a monochrome part of its surface, inside or outside the respective region in the same immediate and effortless way as we can imaginatively insert a red circle in its proper region within a color space.

Following Wittgenstein, we can call an imaginary scene that is supposed to illustrate our use of a concept, a picture of that concept. The philosophical picture of color concepts as delineating a certain region of a color space and of the application of such concepts as an immediate placing of a sample inside or outside such a region cannot do justice to the intricacy of our color concepts and their actual application. This can be seen by paying close attention to the application of such concepts in real life situations.

Fig. 8: Inserting a red circle in its proper region within a color space



First, the picture of placing a thing inside or outside a region of a color space is static insofar as it does not take into account the temporal extension of things. One way of running into trouble when applying the concept *red* is occasioned by a thing changing its chromatic appearance when moved to a different place, due to diverging lighting conditions. Let us assume that our original concept of *red* had been formed ostensively by recourse to samples of red in plain daylight. It is part of our rudimentary color concepts that things do not simply change their colors upon being moved. Accordingly, when faced with a thing that reliably changes its chromatic appearance from red to brown while being moved to-and-fro between two places, we have run into trouble, risking losing our concepts of *red* and *brown* in the face of a situation that makes them inapplicable as they are. The trouble we have run into has the form of a dilemma: Neither saying of a thing which, in plain daylight, we took to be red and which now, indoors, appears to be brown, that it is *just* red or *just* brown, nor saying that it is *both* red *and* brown will do, for both involve a contradiction, either with what we see or within what we say. On the one hand, the thing viewed inside and viewed outdoors looks too different to be attributed one and the same color, while stating that the object simply changes its color from red

to brown contradicts the principle that things do not just change their colors when moved. The predicament is resolved by expanding our concepts of *red* and further colors so as to allow for a distinction between standard lighting conditions in which things exhibit the color they have and deviant lighting conditions in which their color looks different from what it is. The *expanded* color concept thus comes with an inbuilt distinction between *is red* and *looks red*, which can be applied in the kind of situation that beforehand defied its application. It follows from this that the expanded concept of *red* cannot appropriately be visualized as a continuous region in color space. For it essentially involves discontinuous junctures: it is part and parcel of the expanded concept of *red* that its instances are subject to abrupt shifts of chromatic appearance depending on lighting conditions.

Second, even given constant lighting conditions, e.g., plain daylight, the subsumption of a thing under a color concept does not consist in *immediately* placing it, *without further ado*, inside or outside a certain continuous region of the color space as the philosophical picture that holds us captive makes it appear. The principle that things do not abruptly change their color can not only be challenged by a change of lighting conditions, but also by unexpected behavior of things under given lighting conditions. A glistening object such as a bronze pot might *momentarily* have the same appearance as a yellow object, while its manipulation, e.g., rotation, will bring out that its color isn't in fact yellow, but golden. Accordingly, we cannot properly attribute a color to an object viewed *in an instant*, but subsuming a thing under a certain color concept amounts to placing it in the same class with other objects that change chromatic appearance in a similar way when subject to certain kinds of manipulation.[6] Techniques of manipulation which we are used to applying, unreflectively, and which serve to ascertain the real color of a thing are part and parcel of our ordinary application of color concepts and can be viewed as the result of an expansion of rudimentary color concepts in light of the trouble we run into when taking the momentary chromatic appearance of things at face value.

What has been exemplified by color concepts applies to any concept whatsoever. In our ongoing use of any concept we can meet with situations which defy its application and occasion a sort of modification of the concept which is both creative and rational, i.e., retroactively justifiable insofar as it allows us to apply the expanded concept to the situations which rendered its unexpanded

6 Mark Wilson, *Wandering Significance: An Essay on Conceptual Behavior* (Clarendon Press, 2006), 104–06 and 454–67.

version inapplicable. Each mature concept thus involves a series of inbuilt logical junctures which result from the resolution of a certain kind of dilemma to which its previous application gave rise.

The Simulation of Conceptual Activity by Deep Neural Networks

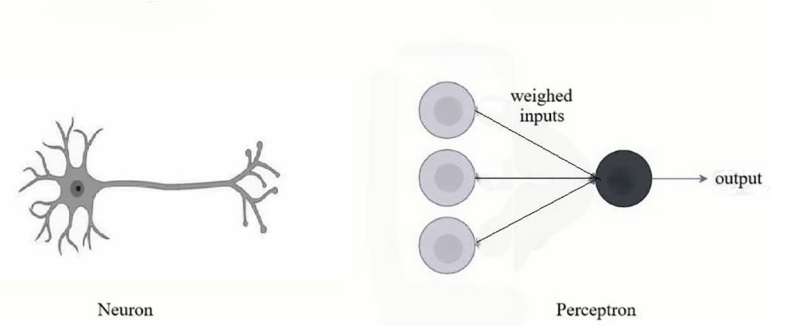
Having shed light on what concepts are and what conceptual activity amounts to in our own—human—case, we now turn to the attempt to build machines that can be trained to exhibit conceptual behavior. We will focus on deep learning, the now-dominant branch of AI research. In contrast to the symbolic approach to AI that had been prevalent for decades, the deep learning approach is subsymbolic: It does not conceive of intelligence first and foremost on the model of rule-governed manipulation of symbols, but on the model of neural activity in the brain. Accordingly, it does not seek to make machines exhibit conceptual behavior by feeding them with detailed instructions about how to manipulate symbols in the face of certain external inputs, but seeks to construct a mechanism that allows them to learn concepts on their own in the course of responding to input in a way that is based on trial and error. This approach might seem promising insofar as it is analogous to how we humans acquire our first concepts, given that infants cannot acquire concepts by following explicit rules or instructions given to them, for in order to understand such rules or instructions they would already have to grasp the concepts involved in their formulation. It should be noted, though, that both the symbolic and the subsymbolic approach to AI involve algorithms, i.e., recipes for step-by-step procedures that yield well-determined results. For in order to learn by trial and error in the course of humanly-supervised training, it needs to be uniquely determined how the machine is supposed to change its own parameters if its response to a certain input doesn't comply with the response human trainers have labelled correct.

It is characteristic of the deep learning approach to view conceptual activity in terms of input–output behavior that is evaluated statistically. The question of what a concept even is and whether it exhibits a certain kind of inner articulation is largely absent. It is assumed from the outset that a device can be granted mastery of a certain concept if its outputs partition inputs that do or do not instantiate the given concept into two classes in a way that is statistically reliable. In that case the machine is said to be able to recognize a pattern. It is thus fair to say that the deep learning approach assumes without further ado that a

concept can be represented by a set of isolated instances, the so-called training set, and that grasp of a concept consists in a reliable responsive disposition that allows to sort sample items into two classes—the class of those which instantiate the concept and the class of those which don't. However, as we have seen, conceptual capacities do not simply consist in the ability to uniquely partition a set of samples into two disjoint subsets, but essentially involve the capacity to *creatively and rationally* extend a concept in light of unforeseen situations which defy placing a sample in one class or the other. One might therefore venture that the machine learning approach to conceptual activity is a non-starter: It can at best result in a model or simulation of conceptual activity, because it bypasses what concepts are right from the start.

We need to take a closer look, though, to substantiate this conjecture. As indicated, the paradigm on which the deep learning approach models intelligent behavior is the brain and its characteristic cells: neurons. A neuron allows electro-chemical signals sent out by other neurons to be received and processed. These signals can have a different importance or weight. The neuron works by summing up its weighed inputs, and if the resulting value exceeds a certain threshold, the neuron "fires," i.e., it has a non-zero output. Otherwise, it does not fire, i.e., its output is zero. The guiding idea of automated pattern recognition as viewed from the perspective of deep learning is to construct networks of artificial neurons which learn to specifically and reliably respond to inputs caused by instances of a certain concept with the output 1, while yielding 0 in all other cases.

Fig. 9: The neuron and the perceptron



Neural networks are supposed to learn to recognize certain patterns in their environment by trial and error, i.e., by a process of adaptation where inappropriate responses to environmental stimuli result in a certain change of weights within the network, while appropriate responses leave it as it is. Even while the initial responses of the network and the changes of weights occasioned by inappropriate responses might be random, the whole setup is such that the network gradually organizes itself so as to make appropriate responses more likely. Frank Rosenblatt's original idea about how to achieve this was to construct an artificial neuron, a so-called perceptron.

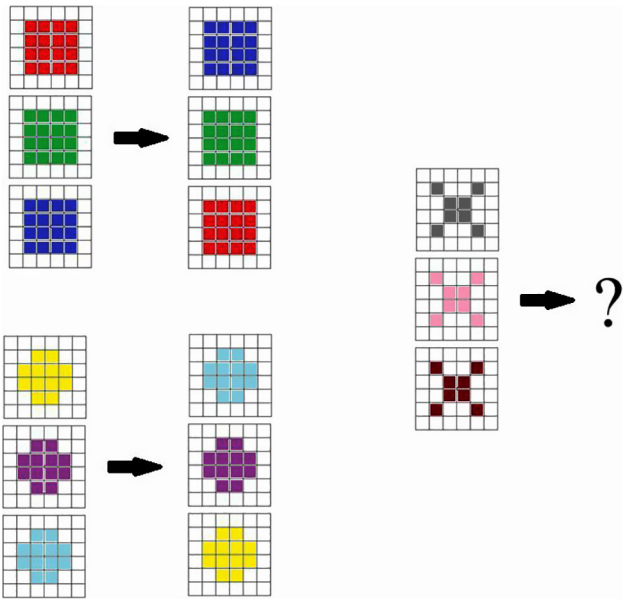
The perceptron gradually acquires the appropriate responsive dispositions by way of trial and error, guided by mathematical approximation techniques implemented in a digital computer endowed with a sensor for environmental stimuli. The sensor might be a camera yielding images that comprise a number of pixels to which a number of inputs or entryways on the part of the perceptron corresponds. Initially, the weights of these inputs and, hence, the output of the perceptron are random. However, this is supposed to be changed through training. The so-called training set might consist of images that have been labelled by humans according to whether the image instantiates a certain concept or not. Whether a neural network can actually be trained to learn to recognize a certain pattern depends not only on its architecture but on the algorithm or recipe that determines how its weights are supposed to be changed if its response to a test sample deviates from the expected result. By repeating the training process time and again the weights are supposed to be gradually changed until the network reliably responds to samples with the appropriate response. While still following what is essentially the same kind of procedure, modern day neural networks have a much more complex architecture than Rosenblatt's perceptron, consisting of multiple layers of artificial neurons. For this reason, they are called *deep* neural networks.[7]

By now, such networks are astonishingly good at certain pattern-recognition tasks. Does this mean that they can be granted conceptual capacities? In order to answer this question, AI researchers have designed benchmarks that are supposed to test the ability to form concepts. A benchmark that is thought

7 On the invention of the perceptron and its relation to deep neural networks, see chapter 9 of Matteo Pasquinelli, *The Eye of the Master: A Social History of Artificial Intelligence* (Verso, 2023). Pasquinelli's book provides a critical history of the AI paradigm as driven by the capitalist attempt to automate labor from the vantage point of historical epistemology.

to be particularly precise and challenging is the so-called "abstraction and reasoning corpus" (ARC) designed by François Chollet.[8] The benchmark consists of tasks that are supposed to test the capacity to form a concept by learning from examples. These examples consist of simple shapes within a grid that illustrate elementary concepts such as *inside*, *square*, or *even*. Abstraction, i.e., the acquisition of a concept, is supposed to be tested by the task of completing further grids involving similar shapes.[9]

Fig. 10: Example of an ARC task. The challenge is to demonstrate grasp of the abstract rule governing the demonstration transformations by completing the test input.



8 See François Chollet, "On the Measure of Intelligence" (2019), https://arxiv.org/abs/1911.01547.

9 See Arseny Moskovichev, Odouard Victor, and Melanie Mitchell, "The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain" (2023), https://arxiv.org/abs/2305.07141.

In December 2024 it was found that certain AI tools score higher at ARC tasks than humans.[10] Chollet and others have suggested that this shows such tools can be credited with the ability to abstract, i.e., to form concepts.[11] Other members of the AI community have been more critical, pointing out that immense human effort had gone into tailoring AI tools that fit the purpose which makes it somewhat difficult to say to what extent the machines should be credited with success and to what extent the credit goes to their human designers.[12]

From a philosophical vantage point, the assumption that deep neural networks can acquire conceptual capacities can be criticized in a more fundamental manner. The misguided assumption underlying attempts at teaching deep neural networks to form concepts is to assume that a concept could somehow be contained in a set of its instances. Teaching a machine to form a concept would accordingly mean making it recognize what is contained in such a training set. The sets comprise a number of isolated items, and training consists in making the network respond to these items in isolation, one at a time, without explicit recourse to the others. Concepts, as we have seen, do not have their place in things, but in the eye of the beholder: a concept is the content of a capacity to recognize a characteristic similarity or continuity between an indefinite number of things. Moreover, a concept is rationally extendable in light of circumstances which at first sight defy its application. Teaching a machine to acquire concepts therefore requires teaching it something *intangible* that *we can do* in the face of sets of items rather than teaching it to recognize something that is supposedly *contained* in such sets.

If training sets don't contain concepts, trying to make a machine recognize what is contained in a training set cannot in principle result in it acquiring a concept. How to then interpret the fact that machines can indeed successfully be trained to respond to samples by sorting them into those that instantiate a certain concept and those that don't, sometimes even more successfully than humans? The first thing to stress in response to this question is that *we* select the training samples in a clear-cut way so as to avoid borderline cases that defy

10 See François Chollet, "OpenAI Breakthrough High Score on ARC-AGI-PUB," *Arc Prize* (blog), December 20, 2024, https://arcprize.org/blog/oai-o3-pub-breakthrough.

11 See Chollet, "OpenAI Breakthrough," and François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers, "ARC Prize 2024: Technical Report" (2025), https://arxiv.org/abs/2412.04604.

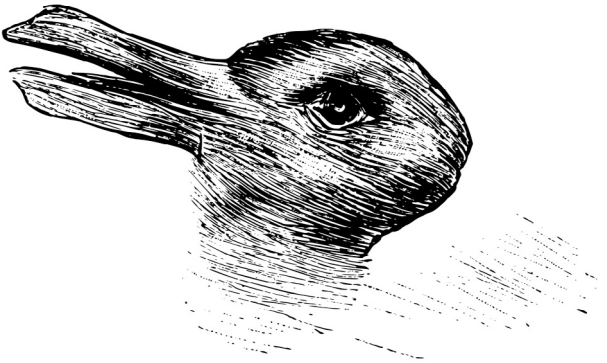
12 See Melanie Mitchell, "Did OpenAI Just Solve Abstract Reasoning?," *AI: A Guide for Thinking Humans* (blog), December 23, 2024, https://aiguide.substack.com/p/did-openai-just-solve-abstract-reasoning.

being put in either of two boxes. When trying to teach a machine the concept *inside*, for instance, it is exposed to figures either fully enclosed by or fully outside of a bounded area, rather than situated at the margin of a half-open form. This seems justified since acquisition of the rudimentary concept *inside* that is applicable to a continuous range of unproblematic cases has to precede its *extension* in light of borderline cases. However, as we shall now see, a neural network that reliably responds to a continuous range of unproblematic cases can neither be credited with mastery of a rudimentary concept nor is it in principle capable of rationally extending a rudimentary concept when faced with deviant situations.

The reason why a neural network that has successfully been trained to reliably respond to a range of unproblematic instances of a concept which are continuous with one another cannot even be credited with mastery of a rudimentary version of that concept is that it isn't able to recognize similarity or resemblance *as such*. Similarity and resemblance are *relations between* samples. A neural network, on the other hand, responds to samples *in isolation*, possibly adapting its weights, while being unable to explicitly compare a new sample to a previous one. The successfully trained network exhibits the same reaction—producing the output 1—on different occasions to different samples that instantiate the same concept, but it does not recognize sameness as such. Mastering a rudimentary concept instead means relating an indefinite range of multiple items to one another as partaking in one and the same trait and thus *viewing them as characteristically similar*. That there is a difference between reacting in the same way to isolated items and recognizing their similarity as such can be illustrated by the shift in appearance of a figure that we recognize as similar to others. The well-known duck-rabbit, for instance, changes it appearance when first seen as a duck and then recognized as a rabbit.

However, the difference matters nonetheless. For it is on account of not being able to compare items to one another and to recognize relevant continuities and discontinuities as such that a neural network is in principle unable to *rationally* extend rudimentary concepts in the face of circumstances which require such extension. The extension of a concept requires two things: Recognizing a novel circumstance as one that makes the concept as hitherto understood inapplicable by threatening its application with contradiction, and modifying the conditions of application of that concept in a way that allows to apply it to such circumstances without running into contradiction. Artificial neural networks are built in a way that prevents them from achieving either of these.

Fig. 11: The duck-rabbit



Such a shift of appearance can only occur to one who can *compare* items to one another, e.g., drawings to certain animals one has seen. There can clearly be a device that responds to two items in the same way without ever comparing them. This difference, however, might appear too subtle to matter when it comes to assessing the ability of humans and machines to recognize recurrent patterns. If performance is all that matters, it might indeed be safe to say that the machine simulation of a rudimentary concept can indeed outperform the human original.

In order to recognize a novel situation as one that requires the concept that is being applied to be extended (or its whole network of weights to be changed, for that matter) the machine would have to be able to compare its response to the new situation with earlier responses and to recognize that the two systematically contradict each other. Remember how the dilemma that gives rise to the extension of our rudimentary color concepts arose from comparing the color of the same object in situations that differ with regard to lighting. An artificial neural network that is being trained cannot compare samples and situations, because it can only respond to one sample at a time, adapting its weights accordingly, while not storing its own reactions to previous samples *as such*. Even if it did store these reactions, it could not detect a contradiction between them, because its only possible outputs are 0 or 1, i.e., yes *or* no, rather than 0 *and* 1.

A neural network might very well be trained to affirmatively respond to red items in standard lighting conditions *as well as* to samples that appear to be brown in green light. This, however, would not show that it now masters an ex-

tended concept of *red*. For in order to do that it would have to be able to distinguish between *one coherent* concept that exhibits a certain internal articulation by virtue of having been extended in response to a dilemma to which its application to a novel kind of circumstance gave rise, and *a mere combination* of *two incoherent* concepts such as *red and blue*. However, the network is built in a way that prevents it from noticing the *logical* difference *in kind* between *red and blue*, on the one hand, and *red in normal light while brown in green light*, on the other.

Why should it not be possible, though, to connect two parts of a neural network in such a way that a contradiction is indeed registered, if the response of one partial network to a thing is 0, while the other's response to the same thing is 1? If the machine is supposed to recognize these two responses as contradicting each other, it would have to recognize them as responses to *the same* thing, rather than just responding twice to what in fact is the same thing without noticing it to be so. Why should a machine not notice sameness and difference, though? Sameness and difference are neither contained in things, nor are they real relations between them such as spatial distance. Envisaging sameness and difference requires *comparing* things, and things do not compare themselves with one another, *we* do. A neural network accordingly cannot learn the concepts of sameness and difference by being exposed to things which *we* recognize to be the same or different. The network also cannot come up with these concepts on its own initiative, because the only thing it can change are the weights between its nodes. Each weight is an isolated numerical value that simply is what it is and thus cannot represent a distinction between itself and something else. If neural networks are constitutively unable to recognize dilemmas to which the application of a concept in a novel kind of situation gives rise, they cannot recognize a situation as one which requires the rational modification of the concept at issue.

That deep neural networks cannot engage in conceptual activity neither means that they couldn't do astonishing things, nor that they couldn't successfully do things we are unable to do or do less successfully. When asking what it is that successfully trained neural networks actually do, we shouldn't forget that their behavior matters to us, because we view it in light of intelligent activities *we* engage in and care for, rather than independently of them. Insofar as their behavior strikes us as meaningful and perhaps astonishing, it is because we assess this behavior, perhaps unwittingly, in a way that is parasitic on ours. In view of our understanding of a concept and a certain range of unproblematic applications that don't give rise to dilemmas, we might marvel at how much better the machine is at putting samples into that range. It is only better

than us, though, at a kind of activity *we* know of, while it doesn't, in the same way litmus paper is better than we are at reliably responding to slightly acidic liquids. It being better than us obviously doesn't mean that litmus paper could be credited with understanding the concept of an acid or that we care for what it does independently of viewing its behavior in light of the concept of acid *we* dispose of.

A Challenge for Our Times: To Distinguish between Use and Abuse of Simulated Thought

It has been argued that deep neural networks as conceived in machine learning can at best simulate the use of concepts rather than actually apply them. It remains to be asked what kinds of use *we* human beings who actually engage in conceptual activity might have for machines that simulate conceptual activity. While automated pattern recognition might turn out to have all kinds of meaningful uses that cannot yet be fully anticipated, it is important to point out that the simulation of our own intelligent activities by way of machines allows both for meaningful uses on our part as well as for abuse. Abuse is inevitable if we attribute to machines intelligent capacities which in fact only we humans possess, and which the machines themselves can only simulate. A simulation per definition deviates from the reality it simulates. However, within a certain limited range of application, we might be struck with how compelling, life-like, and maybe even statistically superior the simulation is with regard to its output. This fascination might make us overlook how poorly the system performs outside of that range of application. Even while deep neural networks might ultimately be more reliable than we are at sorting items within a continuous range of unproblematic cases, they are constitutively unable to recognize and rationally respond to situations which defy being assessed according to a given pattern. For that very reason, deep neural networks cannot be credited with conceptual capacities. Assigning machines tasks that indeed require conceptual capacities that put one in a position to both creatively and rationally respond to tricky cases can only result in failure, and possibly disaster.

A device for the automated recognition of a pattern, X, will not be able to recognize borderline cases which defy being classified as either X or non-X and to rationally respond to such cases by diversifying the pattern that is being sought. Instead, the device will inflexibly stick with putting samples in one of the two boxes it has been trained to recognize. We do not need to elaborate

here on the possible consequences of entrusting devices for automated pattern recognition instead of human administrators with the classification of social affairs.

It might be objected, though, that nothing speaks against training devices for automated pattern recognition to distinguish between clear cases for which automated classification is sufficient and borderline cases which require special—human—attention. This solution, however, is spurious. For it is written on the sleeves of our concepts that we are finite thinkers: our concepts cannot rule out the emergence of unpredictable situations which defy their application and require us to both creatively and rationally modify them in a way that allows us to continue using them. The machines which we might build to simulate conceptual activity inherit the finitude of our concepts and, hence, the partial unpredictability and elusiveness of reality in the face of our attempts to put it into boxes. In consequence, we might very well construct a device for automated pattern recognition that can recognize *a certain determinate kind* of situation as a borderline case that requires special—human—attention. However, no machine simulating the use of concepts can be trained to recognize *all relevant kinds* of borderline cases for what they are. For in order to do that we would have to be able to *conceptually anticipate* all possible situations that *defy* the application of a certain concept, which is a contradiction in terms.

There is no general reason why learning machines that we have trained to simulate conceptual activity by reliably responding to input within a certain standard range of application should not prove to be better than us at putting samples into their proper box. This doesn't make them, rather than us, authorities on the general distinction between unproblematic and borderline cases which require *thoughtful* responses, namely comparisons between cases, recognition of contradictions, and rational modification of our concepts. Ultimately, in the same way as microscopes or telescopes are tools that can help us to see things that we otherwise couldn't see, learning machines are *tools* that might help us to improve our own intelligent activities rather than delegating them to seemingly autonomous non-human agents.

I have argued in this essay that pattern recognition as conceived in deep learning only amounts to the simulation of concepts rather than proper conceptual activity. It is not immediately clear what this implies for the realm of Generative AI, which is not just about machines reliably responding to certain kinds of stimuli but about generating meaningful text, images, videos, etc. when prompted in a certain way. If it could be shown that concepts are involved, in some way or other, in meaningful speech as well as in the creation

of images and the like, it would follow that Generative AI can generally only simulate these activities rather than ever properly and autonomously engage in them.