Eberhard Karls Universität Tübingen
Faculty of Humanities

Thesis submitted in fulfillment of the requirements for the degree
Master of Arts (M. A.) in Philosophy

Artificial Neural Networks and the Reference Class Problem

Supervisor:
JProf. Dr. Alexandra Zinke

Submitted by:
Oliver Buchholz

Submitted on:
September 15, 2020

Contents

1 Introduction **1**

2 The Reference Class Problem **4**

2.1 What Is the Problem? 4

2.2 Are there Solutions? 8

3 Machine Learning and Artificial Neural Networks **16**

3.1 What Is Machine Learning All About? 16

3.1.1 Basic Concepts 17

3.1.2 Empirical Risk Minimization and its Philosophical Ramifications 21

3.2 Artificial Neural Networks 25

3.2.1 Basic Functionality 26

3.2.2 What Makes Them Special? 29

4 Do Artificial Neural Networks Solve the Reference Class Problem? **34**

4.1 The Argument: In Some Cases They Do 34

4.1.1 First Observation: “Big Data” is Related to Narrowness and Precision 35

4.1.2 Second Observation: Empirical Risk Minimization is Related to Epistemic Accuracy 37

4.1.3 Third Observation: There Are No Distributional Assumptions 38

4.1.4 The Artificial Neural Network Approach to the Reference Class Problem 40

4.2 . . . In Many Others, They Don’t: Possible Objections 44

5 Conclusion **48**

References **51**

List of Figures

1 Example of a linear regression 23

2 Depiction of the relationship between overfitting, underfitting, and the capacity of the hypothesis class 24

3 Schematic representation of a perceptron 26

4 Depiction of data obtained from an application of the XOR function 27

5 An example of an artificial neural network 28

6 Depiction of the double-descent risk curve 31

1 Introduction

Over the last decades, advances in the field of machine learning affected our everyday lives through new technology and gave rise to methods that are nowadays commonly and succesfully used in scientific applications. The way in which machine learning achieves its undeniable successes is a peculiar one, for on the one hand, it shares with traditional statistics the use of empirical observations to draw inferences, while on the other hand it shares with computer science the design of algorithms that are not told in advance how to solve a particular task (Wheeler 2016, p. 323). Clearly, this observation prompts a number of questions and perhaps the most urgent one, at least from a methodological point of view, is the one concerning the relation between machine learning and classical statistics, the latter being the philosophically well-scrutinized workhorse of any empirical discipline (see Romeijn 2017 and references therein).

In the following, I intend to contribute to the analysis of this question. Given the scope of this thesis and the breadth of the debate, the only feasible way of doing so consists in focusing on one particular strand of it.
As far as methods of machine learning are concerned, I restrict my attention to *artificial neural networks*. On the one hand, this choice is mainly motivated by the fact that most of the recent breakthroughs involving machine learning methods, ranging from advances in autonomous driving to the detection of exoplanets, were achieved by employing artificial neural networks (Buckner 2019, p. 1). Even the accomplishment that arguably received most public attention, the triumph of AlphaGo in the strategic board game Go over Lee Sedol, one of its strongest players, relied heavily on the use of artificial neural networks (Silver et al. 2016). On the other hand, and despite this astonishing track record, their internal functionality and the way by which they learn from data is still not well understood and remains essentially opaque even for machine learning researchers. This has only recently sparked some interest in the philosophical aspects of artificial neural networks (see e.g. Buckner 2018, Buckner 2019, Schubbach 2019). The literature on the subject can, however, still be considered in its infancy which is another reason for my focus on artificial neural networks within this thesis.

Furthermore, I restrict the following discussion to one particular issue discussed in the philosophy of statistics, namely the so-called *reference class problem*. Introduced presumably in the nineteenth century by the mathematician John Venn, Hans Reichenbach provided the first rigorous treatment of the problem by which he also established it in the philosophical discourse. Briefly put, the problem arises whenever one tries to infer whether a single individual belongs to some target class and, in doing so, one employs information regarding the frequency with which elements of this class are among the elements of another class, the reference class, to which the individual belongs. The difficulty in such situations is to find the right reference class, for it is usually the case that an individual belongs to more than one of them. At its heart, this is the reference class problem that remains one of the greatest challenges classical statistics is faced with.

As a consequence, the question that I try to answer in this text is whether artificial neural networks are, just as classical statistics, plagued by the reference class problem. If so, this would require an explanation for the contradictory observation that the achievements of artificial neural networks were not possible without a high predictive ability, that is, without the ability to successfully generalize from past observations to new events. If, on the contrary, it turns out that artificial neural networks are not subject to the reference class problem, the reasons for this surprising finding would have to be carved out in detail.

In fact, I will argue, artificial neural networks offer some remedy for the reference class problem in a very specific class of situations, while falling prey to it just as classical statistics in many others. I base my argumentation on a thorough analysis of possible solutions to the reference class problem that have been proposed in the literature. I consider the most promising approaches of identifying the right reference class and relate them to artificial neural networks and their specific functionality. My main argument centers around recent insights from machine learning research that reveal a peculiar behavior of artificial neural networks which distinguishes them from other methods of machine learning as well as from classical statistics. This peculiar behavior that arises in situations involving what has become publicly known as “big data”, I contend, allows artificial neural networks to solve the reference class problem in precisely these situations. This, however, directly confines the argument to the aforementioned situations. For this reason, admittedly, there are many other scenarios in which artificial neural networks are of little help in solving the reference class problem, just as classical statistics.

Following the steps of the argumentation that I just mentioned, the remainder of this text is organized as follows: In the second chapter, I analyze the reference class problem and the philosophical debate that evolved from it in detail. The third chapter is meant as an introduction to machine learning and artificial neural networks. First, I outline basic concepts of machine learning; then, I move on to artificial neural networks and point out how they differ from other methods of machine learning. Having provided the relevant background in chapters two and three, I set out my argumentation in chapter four by presenting three preliminary observations and drawing a conclusion afterwards. I also indicate the inherent dialectic of my argumentation by addressing several objections to it in the last part of chapter four. In chapter five, I conclude with a brief summary and discuss possible directions for future work on the topic.

2 The Reference Class Problem

This chapter is concerned with the central philosophical issue that is treated in this thesis: the reference class problem. In section 2.1, I show how it arises naturally in instantiations of the statistical syllogism, a rule of inference that is employed regularly in situations that involve probabilistic information. Afterwards, in section 2.2, I take a look at the philosophical debate that evolved from Reichenbach's initial treatment of the reference class problem and in particular at remedies that were proposed to solve it.

2.1 What Is the Problem?

Although most space in introductory textbooks on logic is devoted to deductive reasoning that guarantees the truth of a given conclusion provided the premises are true (Zoglauer 2016, p. 58), much of our actual reasoning takes quite a different path. In fact, although often only implicitly, we employ a rule of inference that is discussed as the *statistical syllogism* in the literature and takes the following general form:

$$\begin{array}{l} 1'.\ \text{Most } R \text{ are } T. \\ 2'.\ a \text{ is an } R. \\ \hline 3'.\ a \text{ is a } T. \end{array}$$

Clearly, this argument does not license an inference from true premises to a true conclusion with the same assurance as any rule of deductive inference, since it replaces the universal quantifications of the form "*all* R are T" that are characteristic for the premises of deductive arguments by extenuated quantifications according to which only "some", "few" or, in the present example, "*most* R are T". Consequently, the assurance about the inference from the premises to the conclusion seems inherently linked to the quantifier in front of the R in ($1'$).

Formalizing these ideas such that R and T are two arbitrary sets, the quantifier in front of the R is replaced by a statement about the relative frequency of elements of T among elements of R, $\text{freq}(T|R)$, and the assurance about the inference from the premises to the conclusion is interpreted as the probability that the conclusion "a is a T" is true, $P(a \in T)$, yields a precise formulation of the statistical syllogism that is, for example, presented in Thorn (2012, p. 301):

$$\text{(DI)} \qquad \begin{array}{l} 1^*.\ \mathrm{freq}(T|R) = r \\ 2^*.\ a \in R \\ \hline 3^*.\ P(a \in T) = r \end{array}$$

This form of inference, where "a conclusion about the probability of a proposition" as in (3^*) is drawn "based on frequency information" as in (1^*) is also known as a *direct inference* (Thorn 2012, p. 300).

Several aspects are worth mentioning about this formulation. First, observe that the frequency statement in the first premise, (1^*), ranges from zero—"no R is T"—to one—"all R are T"—, which makes the formulation flexible and widely applicable, for it nests various instantiations of the statistical syllogism.

Second, note that the conclusion, (3^*), is formulated in terms of a probability for some proposition being true. At this point, I do not intend to enter the—vast—discourse on how the nature of this probability should be interpreted properly, but rather, I would like to draw the attention to its role within the entire argument (DI), for it is not as straightforward as it might seem. To start, note that in (DI), the conclusion does not follow deductively from the premises, for they "can at most create a presumption in favor of the conclusion, and that presumption can be defeated by contrary information" (Pollock 1990, p. 78). Thus, the statistical syllogism or the direct inference, respectively, are rules of *non-monotonic* or *defeasible* reasoning in which the inference made only remains reasonable as long as no contrary information is introduced via additional premises. Consequently, some authors, like, for instance, McGrew (2001, p. 156) or Wallmann (2017, p. 485), hold that the probability statement in (3^*) indicates the assurance with which the conclusion, $a \in T$, follows from the premises. This view is inspired by the informal presentation of the statistical syllogism above, where, as we have seen, the assurance about the conclusion being true depends on the quantifier in the first premise, ($1'$), that is replaced by the frequency statement in the schema (DI). Thorn (2017, p. 2026) proposes a similar interpretation, stating that given the evidence in (1^*) and (2^*), it is rational for an agent to assign a probability of r to the proposition $a \in T$ being true. In the following, I adopt this interpretation of (3^*) but also Thorn's—at least in my opinion—concise notation, which is why in the schema (DI), the conclusion contains a probability statement that concerns the *transition* from the premises to the conclusion rather than this very conclusion itself.

Finally, there is a third peculiarity of the schema (DI) above: As we have seen, its aim is to infer the probability—in the sense that I just discussed—of some individual a being part of a set T that I denote by *target class* from now on. To do so, it uses the relative frequency of elements of T among another set R, that I denote by *reference class* from now on, as well as the observation that a is an element of R. Now, however, the difficulty consists in choosing the right reference class R, as it might be the case that the individual a possesses several properties and hence belongs to several plausible reference classes, $R_1, R_2, \ldots$, among which the elements of T are represented differently. To make this point concrete, consider the following example presented by Thorn (2012, p. 300):

> "For example, in the case regarding my neighbor's dog, the conclusion that the probability is 0.05 that Flint has fleas is based on my frequency information about the set of dogs. But Flint is a member of numerous reference classes (in addition to the set of dogs), such as the set of smallbreed dogs, the set of dachshunds, the set of brown dogs, etc., and direct inference based on frequency information for the different reference classes may lead to mutually inconsistent conclusions."

This difficulty of finding the right reference class as a basis for an inference in which one assigns a probability to a single case is called the *reference class problem.* According to Hájek (2007, p. 564), the conceptual problem originates in the nineteenth century, introduced by the mathematician John Venn—inventor of the well-known set-theoretic diagrams—, who observes a problem in assigning probabilities to individuals due to the fact that "every individual thing or event has an indefinite number of properties or attributes observable in it, and might therefore be considered as belonging to an indefinite number of different classes of things" (Venn 1876, as cited in Hájek 2007, p. 564).

Subsequently, Reichenbach (1949) was the first to provide a thorough treatment of the problem and to introduce it into the philosophical discourse. Just as Venn did before, Reichenbach (1949, p. 374) observes that

> "[i]f we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the *problem of the reference class.*"

Unlike Venn, however, Reichenbach does not close his investigation at this point but instead proposes a way to solve the problem, that I examine in the next section.

Reichenbach's work sparked a dynamic discourse that resulted in an extensive literature concerning the reference class problem, its implications as well as possible remedies. During this process, the treatment of the problem has gradually become more fine-grained. For instance, Fetzer (1977, p. 185) argues that the problem of the single case, that is, the reference class problem, has to be approached from two different perspectives, one being "the problem of (single case) prediction", the other being "the problem of (single case) explanation" (ibid.). While the former is about "selecting appropriate reference classes for *predicting* singular occurences" (ibid.), the latter is about the same selection with the purpose of "*explaining* singular occurences" (ibid.). Since Reichenbach, according to Fetzer, focuses on the problem of *predicting* single cases and most of the literature follows him in the statement of the problem, I focus on this case, in which one is interested in predicting or inferring the probability of an individual case, as well.

Another important distinction in the treatment of the reference class problem is drawn by Hájek (2007). Just as Fetzer, he argues that in fact there are two different reference class problems, namely a *metaphysical* and an *epistemological* one (Hájek 2007, p. 565). According to Hájek, the former, metaphysical reference class problem "concerns what probabilities are 'out there' " (ibid.) as there should exist some fact or objective truth about the nature of the probability of an individual thing or event. The epistemological reference class problem, on the contrary, arises from the fact that a rational agent can only assign exactly one probability to a single case which begs the question as to "which probabilities form appropriate bases for our inductive inferences" and should ultimately "serve us as guides to life" (ibid.). The intuition behind this reasoning is straightforward: Whenever we are confronted with a situation that involves frequency information, this information is confined to one or several particular reference class(es) and, all of a sudden, we find ourselves in the midst of the reference class problem while being forced to make a decision. Perhaps, before letting Flint move in, Flint's owners were pondering on which dog breed to get based on the criterion of their susceptibility to fleas. Given that they had information on the frequency of fleas among all dogs, smallbreed dogs, and dachshunds, they had to decide based on which information and hence, depending on which reference class, they would like to make their final decision. Granted, in this example, for reasons I will discuss in the context of possible remedies to the reference class problem below, the right reference class seems to be obvious. However, the example nevertheless illustrates that the reference class problem in its epistemological flavor comprises an inherently decision-theoretic component: As human beings, we are—at least for most of the time—agents to whom life not merely *happens*, but who *act* based on decisions. Yet, without committing ourselves to *one* reference class when dealing with frequency information, which amounts to some kind of a "meta-decision" *for* frequency information conditional on a specific reference class and *against* competing information that is conditional on other reference classes, any subsequent decision-making becomes infeasible. Consequently, solutions to the reference class problem that find and justify the reference class we should commit ourselves to are desperately needed. I investigate existing proposals in the following section.

2.2 Are there Solutions?

Since the reference class problem is central to any reasoning that follows the schema (DI) of direct inference, it is of foremost importance for many instances of human reasoning and, even more generally, for any situation in which statistical evidence is used to draw inferences. This is bad news, especially for the empirical sciences where methods of classical statistics are the sole means available to grasp reality at least partly (Poser 2012, p. 55). Consequently, much effort has been put into the quest for an appropriate solution of the problem and indeed, there exist several approaches in the literature that claim to find one—yet they are convincing to a varying degree. In a nutshell, any solution to the reference class problem needs to find an answer to the following question: How to choose the right reference class R in a schema like (DI)? This sounds very much like a decision-theoretic problem and in fact, as we shall see shortly, there are authors who put forward arguments for choosing the right reference class that are motivated by decision theory. Before going into the details, however, note that any solution to the reference class problem needs to consist of two essential ingredients, a strategy for choosing the right reference class and at least one normative criterion that justifies the strategy at hand as well as the fact that the reference class it chooses is really the *right* one.

The first proposal of a solution is due to Reichenbach (1949). Only few lines after the passage cited above, in which he introduces the reference class problem, he proposes that in order to deal with the problem, "[w]e then proceed by considering *the narrowest class for which reliable statistics can be compiled*" (Reichenbach 1949, p. 374). Intuitively, this makes sense.

Recall Flint, the neighbor's dachshund, for which we would like to find out the probability that he has fleas. Now suppose, that we are "only able to make reliable judgments about the frequency of dogs having fleas and about the frequency of dachshunds having fleas" (Thorn 2012, p. 300). Clearly, in this situation, we "should apply the latter frequency judgment in forming a belief about the probability that Flint, a dachshund, has fleas" (ibid.), because every dachshund belongs to the class "dogs", making the class "dachshunds" the narrowest reference class "for which reliable statistics can be compiled" as required by Reichenbach. Another indication for the intuitive appeal of Reichenbach's proposal might be found in the fact that several authors concur with it. A. J. Ayer (1963, as cited in Gillies 2000, p. 816), for instance, confirms the view of Reichenbach when he writes that

> "[t]he rule is that in order to estimate the probability that a particular individual possesses a given property, we are to choose as our class of reference, among those to which the individual belongs, the narrowest class in which the property occurs with an extrapolable frequency."

Note, however, that although Ayer replaces the term "reliable statistics" by "extrapolable frequency", the meaning of the latter concepts as well as that of "narrow" in the context of competing reference classes remains rather vague in Reichenbach's approach. Thus, apart from being intuitive *prima facie*, Reichenbach's approach faces several conceptual difficulties. While some authors, like Hájek (2007, p. 568), criticize and even reject the approach for precisely these reasons, others try to formalize it in order to make the concepts involved more precise and to save Reichenbach's proposal at least from the attacks I just mentioned.

Thorn (2012), for instance, belongs to the latter group of authors. He introduces two principles that are meant to capture Reichenbach's theory, the first being the principle of direct inference (DI) that we have already seen above, the second being the so-called *subset defeat*, which states the conditions under which a direct inference "based on frequency information for a given reference class, is defeated in virtue of frequency information for a subset of that reference class" (Thorn 2012, pp. 300). It takes the following form:

$$\text{(SD)} \qquad \begin{array}{l} \text{(i) } a \in R', \\ \text{(ii) } R' \subseteq R, \text{ and} \\ \text{(iii) } \mathrm{freq}(T|R') \neq r \end{array}$$

In the formulation (SD), Reichenbach's idea of the narrowest reference class is made precise by the subset relation in (ii), which states that some reference class R' is a subset of another reference class R. Furthermore, it clarifies in which cases an instantiation of (DI) is defeated and the narrower reference class should be chosen; namely, whenever there is reliable frequency information regarding the narrower reference class that contradicts the information for the broader one. Let me return to Flint, the dachshund, to illuminate this point: As seen above, Flint, a, is both an element of the set of dogs, R, and the set of dachshunds, R'. Now, assume we possess reliable information that every twentieth dog has fleas, $\text{freq}(T|R) = 0.05$. Given this information, we might infer by (DI) that the probability of Flint, a dog, having fleas is $P(a \in T) = 0.05$. Next, assume we get to know that every fortieth dachshund has fleas, that is, $\text{freq}(T|R') = 0.025$. Following Reichenbach, we should go with the narrower reference class, "dachshunds", and additionally, the initial direct inference would be defeated by an instance of (SD), since all conditions (i) to (iii) would be fulfilled, resulting in the new inference that $P(a \in T) = 0.025$. Finally, let us take Reichenbach at his words and consider the narrowest possible reference class, the singleton containing Flint only. Obviously, the relative frequency of elements of T among the singleton containing Flint will always be zero or one, since, for instance, Flint either has fleas or not. This, however, is not really insightful, for "all interesting instances of direct inference would be defeated", which is why the situation is considered "a paradigmatic example of the problem of Uninformative Statistics" (Thorn 2012, p. 303).

Thus, although the schema (SD) formalizes Reichenbach's ideas in a concise way and defines the narrowest reference class in terms of the subset relation, a first shortcoming is easily identified. Next, we have to return to Hájek (2007) and its criticism of Reichenbach's solution, since he mentions a further issue concerning the concept of the "narrowest reference class", that remains problematic even when it is formalized in terms of the subset relation. It arises when trying to compare the narrowness of different reference classes that overlap only partly, since, as Hájek (2007, p. 568) observes, reference classes cannot "be *totally ordered* according to their narrowness". For instance, in the case of Flint, who belongs to the class of dogs as well as to the class of dachshunds, it was straightforward to identify the narrowest reference class according to the subset relation. Now, however, consider a situation in which there are only reliable statistics regarding the frequency of fleas affection among mammals that weigh less than ten kilograms and regarding the frequency of fleas affection among brown mammals. How to proceed in this situation to infer the probability of Flint, a brown mammal of less than ten kilograms, having fleas? Obviously, each of the classes is narrower than the class of all mammals, but there is no reliable information as to which of them should be considered as the narrowest reference class. Furthermore, it would be a mistake to judge them as equally narrow, since "the mere fact that they each refine that class through the application of one further predicate [...] is by itself no reason" (Hájek 2007, p. 569) to do so.

In addition to his criticism of the notion of "narrowness", Hájek (2007) stresses that the concept of "reliable statistics" is not made precise by Reichenbach. Indeed, it still remains unclear within the formulation (SD) for it only presupposes that there is *some* frequency information without specifying it any further. By means of two observations, Hájek (2007, p. 568) reveals this conceptual weakness: First, he argues that "reliable" is a vague concept *per se* that cannot be pinned down employing ideas of classical statistics such as a sufficiently large sample size or unbiasedness, let alone only one of them. Second, and even worse, the notion of a "reliable statistic" might in fact be context-dependent and "sensitive to pragmatic considerations such as the weighing of utilities" (Hájek 2007, p. 568). This point, that once more stresses the pronounced decision-theoretic character of the reference class problem, becomes intuitively obvious once we compare a situation in which "the formation of white dwarves" (ibid.) is investigated to one in which "the safety of a new drug" (ibid.) is tested. In the former situation, one could imagine that much emphasis is put on accuracy in scientific theorizing. Therefore, a scientist might gather all evidence available, ending up with a large amount of properties associated with each white dwarf and a very narrow reference class which leads to an inference in the schema (DI) that is both very specific and relatively robust to subset defeat. On the contrary, in the latter situation, in which the safety of a new drug is tested, a high risk for human lives is involved and, consequently, most emphasis is likely put on its minimization. In this context, it might be a bad idea to choose a very narrow reference class, for it would restrict the investigation to individuals with a specific combination of characteristics while excluding others whose susceptibility to side-effects of the drug will remain unknown.

Hájek's analysis casts serious doubt on Reichenbach's solution to the reference class problem and the question as to whether it consists of a normative ideal that states what is the right reference class as well as a decision strategy that is sufficient to actually choose it. What then, if any, is his own proposal? Recall, that he distinguishes a metaphysical and an epistemological reference class problem which he approaches separately. For the first one, that is concerned with the probabilities as they are "out there", Hájek (2007, p. 582) suggests that "rather than try to solve the reference class problem" we should "*dissolve* it". How does he reach this surprising conclusion? According to Hájek (2007, p. 580), whenever "we seek unconditional, single-case probabilities we keep finding conditional probabilities instead", that is, all probabilities that exist are in fact reference class-dependent and hence conditional rather than unconditional probabilities.[1] Consequently, following this line of argumentation, it is impossible and even inconceivable to come up with a solution to the metaphysical reference class problem, for the basic probabilities "out there" are always conditional ones at heart. Unfortunately, after this insight, Hájek does not proceed by solving the epistemological reference class problem, but rather, he acknowledges that it still remains an open question which probabilities "should underpin our inductive reasonings and decisions" (Hájek 2007, p. 583). I will focus on this latter problem for the rest of this text, leaving aside the question whether Hájek is right about the metaphysical reference class problem and the claim that conditional probabilities form the proper basis of probability theory. For now, a final remark regarding Hájek's article seems appropriate: Albeit reaching a rather negative conclusion, the author achieves two objectives from which the entire debate benefits greatly. First, he clearly reveals the shortcomings of Reichenbach's initial proposal, thereby guiding later work towards these aspects that require further refinement. Second, he shows—by an argumentation that is beyond the scope of this text—that all interpretations of probability fall prey to the reference class problem. The importance of this result cannot be overestimated as it implies that there is no escape from the reference class problem by simply arguing that it arises from the frequentist interpretation of probability put forward by Venn and Reichenbach. As a consequence, the result invalidates all solutions to the reference class problem that rely on a different interpretation of probability and, furthermore, it licences a certain degree of inattentiveness regarding the definition and interpretation of the frequency and probability statements in schema (DI) above.

[1] Indeed, Hájek (2007) goes beyond the reference class problem and argues that conditional instead of unconditional probabilities should be regarded as the "proper primitive of probability theory", that builds conditional from unconditional probabilities in Kolmogorov's axiomatization until now.

As for the refinements of Reichenbach's proposal, the contributions of Thorn (2017) and Wallmann (2017) have to be mentioned, since they formalize the reference class problem and approaches towards its solution even further than Thorn (2012), thereby particularly addressing criticism of the vague concepts "narrow" and "reliable". Thorn (2017), for instance, is the first to justify the recommendation to choose the narrowest reference class using a clear normative criterion. To do so, he makes use of decision theory and puts forward an epistemic utility argument that is summarized concisely in Wallmann (2017, pp. 489). It proceeds as follows: An agent's goal in choosing the right reference class for a direct inference such as (DI) is to maximize *epistemic accuracy*, that is, to achieve the lowest possible difference between the inferred probability $P(a \in T)$ and the true value. Thorn (2017, p. 2029) employs so-called "proper scoring rules" to measure epistemic accuracy and shows, that accuracy is maximized if and only if frequency information for the narrowest reference class—in terms of the subset relation—is used in the direct inference.[2] A straightforward way to trace Thorn's use of proper scoring rules, taken by Wallmann (2017, pp. 489), is to consider a situation in which there is an individual, a, along with a conjunction of all properties it is known to have, S_c. Now, according to the schema (DI), all other individuals $d_1, \ldots, d_n$ that possess the same properties S_c should be assigned the same probability $P(d_i \in T) = v_i = v$. In a next step, the truth value of the proposition $d_i \in T$ is defined as $V(d_i \in T) \in \{0, 1\}$ and epistemic inaccuracy is measured by the average squared difference between predicted values, v, and true values, $V(d_i \in T)$, as follows:

$$S = \frac{1}{n} \sum_{i=1}^{n} (v - V(d_i \in T))^2 \tag{1}$$

This expression—as well as all proper scoring rules—then is minimized if and only if $v = \text{freq}(T|S_c)$ (ibid.). Thus, as stated above, inferring the probability $P(d_i \in T)$ in line with frequency information for the narrowest reference class maximizes epistemic accuracy.

Interestingly, this result even holds for cases in which the "conflict of narrowness and precision" (Wallmann 2017, p. 485) arises—cases, in which the frequency information for the narrowest reference class is not precise, while there is precise frequency information for a broader one. From a sta-

[2] A formal definition of proper scoring rules and several hints to the literature on the topic are provided in Thorn (2017, p. 2029).

tistical point of view, this is a very intuitive situation, since narrower reference classes have less members than broader ones, resulting in *estimated* frequency information that is based on a smaller sample and thus, by simple statistical arguments, less precise. Thorn (2017, Theorem 3), however, shows that according to the normative guideline of maximizing accuracy, one should still choose the narrowest reference class along with "statements of *expected* frequency" which are "the proper statistical premises for direct inference" (Thorn 2017, p. 2034). The schema (DI) introduced above would take on the following form in the case of imprecise information for the narrower reference class, assuming for the moment that $R' \subseteq R$ is the narrowest reference class:

$$(\mathrm{DI}_{\mathrm{exp}}) \qquad \begin{array}{l} 1^{*}.\ E[\mathrm{freq}(T|R')] = r \\ 2^{*}.\ a \in R' \\ \hline 3^{*}.\ P(a \in T) = r \end{array}$$

Here, E denotes the operator for the expected value that is applied to the imprecise frequency information for class R' and that replaces the known frequency information in the initial schema (DI). Clearly, the next question must concern the calculation of this expected frequency. Thorn (2017, p. 2034) proposes to do so by computing "probability weighted averages of frequencies" of the following form:

$$E[\mathrm{freq}(T|R')] = \sum_{i=1}^{n} v_i \times P(\mathrm{freq}(T|R') = v_i) \tag{2}$$

Here, the v_i are meant to reflect the imprecise frequency information for the reference class R' in the sense that $\mathrm{freq}(T|R') = v_1 \vee \cdots \vee \ \mathrm{freq}(T|R') = v_n$. The weights $P(\mathrm{freq}(T|R') = v_i)$ are the probabilities that $\mathrm{freq}(T|R')$ has value v_i, which Thorn arrives at via another application of schema (DI), a "meta direct inference" (Wallmann 2017, pp. 491) involving "relative frequencies in arbitrary subsets of the broader reference class" (Wallmann 2017, pp. 493), that is, by considering $\mathrm{freq}(T|S)$ where $S \subseteq R$.[3]

As we have seen, Thorn (2017) provides a considerable refinement of Reichenbach's theory, first by defining a norm—epistemic accuracy—that justifies choosing the narrowest reference class, second by developing a decision rule—maximizing epistemic accuracy—to actually identify and choose

[3] Wallmann (2017, pp. 490) provides a comprehensive yet accessible presentation of the entire and rather cumbersome procedure that is beyond the scope of the present discussion.

this reference class and third, finally, by developing a method to deal with instantiations of ($\mathrm{DI}_{\mathrm{exp}}$) in which the frequency information used in a direct inference is of statistical nature, as is the case in many real-world contexts.[4]

Acknowledging the first and second of Thorn's contributions, Wallmann (2017) suggests to improve the third aspect, that is, the computation of expected frequencies. The argument proceeds from the observation that the weights $P(\mathrm{freq}(T|R') = v_i)$ in (2) are inaccurate, since they are obtained from *arbitrary* subsets $S \subseteq R$ which causes them to cluster around the value of $\mathrm{freq}(T|R)$ (Wallmann 2017, p. 491). To obtain accurate weights, p_i^*, one should consider *exceptional* subsets in the sense that they exhibit a probabilistic dependence with the target class T instead of arbitrary ones. Wallmann (2017, p. 496) then calls "distributions that describe how frequencies in sub-reference classes [...] are distributed *natural distributions*" and thus, at least for him, the reference class problem boils down to finding these natural distributions, for they allow an agent to compute the expected frequency for the narrowest reference class which in turn yields the maximum accuracy when inferring a single-case probability such as $P(a \in T)$ in schema ($\mathrm{DI}_{\mathrm{exp}}$).

There cannot be any doubt that we have come a long way, tracing the proposed solutions to the reference class problem starting with Reichenbach and then, very much in a cone-like movement, reaching the contemporary debate. As far as I can see, the contributions by Thorn (2012, 2017, 2019) and Wallmann (2017) provide the most elaborate account of Reichenbach's initial ideas, since they are able to dispel much of the criticism concerning the precise meaning of "narrow" and "reliable". Granted, one might question whether the definition of accuracy in terms of proper scoring rules as used by the authors I just mentioned is the appropriate one and—with Hájek in mind—whether accuracy is the appropriate normative criterion altogether. But given one accepts this stipulation, the attention should be directed towards the identification of natural distributions for the narrowest reference class available when searching for a remedy to the reference class problem.

[4] Additionally, Thorn (2012) addresses the problem of uninformative statistics that is related to the singleton reference class and proposes a remedy to tackle the case of partly overlapping reference classes in Thorn (2019).

3 Machine Learning and Artificial Neural Networks

While the previous chapter was concerned with setting the stage for the philosophical considerations in this thesis, the present one deals with machine learning and artificial neural networks. Since the latter require at least a basic understanding of machine learning, section 3.1 is meant to serve as a short introduction to this topic, in which I outline the most important general principles by which it is governed. Then, in section 3.2, I zoom in on the machine learning methodology that is of primary interest for the following argumentation, that is, on artificial neural networks. In particular, I describe their functionality and show which characteristics make them stand out as peculiar against other machine learning methods.

3.1 What Is Machine Learning All About?

When confronted with the field of machine learning and the question what it is all about for the first time, one straightforward approach—especially for the philosopher—might consist in a conceptual analysis: What is *learning*, or, at least, what does it mean in the present context? And what does it mean to insinuate that the process of learning is performed by a *machine*? The latter question can be answered by observing that the term "machine" serves to qualify the domain of the subject, in the sense that it "does not study the process of learning in living organisms" (von Luxburg and Schölkopf 2011, p. 651). Instead, the focus is on "automated learning" performed by machines, namely by computers (Shalev-Shwartz and Ben-David 2016, p. 19). Consequently, the former question can be rephrased by asking what kind of learning it is that a computer performs. More or less detailed answers can be found in the machine learning literature that nonetheless converge on the main aspects: First, that learning must be conceived of as a *process* in which, second, "general rules" are inferred "by observing examples" (von Luxburg and Schölkopf 2011, p. 651) or, put differently, "experience" is converted "into expertise or knowledge" (Shalev-Shwartz and Ben-David 2016, p. 19). There is no doubt that concepts such as "observation", "experience" or "knowledge" are of foremost philosophical importance on their own—in fact, entire branches of philosophy analyze them as their objects of investigation—, yet I would like to refrain from this kind of discussion and instead mention a further, more detailed definition of learning that is presented in one of the standard textbooks of machine learning (Mitchell 1997, as cited in Goodfellow et al. 2016, p. 97):

> “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

Once more, the understanding of learning as a process is—at least implicitly—stated by mentioning the improvement that a learning computer program exhibits over time. Furthermore, experience plays a central role, just as in the aforementioned definitions. Finally—and here is where the definition goes beyond those that were mentioned above—, learning seems to depend highly on the given context as it is only defined relative to some performance measure and some specific task. Overall, the definition provides an ideal stepping stone to get more precise and explore the concepts involved in greater detail.

3.1.1 Basic Concepts

Let us start with the first component brought up in the definition and investigate the nature of the experience that is central to machine learning. Very briefly, experience in the context of machine learning equates to a *dataset*, since this serves as the input to any learning computer program or algorithm (Goodfellow et al. 2016, pp. 97). A dataset is best understood as a collection of several examples or *observations* each of which in turn consists of several *features* that have been obtained from some object or event (ibid.). A common and concise way of capturing a dataset is by means of a *design matrix*. This is a matrix $\mathbf{X}$ that contains the observations in the dataset as its rows and that has one column for each feature (ibid.). Following a notation that is used commonly in the literature, I will denote the number of observations—also known as the sample size—by n and the number of features—also known as the *dimension* of the data—by d, such that each observation is a vector $\mathbf{x}_i \in \mathbb{R}^d$ and for the design matrix it holds that $\mathbf{X} \in \mathbb{R}^{n \times d}$ with elements $x_{i,j},\ i = 1, \dots, n,\ j = 1, \dots, d$.

Note, that at first blush, this conception of representing objects in a dataset seems rather restrictive: How, if at all, should we deal with objects like texts or images that are not numbers and thus cannot be readily captured in a design matrix? Fortunately, there are ways to circumvent this constraint by defining the features in the dataset in a clever way. For example, images can be represented such that each pixel corresponds to one feature containing a numeric value for the color, either measured as an intensity or according to the CMYK/RGB color models (Hastie et al. 2009, p. 4). Texts, on the other hand, are commonly represented such that each feature corresponds to a certain word, counting the number of its occurences within the text.

Clearly, this means that the number of features d in a dataset might be high, in some cases even *very* high such that $d \gg n$. This issue is discussed under the headline of *high-dimensional data* in the literature and it will come up again in the discussion below. For the moment, note that the features included in a dataset are somehow related to properties or characteristics associated with the objects or events that constitute the observations in the dataset and might consequently provide a link to the analysis of the reference class problem above.

An important distinction in machine learning that is linked directly to the kind of experience from which a computer is allowed to learn is that between *supervised* and *unsupervised* machine learning (see e.g. Shalev-Shwartz and Ben-David 2016, pp. 22). While the design matrix is all there is in the unsupervised case, it is complemented with a vector $\mathbf{y} \in \mathbb{R}^n$ of *labels* or *targets* associated with each observation in the supervised case. The focus will be on the latter type of machine learning in the following and often, I will refer to the design matrix plus the vector of labels rather loosely as "the data". A central assumption that distinguishes machine learning from classical statistics then concerns the generation of the data: While classical statistics usually presupposes a *specific* underlying probability distribution from which the data was sampled in order to derive the properties of estimators or hypothesis tests, machine learning only acknowledges that *some* probability distribution generated the data without specifying it any further.[5] Additionally, it is assumed that some unknown and correct labeling function f that is "out there" relates each observation to its corresponding label in the way that $f\colon \mathbb{R}^d \to \mathbb{R},\ \mathbf{x}_i \mapsto f(\mathbf{x}_i) = y_i$ (Shalev-Shwartz and Ben-David 2016, p. 34).

Before proceeding by carving out the specific tasks tackled by methods of machine learning, let me illustrate the preceding discussion with an example. To do so, recall Flint, the dachshund, and suppose that we are interested in compiling a dataset that consists of information regarding six features, namely age, breed, coat type, color, weight, and height for Flint and all other dogs in the neighborhood as well as a binary label indicating whether a given dog has fleas or not. Assuming there are 12 dogs in the neighborhood—Flint included—that are ordered alphabetically in the dataset, we would end up with a design matrix $\mathbf{X} \in \mathbb{R}^{12\times 6}$ in which, for

[5] The only assumption commonly applied is that each observation is sampled from the *identical* yet unknown probability distribution and *independently* of all other observations, which is why it is called iid (independent and identically distributed) assumption (von Luxburg and Schölkopf 2011, p. 653).

instance, the first element in the first row, $x_{1,1}$, would give the age of the dog that is ordered first. Similarly, the first element of the target vector $\mathbf{y} \in \mathbb{R}^{12}$, y_1, would indicate whether the dog that is ordered first has fleas or not.

The next question one might feel inclined to ask concerns the "class of tasks" mentioned in the definition above that is performed by a learning machine: What is the goal of machine learning and how does it incorporate experience, that is, the data, in order to achieve it? First, and perhaps surprisingly, the process of learning itself is neither the task nor the goal of machine learning, but rather the "means of attaining the ability to perform the task" (Goodfellow et al. 2016, p. 97). Thus, keeping in mind the idea introduced above that machine learning is about inferring general rules from experience, we are in a position to discern the task of machine learning from other aspects that were already discussed: By now we know that experience enters a machine learning algorithm as input in the form of a dataset and furthermore, we know that the process of learning converts this input into a general rule, the output of the machine learning algorithm. Consequently, the task of machine learning is to come up with a general rule that is obtained from existing observations and that is able to generalize "to previously unseen, new examples" (von Luxburg and Schölkopf 2011, p. 651). Obviously, this must sound familiar to the philosopher and indeed, machine learning can be considered as an instantiation of inductive inference, a property that it shares with classical statistics. However, unlike in classical statistics, the focus of machine learning is not on statistical generalization, that is, on drawing a conclusion about some underlying population based on a limited yet representative sample, but, on the contrary, on predicting the label for a new and previously unseen observation (Hastie et al. 2009, p. 1). Now, to accomplish this task of predicting new labels, the general rule a machine learning method seeks to infer from the data is the labeling function introduced above, since it constitutes the true mechanism that links the observations with their corresponding labels. The output of a learning algorithm therefore consists of a *prediction rule h* that also maps an observation to a label and that tries to come as close as possible to the true labeling function. In a next step, that will complete the close reading of the definition from above, we will have to discuss the precise meaning of "as close as possible", an issue that concerns the "performance measure" mentioned in the definition and that will ultimately reveal *how* the prediction rule is inferred from the data.

However, before doing so, I would like to emphasize another important distinction in machine learning, this time related to the type of task that is performed by a learning algorithm: On the one hand, one refers to the task performed by an algorithm as *classification* when the labels associated with the input data are not real numbers, but specify which of k distinct categories each observations belongs to. In this case, the true labeling function is given by $f\colon \mathbb{R}^d \to \{1, \dots, k\}$ and the machine learning or classification algorithm tries to infer a prediction rule h that is defined accordingly. On the other hand, cases in which the labels are real numbers and the inferred prediction rule consequently has the form $h\colon \mathbb{R}^d \to \mathbb{R}$ are referred to as *regression* (Goodfellow et al. 2016, pp. 98). For instance, applying a machine learning algorithm to the example from above would be a special case of classification, namely binary classification, since the labels associated with the observations in the dataset can take on two different values, "dog has fleas", which might be encoded as "1", and "dog has no fleas", which might be encoded as "0". In this case, the output of a machine learning algorithm would be a rule $h\colon \mathbb{R}^6 \to \{0, 1\}$ that predicts whether a dog not included in the data has fleas, once it is supplied with information about its age, breed, coat type, color, weight, and height as input.

After discussing the kind of experience that serves as an input to machine learning algorithms and the type of tasks they accomplish, we now have to investigate the performance measures used to assess the specific degree of this accomplishment. As already mentioned, the goal of machine learning is to come up with a general rule that is able to predict future labels and mimics the true underlying labeling function as close as possible. Thus, a straightforward way to measure the performance of an algorithm consists in considering the differences between its predictions and actual observations. This idea is captured by the concept of a *loss function* that reports the "cost" associated with a wrong prediction. For instance, in a classification task as presented in the example above, a straightforward loss function is the so-called 0-1-loss: misclassifying observation $\mathbf{x}_i$ as $h(\mathbf{x}_i)$ when its true label is $f(\mathbf{x}_i) = y_i$ yields a loss of one, while a correct classification $h(\mathbf{x}_i) = y_i$ yields a loss of zero (von Luxburg and Schölkopf 2011, p. 654):

$$\ell(\mathbf{x}_i, y_i, h(\mathbf{x}_i)) := \begin{cases} 1 & \text{if } h(\mathbf{x}_i) \neq y_i \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \dots, n \tag{3}$$

Note, however, that in general, it is not straightforward to come up with a loss function that exactly matches the task at hand and “that corresponds well to the desired behavior of the system” (Goodfellow et al. 2016, p. 102). Consequently, the choice of an adequate loss function usually requires careful consideration.

At this point, one might wonder how the value of a loss function can be computed in practice, given that we are ultimately interested in the performance of a prediction rule on data that it has not seen before. To this end, a simple yet elegant workaround has been developed in the literature: The overall dataset available is split into one part the machine learning algorithm is allowed to access, the *training set*, and one part that the algorithm cannot access and that is solely used to measure its performance on unseen data, the *test set* (ibid.).

Finally, one might be interested in comparing the performance of two different prediction rules, say h_1 and h_2. To this end, the *risk* of a prediction rule h, $R_n(h)$, is defined as the average loss over a sample of n observations,

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, y_i, h(\mathbf{x}_i)), \tag{4}$$

where as before, $\mathbf{x}_i \in \mathbb{R}^d$ is an observation consisting of d features, $y_i \in \mathbb{R}$ is the corresponding label and $h(\mathbf{x}_i)$ is the label predicted by prediction rule h (von Luxburg and Schölkopf 2011, p. 654). In this setting, we would clearly prefer prediction rule h_1 to its competitor h_2 if its risk is smaller, that is, if $R_n(h_1) < R_n(h_2)$.

The last observation also hints towards an answer to the question as to how a machine learning algorithm infers a prediction rule from the data: by following the decision rule to output the prediction rule with the smallest risk. This notion is formalized within the framework of *empirical risk minimization* (ERM) that is of foremost importance in machine learning and that I will investigate in greater detail in the following section.

3.1.2 Empirical Risk Minimization and its Philosophical Ramifications

The ERM framework embodies a straightforward decision rule to infer predictors from data that is already summarized in its title: An algorithm ought to choose the prediction rule h for which the empirical risk or training error, that is, the risk that can be computed from the training set accessible to the algorithm, is minimized (Shalev-Shwartz and Ben-David 2016, p. 35). Intuitively, this makes sense, for the training set is the window through which an algorithm "sees" the world and it seems reasonable to infer a prediction rule that performs well on this part of it.

Any chosen prediction rule h should furthermore exhibit considerable predictive ability in the sense that in addition to minimizing the training error, the test error is small as well. Thus in sum, "[t]he factors determining how well a machine learning algorithm will perform are its ability to (1) [m]ake the training error small" and "(2) [m]ake the gap between training and test error small" (Goodfellow et al. 2016, p. 109). Two central challenges in machine learning can be related directly to these factors, namely *underfitting* and *overfitting*. Underfitting occurs, when a prediction rule is not able to achieve a sufficiently low training error and hence lacks fit to the data. Overfitting, on the other hand, occurs when a prediction rule achieves a very close fit to the data and, consequently, a very low training error (ibid.). Although this might seem beneficial *prima facie*, a closer look reveals that a near-perfect fit to the data makes a prediction rule rather inflexible to correctly predict new labels, resulting in a large gap between training and test error. Beyond the purely methodological literature in machine learning and statistics, the problem of over- and underfitting has also attracted the attention of philosophers. For instance, though in the context of scientific theories and their fit with the evidence, Hitchcock and Sober (2004, p. 3) refer to overfitting the data as a "methodological sin [...] because it undermines the goal of predictive accuracy." Thus, whenever the chosen goal is predictive accuracy, overfitting the evidence, that is, the existing data, should be avoided. Of course, one might question the goal of predictive accuracy in the first place, especially in the case of scientific theorizing that is discussed by Hitchcock and Sober (2004). In the case of machine learning, however, we are after *prediction* rules, so it seems reasonable to take predictive accuracy as their primary goal. As a consequence, the problem of over- and underfitting becomes inescapable.

The standard way to balance under- and overfitting is to prespecify a *hypothesis class* $\mathcal{H}$ from which the machine learning algorithm is allowed to choose the prediction rule h, such that $h \in \mathcal{H}$ (Shalev-Shwartz and Ben-David 2016, p. 36). The intuition is as follows: If we choose a suitable hypothesis class for some given problem, for instance based on prior knowledge about the nature of the data, we might avoid underfitting due to the fact that all prediction rules in that class will fit the data rather well. Additionally, we might also avoid overfitting, since we actively restrict the algorithm's possible outputs to the elements of that class, thereby excluding prediction rules that allow a near-perfect fit to the data.[6] As an example, consider the case of a simple linear regression, a method that is commonly applied, for instance, in the social sciences, and that can be conceived of as a simplistic machine learning algorithm (Goodfellow et al. 2016, pp. 105): Given a training set of data, $\mathbf{X} \in \mathbb{R}^{n \times d}$, the goal of the algorithm is to find a prediction rule, $h\colon \mathbb{R}^d \to \mathbb{R}$, that minimizes the training error as measured by the squared distances between the regression line and the observations, a situation that is depicted for the case where $n = 10$ and $d = 1$ in figure 1. However, since we are dealing with a *linear* regression, the hypothesis class from which the algorithm is allowed to select a solution only consists of linear functions mapping the observations, $\mathbf{x}_i \in \mathbb{R}^d$, to their corresponding labels, $y_i \in \mathbb{R}$, that is, we have $\mathcal{H} = \{h\colon y_i = x_{i,1}w_1 + \cdots + x_{i,d}w_d,\ i = 1, \ldots, n\}$. As a consequence, overfitting will be prevented by the linear prediction rule that the algorithm outputs, because it cannot perfectly fit the data as might be the case with a high-degree polynomial. Still, underfitting will be prevented as well given that there is some indication for a linear relationship in the data that justifies the choice of the hypothesis class.

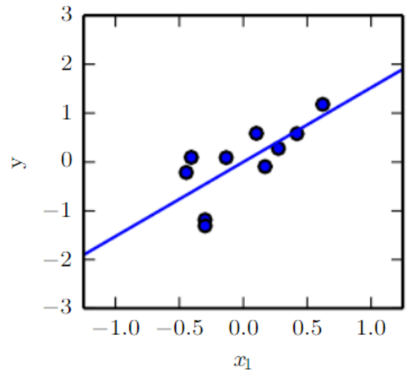

Figure 1: Example of a linear regression where the dimension of the data is $d = 1$ and the sample size is $n = 10$ (Goodfellow et al. 2016, p. 107).

In general, it holds—informally at least—that the "larger" the hypothesis class, the higher an algorithm's *capacity*, that is, its ability to fit a variety of different functions (Goodfellow et al. 2016, p. 110).[7] As we have seen, there is a close relationship between over- and underfitting as well as the hypoth-

[6] Note, that a near-perfect fit to the data is enough for overfitting to occur, since "[r]eal data are almost always noisy" and hence, prediction is "rarely, if ever, a matter of achieving *perfect* fit to data" (Hitchcock and Sober 2004, p. 10).

[7] The quotation marks are meant to indicate the informality of the statement, for the capacity is linked only indirectly to the actual size of $\mathcal{H}$ as measured by the number of its elements.

esis class at hand and in fact, it is the capacity that links these concepts. A concise overview over how this link arises is given in figure 2. As becomes obvious, the training risk decreases with increasing capacity of the hypothesis class, such that at some point, underfitting can be overcome. In fact, there is a "sweet spot" at which fit to the training data and susceptibility to overfitting are in balance.

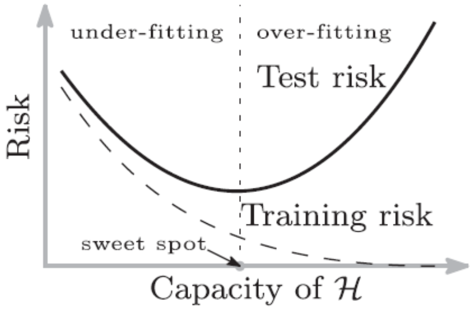

Figure 2: Curves for training risk (dashed line) and test risk (solid line) depicting the relationship between overfitting, underfitting, and the capacity of the hypothesis class $\mathcal{H}$ (Belkin et al. 2019, p. 15850).

Yet, beyond that "sweet spot", the fit to the training data is just too close and a successful prediction of unseen labels turns infeasible, giving rise to a high test error and, in particular, a large gap between training and test error that corresponds to overfitting. I will bring up this topic again in the discussion below, since, as it turns out, artificial neural networks behave rather peculiarly against this background.

Before, I would like to pause for a moment and hint at some further issues mentioned in the preceding analysis that are of genuine philosophical importance. First, and beyond the aforementioned fact that the process of learning from experience makes machine learning an instantiation of inductive inference, several authors, philosophers and machine learning researchers alike, have pointed out that an algorithm's selection of some prediction rule from a predefined hypothesis class can be interpreted as an application of Popper's idea of falsification.[8] This is because the algorithm goes through all functions, that is, through all hypotheses in the hypothesis class and "falsifies" or rejects all those with a deficient fit to the data. At least in my opinion, this makes machine learning an exciting object of investigation in the context of philosophy of science, for it seems—and the literature apparently supports this view—that two paradigms of scientific method, namely the principle of induction and falsificationism coincide within its algorithms.

[8] See Corfield et al. (2009) for a rigorous and concise treatment and Harman and Kulkarni (2007) for a book-length investigation.

Second, recall that as we have seen above, anyone applying machine learning methods needs to come up with a suitable loss function to measure the method's performance and, additionally, with a hypothesis class that avoids overfitting as well as a computationally infeasible search across all functions there are to come up with a prediction rule. These are inherently *normative* questions, at least from an epistemological point of view, since ultimately, they center on the issue as to whether one ought to prefer some loss function and hypothesis class over others, perhaps because the former license a higher credence in the algorithm's predictions than the latter do. In addition, the decisions the user of machine learning algorithms needs to make reveal the relevance of human involvement, especially via prior knowledge, for a process that is allegedly performed by a machine alone. In this way, the behavior of algorithms is influenced or biased towards some specific goal, be it the minimization of a particular loss function or the search of a prediction rule among linear functions only. It is for this reason that human decisions entering the machine learning process are referred to as *inductive bias* in the literature (Shalev-Shwartz and Ben-David 2016, p. 37).

3.2 Artificial Neural Networks

The preceding analysis of central concepts in machine learning enables us to approach artificial neural networks in a next step. Although they seem to represent a relatively recent methodology, especially under the heading of "deep learning", their underlying idea dates back as far as to the 1940s, according to the historical overview provided in Goodfellow et al. (2016, pp. 12).[9] While artificial neural networks have always been loosely motivated by "the behavior of neurons and synapses at some level of abstraction" (Buckner 2019, p. 2)—hence their name—their recent surge in popularity mainly stems from technological advances under the headline of "big data" that permit the measurement and processing of ever-increasing amounts of data (Goodfellow et al. 2016, pp. 18). In this context, for which artificial neural networks are suited particularly well, a large number of characteristics can be associated with each single object or event that constitutes one observation in the data (Goodfellow et al. 2016, pp. 12, 152). Consequently, datasets used in machine learning applications nowadays often belong to the setting of high-dimensional data that I briefly mentioned above, that is, to a setting in which the number of characteristics associated with each observation exceeds the overall number of observations. In light of our discussion of the reference class problem above, this is an interesting situation,

[9] For an in-depth historical overview, Schmidhuber (2015) is an excellent reference.

for the high number of characteristics might be linked to the concept of a narrow reference class, while the comparatively low number of observations might have implications for the reliability of statistics computed from them. Before investigating these links in greater detail, I would like to outline the basic functionality of artificial neural networks first.

3.2.1 Basic Functionality

The basic building-block of artificial neural networks is the *perceptron*, a simplistic machine learning algorithm as depicted in figure 3. The perceptron takes d features of an observation $\mathbf{x}_i \in \mathbb{R}^d$ as input. Then, it computes a weighted sum of the inputs and their corresponding weights w_j, $j = 1, \ldots, d$ and, finally, outputs the value of a so-called *activation function* that depends on the weighted sum of inputs. Consequently, the weights indicate the strength of the relationship between input and output (Harman and Kulkarni 2007, p. 79). The learning process of a perceptron then proceeds as follows: First, the weights are initialized with random values. Next, the features of one observation are fed into the perceptron and the resulting output is compared to the label that was observed in reality. If the output corresponds to the true label, no change is made to the weights, otherwise they are altered slightly and the next observation is input into the perceptron. This procedure is repeated until the empirical risk cannot be decreased any further and the collection of weights that achieves the lowest risk is the prediction rule selected by the algorithm (ibid.).

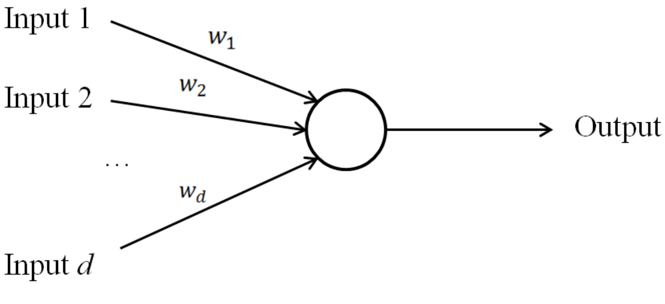

Figure 3: Schematic representation of a perceptron with d inputs and corresponding weights w_j, $j = 1, \ldots, d$ (adapted from Harman and Kulkarni 2007, p. 78).

Note, that the hypothesis class defined by a perceptron is the same as the one discussed above for the case of a linear regression, as the algorithm chooses a prediction rule that is a linear function of the inputs in any case. This fact and the corresponding inability of the perceptron to fit a variety of data patterns was already noted by Minsky and Papert (1969) in an example that has become a classic by now and is reproduced in Harman and Kulkarni (2007, pp. 80) as well as in Goodfellow et al. (2016, pp. 167). It proceeds as follows: Consider the XOR function that applies the exclusive "or" to observations consisting of two binary features, $\mathbf{x}_i = (x_{i,1}, x_{i,2})' \in \{0,1\} \times \{0,1\}$. Thus, the function returns the value "1" when exactly one feature of an observation equals "1" and "0" otherwise. This XOR function is the true labeling function f that we would like to learn from the data using a perceptron. Now, assume we were able to compile a dataset of *four observations* and *two features* that can serve as the training set for the perceptron from which it can try to infer a prediction rule h that comes as close as possible to the XOR function. As we have seen above, this dataset can be represented in a (4×2)-matrix as follows:

$$\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{5}$$

Furthermore, we can depict the observations and the corresponding labels in a two-dimensional coordinate system as shown in figure 4. It is this figure that reveals the problem that arises when trying to infer the XOR function using a perceptron: No prediction rule that is a linear function can predict the outputs of the XOR function correctly.

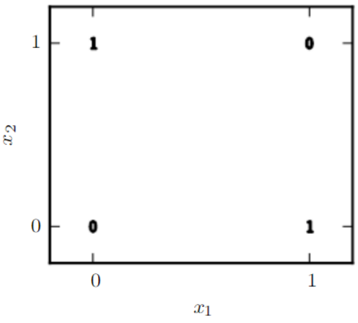

Figure 4: Depiction of data obtained from an application of the XOR function (Goodfellow et al. 2016, p. 169).

This drawback of the perceptron led machine learning researchers to build more flexible algorithms by combining several perceptrons to so-called multilayer perceptrons, or, put differently: to artificial neural networks. Essentially, the term "multilayer perceptron" already captures their entire functionality, since it indicates that several perceptrons are combined to multiple layers, the number of which is usually denoted by $t = 1, \dots, T$. This is why an artificial neural network can be conceived of as a graph that consists of a set of nodes, V, the neurons, and a set of edges, E, linking the output of one neuron to the input of another one by means of weights as shown in figure 3 above (Shalev-Shwartz and Ben-David 2016, p. 269). Within each neuron, the weighted sum of inputs is processed by the activation function and, subsequently, leaves the neuron as an input to the next one. Two important terms in this context are a network's *depth* and *width*, the latter referring to the maximum number of neurons encountered across all layers of the network, the former referring to the number T of layers in the network.[10] For instance, an artificial neural network with depth two and width five is depicted in figure 5. Here, the ith neuron of the tth layer is denoted by $v_{t,i}$, V_0 is the input layer that takes the data as its input and V_2 is the output layer that produces the final output. The layer V_1 is referred to as "hidden layer", since its outputs are only processed to the next layer and never revealed to the user of the network (ibid.).

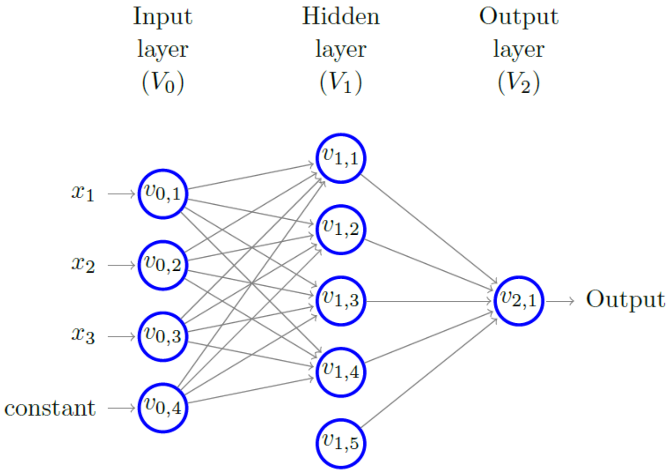

Figure 5: An example of an artificial neural network with depth two—the input layer is not considered for the depth—and width five (Shalev-Shwartz and Ben-David 2016, p. 270).

Clearly, an artificial neural network nicely fits into the general framework of machine learning introduced above. First, there is data that serves as input to the first layer of the network. Next, the algorithm tries to come up with a prediction rule producing outputs that match the actual labels associated with the observations as close as possible. However, while the process of finding a prediction rule was straightforward to describe in the case of a single perceptron, it becomes increasingly cumbersome with every layer that is added to an artificial neural network. Recall, that for the single perceptron, the prespecified hypothesis class consists of linear functions and,

[10] This is the origin of the term "deep learning", since the strategy in this field is to build networks with an enormous number of layers, that is, with a very high depth (Goodfellow et al. 2016, p. 165).

consequently, has a rather limited capacity. For artificial neural networks, on the contrary, the prespecified hypothesis class consists of highly complex functions that arise from the successive application of chained linear functions: The input to $v_{1,i}$ arises from a linear function, its output is processed by a linear function feeding it into $v_{2,i}$ and so on up until $v_{T,i}$. Since it is infeasible to specify the set of functions from which the algorithm is allowed to choose the final prediction rule explicitly—as in the case of linear functions—, one usually defines it implicitly by specifying an *architecture* for the network, consisting of the number of edges and nodes as well as the activation function at the nodes, which nevertheless "is not a trivial job" (Schubbach 2019, p. 7). Then, the algorithm performs the learning process similarly to the case of the single perceptron, that is, it chooses the weights by minimizing the deviation between the network's output and the actual labels.

Apart from the mere description of the basic functionality of artificial neural networks, there is one aspect that cannot be stressed enough and that has in fact already been subject to philosophical investigation in Schubbach (2019): The higher the number of layers in a network, the more complex and intransparent becomes the process by which the algorithm comes up with its output, the final configuration of weights that represents the chosen prediction rule. "Although we do get an output, we do neither know how this output was computed nor why it is this output and no other" (Schubbach 2019, p. 8). Just considering the number of possible interactions of weights across layers as well as the number of outputs within hidden layers that remain entirely obscure greatly supports this argumentation. And there are further properties that make artificial neural networks stand out as peculiar among methods of machine learning, both from a methodological and a philosophical perspective. They are the subject of the subsequent section.

3.2.2 What Makes Them Special?

Until now, we have seen that methods of machine learning are regularly confronted with the problem of over- and underfitting, since their ultimate goal is to learn a prediction rule from existing data that generalizes well to new and unseen data in the sense that it achieves a high predictive accuracy. An observation that has been made in recent machine learning research and that has been confirmed repeatedly, however, is the following: Artificial neural networks perform particularly well in the high-dimensional setting mentioned above, that is, “when the number of parameters [or features] is significantly larger than the amount of training data” (Neyshabur et al. 2017, p. 5947). In this setting, they are able—and in practice deliberately trained to—exactly fit the training data, thereby achieving zero training error (Belkin et al. 2019, p. 15849). With our preceding discussion of central ideas in machine learning in mind, one might take this behavior as a clear-cut indication for overfitting and a poor predictive performance of artificial neural networks on new data. However, as several authors show in empirical experiments, artificial neural networks “exhibit a remarkably small difference between training and test performance” (Zhang et al. 2017, p. 1) and, consequently, “good generalization behavior” (Neyshabur et al. 2017, p. 5947). Clearly, this seems peculiar as it is at odds with the standard theoretical framework of machine learning, especially regarding its treatment of the under- versus overfitting problem and the conventional wisdom presented in standard textbooks that “a model with zero training error is overfit to the training data and will typically generalize poorly” (Hastie et al. 2009, p. 221). Note, that these observations are also at odds with the philosophical view of overfitting existing evidence as a “methodological sin” that violates the goal of predictive accuracy—it seems as artificial neural networks were able to fit the evidence perfectly while preserving the flexibility that is necessary to predict new data correctly.

Thus, apparently, the case of artificial neural networks is not appropriately captured by the depiction in figure 2 where, as we have seen, an algorithm’s ability to predict new data diminishes with increasing capacity of the underlying hypothesis class, that is, as soon as the algorithm becomes subject to overfitting. As a consequence, Belkin et al. (2019) propose and empirically confirm an alternative framework that combines the traditional context of under- and overfitting—the “classical” regime according to the authors—with the specific behavior of artificial neural networks—the “modern” interpolating regime. The main feature of their framework is what the authors refer to as the *double-descent risk curve* depicted in figure 6. As becomes evident, this curve corresponds to the classical U-shaped curve depicted in figure 2 above, as long as an algorithm’s capacity is below the so-called interpolation threshold. This threshold marks the point beyond which an algorithm achieves zero training risk or, in other words, interpolates, that is, perfectly fits the training data. Now, while prediction rules obtained directly at the threshold generally exhibit a high test risk as shown in the figure, indicating a low predictive accuracy, Belkin et al. (2019, p. 15850) “show that increasing the function class capacity beyond this point leads to decreasing risk, typically going below the risk achieved at the sweet spot in the 'classical' regime." This means that large or deep artificial neural networks with a complex architecture involving many layers and incorporating a high number of features as inputs are suited particularly well for any kind of prediction task.

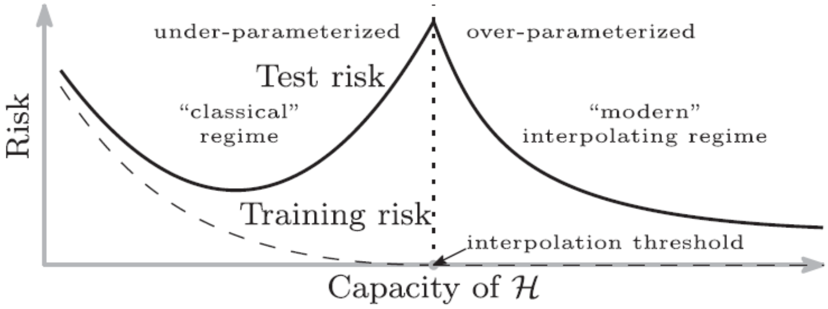

Figure 6: Curves for training risk (dashed line) and test risk (solid line) depicting the double-descent risk curve, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from the "modern" interpolating regime, separated by the interpolation threshold. (Belkin et al. 2019, p. 15850).

What, then, is the mechanism behind this observation? In fact, apart from the empirical confirmation of the double-descent risk curve, there is little insight, let alone consensus in the literature as to what might be the driving force behind the behavior of artificial neural networks: "until recently it has been difficult to explain why they work so well" (Buckner 2018, p. 5355). Most notably, the combination of severe overfitting and high generalization performance that is at the heart of the double-descent framework proposed by Belkin et al. (2019) still represents "a phenomenon that remains largely unexplained" (Neyshabur et al. 2017, p. 8). A possible key to unlock at least tentative answers to the preceding questions seems to be the procedure by which the weights of an artificial neural network are chosen.[11] The starting point of the analysis concerns "the precise relationship between optimization and *implicit regularization*" (ibid., emphasis added) in the determination of a network's weights, a direction investigated by several other authors as well (Belkin et al. 2019, Neyshabur et al. 2015, Zhang et al. 2017). The idea—for which all of the mentioned authors gather experimental evidence—is that the procedure that determines a network's weights seems to possess an implicit preference for "simple" prediction rules, as it often ends up with solutions of a rather small "complexity" even though

[11] For the sake of completeness, note that this procedure consists in an optimization algorithm that is called *stochastic gradient descent* (SGD) with *backpropagation*. A detailed treatment can be found in Goodfellow et al. (2016, Ch. 5.9 and 6.5).

the network's overall architecture is very complex. This is why the behavior is called "implicit regularization", since it is not explicitly built into the algorithm and it performs *regularization*, that is, a "modification [. . .] to a learning algorithm that is intended to reduce its generalization error [the error made when predicting new data] but not its training error" (Goodfellow et al. 2016, p. 117). This implicit regularization, albeit being considered a "major issue still left unresolved" (Neyshabur et al. 2017, p. 8), can explain why vastly complex artificial neural networks with zero training error do not overfit, but rather achieve a high predictive accuracy as suggested by the double-descent graph above.

Now, the final question that remains to be answered concerns the specific nature of "simplicity" or "complexity" as discussed in the context of implicit regularization that is at work in the determination of a network's weights. Clearly, it cannot be related to the overall complexity of the chosen hypothesis class, for the latter depends on the network architecture and will necessarily be high for artificial neural networks beyond the interpolation threshold. So, in which regard do the weights of resulting prediction rules tend to be simple? The link to the notion of simplicity that is relevant in this context is the insight that all weights of an artificial neural network can be stacked in a large vector whose "size" can be computed subsequently. This is usually achieved by employing a specific function, called *norm*, that takes the elements of a vector—the weights of a network in the case at hand—as its input and yields a real number, a vector's size or norm, as its output (Goodfellow et al. 2016, p. 37). Thus, when stating that the procedure selecting a network's weights exhibits an implicit regularization in the sense that it often yields simple solutions, this means that it "will often converge to the solution with minimum norm" (Zhang et al. 2017, p. 9). In this situation, many weights will be of a rather small magnitude while others will even end up with a value of zero, indicating a limited or even non-existent connection between the corresponding nodes of the network (Neyshabur et al. 2015, p. 6).

After this discussion full of technical details regarding artificial neural networks and their behavior that distinguishes them from other machine learning methods, let me once again stress its philosophical implications: We have seen that artificial neural networks represent a methodology that is able to overcome the classical problem of overfitting and instead achieves both a near-perfect fit to existing evidence and a high predictive accuracy. This behavior hinges on the number of inputs to the network being high which means that the dataset at hand needs to contain a high number of properties associated with each observation. Then, by means of an implicit regularization, the algorithm that selects the network's final configuration decides by how much each property should be weighted and which of them should not be considered in the final configuration that is used to compute predictions. Although the precise functionality of the implicit regularization remains opaque, the entire setting provides an exciting bridge across which we can return to the reference class problem introduced in section 2.

4 Do Artificial Neural Networks Solve the Reference Class Problem?

Having examined the reference class problem and central ideas of machine learning separately, the latter in particular in the context of artificial neural networks, the crucial step to answer the question overarching this text consists in analyzing the combination of both subjects: How, if any, are artificial neural networks suited to deal with the reference class problem that plagues classical statistics? In section 4.1, I argue that, in fact, artificial neural networks remedy specific instantiations of the reference class problem to a certain degree. In section 4.2, however, I try to show that this result comes with significant qualifications by presenting several objections to this line of argumentation, some of which can be only partially refuted.

4.1 The Argument: In Some Cases They Do ...

The discussion of the reference class problem revealed that when trying to infer whether an individual belongs to a particular target class, one ought to base the inference on the narrowest reference class available, at least if the goal is epistemic accuracy. This holds even in cases in which frequency information for this class is imprecise, for it might be replaced with information regarding the *expected* frequency. Consequently, much of the recent literature focuses on the correct computation of expected frequencies when trying to maximize accuracy in order to overcome the reference class problem. This is, as we shall see shortly, where artificial neural networks come into play.

Before, let us briefly reflect on how to conceive of the reference class problem's general structure within the schema of direct inference, be it (DI) or ($\mathrm{DI}_{\mathrm{exp}}$), from a machine learning perspective. First, note that the entire setup resembles a setting of binary classification: We would like to infer or predict, whether some individual, a, is a member of a given target class, T, so we might attach a binary label $y \in \{0, 1\}$ to individual a that takes on the value "1" if $a \in T$ and the value "0" otherwise. Furthermore, to enable such a prediction, some machine learning algorithm has to come up with a general prediction rule, h, based on an existing dataset, $\mathbf{X} \in \mathbb{R}^{n \times d}$. Clearly, the data from which an algorithm learns the prediction rule can be interpreted along the lines of existing frequency information in the premises of (DI) or ($\mathrm{DI}_{\mathrm{exp}}$). In particular, note that there seems to be a close analogy between the concept of a reference class and the structure of the data that is used as an input for machine learning algorithms, since in the former, by definition, each element has a number of properties such as “dog”, “smallbreed dog” or “wire-haired dachshund of less than ten kilograms”, while in the latter, very similarly, each observation, $\mathbf{x}_i$, is associated with a range of d properties. Finally, as outlined above, once obtained from the data, the prediction rule then *predicts* the label for a new and previously unseen observation, for instance the individual a, by mapping its properties to the binary label that indicates its belonging to the target class, such that $h\colon \mathbb{R}^d \to \{0, 1\},\ a \mapsto h(a) = y$.

With this setup in mind and by means of three preliminary observations, I will now illustrate how an application of artificial neural networks resembles approaches that have been proposed in the literature to tackle the reference class problem.

4.1.1 First Observation: “Big Data” is Related to Narrowness and Precision

The first observation is rather straightforward and concerns the nature of the data that is used as an input to artificial neural networks and machine learning algorithms in general. We have seen that when making direct inferences, information about the frequency of elements of a target class among the elements of a reference class, that ought to be as narrow as possible, plays a central role. A problem arises, when the normative guideline of choosing the narrowest reference class conflicts with the precision of the frequency information for that class—the “conflict of narrowness and precision” (Wallmann 2017, p. 485) mentioned above, that led to the debate about the correct computation of expected frequencies. Artificial neural networks, however, although resting upon ideas from the first half of the last century, can be considered a relatively new technology that gained its applicability mainly from what Wheeler (2016, p. 330) refers to as “the era of big data”. As he points out, “big” can be interpreted along two dimensions that come to mind very naturally within the structure of the data used in machine learning: Recall, that it is governed by the number of observations, n, on the one hand, and, on the other hand, by the number of features or properties, d, both of which could be identified as important for the functionality of artificial neural networks above.

Now first, the sheer number of observations in datasets that are handled nowadays is vast. Thus, while classical statistics is in large part concerned with assessing the significance and precision of inferences made from a restricted sample, either by analyzing the small-sample behavior of estimators or by employing—relatively conservative—asymptotic arguments, “we are now routinely handling population datasets directly or sample sizes so immense [...] that they behave like population data” (Wheeler 2016, p. 330). As a consequence, the considerations regarding the precision of inferences in classical statistics do not, or at least to a far lesser extent, carry over to applications of machine learning, since in this case, the representativeness of a given sample for the entire population is almost guaranteed solely based on the size of the sample.

Second, most datasets nowadays belong to the high-dimensional setting outlined above, where in addition to a large number of observations, n, each of them is associated with a—possibly much higher—number of properties, d (Wheeler 2016, p. 327). This especially holds for applications of artificial neural networks, since, as shown above, they are particularly suited for these kinds of settings that typically occur in applications involving text, speech or images. In fact, the discussion of the double-descent framework and its intuitive depiction in figure 6 revealed that artificial neural networks, contrary to other machine learning methods, perform best in situations in which the data contains a very high number of features. In the context of the reference class problem, where any further predicate that is added to the definition of a set yields a narrower reference class, this means that artificial neural networks regularly deal with *very* narrow reference classes and that, moreover, they have a high ability to do so.

Consequently, “big data”, understood as the two-dimensional concept I just described, provides a promising basis to approach the reference class problem employing artificial neural networks, for it addresses both components of Reichenbach’s initial idea: to choose a reference class that is narrow and for which reliable statistics are available. Furthermore, the conflict between narrowness and precision is sidestepped by the fact that even for narrow reference classes, the total number of members and hence the precision of information obtained from a sample is high. This also ameliorates the concerns regarding the correct computation of expected frequencies that I discussed above.

At this point, one might object that many of the observations made in this paragraph are not confined exlusively to the case of artificial neural networks and might instead be part of an argumentation for using machine learning methods in general to approach the reference class problem. Although this is a legitimate objection, I would like to underscore the fit of artificial neural networks to situations involving high-dimensional data that makes them stand out against other methods of machine learning as particularly suited to deal with the reference class problem. By the end of the chapter, the reader hopefully acknowledges that I convincingly made a case for artificial neural networks, if any, being the appropriate machine learning method in the context of the reference class problem. As an intermediate step, however, let us examine another more general observation.

4.1.2 Second Observation: Empirical Risk Minimization is Related to Epistemic Accuracy

As pointed out in the discussion of the reference class problem, authors like Thorn (2017) and Wallmann (2017) set out their argumentation under the normative premise that the reference class ought to be chosen such that epistemic accuracy is maximized. While Thorn (2017) measures epistemic accuracy in terms of various proper scoring rules, Wallmann (2017) illustrates the argument using the average squared difference between inferred values and true values as shown in equation (1). In both cases, accuracy is maximized if the chosen reference class leads to the smallest deviation of predictions from true values. In the case of the average squared difference, this becomes particularly obvious by the fact that it is bounded below by zero—the situation in which the inferences for all individuals correspond to the true values—and gets higher as the deviation of predictions from true values increases.

At this point, I would like to draw the attention to an interestig analogy that arises in the context of machine learning. We have seen that in order to learn a general prediction rule from a sample of training data, machine learning algorithms operating withing the ERM framework follow the decision rule of choosing the prediction rule that results in the lowest possible training risk. This risk is computed as shown in equation (4), that is, as the average over some loss function evaluated at each observation within the training data. Furthermore, the rationale of a loss function is to capture the deviation between a prediction made by the general prediction rule that an algorithm inferred and the associated true observation. With this in mind, let us reconsider expression (4) for the risk of a machine learning prediction and expression (1) for the average squared difference used by Wallmann (2017, p. 489) in his discussion of the reference class problem. There is no doubt that both expressions are structurally similar, since both of them consist of an average that is taken over a range of n individuals and some measure specifying the cost that occurs in the case of a wrong prediction.

Indeed, the squared difference employed by Wallmann is an instantiation of a loss function that is commonly applied in machine learning in the context of regression problems and that is usually referred to as *square loss* in the literature (Shalev-Shwartz and Ben-David 2016, pp. 48). The similar structure of the expressions hints at an observation that might seem even more striking: Both of them are linked by a common goal they try to achieve.

While the normative criterion guiding the argumentation in Thorn (2017) and Wallmann (2017) is to maximize epistemic accuracy, the goal of machine learning algorithms is empirical risk minimization and hence, accuracy maximization as well (Goodfellow et al. 2016, pp. 101). Note, however, that there is a subtlety to address before drawing a conclusion from this observation of converging goals: Thorn's proof that frequency information for the narrowest reference class maximizes epistemic accuracy relies on the computation of differences between predictions and true values, but the idea of the ERM framework is to infer the prediction rule that maximizes accuracy within the training sample first and to use it to compute actual predictions only *afterwards*, within the test sample—so can we say anything about the test risk that is obtained from the differences between actual predictions and true values in the test sample, if machine learning algorithms are solely confined to actively minimizing the training risk? As it turns out, we can, for we have seen above that besides minimizing the training risk, machine learning algorithms should also "[m]ake the gap between training and test error small" (Goodfellow et al. 2016, p. 109), either by balancing over- and underfitting succesfully as depicted in figure 2 or—in the case of artificial neural networks—by going beyond the interpolation threshold as depicted in figure 6. Thus, maximization of accuracy in solutions to the reference class problem and in machine learning can in fact be regarded as conceptually similar. Consequently, machine learning algorithms seem appropriate to approach the reference class problem, for their goals converge with those of other solutions put forward in the literature.

4.1.3 Third Observation: There Are No Distributional Assumptions

The third and final preliminary observation once again concerns the debate regarding the correct computation of expected frequencies in the recent literature. Above, I argued that the "era of big data" generally weakens concerns about imprecise frequency information for narrow reference classes that would otherwise require the calculation of expected frequencies. Apart from this observation, however, recall that the philosophical discourse centers around the question as to how the weights used to compute expected frequencies ought to be determined. In this context, a particular emphasis was put on the—implicit or explicit—assumptions about the probability distribution from which the weights are obtained. Especially Wallmann (2017, p. 496) criticizes the strategy to obtain the weights from arbitrary subsets of the broader reference class employed by Thorn (2017) and suggests to use "'exceptional' subsets" instead of arbitrary ones. According to his argumentation, that I also outlined above, these subsets would be more appropriate for inferring the weights, since they are related to the target class one is interested in. As an example, Wallmann (2017, p. 496) considers the target class of people who smoke and argues that smoking rates "vary strongly with gender, age, education, poverty status and many more", thereby alluding to a relation between some particular subsets of a broader reference class and the target class. However, the precise nature of this relation remains unclear, for he mentions a causal relevance of the exceptional subsets to the target class and proceeds by concluding that hence, "they are probabilistically dependent" (ibid.). Although this direction of the argument is certainly true, the opposite direction does not hold: We are not licensed to infer a causal relation between some subset of a reference class and the target class merely from observing some probabilistic dependence—"correlation does not imply causation" as the saying goes. Yet, to follow Wallmann's proposal would require an additional rule of inference that is able to identify causal relations and thus the "'exceptional' subsets" he refers to. In his article, he seems to largely neglect this aspect, because the only comment defending his proposal against this latter objection is that "we consider certain classes of individuals, because we *believe* that they are causally related to the target class" (ibid., own emphasis). The discussion as to whether a formal method of causal inference would license more than a mere belief in a causal relation or whether it would serve as a justification of this belief is beyond the scope of this text. Suffice it to say that at least *prima facie*, an objection against selecting arbitrary classes of individuals that is based on some not otherwise specified belief as the more appropriate decision rule seems rather shaky.

In sum, the issue that existing solutions to the reference class problem are facing can be described concisely as follows:

> "[E]xpected accuracy is relative to the distribution employed to calculate the expected accuracy. Maximising expected accuracy is only a legitimate aim if the distribution is empirically accurate, i.e., if it

> matches the relative frequencies in the world" (Wallmann 2017, p. 490).

It is this observation that creates another opportunity for artificial neural networks to enter the debate as a possible remedy to the reference class problem. As mentioned above, one central difference between machine learning and classical statistics consists in the fact that the former remains entirely agnostic about the probability distribution from which the data was sampled (von Luxburg and Schölkopf 2011, p. 653). In other words, this means that within the ERM framework, machine learning algorithms only pursue the goal of minimizing empirical risk without making any prior distributional assumptions and without assuming or trying to infer probabilistic or causal relationships within the data. As a consequence, there is no distribution employed in their operation and thus, they sidestep the problem discussed in this section.

4.1.4 The Artificial Neural Network Approach to the Reference Class Problem

After three preliminary observations regarding the intersection of the reference class problem and artificial neural networks, some of which concern the field of machine learning as a whole, I would like to put together the different and still distinct pieces that I worked out in the course of this text. The result, I hope, will reveal that there are in fact cases, in which artificial neural networks offer a solution to the reference class problem.

As a first step, consider the reference class problem and, more generally, the schema (DI) of direct inference from the machine learning perspective as described above. Now, assume that we employ an artificial neural network to approach the situation at hand. While the schema (DI) tells us to begin the direct inference with a premise regarding frequency information about the occurrences of elements of the target class T we are interested in among elements of the reference class R that we have to choose, the outline on artificial neural networks revealed that in their case, everything starts with input data given in a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. At this point, we might invoke further insights about the functionality of artificial neural networks and, in particular, the findings made in the first observation above: Artificial neural networks perform best in a high-dimensional setting—this was indicated by the double-descent framework—and the "era of big data" regularly brings about datasets that belong to precisely this setting—this was an insight of the observation made in section 4.1.1. Consequently, it seems reasonable to assume that $d \gg n$ for the case at hand, that is, that it also belongs to a high-dimensional setting.

Next, we have to examine the structure of the input data more closely. Recall, that an artificial neural network processes the data observation by observation, which is why one observation, $\mathbf{x}_i \in \mathbb{R}^d$, at a time enters the network's input layer. We have seen that the dimension d indicates the number of features associated with each observation $\mathbf{x}_i$, so each observation might be interpreted as possessing d different properties or characteristics. Bearing in mind the first preliminary observation, namely that d is likely to be high and that a reference class gets narrower with each predicate that is added to its definition, we can conclude that an artificial neural network starts the whole prediction exercise with the narrowest reference class possible that is defined by a high number d of properties. Thus, this very first step is in line with the recommendation by Thorn (2017) and Wallmann (2017) to use—frequency—information for the narrowest reference class available.

Before proceeding, let me briefly address an objection that might be raised at this point to bring the argument to a halt right from the beginning. I have argued that each observation $\mathbf{x}_i$ is associated with d different characteristics. These might be seen from at least two different angles. To explore them, recall the example from above where we considered a design matrix $\mathbf{X} \in \mathbb{R}^{12\times 6}$, that contains information on age, breed, coat type, color, weight, and height for Flint and eleven other dogs in the neighborhood. Now, the first perspective would be to hold that each dog in the dataset is associated with *the same* features, namely those six that I just mentioned: Flint possesses the property "age" and so do all other dogs. The second perspective would be that albeit the *names* of the features—"age" and so on—are the same for all dogs, the particular *instantiations* are not: For instance, Flint might be of the same age as Clint, but all other dogs might be of a different age. The two perspectives thus raise the question as to which of them is the one that is implicitly built into the operation of artificial neural networks and, furthermore, whether the first perspective really gives rise to a reference class as regularly conceived of in the literature. Indeed, what defines a reference class in the context of characteristics such as "age" or "height"—and opposed to characteristics such as "dachshund"—is not the property of possessing the characteristic or not, but its specific instantiation: While it is uninformative to distinguish the reference classes $R_1 := \{x\colon x \text{ is a dog}\}$ and $R_2 := \{x\colon x \text{ is a dog that has an age}\}$, this is not true for a comparison of R_1 and $R_2' := \{x\colon x \text{ is a dog that is less than ten years old}\}$, for in the latter case it holds that $R'_2 \subseteq R_1$, while in the former case we have $R_1 = R_2$. I contend that it is the second perspective that artificial neural networks take when processing the data. This is because they learn from *any* structure the data possesses, in particular from the features' different instantiations. A way to illuminate this point is to think of transforming the data such that each column in the new design matrix corresponds to a specific instantiation of the properties contained in the original design matrix and contains the value "1" if an observation has this property and "0" otherwise. So instead of the feature "age" in the original design matrix, the new design matrix would have features "nine years old", "ten years old" and so on. Clearly, it could be used as an input to an artificial neural network as well.

So let us return to the main line of the argument. Given the input data $\mathbf{X} \in \mathbb{R}^{n \times d}$, the artificial neural network starts the process of learning from it. As outlined above, this means that all weights within the network are chosen such that the empirical risk is minimized, that is, in a way in which the predictions computed from a particular configuration of weights are as close as possible to the true values within the training sample. At this point, I would like to underscore the importance of the second preliminary observation made in section 4.1.2: Very much in the spirit of the suggestion to maximize epistemic accuracy when approaching the reference class problem, artificial neural networks operate within the ERM framework and hence maximize accuracy as well. We have seen, for instance in figure 6, that artificial neural networks typically achieve perfect accuracy in the training sample, that is, the training error is zero. In the context of the reference class problem, this means that artificial neural networks are able to exploit the information regarding the different characteristics in the data to an extent that allows them to correctly predict for each observation whether $\mathbf{x}_i \in T$ or not. The configuration of weights that allows them to do this is the general prediction rule h that can be used to predict new observations.

Before proceeding to the final step that concludes my argument, I would like to clarify one aspect in the artificial neural network approach to the reference class problem and its relation to the schema (DI). In the schema (DI), existing data and the observation that a new individual belongs to the chosen reference class are used as premises to infer the probability with which the individual also belongs to the target class. In the artificial neural network approach, however, the learning process takes place prior to any actual prediction. This process might thus be interpreted as several instantiations of (DI) in which the algorithm iteratively observes an individual along with its characteristics in the training sample, issues a prediction for it and evaluates it against the true realization. Only upon completion of this process, the network's weights are fixed and constitute the prediction rule h that can be used to compute predictions for new observations. Note, that this implies that once confronted with a new observation, the network does not explicitly evaluate its different characteristics in order to assign it to the reference class that yields the most precise prediction for its belonging to the target class. Instead, the rationale is that all information regarding the different properties and their relevance for an individual being member of the target class is *learned* from the data and hence *implicitly embodied* in the final configurations of weights.

The crucial part that makes artificial neural networks stand out as particularly suited to deal with the reference class problem against other methods of machine learning is the following: the precise way in which they exploit the information regarding the different characteristics in the data during the learning process. The discussion of aspects that distinguish artificial neural networks from other methods of machine learning allowed us to identify the central role of implicit regularization that takes place in the determination of a network's weights. We have seen that it is this particular feature of artificial neural networks that prevents them from overfitting, ensures their high predictive ability and generally yields a final prediction rule h that is simple in the sense that the network's weights have a small norm. Thus, by means of implicit regularization, the impact that each of the d inputs to the network—the characteristics of the observations—should exhibit on the network's output—the prediction whether an individual belongs to a target class or not—is determined following the normative guideline of maximizing accuracy. This is a process in which some weights are assigned a high value, since the impact of the associated input on the final output is high, and others are assigned a low value—maybe even zero—, since the associated input does not contribute to the prediction of the output. Important properties are considered important for the prediction, less important properties are considered less important or are neglected altogether. As a consequence, the process of implicit regularization can be conceived of as a decision process that selects the reference class that is most suitable to predict with maximal accuracy whether a new and previously unseen individual is an element of a target class or not. So in summary, one can say that situations in which artificial neural networks are regularly employed, that is, situations involving "big data", allow them to use precise data for very narrow reference classes and to incorporate the information in a combination of weights that maximizes their predictive accuracy. This is why artificial neural networks are suited to deal with the reference class problem and might in fact be a remedy to it in these situations.

4.2 . . . In Many Others, They Don't: Possible Objections

While I addressed several minor objections that might be raised against my argumentation directly throughout the text, there are more severe ones that require an own section devoted to their examination—this is the purpose of the subsequent paragraphs.

First, we have to address the issue of accuracy. Undeniably, it played a central role thus far by serving as a normative criterion guiding Thorn (2017) and Wallmann (2017) in their treatment of the reference class problem and machine learning algorithms in their choice of a general prediction rule from an entire hypothesis class in the ERM framework. Consequently, I use the concept of accuracy in my argumentation to establish a link between the philosophical literature and the behavior of artificial neural networks. However, there are at least two drawbacks to this strategy. The first one is mentioned by Wallmann (2017, p. 489) who observes that the result derived by Thorn (2017) that using frequency information for the narrowest reference class available maximizes epistemic accuracy only holds as long as the latter is measured by proper scoring rules. Consequently, simply using another measure of accuracy that does not conform to the definition of a proper scoring rule might invalidate Thorn's entire reasoning that I somehow incorporated into my own argumentation. Yet—and this is my reply to the objection—proper scoring rules come in various forms that both correspond to several common loss functions used in machine learning and are widely applicable to situations in which accuracy needs to be measured. When reasoning about accuracy, it therefore seems reasonable to do so in terms of proper scoring rules.

The second drawback to my strategy of using the concept of accuracy as an analogy between treatments of the reference class problem and machine learning, however, is more serious. Recall the observation by Hájek (2007, p. 568) mentioned above that the answer to the question as to what is the right reference class might be a matter of context, depending, for instance, on "the weighing of utilities". This point then challenges the concept of accuracy altogether, regardless of the specific measure that is used to capture it: What if accuracy is not the normative criterion by which our choice of a reference class ought to be guided? At this point, I must confess that I do not have an appropriate strategy to attenuate this objection. I acknowledge the relevance of questioning the concept of accuracy, but neither do I see a convincing alternative to it nor do I see how possible alternatives will not rely on the concept of accuracy as well, in some way or another. Thus, my use of the concept of accuracy can be seen as a working hypothesis that I employ in lack of a more appropriate concept.

Another objection that concerns both the argumentations by Thorn (2017) and Wallmann (2017) as well as my own can arise from questioning the nature of what is in fact predicted making use of the chosen reference class R. Clearly, we are interested in predicting whether an individual a is an element of a target class T, $a \in T$, based on information involving R. But does an argumentation in favor of maximizing accuracy address this type of single-case prediction properly? Recall, that both the accuracy measure employed by Wallmann shown in equation (1) and the expression for the training or test risk of a machine learning algorithm in equation (4) run over n observations instead of only one. This observation leads Wallmann (2017, p. 490) to question whether the long-run interpretation of the accuracy measures I just mentioned is even relevant for the short run. He tries to find an answer by referring to Pollock (2011, p. 349) who states that in the context of single-case versus long-run prediction

> "[p]eople sometimes protest at this point that they are not interested in the general case. They are concerned with some inference they are only going to make once. They want to know why they should reason this way in the single case. But all cases are single cases. If you reason in this way in single cases, you will tend to get them right.

Thus, following Pollock's suggestion, one should not treat single-cases in a special way, for many of them will eventually add up to a long run. Wallmann (2017) seems to embrace this approach largely unchallenged. I am, however, skeptical about the argument that many single cases are easily equated with one long run. After all, it might be conceptually impossible, due to ontological reasons, to predict anything about a previously unseen individual. Yet I acknowledge that when facing a situation in which a single-case prediction ought to be made, evidence obtained from known individuals is likely to be the best basis for it. So again, as in the case of accuracy measures, the strategy of deriving single-case predictions from long-run considerations is certainly not without shortcomings, but it is—at least as far I can see—the best strategy available.

While the first two objections concerned the general setup of the argumentation, the last two that follow focus more explicitly on the role of artificial neural networks and the data they use as their input. Above, I argued that artificial neural networks are able to approach and solve the reference class problem because they are structured such that they always start their processing of the data with the narrowest reference class possible, since d features enter the input layer. Subsequently, the relevance of the features for the network's output is assessed and embodied—via implicit regularization—in the final combination of weights that yields the highest predictive accuracy. Yet this line of argumentation highly understates the role of human individuals and especially their prior knowledge, for instance regarding the data. Recall the concept of inductive bias, that I mentioned when discussing basic concepts of machine learning. The bottom line of the concept is that machine learning is only possible when human expertise is part of the process and biases an algorithm's quest for a general prediction rule, that would otherwise become infeasible, in a particular direction. Thus, it is a human agent who defines a hypothesis class $\mathcal{H}$ from which an algorithm chooses the prediction rule h. As we have seen, in the case of artificial neural networks, this means that the network's architecture is specified *before* the algorithm gets to learn anything. Furthermore, it is a human decision as to what data is used as an input to a machine learning algorithm. Very likely, this also entails the human decision as to what phenomena should be measured and thus captured in the form of data in the first place. As becomes evident, the claim that there are situations in which artificial neural networks are able to solve the reference class problem seems exaggerated and is certainly misleading. Thus, a more reasonable claim would be to state that *given an appropriate amount of data and an architecture of sufficient complexity*, there are situations in which artificial neural networks are able to solve the reference class problem.

Unfortunately, the last statement directly raises another issue: What is an appropriate amount of data? So far, I have confined my argumentation to situations involving high-dimensional data and, more generally, to data that is representative for "the era of big data". I did so because the discussion of artificial neural networks revealed that they are applied regularly to and perform best in these situations, but also because high-dimensional data that contains many features for each individual observation squares well with the idea of preferring narrow reference classes over broader ones. Having said that, let us return to Flint, the dachshund, for which we are still pondering whether he has fleas or not. There is no doubt that this example does not belong to a “big data” setting, neither with respect to the number of observations involved, nor regarding the characteristics associated with each observation. Nevertheless, it is a realistic example for the type of inference we make day after day. So how do artificial neural networks contribute to solving the reference class problem in this kind of situations? My rather disappointing answer is that they are of no help in these situations. This is because the process of learning, that is, the process of finding the optimal combination of weights in a network that is sufficiently complex to achieve a high predictive ability, requires amounts of data that are beyond human grasp and thus beyond situations of everyday reasoning. Furthermore, the double-descent framework that is central to explaining the success of artificial neural networks relies heavily on data that contains both a high number of observations and a high number of features. As it turns out, this is the qualification that discerns situations in which artificial neural networks might solve the reference class problem from many others in which they do not.

5 Conclusion

In the course of this thesis, I tried to shed light on the relation between machine learning and classical statistics. In order to turn this into a feasible undertaking, I confined the investigation to the question as to whether one particular method of machine learning, namely artificial neural networks, is subject to one particular problem of classical statistics, namely the reference class problem.

In a first step, the analysis of the reference class problem and solutions to it that have been proposed in the literature revealed that, most importantly, one ought to choose the narrowest reference class for which frequency information is available when trying to maximize epistemic accuracy. Recent results, especially those by Thorn (2017) and Wallmann (2017), indicate that this even holds in the case of imprecise frequency information.

In a second step, we have seen that artificial neural networks differ from other methods of machine learning mainly because they are not subject to overfitting: In most cases, they perfectly fit the training data while being still flexible enough to exhibit a high predictive accuracy when confronted with previously unseen data. Furthermore, they perform best in situations involving high-dimensional or "big data" and their behavior is best illuminated by the double-descent framework introduced by Belkin et al. (2019).

My subsequent argumentation made the attempt to synthesize the findings from the first and the second step. I argued that the concepts "narrowness" and "reliability" that play a central role in the debate concerning the reference class problem are related to "big data" that is processed by artificial neural networks. The key insight in this context, for instance put forward by Wheeler (2016), was to conceive of "big data" as a two-dimensional concept that involves both a high number of observations and a high number of features.

Furthermore, I pointed out that the normative guideline of maximizing epistemic accuracy when addressing the reference class problem is analogous to the ERM framework in machine learning. This is because machine learning algorithms operating within the ERM framework follow the decision rule of choosing the prediction rule that results in the lowest possible training risk and hence, in the highest possible accuracy.

The latter observations together with the insights into the specific functionality of artificial neural networks allowed me to conclude that there are in fact situations in which artificial neural networks are able to overcome the reference class problem. These situations necessarily involve "big data", such that a high number of features enters a network's input layer. Thus, they start with the narrowest reference class available and with the goal of maximizing accuracy. Then, the algorithm that determines the final configuration of weights achieves this goal via an implicit regularization that is unique to artificial neural networks and prevents them from overfitting.

Clearly, this line of argumentation is not without objections. For this reason, I mentioned the—from my point of view—most serious ones in the last part of the text. At least two objections arise from the concept of accuracy that plays a central role in my argumentation and also in the general debate regarding the reference class problem: First, one might question the measurement of accuracy using proper scoring rules and second, one might question the concept of accuracy altogether as the right normative guideline when trying to solve the reference class problem. Although I acknowledge the relevance of questioning the concept of accuracy, I hold that it is at least the most feasible guideline and can thus be seen as a working hypothesis in the course of this thesis.

Another serious objection concerns the relation between human involvement and the agency of artificial neural networks or machine learning methods in general. My argumentation might have created the impression that, in specific situations, artificial neural networks can solve the reference class problem *on their own*. This, however, masks the relevance of human involvement during the entire process, be it the choice of an architecture for the network or that of the input data. Consequently, it is a more sensible formulation that artificial neural networks can *aid* solving the reference class problem given an appropriate network architecture and appropriate input data.

Finally, the question as to what constitutes "appropriate input data" led us back to Flint and the inference whether he has fleas or not. Here, a final objection and, in fact, a serious qualification to my argumentation was that in many everyday situations, artificial neural networks are of no help in solving the reference class problem. Thus, in these situations, artificial neural networks fall prey to the reference class problem just as classical statistics.

What does this result imply for the overall relation between classical statistics and methods of machine learning? After all, there seem to exist some very specific situations in which particular methods of machine learning go beyond classical statistics. However, the requirements for these situations are high and of little relevance for our everyday reasoning. Additionally, the means by which methods of machine learning achieve to go beyond classical statistics remain opaque. On the one hand, the role of human involvement is an issue that is largely neglected, both in machine learning research and in the philosophical literature. On the other hand, methods of machine learning—and especially artificial neural networks—are inherently opaque and "barely explainable" (Schubbach 2019, p. 1). For instance, we have seen that machine learning researchers found *some* implicit regularization taking place in the determination of a network's weights. But we have also seen that this implicit regularization is considered a "major issue still left unresolved" (Neyshabur et al. 2017, p. 8). As a consequence, the present discussion of artificial neural networks and the reference class problem, that was intended to shed light on the relation between machine learning and classical statistics, also reveals avenues for future work. In particular, the role of human involvement in the machine learning process, that is, the interplay of human and machine agency is a philosophically relevant topic. Furthermore, an important area for future research arises from the question as to how methods of machine learning might become more explainable. Although this would not make them more applicable to everyday situations such as Flint's case, a philosophical investigation of how to make the functionality of machine learning methods and the way by which they arrive at their predictions more explainable might help to overcome much of their current opacity.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854.

Buckner, C. (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese*, 195(12):5339–5372.

Buckner, C. (2019). Deep Learning: A Philosophical Introduction. *Philosophy Compass*, 14(e12625).

Corfield, D., Schölkopf, B., and Vapnik, V. (2009). Falsificationism and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions. *Journal for General Philosophy of Science*, 40(1):51–58.

Fetzer, J. H. (1977). Reichenbach, Reference Classes, and Single Case 'Probabilities'. *Synthese*, 34:185–217.

Gillies, D. (2000). Varieties of Propensity. *The British Journal for the Philosophy of Science*, 51(4):807–835.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

Hájek, A. (2007). The Reference Class Problem Is Your Problem Too. *Synthese*, 156(3):563–585.

Harman, G. and Kulkarni, S. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press, Cambridge, MA.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition.

Hitchcock, C. and Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, 55(1):1–34.

McGrew, T. (2001). Direct Inference and the Problem of Induction. *The Monist*, 84(2):153–178.

Minsky, M. L. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge, MA.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring Generalization in Deep Learning. In Guyon, I., Luxburg, U. v., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5947–5956.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *International Conference on Learning Representations.*

Pollock, J. L. (1990). *Nomic Probability and the Foundations of Induction.* Oxford University Press, New York.

Pollock, J. L. (2011). Reasoning Defeasibly About Probabilities. *Synthese*, 181(2):317–352.

Poser, H. (2012). *Wissenschaftstheorie: Eine philosophische Einführung.* Reclam, Stuttgart, 2nd edition.

Reichenbach, H. (1949). *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability.* University of California Press, Berkeley and Los Angeles, CA, 2nd edition.

Romeijn, J.-W. (2017). Philosophy of Statistics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy.*

Schmidhuber, J. (2015). Deep Learning in Neural Networks: an Overview. *Neural Networks*, 61:85–117.

Schubbach, A. (2019). Judging Machines: Philosophical Aspects of Deep Learning. *Synthese*, 134(6245):1–21.

Shalev-Shwartz, S. and Ben-David, S. (2016). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, New York, 1st edition.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, George, Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587):484–489.

Thorn, P. D. (2012). Two Problems of Direct Inference. *Erkenntnis*, 76(3):299–318.

Thorn, P. D. (2017). On the Preference for More Specific Reference Classes. *Synthese*, 194(6):2025–2051.

Thorn, P. D. (2019). A Formal Solution to Reichenbach's Reference Class Problem. *Dialectica*, 73(3):349–366.

von Luxburg, U. and Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier, Amsterdam and Boston.

Wallmann, C. (2017). A Bayesian Solution to the Conflict of Narrowness and Precision in Direct Inference. *Journal for General Philosophy of Science*, 48(3):485–500.

Wheeler, G. (2016). Machine Epistemology and Big Data. In McIntyre, L. and Rosenberg, A., editors, *The Routledge Companion to Philosophy of Social Science*, pages 321–329. Routledge, London and New York, NY.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding Deep Learning Requires Rethinking Generalization. *International Conference on Learning Representations.*

Zoglauer, T. (2016). *Einführung in die formale Logik für Philosophen.* Vandenhoeck & Ruprecht, Göttingen, 5th edition.